Generalization at the Edge of Stability

Mario Tuci[1,2,*] **Caner Korkmaz**[2,*] **Umut Şimşekli**[1,†] **Tolga Birdal**[2,†]

[1]INRIA, CNRS, École Normale Supérieure [2]Imperial College London,
mario.tuci@inria.fr, c.korkmaz23@imperial.ac.uk,
umut.simsekli@inria.fr, tbirdal@imperial.ac.uk

Abstract

Training modern neural networks often relies on large learning rates, operating at the *edge of stability*, where the optimization dynamics exhibit oscillatory and *chaotic behavior*. Empirically, this regime often yields improved generalization performance, yet the underlying mechanism remains poorly understood. In this work, we represent stochastic optimizers as *random dynamical systems*, which often converge to a *fractal attractor set* (rather than a point) with a smaller intrinsic dimension. Building on this connection and inspired by Lyapunov dimension theory, we introduce a novel notion of dimension, coined the 'sharpness dimension', and prove a generalization bound based on this dimension. Our results show that generalization in the chaotic regime depends on the *complete* Hessian spectrum and the structure of its *partial determinants*, highlighting a complexity that cannot be captured by the trace or spectral norm considered in prior work. Experiments across various MLPs and transformers validate our theory while also providing new insights into the recently observed phenomenon of grokking.

Understanding why large, overparameterized networks trained by gradient-based methods generalize remains one of the central open problems in machine learning. Recent empirical evidence has revealed a phenomenon that challenges classical convex optimization theory. Cohen et al. [1] observed that when training neural networks with gradient descent of fixed learning rate η, the largest eigenvalue of the loss Hessian often oscillates around, and frequently exceeds $2/\eta$, even as the training loss continues to decay. This behavior, termed the *edge of stability* (EoS), has generated considerable interest (see Sec. 1), since the threshold $2/\eta$ implies instability and divergence for quadratic objectives [2]. [3] demonstrated that exceeding the threshold of $\frac{2}{\eta}$ is sufficient to induce chaotic training dynamics. Additionally, in the chaotic regime the optimizer will not settle at a single point, rather it explores a bounded, typically fractal like set [4]. This raises a fundamental question:

How can generalization be explained in the regime that is locally unstable and potentially chaotic?

A natural response to this puzzle is to examine the local geometry of the loss landscape: the Hessian, which encodes local curvature, has long been viewed as a key lens for understanding generalization [5, 6], motivating a large body of work based on *pointwise* notions of *sharpness* and *flatness* (see Sec. 1). Despite its appeal and empirical successes, this viewpoint has been challenged. It is now well understood that sharpness-based criteria (e.g., Hessian trace) are neither necessary nor sufficient for good generalization: there exist flat minima that generalize poorly and sharp that do well [7–9]. Consequently, characterizing generalization at the EoS through pointwise analysis of individual solutions may be inadequate. For practical learning rates, training dynamics are expected to exhibit chaotic behavior [4], wherein trajectories display sensitive dependence on initialization. In this regime, generalization performance should be attributed not to the properties of any single solution, but rather to the geometry of the entire solution set explored by the optimizer in the long term.

[*,†] Authors contributed equally.

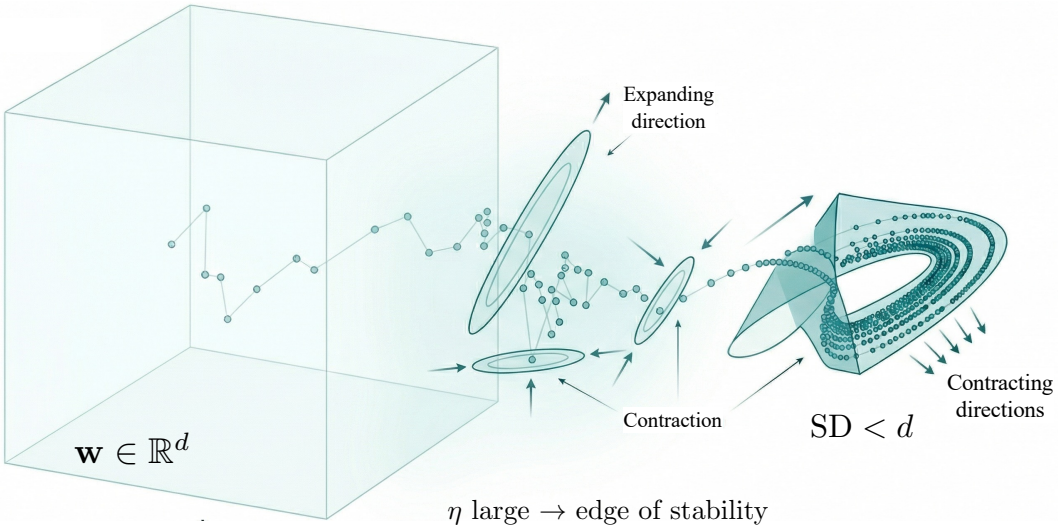

Figure 1: **Generalization at the Edge of Stability (EoS).** Modeling stochastic optimization as a random dynamical system (RDS), we show that at EoS the leading sharpness satisfies $\lambda_1 > 0$, implying expansion along at least one direction. The fundamental balance between expansion and contraction implies that the effective dimensionality of the dynamics, measured by our **Sharpness Dimension** (SD), is strictly smaller than the ambient parameter space: $\mathrm{SD} < d$. We prove that the **worst-case generalization error is governed by SD rather than the parameter count.**

Contributions. We introduce a framework for studying generalization at the edge of chaos where modern deep networks operate [10]. In particular, we contribute the following:

- **Attractor-Centric Framework**: Modeling stochastic optimization as a random dynamical system (RDS), we shift the study of generalization at EoS from isolated parameter vectors to the **geometric properties of the (random) attractor**.
- **RDS Sharpness & Sharpness Dimension** (SD): We propose **two new complexity measures**, RDS Sharpness & SD, derived not from the trajectory but from expansion and contraction rates that characterizes the attractor's geometry.
- **Generalization Bound**: We provide **a new bound on the worst case generalization error**, rigorously linking it to the fractal dimension of the Random Attractor measured as SD.
- **Empirical Validation**: We explain how to **compute SD efficiently** and deploy it in quantifying generalization across multilayer perceptrons and **recent transformers** (GPT-2 [11]) as well as to study the recently introduced paradigm of *grokking*, delayed and sudden generalization [12, 13].

As illustrated in Fig. 1, our results show that at the EoS, the coexistence of expanding and contracting directions gives rise to an attractor with provably lower intrinsic dimension. The attractor's effective dimension is controlled by SD and is strictly smaller than the dimension of the ambient parameter space. We further prove that the worst-case generalization error under certain assumptions admits a bound in terms of SD, providing a principled explanation for why overparameterized models can generalize in ways not captured by classical complexity measures. Our code is available under: https://circle-group.github.io/research/GATES.

1 Related Work

Generalization bounds. Recent work has established strong empirical and theoretical connections between generalization and geometric and topological complexity measures derived from the *optimization trajectory* [14–18]. In particular, it has been shown that weight iterates sampled *after* initial convergence encode critical information about generalization. Topological summaries such as the 'α-weighted lifetime sum' exhibit consistent correlations with the generalization gap across training runs [17, 18]. These findings suggest that the stochastic fluctuations observed during late-stage training are not merely noise, but reflect the optimizer exploring a structured geometric object. Differently, [19] showed that under contraction, stochastic GD (SGD) admits an invariant measure supported on a fractal set and linked its geometric complexity to generalization. However, this theory relies on contraction and therefore does not apply at the EoS.

Hessian and generalization. Since [20], it has been conjectured that flatter minima generalize better, motivating extensive work on flatness and sharpness [5, 7, 21–29]. Despite this effort, there is no universally accepted definition of flatness, and most practical surrogates rely on second-order cues such as the Hessian trace [9, 30]. In parallel, optimization methods designed to favor flat minima showed strong empirical gains in generalization [31–35]. Theoretically, minimizing the Hessian trace was shown to select the true solution in low-rank matrix recovery [36], extended to deep networks [37], supported empirically for large language models [38], and linked to stability [39]. More recently, generalization bounds have been obtained in terms of gradient norms [40, 41], which reduce to the Hessian trace under suitable conditions. Despite these advances, recent results show that trace-based flatness alone does not guarantee good generalization [9].

Edge of Stability (EoS) and chaos. The dynamics of deep network training have been widely studied, with early work documenting rapid changes in the local loss landscape during the initial phase of optimization [5, 30, 42], and later characterizing it for gradient descent by Cohen et al. [1]. A growing literature has since been interested in EoS phenomena [43–49]. In particular, [39, 45] showed that the existence of a forward-invariant set prevents divergence, and that for $\tanh$ networks such a set exists, explaining stability even at large learning rates in the EoS regime. Related phenomena have also been observed for SGD via the notion of mini-batch sharpness [50]. While implicit dynamical stability of SGD was investigated by [51] in the case of two-layer neural networks, complementary perspectives reveal connections to chaos [3, 52]. In particular, [3] shows that sustained criticality of the top Hessian eigenvalue (EoS) is sufficient to induce chaos, and [53] studies a cubic model to characterize the boundary between EoS and divergence. Existing analyses of generalization near EoS have primarily considered optimization below the EoS threshold. This work examines how the data distribution shapes generalization [54] and contrasts GD with SGD through stability analyses around interpolating minima [51]. We instead study the locally unstable dynamics at EoS itself, connecting the geometry of the resulting attractor to generalization

Our work connects Edge of Stability, chaotic dynamics, Hessian-based generalization bounds via rigorous generalization guarantees through the lens of random dynamical systems.

2 Preliminaries

Learning setup. We consider supervised learning with parameter vector $w \in \mathbb{R}^d$ and population risk $\mathcal{R}(w) := \mathbb{E}_{Z\sim\mu}[\ell(w, Z)]$, where μ is an unknown data distribution and $\ell : \mathbb{R}^d \times \mathcal{Z}$ is the composed loss function. In practice, training proceeds by minimizing the empirical risk $\widehat{\mathcal{R}}_S(w) := \frac{1}{n}\sum_{i=1}^n \ell(w, Z_i)$ over a dataset $S = \{Z_i\}_{i=1}^n \sim \mu_z^{\otimes n}$ using stochastic gradient methods. Our analysis focuses on the asymptotic behavior of the optimization dynamics rather than on a single training iterate. Therefore we are interested in the worst-case generalization gap

$$\mathcal{G}(\mathcal{A}_S(\omega)) := \sup_{w\in\mathcal{A}_S(\omega)} \mathcal{R}(w) - \widehat{\mathcal{R}}_S(w). \tag{1}$$

Here, $\mathcal{A}_S(\omega)$ denotes a dataset-dependent random set, where $S \in \mathcal{Z}^n$ is the dataset and ω is an independent random variable capturing algorithmic noise (e.g., minibatch sampling). In particular, $\mathcal{A}(\omega)$ denotes the *random attractor* associated with the optimizer, to be precised in Dfn 2.2.

Random dynamical systems. To have a rigorous consideration of random attractors as sets, we represent stochastic optimization algorithms as discrete-time *random dynamical systems* (RDS) according to Arnold [55]. While the definition is abstract, it will follow with a concrete example for SGD, represented in the RDS form. Note that a similar formalism was introduced to study the stability of SGD in Chemnitz and Engel [53].

Definition 2.1 (Random Dynamical System)**.** A *discrete-time random dynamical system (RDS)* on $\mathbb{R}^d$ is a tuple $(\Omega, \mathcal{F}, \mathbb{P}, \theta, \phi)$ consisting of the following components:

1. A **metric dynamical system** $(\Omega, \mathcal{F}, \mathbb{P}, \theta)$, where $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space and $\theta : \Omega \to \Omega$ is an invertible, measure-preserving, and ergodic transformation[2] such that the map $(t, \omega) \mapsto \theta^t(\omega)$ defines a $\mathbb{Z}$-action on Ω. That is, the family of maps $\{\theta^t\}_{t\in\mathbb{Z}}$ satisfies
$$\theta^0 = \mathrm{id}_\Omega{}^3, \qquad \theta^{t+s} = \theta^t \circ \theta^s \quad \text{for all } t, s \in \mathbb{Z},$$
where we use the notation $\theta\omega := \theta(\omega)$.

[2] A transformation θ is *measure-preserving* if $\mathbb{P}(\theta^{-1}A) = \mathbb{P}(A)$ for all $A \in \mathcal{F}$, and *ergodic* if any invariant set $A = \theta^{-1}A$ satisfies $\mathbb{P}(A) \in \{0, 1\}$. [3] For an arbitrary set A, $\mathrm{id}_A : A \to A$ denotes the identity map defined by $\mathrm{id}_A(x) = x$ for all $x \in A$.

2. A **measurable cocycle** $\phi : \mathbb{N}_0 \times \Omega \times \mathbb{R}^d \to \mathbb{R}^d$ over θ, which is $(\mathcal{B}(\mathbb{N}_0) \otimes \mathcal{F} \otimes \mathcal{B}(\mathbb{R}^d), \mathcal{B}(\mathbb{R}^d))$[4] -measurable and satisfies the *cocycle property* for all $t, s \in \mathbb{N}_0$, $\omega \in \Omega$, and $w \in \mathbb{R}^d$:

$$\phi(0, \omega, \cdot) = \mathrm{Id}_{\mathbb{R}^d}, \quad \phi(t + s, \omega, w) = \phi(t, \theta^s \omega, \phi(s, \omega, w)), \tag{2a}$$

An RDS is said to be k times *continuously differentiable* (or C^k) if the mapping $w \mapsto \phi(t, \omega, w)$ is C^k for all $t \in \mathbb{N}_0$ and $\omega \in \Omega$.

Let us provide a more intuitive explanation for the above definition by considering SGD as a special case. Given a dataset $S = \{Z_1, \ldots, Z_n\} \in \mathcal{Z}^n$, and $\Omega_t \subset \{1, \ldots, n\}$ being the minibatch set with $|\Omega_t| = b$ being the batch-size. Since there are only finitely many possible choices of minibatches, i.e., $\binom{n}{b}$ many, we can enumerate all the minibatches such as $\{\Omega^{(1)}, \ldots, \Omega^{(\binom{n}{b})}\}$ s.t. $\Omega_t = \Omega^{(j)}$ for some j. Hence SGD can be written in the following form

$$w_{t+1} = w_t - \eta \left(\frac{1}{b} \sum_{i \in \Omega^{(j_t)}} \nabla \ell(w_t, Z_i) \right), \tag{3}$$

where j_t is randomly drawn from the set $\{1, \ldots, \binom{n}{b}\}$.

Going back to Definition 2.1, the random event ω corresponds to the *algorithmic randomness*. In the case of SGD, the only source of algorithmic randomness is the choice of minibatches at every iteration. Hence, for SGD, $\omega = \{\ldots, j_{-t}, \ldots j_{-1}, j_0, j_1, \ldots, j_t, \ldots\}$ will encapsulate the infinite sequence of minibatch indices that are drawn through optimization[5].

Given a sequence of minibatch indices ω, we can define the *RDS map* ϕ, which essentially corresponds to the algorithm update rule. For $t = 1$, we have the following form:

$$\phi(1, \omega, w) := w - \eta \left(\frac{1}{b} \sum_{i \in \Omega^{(j_0)}} \nabla \ell(w, Z_i) \right), \tag{4}$$

which coincides with (3), but only for $t = 1$.

At iteration t, the stochastic gradient is computed over the minibatch with index j_{t-1}. Hence, to define the RDS map for a general t, i.e. $\phi(t, \omega, w)$, we additionally need to extract the minibatch index j_{t-1} from the infinite sequence ω. This operation will be done by the *metric dynamical system* θ. For this example[6], we set $\theta : \mathbb{Z} \times \Omega \to \Omega$ to the so-called the 'left-shift operator', $(\theta\omega)_k := j_{k+1}$, which takes an infinite sequence ω and shifts its elements by one coordinate, and returns the resulting infinite sequence. Iterating this operator t times would shift the coordinates t times: i.e., $(\theta^t \omega)_k := j_{k+t}$.

Given all the ingredients, by (2a) we can define our RDS for a general t:

$$\phi(t, \omega, w) = \phi(1, \theta^{t-1}\omega, \phi(t - 1, \omega, w)), \tag{5}$$

where $\phi(1, \cdot, \cdot)$ is defined in (4). It is now easy to verify that (5) exactly recovers the recursion given in (3). Finally, we observe that the system satisfies the cocycle property:

$$\phi(t, \omega, w) = \phi(1, \theta^{t-1}\omega, \cdot) \circ \cdots \circ \phi(1, \omega, w), \forall t \in \mathbb{N}_0.$$

Intuitively, it ensures that evolving a state for $t + s$ steps is equivalent to evolving it for s steps and then resuming for t more steps using the remaining noise history $\theta^s\omega$. While denoting an optimization algorithm with such abstract notions might seem rather unorthodox, thanks to this formal connection, we will be able to access the rich toolbox of random dynamical systems theory.

Random Attractors. Dynamics driven by stochastic optimization with persistent noise (e.g. constant learning rate SGD) generally do not converge to a single location. We are interested in the set in which an RDS settles.

Definition 2.2 (Pullback Random Attractor)**.** Let $(\Omega, \mathcal{F}, \mathbb{P}, \theta, \phi)$ be a discrete-time RDS according to Dfn. 2.1. A mapping $\omega \mapsto \mathcal{A}(\omega)$ from Ω into the space of non-empty compact subsets of $\mathbb{R}^d$ (denoted $\mathrm{comp}(\mathbb{R}^d)$) is called a *pullback random attractor* if it is $(\mathcal{F}, \mathcal{B}(\mathrm{comp}(\mathbb{R}^d)))$-measurable and satisfies the following two properties for $\mathbb{P}$-almost all $\omega \in \Omega$:

[4] Let $\mathcal{B}(\mathbb{R}^d)$ or $\mathcal{B}(\mathbb{N}_0)$ denote the corresponding Borel σ-algebras. [5] For technical reasons, we need to consider a doubly infinite sequence that also takes into account for negative times $-t$. [6] For most of the stochastic optimizers one can use the same metric dynamical system as long as the algorithmic randomness is only coming from random minibatches.

1. **Invariance:** $\phi(t, \omega, \mathcal{A}(\omega)) = \mathcal{A}(\theta^t \omega)$ for all $t \in \mathbb{N}_0$.
2. **Pullback Attraction:** For every deterministic bounded set $B \subset \mathbb{R}^d$:

$$\lim_{t \to \infty} \operatorname{dist}\big(\phi(t, \theta^{-t}\omega, B), \mathcal{A}(\omega)\big) = 0, \tag{6}$$

where $\operatorname{dist}(X, Y) := \sup_{x \in X} \inf_{y \in Y} \|x - y\|$ is the Hausdorff semi-distance. Further the image of a function is defined as follows. If $f : X \to Y$ is a map and $A \subset X$, then the image of A under f is defined as $f(A) := \{\, f(x) \in Y \ : \ x \in A \,\}$.

In stochastic systems, attraction must be understood in a *pullback* rather than forward sense. Because fresh noise is continually injected, trajectories that come close can later separate, so pathwise forward convergence generally fails. The pullback viewpoint instead fixes a noise realization and examines the state at time $t = 0$ obtained by initializing the system in the remote past. The resulting pullback attractor is a noise-conditioned "snapshot" of the asymptotic state, representing the set to which all past histories converge at the present time. This is the central object of our analysis.

3 Theoretical Results

Sharpness. Recent work introduced the notion of sharpness in terms of the Hessian of the empirical risk [1]. We introduce an alternative notion of sharpness that extends to general random dynamical systems and provides an intuitive understanding of how the two notions are related.

Definition 3.1 (RDS Sharpness): Let $(\Omega, \mathcal{F}, \mathbb{P}, \theta, \phi)$ be a C^1 random dynamical system on $\mathbb{R}^d$. For a fixed state $w \in \mathbb{R}^d$, let $D\phi(1, \omega, w) \in \mathbb{R}^{d \times d}$ denote the Jacobian (Fréchet derivative) of the map $x \mapsto \phi(1, \omega, x)$ evaluated at $x = w$. Let $\mathcal{A}(\omega)$ be random compact set. We define the **RDS Sharpness of Order** k as the expected log-variation of the k-th singular value:

$$\lambda_k := \mathbb{E}\left[\sup_{w \in \mathcal{A}(\omega)} \ln \sigma_k(\omega, w)\right] \quad \forall k \in \{1, \ldots, d\}, \tag{7}$$

where $\sigma_1(\omega, w) \geq \cdots \geq \sigma_d(\omega, w)$ are the singular values of $D\phi(1, \omega, w)$, assuming integrability holds.

The RDS sharpness introduced in Dfn. 3 is algorithm-dependent, as it is defined through the optimization rule. For SGD, it is closely related to the notion of sharpness studied at the stochastic EoS [50]. The next example, shows how Dfn. 3 relates to the classical notion of sharpness introduced in Cohen et al. [1], which is defined in terms of the largest eigenvalue of the Hessian.

Example 3.1 (GD Sharpness):
We can interpret GD as a deterministic instance of a random dynamical system where the algorithmic randomness ω is constant. Let $\widehat{\mathcal{R}}_S(w)$ denote the empirical risk for a fixed dataset $S \in \mathcal{Z}^n$ and $w \in \mathbb{R}^d$; the discrete-time update map ϕ is defined as:

$$\phi(1, w) = w - \eta \nabla \widehat{\mathcal{R}}_S(w). \tag{8}$$

The local stability of this system is governed by the Jacobian $D\phi(1, w) = I - \eta \nabla^2 \widehat{\mathcal{R}}_S(w)$. In the optimization literature [1], sharpness is traditionally defined as the largest eigenvalue the Hessian $\nabla^2 \widehat{\mathcal{R}}_S(w)$.

If we interpret our sharpness definition (see Dfn. 3) locally over point (e.g. $\mathcal{A}(\omega) = \{w\}$), we obtain, that the sharpness of order 1 corresponds to $\lambda_1 = \ln(\|I - \eta \nabla^2 \widehat{\mathcal{R}}_S(w)\|)$, which exhibits related information to the largest eigenvalue of the hessian. Indeed in the EoS regime, we have $\lambda_1 \geq 0$.

Edge of Stability and Chaos. To build intuition for the relationship between our notion of RDS-sharpness of order 1 (Dfn. 3), the classical sharpness of Cohen et al. [1], and the onset of chaos, we consider the linearized framework in Ly and Gong [3], and examine GD in the EoS regime through its *sensitivity to initial conditions*. Let $\phi(K, w)$ denote the K-step GD cocycle (cf. (8)) and consider a reference trajectory $x_t = \phi(t, w_0) = \phi(1, w_{t-1})$. To track the propagation of an infinitesimal perturbation δw_t, we linearize the dynamics:

$$\delta w_{t+1} := \phi(1, w_t + \delta w_t) - \phi(1, w_t) \approx D\phi(1, w_t)\delta w_t,$$

where $D\phi(1, w_t) = I - \eta\nabla^2\widehat{\mathcal{R}}_S(w_t)$ is the *one-step Jacobian* (cf. Ex. 3.1). The local exponential growth rate in the direction $v_t = \delta w_t/\|\delta w_t\|$ is given by:

$$\log\left(\|\delta w_{t+1}\|/\|\delta w_t\|\right) = \log\|(I - \eta\nabla^2\widehat{\mathcal{R}}_S(w_t))v_t\|. \tag{9}$$

Let $\{(\alpha_i(t), u_i(t))\}_{i=1}^d$ denote the eigenpairs of the Hessian $\nabla^2\widehat{\mathcal{R}}_S(w_t)$. If the perturbation direction v_t is aligned with an eigenvector $u_i(t)$, the one-step growth factor is $|1 - \eta\alpha_i(t)|$. In the EoS regime, the maximum eigenvalue $\alpha_{\max}(t)$ typically exceeds the stability threshold $2/\eta$. When $\alpha_{\max}(t) > 2/\eta$, the growth factor satisfies $|1-\eta\alpha_{\max}(t)| > 1$, or equivalently, $\ln|1-\eta\alpha_{\max}(t)| > 0$. In the context of Ex. 3.1 this would make the singular value σ_1 in (7) larger than 0, hence we obtain $\lambda_1 > 0$ for this system.

If this condition holds on average along the trajectory, the top Lyapunov exponent $\Lambda_1 := \lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\ln\|D\phi(1, x_t)v_t\|$ becomes positive Arnold [55, Lemma 3.2.2, p. 113]. In dynamical systems, $\Lambda_1 > 0$ implies that trajectories diverge from each other exponentially (though the system might not be divergent), providing a sign of *deterministic chaos* within the EoS oscillations. Hence, we argue that EoS emerges when the system is chaotic with $\lambda_1 > 0$ and design theoretical tools for this challenging setup.

Existence of Random Attractor. Next, we characterize the existence of the random attractor.

Proposition 3.2 (Existence of the Random Pullback Attractor [56])**.** *Let $(\Omega, \mathcal{F}, \mathbb{P}, \theta, \phi)$ be a C^0 random dynamical system on $\mathbb{R}^d$. Suppose there exists a bounded random set $K(\omega)$ that is **pullback absorbing** for $\mathbb{P}$-almost every $\omega \in \Omega$; that is, for every deterministic bounded set $D \subset \mathbb{R}^d$, there exists a time $T(D, \omega) \geq 0$ such that $\phi(t, \theta^{-t}\omega, D) \subset K(\omega) \quad \forall t \geq T(D, \omega)$. Then, there exists a unique, compact and measurable [7] random attractor $\mathcal{A}(\omega)$ satisfying Dfn. 2.2.*

The existence of a pullback absorbing set $K(\omega)$ means that, for a fixed noise realization ω, all trajectories eventually enter a bounded region of the state space, provided the system evolves from the distant past. The random pullback attractor $\mathcal{A}(\omega)$ is the smallest invariant set inside this region. It contains exactly the states that can be reached asymptotically under the fixed noise realization ω. Indeed, this demonstrates that local instability, when coupled with global dissipativity, leads to the emergence of a compact pullback attractor. An intuitive explanation of this mechanism, for example in the case of neural networks using $\tanh$ as activation function, is given by Ahn et al. [45].

Complexity measure of random attractors. Given the random pullback attractor exists, we ask: *Can we quantify its complexity, particularly in the EoS regime, where we expect chaos?*

Definition 3.3 (Sharpness Dimension (SD)): Let $(\Omega, \mathcal{F}, \mathbb{P}, \theta, \phi)$ be a C^1 discrete random dynamical system on $\mathbb{R}^d$. Further let $\mathcal{A}(\omega) \subset \mathbb{R}^d$ be an almost surely compact random set. As in Dfn. 3, let λ_k be the RDS sharpness of order k, for $k = 1, \ldots, d$. Set

$$j^* := \max\left\{i \in \{1, \ldots, d\} : \sum_{k=1}^{i}\lambda_k \geq 0\right\}, \tag{10}$$

with the convention $j^* = 0$ if $\lambda_1 < 0$. The **Sharpness Dimension of** $\mathcal{A}(\omega)$ is defined as

$$\dim_{\mathrm{S}}\mathcal{A} := \begin{cases} j^* + \frac{\sum_{i=1}^{j^*}\lambda_i}{|\lambda_{j^*+1}|}, & \text{if } 1 \leq j^* < d, \\ d, & \text{if } j^* = d, \\ 0, & \text{if } \lambda_1 < 0. \end{cases} \tag{11}$$

Similar notions such as Kaplan-Yorke and Lyapunov dimensions have been considered in [57–59]. SD measures the effective number of expanding directions of the dynamics on the attractor before global contraction dominates. It is defined from the ordered global sharpness indices $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$, which quantify worst-case logarithmic stretching rates along principal directions. Let j^* be the largest index s.t. $\sum_{k=1}^{j^*}\lambda_k \geq 0$; then j^* is the maximal dimension volumes do not contract. In the case of a proper set, SD is strictly smaller than the ambient dimension in the EoS regime where $\lambda_1 > 0$.

Generalization Bounds. We are now ready to formulate our main theorem and establish the connection between SD and the worst-case generalization gap over the random attractor. We assume

[7] Measurable with respect to the **past σ-algebra** $\mathcal{F}_{-\infty}^0 := \sigma\{\theta^t\omega : t \leq 0\}$.

the *boundedness* (App. B.1) and *Lipschitz continuity of the loss* (App. B.2) as well as *regular random dynamics* (App. B.3). The first two of which are standard practice in the literature [14, 15, 17, 60]. We make these assumptions precise in App. B and provide the proof of our main theorem in App. C.1.

Theorem 3.4 (Generalization via Sharpness Dimension):
Let $S = \{z_1, \dots, z_n\} \sim \mu_z^{\otimes n}$ be a dataset of size n. Let $(\Omega, \mathcal{F}, \mathbb{P}, \theta, \phi)$ be a discrete-time RDS according to Dfn 2.1 such that Assump. B.3 holds. Under Assumps. B.1 and B.2, there exists a constant $C > 0$ s.t. with probability at least $1 - \zeta - \gamma$ over the joint draw $(S, \omega) \sim \mu_z^{\otimes n} \otimes \mathbb{P}$, there exists $\delta_{n,\gamma} > 0$ such that for all $0 < \delta < \delta_{n,\gamma}$,

$$\mathcal{G}_S(\mathcal{A}(\omega)) \leq 2L\delta + 2B\sqrt{\frac{4\ \dim_{\mathrm{S}} \mathcal{A}_S\ \log(1/\delta)}{n}} + \frac{I_\infty(\mathcal{A}_S(\omega), S) + \log(1/\zeta)}{\sqrt{n}} + \frac{CB^2}{\sqrt{n}}.$$

Recall that $\mathcal{G}_S(\mathcal{A}(\omega))$ denotes the worst-case generalization gap (Eq. 1) and $I_\infty(\mathcal{A}_S(\omega), S)$ (Dfn. A.7) denotes the total MI between the random pullback attractor $\mathcal{A}_S(\omega)$ and S.

Previous works establish general bounds for broad classes of random sets [16, 18, 60]. In contrast, we exploit the additional geometric structure of an attractor. The proof first covers the random attractor $\mathcal{A}_S(\omega)$ by small balls and uses a uniform Taylor expansion to enclose the image of each ball in a translated derivative ellipsoid plus a quadratic remainder. Crucially, the invariance property of the random attractor allows these local covering estimates to be iterated along the random dynamics. Ellipsoid-covering estimates, combined with ergodic and subadditive arguments, then convert the growth of the leading singular values and the contraction of the next singular value into the Sharpness Dimension bound. Finally, monotonicity extends the estimate from the constructed sequence of radii to arbitrary vanishing scales, yielding the almost-sure upper bound on the attractor's Minkowski dimension and hence a generalization bound.

Thm 3.4 rigorously links the generalization gap to the stability of training: at EoS ($\dim_{\mathrm{S}} \mathcal{A}_S < d$), the gap is governed by the global geometry of the attractor rather than by any isolated solution. In particular, the observed expansion along at least one direction with contraction in other directions implies that the attractor is confined to a proper subset of the ambient space, of dimension strictly smaller than d. SD, $\dim_{\mathrm{S}} \mathcal{A}_S$, quantifies this by capturing the spectral balance between the expanding and contracting directions.

The mutual information (MI) term I_∞ used to handle the dependence of the dataset S and the attractor $\mathcal{A}_S(\omega)$, can be avoided by using the recent 'set stability' framework of [18]. We do so and provide another generalization bound without the MI term in Thm. C.6.

4 Empirical Results

Numerical implementation. Unlike trajectory-based approaches [14, 15, 17], our generalization bound neither requires access to the training trajectory nor necessitates tools of topological data analysis. Instead, our computational bottleneck is in estimating the eigenvalue spectrum of the Hessian of a parameter near convergence (App. 5 for a formal approximation). For small-scale networks, this can be achieved via (randomized) GPU-parallelized singular value decomposition (SVD) [61]. Yet, the quadratic memory complexity of the Hessian renders such methods intractable as model size grows. This challenge is exacerbated by the fact that we require access not just to the leading eigenvalues, but to the entire spectrum, dominated by near-zero modes, precisely the regime where Krylov-methods such as Lanczos iterations [62] become ineffective. As a remedy, we adopt *stochastic Lanczos quadrature* (SLQ) [63, 64], a scalable spectral technique that has recently been applied to (moderately sized) neural networks [65, 66]. Unlike classical SLQ where all runs probe a fixed matrix–vector product (HVP), to estimate SD directly in expectation, we vary the operator across runs by computing HVPs on independently sampled minibatches. The operator is held fixed within each Lanczos run. See App. 5 for a complexity analysis.

Metrics. We assess the **correlation** between various notions of complexity and generalization error by using Kendall's coefficients (KC) [67] as well as their "granulated" versions (GC) [68]. While the classical KC (denoted τ) measures correlation between two quantities, it may fail to capture their causal relationship. Instead, one GC is defined for each hyperparameter (i.e., ψ_{LR} for η and ψ_{BS} for b); it measures correlation when only this hyperparameter is varying. Note that scaling the constant B in Thm. 3.4 does not affect the observed correlation between generalization and

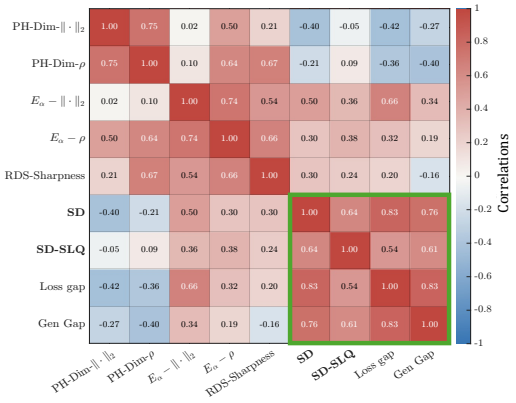

Figure 2: Correlations between various generalization indices and the empirical generalization gap on our small 3-layer MLP trained on the MNIST dataset. The region indicated in green shows that our proposed Sharpness Dimension (SD) better predicts generalization.

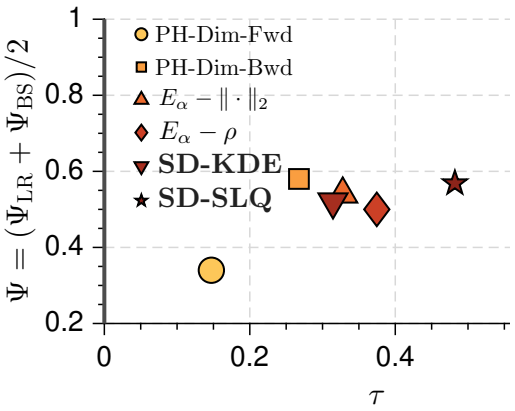

Figure 3: Kendall coefficients and their average granulated variant for the same 5-layer MLP trained using various batch sizes and learning rates. Our estimate further uses the SLQ approximation and its density-estimating variant, SD-KDE.

topological complexities. As target metrics, we compute the *generalization gap* (Gen. Gap) and *loss gap* as the absolute difference in accuracy and loss, respectively, both between training and test data.

Baseline notions of complexity. As a proxy to generalization, we use both Euclidean ($\|\cdot\|$) and data-dependent (ρ) persistent homology (PH) dimensions (PH-Dim) [15, 16], and α-weighted lifetime sums [17] (E_α) using a simulated trajectory of 5,000 iterations (PH-Dim-Fwd). We also consider the last 5,000 training iterations without trajectory simulation (PH-Dim-Bwd). We further include our *RDS sharpness* as λ_1, the leading SV of $I - \eta\nabla^2\widehat{\mathcal{R}}_S(w)$ and our SD. In grokking settings, we also track *Hessian trace* ($\text{tr}(H)$) and the magnitude of its largest eigenvalue (*sharpness*), obtained as the absolute value of the largest Hessian eigenvalue. When $\eta \cdot \text{sharpness} > 1$, the two quantities correspond; otherwise, RDS sharpness is determined by a non-positive Hessian eigenvalue.

4.1 Analysis

Generalization in MLPs. We begin by experimenting on classical learning scenarios with MLPs trained on MNIST across two scales. In both cases, we vary the learning rate (lr, η) and batch size (b) over a 5×5 grid of hyperparameters. In the small 3-layer setting, the full Hessian spectrum is tractable, allowing us to estimate the expectation in our Sharpness Dimension $\dim_{\text{S}} \mathcal{A}_S$ by sampling random minibatches, computing parameter Hessians, and taking the SVD of $I - \eta\nabla^2\widehat{\mathcal{R}}_S(w)$. The RDS sharpness values λ_k are then estimated as averages of the log SVs and SD is computed following Definition 3.3. This setting also allows a direct comparison between exact SD and its SLQ approximation. In the larger 5-layer setting, exact SVD is no longer feasible, so we report only the SLQ-based estimators SD-SLQ (histogram-based) and SD-KDE (Gaussian kernel-smoothed). The exact architectures, optimization settings, and estimator configuration is reported in App. D.

Fig. 2 and Fig. 3 show that this pattern is consistent across both MLP scales. In the 3-layer setting, both Euclidean and data-dependent PH-Dims fail to capture the generalization behavior, while the corresponding E_α quantities show only weak positive correlation with the generalization gap. The top RDS-Sharpness λ_1 alone also exhibits weak negative correlation, suggesting that the largest singular value by itself is not indicative of generalization in this setting. By contrast, both SD and its SLQ-based approximation achieve substantially stronger correlations, with full Hessian-spectrum SD performing best. In the larger 5-layer setting, backward PH-Dim, forward E_α's, SD-SLQ, and SD-KDE attain similar average GKC values, with backward PH-Dim performing best and SD-SLQ following closely; however, when comparing Kendall's τ over the full hyperparameter grid, SD-SLQ outperforms all competing indicators. Overall, these results suggest that trajectory-based topological indices become less reliable at the edge of stability, where $I - \eta\nabla^2\widehat{\mathcal{R}}_S(w)$ is not contractive on average, whereas SD and its scalable approximations more accurately capture generalization across scales.

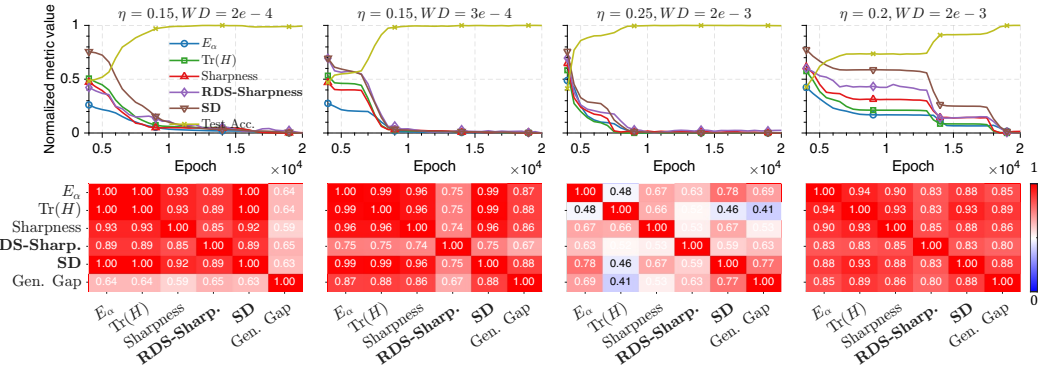

Figure 4: **Grokking analysis** for different LRs (η), weight decay (WD) and seeds across two architectures: (**top row**) 2-layer MLP. (**second row**) 2-layer MLP. Note that the *suddenness* of the grokking behavior is best captured in the complexity measures we introduce: RDS-Sharpness and Sharpness Dimension (SD).

Studying grokking. Next, we study the phenomenon of *grokking*, delayed and sudden generalization [12, 13, 69] through our EoS framework. We consider the task of *arithmetic modulo 97*, a family of supervised learning tasks where two one-hot encoded inputs representing integers $a, b < p$ are used to predict the target $y = a * b \mod p$. $*$ is a binary operation and p is a prime. In all our experiments, we use addition as the binary operation. The dataset size is defined as the percentage of the 97^2 possible input pairs used for training, with the remainder used for testing as in [69] and [12]. We use a 40%/60% train/test split.

To induce and observe grokking, we use a 2-layer MLP with 32 hidden features and GELU activation, trained using SGD with momentum and WD for 20k epochs. In all evaluations presented in Fig. 4, training acc. exceeds 90% and reaches 100% early, while test acc. continues to increase, a.k.a. *grok*. We report complexity measures for 100 uniformly spaced checkpoints as well as their intra-correlations and correlations with the Gen. Gap in Fig. 4.

Fig. 4 shows that correlations our SD achieves with the generalization gap are comparable to or higher than those of E_α, $\mathrm{tr}(H)$, and Sharpness, with our other measure, RDS sharpness, performing worse than SD for all the cases but the first. We observe that the Sharpness and $\mathrm{tr}(H)$ are less stable for certain hyperparameter choices, performing worse in one of the configurations. E_α shows stable correlations when using the training trajectories, whereas SD performs similarly or better even with only the model weights from the current epoch. These results indicate that SD captures information about training dynamics and the underlying attractor without requiring access to trajectory information like E_α. In addition, during the grokking phase transition occurs [70] simultaneously with changes in the structure of the Hessian spectrum.

Grokking for the 3-layer MLP. Fig 5 presents another grokking experiment, with identical seeds and hyper-parameters for a 3-layer MLP with 32 hidden features instead. We use 100 uniformly spaced checkpoints and ReLU activation, trained with SGD using only WD. In particular, we observe an interesting phenomenon in the first plot, where our dimension increases while test accuracy increases slowly and decreases sharply with the test accuracy (grokking), and then increases again afterwards when test accuracy reaches 100%. In contrast, the other measures decrease monotonically. This behaviour aligns with the theoretical motivation of our dimension, which targets the edge-of-stability regime, and accurately reflects the grokking phase transitions. A similar phase transition is visible in Plot 3, while Plot 4 reflects the grokking experiment in fig. 4. In Plot 2, focusing on the sharp grokking region reveals a similarly sharp decrease in SD.

Generalization in modern transformers: the case of GPT-2. To assess whether the same behavior persists at a larger scale, we consider GPT-2 [11] (124M parameters) trained from scratch on WikiText-2 [71]. We evaluate SGD and SGD with momentum on a 5×5 hyperparameter grid over LR and effective batch size, and AdamW [72] on a $4 \times 3 \times 3$ hyperparameter grid over LR, batch size, and weight decay. We report the correlations between the resulting complexity measures and the empirical loss gap and 0–1 loss gap. For this experiment we additionally report SD-PS, an equal-mass pseudo-spectrum estimator computed from the smoothed SLQ density; see App. D for details. Fig. 6 indicates that the qualitative conclusions remain consistent with our smaller-scale experiments, with

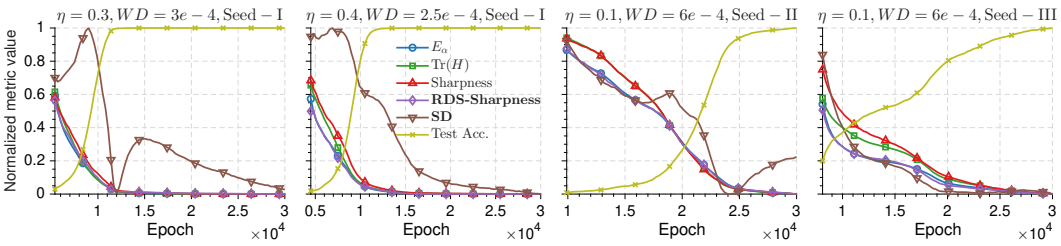

Figure 5: **Grokking analysis** for a 3-layer MLP with ReLU activation and no momentum. Note that the *suddenness* of the grokking behavior is best captured in the complexity measures we introduce: RDS-Sharpness and Sharpness Dimension (SD).

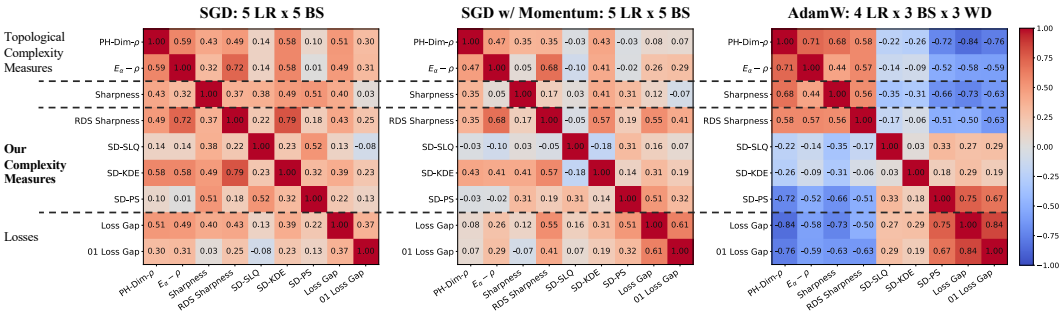

Figure 6: **Correlation Matrices Corresponding to GPT-2 Trained on WikiText2-Dataset** for different LRs, batch sizes (BS) and weight decay values (WD), across three different optimizers: SGD, SGD with momentum and AdamW.

the strongest distinction appearing for AdamW. In particular, classical sharpness does not provide a reliable indicator of generalization, and under AdamW, it shows a strong negative correlation with the loss gaps. By contrast, the RDS-based quantities remain informative overall, with the SD variants, especially the smoothed estimates SD-KDE and SD-PS, showing the most consistent positive correlations across optimizers. PH-Dim and E_α also provide useful signals in some settings, but their behavior is less uniform. Overall, these results suggest that, in this transformer setting as well, generalization is better captured by SD and related RDS-based quantities than by classical sharpness.

Runtime. The main additional cost of our method comes from Hessian-spectrum evaluation. The SLQ estimation takes about 5 minutes for 3-layer MNIST models, 1 hour for 5-layer MNIST models, and for GPT-2, the corresponding analysis takes about 3 hours per model on average. Further experimental and runtime details are reported in App. D.

5 Conclusion

Training neural networks at the edge of numerical stability challenges classical generalization theories that assume convergence to a single solution. In this work, we show that this regime is instead governed by the **geometry of a random pullback attractor** induced by stochastic optimization viewed as a random dynamical system. Building on this, we introduce the Sharpness Dimension (SD), a Lyapunov-inspired spectral complexity measure that captures the effective dimensionality of expanding directions on the attractor. We prove a worst-case generalization bound over the entire attractor, establishing SD as a principled quantity linking chaotic optimization dynamics to generalization. Empirically, SD reliably predicts generalization gaps across optimization regimes both in small and large networks, correlates with grokking phase transitions, and remains computable at scale via stochastic Lanczos quadrature.

Limitations & future work. Currently, we are required to estimate the full Hessian spectrum, which, despite scalable approximations, remains computationally demanding for very large models. We plan to study the impact of a larger class of approximations on the estimation quality in the future. Future work also involves tightening the theory under weaker regularity assumptions, extending it to adaptive optimizers. An important direction for future work is to characterize the Sharpness Dimension in specific settings. Its dependence on both the optimization algorithm and the data distribution provides a natural basis for theoretically analyzing how these factors determine the effective dimension.

Acknowledgments. The authors are grateful for the support of the Excellence Strategy of local and state governments. T. B. was supported by a UKRI Future Leaders Fellowship (MR/Y018818/1). The authors acknowledge support from the UK AI Research Resource (AIRR Isambard AI) through grant 0251-4584-0945-1 - TopoFound. U.Ş. is partially supported by the French government under the management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute). M.T and U.Ş. are partially supported by the European Research Council Starting Grant DYNASTY – 101039676.

References

[1] Jeremy M Cohen, Simran Kaur, Yuanzhi Li, J Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. *arXiv preprint arXiv:2103.00065*, 2021.

[2] Avrajit Ghosh, Bai Cong, Rio Yokota, Saiprasad Ravishankar, Rongrong Wang, Molei Tao, Mohammad Emtiyaz Khan, and Thomas Möllenhoff. Variational learning finds flatter solutions at the edge of stability. *arXiv preprint arXiv:2506.12903*, 2025.

[3] Andrew Ly and Pulin Gong. Optimization on multifractal loss landscapes explains a diverse range of geometrical and dynamical properties of deep learning. *Nature Communications*, 16 (1):3252, 2025.

[4] Dayal Singh Kalra, Tianyu He, and Maissam Barkeshli. Universal sharpness dynamics in neural network training: Fixed point analysis, edge of stability, and route to chaos. *arXiv e-prints*, pages arXiv–2311, 2023.

[5] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

[6] Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. *arXiv preprint arXiv:1912.02178*, 2019.

[7] Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *International Conference on Machine Learning*, pages 1019–1028. PMLR, 2017.

[8] Simran Kaur, Jeremy Cohen, and Zachary Chase Lipton. On the maximum hessian eigenvalue and generalization. In *Proceedings on*, pages 51–65. PMLR, 2023.

[9] Kaiyue Wen, Zhiyuan Li, and Tengyu Ma. Sharpness minimization algorithms do not only minimize sharpness to achieve better generalization. *Advances in Neural Information Processing Systems*, 36:1024–1035, 2023.

[10] Yuhang Cai, Haofeng Huang, Haodong Wen, Deyi Liu, Yiyuan Ma, and Kaifeng Lyu. Does LLM pre-training typically occur at the edge of stability? In *Workshop on Scientific Methods for Understanding Deep Learning*, 2026. URL `https://openreview.net/forum?id=QSb05IuPsy`.

[11] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[12] Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177*, 2022.

[13] Lucas Prieto, Melih Barsbey, Pedro A. M. Mediano, and Tolga Birdal. Grokking at the edge of numerical stability. In *The Thirteenth International Conference on Learning Representations*, 2025.

[14] Umut Simsekli, Ozan Sener, George Deligiannidis, and Murat A Erdogdu. Hausdorff dimension, heavy tails, and generalization in neural networks. *Advances in Neural Information Processing Systems*, 33:5138–5151, 2020.

[15] Tolga Birdal, Aaron Lou, Leonidas J Guibas, and Umut Simsekli. Intrinsic dimension, persistent homology and generalization in neural networks. *Advances in Neural Information Processing Systems*, 34:6776–6789, 2021.

[16] Benjamin Dupuis, George Deligiannidis, and Umut Simsekli. Generalization bounds using data-dependent fractal dimensions. In *International conference on machine learning*, pages 8922–8968. PMLR, 2023.

[17] Rayna Andreeva, Benjamin Dupuis, Rik Sarkar, Tolga Birdal, and Umut Simsekli. Topological generalization bounds for discrete-time stochastic optimization algorithms. *Advances in Neural Information Processing Systems*, 37, 2024.

[18] Mario Tuci, Lennart Bastian, Benjamin Dupuis, Nassir Navab, Tolga Birdal, and Umut Şimşekli. Mutual information free topological generalization bounds via stability. *arXiv preprint arXiv:2507.06775*, 2025.

[19] Alexander Camuto, George Deligiannidis, Murat A Erdogdu, Mert Gurbuzbalaban, Umut Simsekli, and Lingjiong Zhu. Fractal structure and generalization properties of stochastic optimization algorithms. *Advances in neural information processing systems*, 34:18774–18788, 2021.

[20] Sepp Hochreiter and Jürgen Schmidhuber. Simplifying neural nets by discovering flat minima. *Advances in neural information processing systems*, 7, 1994.

[21] Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring generalization in deep learning. *Advances in neural information processing systems*, 30, 2017.

[22] Levent Sagun, Utku Evci, V Ugur Guney, Yann Dauphin, and Leon Bottou. Empirical analysis of the hessian of over-parametrized neural networks. *arXiv preprint arXiv:1706.04454*, 2017.

[23] Zhewei Yao, Amir Gholami, Qi Lei, Kurt Keutzer, and Michael W Mahoney. Hessian-based analysis of large batch training and robustness to adversaries. *Advances in Neural Information Processing Systems*, 31, 2018.

[24] Pratik Chaudhari, Anna Choromanska, Stefano Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Jennifer Chayes, Levent Sagun, and Riccardo Zecchina. Entropy-sgd: Biasing gradient descent into wide valleys. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12): 124018, 2019.

[25] Umut Simsekli, Levent Sagun, and Mert Gurbuzbalaban. A tail-index analysis of stochastic gradient noise in deep neural networks. In *International Conference on Machine Learning*, pages 5827–5837. PMLR, 2019.

[26] Thanh Huy Nguyen, Umut Simsekli, Mert Gurbuzbalaban, and Gaël Richard. First exit time analysis of stochastic gradient descent under heavy-tailed gradient noise. *Advances in neural information processing systems*, 32, 2019.

[27] Rotem Mulayoff and Tomer Michaeli. Unique properties of flat minima in deep networks. In *International conference on machine learning*, pages 7108–7118. PMLR, 2020.

[28] Yusuke Tsuzuku, Issei Sato, and Masashi Sugiyama. Normalized flat minima: Exploring scale invariant definition of flat minima for neural networks using pac-bayesian analysis. In *International Conference on Machine Learning*, pages 9636–9647. PMLR, 2020.

[29] Kwangjun Ahn, Ali Jadbabaie, and Suvrit Sra. How to escape sharp minima with random perturbations. *arXiv preprint arXiv:2305.15659*, 2023.

[30] Stanislaw Jastrzebski, Maciej Szymczak, Stanislav Fort, Devansh Arpit, Jacek Tabor, Kyunghyun Cho, and Krzysztof Geras. The break-even point on optimization trajectories of deep neural networks. *arXiv preprint arXiv:2002.09572*, 2020.

[31] Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*, 2018.

[32] Dongxian Wu, Shu-Tao Xia, and Yisen Wang. Adversarial weight perturbation helps robust generalization. *Advances in neural information processing systems*, 33:2958–2969, 2020.

[33] Yaowei Zheng, Richong Zhang, and Yongyi Mao. Regularizing neural networks via adversarial model perturbation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8156–8165, 2021.

[34] Jean Kaddour, Linqing Liu, Ricardo Silva, and Matt J Kusner. When do flat minima optimizers work? *Advances in Neural Information Processing Systems*, 35:16577–16595, 2022.

[35] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*, 2020.

[36] Lijun Ding, Dmitriy Drusvyatskiy, Maryam Fazel, and Zaid Harchaoui. Flat minima generalize for low-rank matrix recovery. *Information and Inference: A Journal of the IMA*, 2024.

[37] Khashayar Gatmiry, Zhiyuan Li, Tengyu Ma, Sashank Reddi, Stefanie Jegelka, and Ching-Yao Chuang. What is the inductive bias of flatness regularization? a study of deep matrix factorization models. *Advances in Neural Information Processing Systems*, 36:28040–28052, 2023.

[38] Hong Liu, Sang Michael Xie, Zhiyuan Li, and Tengyu Ma. Same pre-training loss, better downstream: Implicit bias matters for language models. In *International Conference on Machine Learning*, pages 22188–22214. PMLR, 2023.

[39] Chao Ma and Lexing Ying. On linear stability of sgd and input-smoothness of neural networks. *Advances in Neural Information Processing Systems*, 34:16805–16817, 2021.

[40] Maxime Haddouche, Paul Viallard, Umut Simsekli, and Benjamin Guedj. A pac-bayesian link between generalisation and flat minima. *arXiv preprint arXiv:2402.08508*, 2024.

[41] Eugenio Clerico, Tyler Farghly, George Deligiannidis, Benjamin Guedj, and Arnaud Doucet. Generalisation under gradient descent via deterministic pac-bayes. *arXiv preprint arXiv:2209.02525*, 2022.

[42] Chen Xing, Devansh Arpit, Christos Tsirigotis, and Yoshua Bengio. A walk with sgd. *arXiv preprint arXiv:1802.08770*, 2018.

[43] Sanjeev Arora, Zhiyuan Li, and Abhishek Panigrahi. Understanding gradient descent on the edge of stability in deep learning. In *International Conference on Machine Learning*, pages 948–1024. PMLR, 2022.

[44] Kwangjun Ahn, Sébastien Bubeck, Sinho Chewi, Yin Tat Lee, Felipe Suarez, and Yi Zhang. Learning threshold neurons via edge of stability. *Advances in Neural Information Processing Systems*, 36:19540–19569, 2023.

[45] Kwangjun Ahn, Jingzhao Zhang, and Suvrit Sra. Understanding the unstable convergence of gradient descent. In *International conference on machine learning*. PMLR, 2022.

[46] Zixuan Wang, Zhouzi Li, and Jian Li. Analyzing sharpness along gd trajectory: Progressive sharpening and edge of stability. *Advances in Neural Information Processing Systems*, 35: 9983–9994, 2022.

[47] Lei Chen and Joan Bruna. Beyond the edge of stability via two-step gradient updates. In *International Conference on Machine Learning*, pages 4330–4391. PMLR, 2023.

[48] Alex Damian, Eshaan Nichani, and Jason D Lee. Self-stabilization: The implicit bias of gradient descent at the edge of stability. *arXiv preprint arXiv:2209.15594*, 2022.

[49] Xingyu Zhu, Zixuan Wang, Xiang Wang, Mo Zhou, and Rong Ge. Understanding edge-of-stability training dynamics with a minimalist example. *arXiv preprint arXiv:2210.03294*, 2022.

[50] Arseniy Andreyev and Pierfrancesco Beneventano. Edge of stochastic stability: Revisiting the edge of stability for sgd. *arXiv preprint arXiv:2412.20553*, 2024.

[51] Lei Wu and Weijie J Su. The implicit regularization of dynamical stability in stochastic gradient descent. In *International Conference on Machine Learning*, pages 37656–37684. PMLR, 2023.

[52] Michele Sasdelli, Thalaiyasingam Ajanthan, Tat-Jun Chin, and Gustavo Carneiro. A chaos theory approach to understand neural network optimization. In *2021 Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–10. IEEE, 2021.

[53] Dennis Chemnitz and Maximilian Engel. Characterizing dynamical stability of stochastic gradient descent in overparameterized learning. *Journal of Machine Learning Research*, 26 (134):1–46, 2025.

[54] Tongtong Liang, Alexander Cloninger, Rahul Parhi, and Yu-Xiang Wang. Generalization below the edge of stability: The role of data geometry. In *International Conference on Learning Representations*, volume 2026, pages 72969–73017, 2026.

[55] Ludwig Arnold. Random dynamical systems. In *Dynamical Systems: Lectures Given at the 2nd Session of the Centro Internazionale Matematico Estivo (CIME) held in Montecatini Terme, Italy, June 13–22, 1994*. Springer, 2006.

[56] Hans Crauel, Arnaud Debussche, and Franco Flandoli. Random attractors. *Journal of Dynamics and Differential Equations*, 9(2):307–341, 1997.

[57] James L Kaplan and James A Yorke. Chaotic behavior of multidimensional difference equations. In *Functional Differential Equations and Approximation of Fixed Points: Proceedings, Bonn, July 1978*. Springer, 2006.

[58] Brian R Hunt. Maximum local lyapunov dimension bounds the box dimension of chaotic attractors. *Nonlinearity*, 9(4):845, 1996.

[59] De-Jun Feng and Károly Simon. Dimension estimates for iterated function systems and repellers. part ii. *Ergodic Theory and Dynamical Systems*, 42(11):3357–3392, 2022.

[60] Benjamin Dupuis, Paul Viallard, George Deligiannidis, and Umut Simsekli. Uniform generalization bounds on data-dependent hypothesis sets via PAC-bayesian theory on random sets. *Journal of Machine Learning Research*, 25(409), 2024.

[61] Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53 (2):217–288, 2011.

[62] Cornelius Lanczos. An iteration method for the solution of the eigenvalue problem of linear differential and integral operators. *Journal of research of the National Bureau of Standards*, 45 (4):255–282, 1950.

[63] Lin Lin, Yousef Saad, and Chao Yang. Approximating spectral densities of large matrices. *SIAM review*, 58(1):34–65, 2016.

[64] Gene H Golub and John H Welsch. Calculation of gauss quadrature rules. *Mathematics of computation*, 23, 1969.

[65] Behrooz Ghorbani, Shankar Krishnan, and Ying Xiao. An investigation into neural net optimization via hessian eigenvalue density. In *International Conference on Machine Learning*, pages 2232–2241. PMLR, 2019.

[66] Vardan Papyan. The full spectrum of deepnet hessians at scale: Dynamics with sgd training and sample size. *arXiv preprint arXiv:1811.07062*, 2018.

[67] Maurice G. Kendall. A new reasure of rank correlation. *Biometrika*, 1938.

[68] Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic Generalization Measures and Where to Find Them. *ICLR 2020*, December 2019.

[69] Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. In *The Eleventh International Conference on Learning Representations*, 2023.

[70] Noa Rubin, Inbar Seroussi, and Zohar Ringel. Grokking as a first order phase transition in two layer networks. In *The Twelfth International Conference on Learning Representations*, 2024.

[71] Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models, 2016.

[72] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. URL `https://arxiv.org/abs/1711.05101`.

[73] Hans Crauel and Franco Flandoli. Attractors for random dynamical systems. *Probability Theory and Related Fields*, 100(3):365–393, 1994.

[74] Harald A Posch, William G Hoover, and Franz J Vesely. Canonical dynamics of the nosé oscillator: Stability, order, and chaos. *Physical review A*, 33(6):4253, 1986.

[75] Vladimir Bogachev. *Measure theory*. Springer, 2007.

[76] Tim Van Erven and Peter Harremos. Rényi divergence and kullback-leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.

[77] Liam Hodgkinson, Umut Simsekli, Rajiv Khanna, and Michael Mahoney. Generalization bounds using lower tail exponents in stochastic optimizers. In *International Conference on Machine Learning*, pages 8774–8795. PMLR, 2022.

[78] Dylan J Foster, Spencer Greenberg, Satyen Kale, Haipeng Luo, Mehryar Mohri, and Karthik Sridharan. Hypothesis set stability and generalization. *Advances in Neural Information Processing Systems*, 32, 2019.

[79] Ilya Molchanov. *Theory of Random Sets*. Number 87 in Probability Theory and Stochastic Modeling. Springer, second edition edition, 2017.

[80] David Yunis. The birkhoff ergodic theorem with applications. *The University of Chicago. Disponıvel em*, 2017.

[81] Yushun Zhang, Congliang Chen, Tian Ding, Ziniu Li, Ruoyu Sun, and Zhiquan Luo. Why transformers need adam: A hessian perspective. *Advances in neural information processing systems*, 37:131786–131823, 2024.

Appendix

We provide here the technical details and full proofs omitted from the main paper, along with supplementary experiments that further support our findings. The appendix is structured as follows:

- **Appendix A** introduces additional background material from random dynamical systems, random attractors, and worst-case generalization bounds.
- **Appendix B** presents the assumptions omitted from the main paper.
- **Appendix C** presents the complete proofs of all theoretical results stated in the main text, as well as several supplementary theoretical developments.
- **Appendix D** describes the experimental setups and implementation details required to reproduce our empirical results.
- **Appendix E** We provide additional empirical results.

A Theoretical Background

This section provides the theorems and technical background necessary for our main results.

A.1 Random Attractor

The notion of an *attractor* is one of the fundamental concepts in dynamical systems theory. While this concept has been central in the study of deterministic systems for several decades, its extension to stochastic systems is more subtle.

In stochastic systems (like SGD), the "noise" never stops kicking the particle. Therefore, the attractor is likely not to be a single static point. Instead, it is a *random set*, a moving target that fluctuates in shape and position over time, driven by the specific realization of the noise. The theory of attractors has been thoroughly developed over the past decades [55, 73]. Unlike the autonomous deterministic case, Random Dynamical Systems (RDS) are inherently non-autonomous and exhibit multiple notions of attraction, such as pullback and forward attraction. All definitions rely on the concept of a compact, invariant, random set.

Definition A.1 (**Compact and Invariant Random Sets**)**.** Let ϕ be an RDS on a Polish space X. A mapping $A : \Omega \to \mathcal{P}(X)$ is called a **compact random set** if:

1. **Measurability:** The indicator function $\omega \mapsto \mathbf{1}_{A(\omega)}(x)$ is measurable for each fixed $x \in X$.
2. **Compactness:** $A(\omega)$ is **compact** (i.e., closed and bounded) for $\mathbb{P}$-almost every $\omega \in \Omega$.

The set A is said to be an **invariant compact random set** if additionally it satisfies ϕ**-Invariance**:

$$\phi(t, \omega, A(\omega)) = A(\theta^t \omega), \quad \forall t \in \mathbb{T}^+, \quad \mathbb{P}\text{-a.s.}$$

Definition A.2 (**Modes of Random Attraction**)**.** Let (θ, ϕ) be an RDS on a Polish space X, and let A be an invariant compact random set according to Dfn. A.1. Let $\mathcal{S}$ be a collection of bounded subsets of X (e.g., bounded initializations). We denote $\operatorname{dist}(E, F) := \sup_{x \in E} \inf_{y \in F} d(x, y)$. A is called:

1. **A Random Pullback Attractor** if for all $D \in \mathcal{S}$:
$$\lim_{t \to \infty} \operatorname{dist}\left(\phi(t, \theta^{-t}\omega)D, A(\omega)\right) = 0 \quad \text{a.s.}$$
Intuition: This asks: "If I started training infinitely long ago ($t \to -\infty$) with noise $\theta^{-t}\omega$, where would I be *right now* ($t = 0$)?"
2. **A Random Forward Attractor** if for all $D \in \mathcal{S}$:
$$\lim_{t \to \infty} \operatorname{dist}\left(\phi(t, \omega)D, A(\theta^t \omega)\right) = 0 \quad \text{a.s.}$$
Intuition: This asks: "If I start training *now* ($t = 0$), where will I be in the distant future?"
SGD Context: This describes the physical trajectory of a specific training run.
3. **A Weak (Point) Attractor** if the convergence holds in probability rather than almost surely, (for singleton sets $D = \{y\}$).

Fig. 7 complements our definitions and results with an example. Indeed in the main write up we focus on the notion of random pullback attractor.

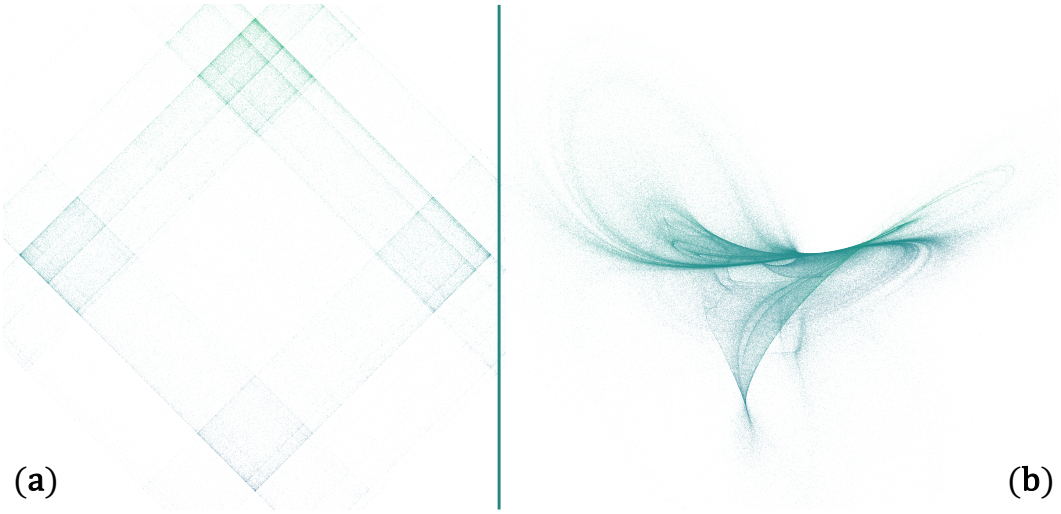

Figure 7: **(a) Fractal Pullback Attractor at the Edge of Stability.** Visualization of the random snapshot attractor $\mathcal{A}(\zeta)$ generated by Stochastic Gradient Descent (SGD) on a non-linear function $L(x) = \frac{1}{2}\|x\|^2 - A\prod_{i=1}^{2}\cos(kx_i)$ ($A = 2.0, k = 4.0$). The system is evolved for $T = 250$ steps with a learning rate $\eta = 0.15$ placing the dynamics at the edge of stability. The figure illustrates the collapse of 6×10^5 particles onto a fractal skeleton under a shared noise realization $\zeta = \{\xi_t\}_{t=1}^{T}$, where $\xi_t \sim \mathcal{N}(0, \sigma^2 I)$ with $\sigma = 0.1$. The intricate filaments emerge from the recursive stretching and folding of the state space, characteristic of chaotic synchronization in random dynamical systems (RDS). **(b) 3D chaotic attractor of an RDS.** Illustration of a chaotic random pullback attractor using a stochastic generalized Nosé-Hoover system [74].

A.2 Fractal Dimensions

We start by introducing the notion of a covering, which plays a central role throughout this paper and is especially important in the proofs of our main results.

Definition A.3 (Covering of a set)**.** Let $X \subseteq \mathbb{R}^d$ be a set. A δ-cover of X is a family $(U_i)_i$ of sets with diameter at most δ such that

$$X \subseteq \bigcup_i U_i.$$

If X is bounded, a δ-cover of minimal cardinality is called a *minimal cover*, and its cardinality is called the *covering number*, denoted by $\mathcal{N}(X, \delta)$.

Based on Dfn. A.3, we are now able to define the upper Minkowski dimension, which will later be bounded using our new notion of fractal dimension.

Definition A.4 (Minkowski dimensions)**.** Let $X \subset \mathbb{R}^d$ be a bounded set. The lower and upper Minkowski (or box) dimensions of X are defined as

$$\underline{\dim}_{\mathrm{B}}(X) := \liminf_{\delta \to 0} \frac{\log \mathcal{N}(X, \delta)}{\log(1/\delta)}, \qquad \overline{\dim}_{\mathrm{B}}(X) := \limsup_{\delta \to 0} \frac{\log \mathcal{N}(X, \delta)}{\log(1/\delta)}.$$

If the two quantities agree, the resulting limit is called the *Minkowski dimension* of X, written $\dim_{\mathrm{B}}(X)$.

A.3 Data-dependent worst-case generalization bounds

Before presenting the proofs of our main results, we state the following technical assumption, lemmata and preperation which are necessray to formulate our main result.

Recent advances in topological generalization bounds rely on data-dependent worst-case generalization bounds, leveraging PAC-Bayesian theory on random sets [60] or the stability-based framework of Tuci et al. [18].

Theorem A.5 (Egoroff's theorem)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let f and $(f_n)_{n\in\mathbb{N}}$ be measurable functions on Ω. Suppose that*

$$f_n(x) \longrightarrow f(x) \quad \textit{for almost all } x \in \Omega.$$

Then, for any $\varepsilon > 0$, there exists a measurable set $\Omega_\varepsilon \in \mathcal{F}$ such that

$$\mathbb{P}(\Omega_\varepsilon) \geq 1 - \varepsilon,$$

and the convergence of f_n to f is uniform on Ω_ε.

This theorem allows us to relate the notion of covering (see Dfn. A.3) to the Minkowski dimension (see Dfn. A.4). Similar ideas have been used in previous work [14, 16, 60].

Corollary A.6 (Uniform control of covering numbers)**.** *Let $(\Omega_U, \mathcal{F}_U, \mathbb{P}_U)$ be the probability space supporting the random variable U, and let $(\mathcal{Z}, \mathcal{F})$ denote the data space. Assume that the random set $\mathcal{W}_{S,U}$ is almost surely of finite diameter. Then, for any $\gamma > 0$, there exists a measurable set $\Omega_\gamma \subseteq \mathcal{F}^{\otimes n} \otimes \mathcal{F}_U$ with*

$$\mu^{\otimes n} \otimes \mathbb{P}_U(\Omega_\gamma) \geq 1 - \gamma,$$

such that, on Ω_γ, the following holds: there exists $\delta_{n,\gamma} > 0$ for which, for all $0 < \delta < \delta_{n,\gamma}$,

$$\log\big|\mathcal{N}(\mathcal{W}_{S,U}, \delta)\big| \;\leq\; 2\,\overline{\dim}_{\mathrm{B}}(\mathcal{W}_{S,U})\,\log\big(\tfrac{1}{\delta}\big)\,.$$

Proof. Let $(\Omega_U, \mathcal{F}_U, \mathbb{P}_U)$ denote the probability space supporting the auxiliary random variable U, and let $\mathcal{F}$ denote the σ-algebra associated with the data space $\mathcal{Z}$. Assume that the random set $\mathcal{W}_{S,U}$ is almost surely of finite diameter.

By definition of the upper Minkowski (box-counting) dimension, we have $\mu^{\otimes n} \otimes \mathbb{P}_U$-almost surely

$$\dim_{\mathrm{B}}(\mathcal{W}_{S,U}) := \limsup_{\delta \to 0} \frac{\log\big|\mathcal{N}(\mathcal{W}_{S,U}, \delta)\big|}{\log(1/\delta)} = \lim_{\delta \to 0} f_\delta(\mathcal{W}_{S,U}),$$

where we define

$$f_\delta(\mathcal{W}_{S,U}) := \sup_{0 < r < \delta} \frac{\log\big|\mathcal{N}(\mathcal{W}_{S,U}, r)\big|}{\log(1/r)}.$$

Let $(\delta_k)_{k \geq 0}$ be a decreasing sequence of positive numbers in $(0, 1)$ such that $\delta_k \to 0$, and fix $\gamma > 0$. By Egoroff's theorem (see 75), there exists a measurable set

$$\Omega_\gamma \in \mathcal{F}^{\otimes n} \otimes \mathcal{F}_U$$

such that

$$\mu^{\otimes n} \otimes \mathbb{P}_U(\Omega_\gamma) \geq 1 - \gamma,$$

and such that the convergence

$$f_{\delta_k}(\mathcal{W}_{S,U}) \longrightarrow \dim_{\mathrm{B}}(\mathcal{W}_{S,U})$$

is uniform on Ω_γ.

Consequently, there exists an index $k_{\gamma,n} \geq 0$ such that for all $k \geq k_{\gamma,n}$ and all $(S, U) \in \Omega_\gamma$,

$$f_{\delta_k}(\mathcal{W}_{S,U}) \leq 2\ \dim_{\mathrm{B}}(\mathcal{W}_{S,U}).$$

Therefore, for any $0 < \delta < \delta_k$ and any $(S, U) \in \Omega_\gamma$, we obtain

$$\log\big|\mathcal{N}(\mathcal{W}_{S,U}, \delta)\big| \leq 2\ \dim_{\mathrm{B}}(\mathcal{W}_{S,U})\ \log\big(\tfrac{1}{\delta}\big)\,,$$

which concludes the proof. □

The framework of Dupuis et al. [60] is based on information-theoretic quantities. In particular, we provide below a precise definition of the total mutual information term appearing in our main theoretical results; see [76, 77] for further background.

Definition A.7 (Total mutual information)**.** Let X and Y be two random elements defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ (note that the codomains of X and Y may be distinct). We define the total mutual information between X and Y by

$$I_\infty(X, Y) = \log \sup_{A \in \mathcal{A}_{X,Y}} \frac{\mathbb{P}_{X,Y}(A)}{\mathbb{P}_X \otimes \mathbb{P}_Y(A)}.$$

Generalization Bounds via Mutual Information. Many studies have employed information-theoretic techniques, particularly within the "fractal-based" literature [14, 15]. A unifying perspective recently emerged with the PAC-Bayesian theory for random sets [60], which was recently employed by Andreeva et al. [17] to establish generalization bounds based on novel topological complexity measures. Informally, all these bounds are of the following form[8]:

$$\sup_{w\in\mathcal{W}_{S,U}}\big(\mathcal{R}(w)-\widehat{\mathcal{R}}_S(w)\big)\lesssim\sqrt{\frac{\mathbf{C}(\mathcal{W}_{S,U})+\mathrm{IT}+\log(1/\zeta)}{n}},\tag{12}$$

with probability at least $1-\zeta$. The term IT is an *information-theoretic* (IT) term, typically the *total* mutual information between the dataset S and the set $\mathcal{W}_{S,U}$ The aforementioned bounds differ in the choice of complexity measure $\mathbf{C}(\mathcal{W}_{S,U})$, but all include an IT term.

By adapting Corollary 33 of Dupuis et al. [60] to our framework, we derive the following Corollary

Corollary A.8. *Assume that $\ell(w,z)$ is L-Lipschitz in w, bounded, and that $\mathcal{W}_{S,U}$ is almost surely bounded. Then there exists a constant $C>0$ such that, for any $\lambda,\delta>0$, with probability at least $1-\zeta$ under the joint law of (S,U), we have*

$$\sup_{w\in\mathcal{W}_{S,U}}\Big(\mathcal{R}(w)-\widehat{\mathcal{R}}_S(w)\Big)\le 2L\delta+2B\sqrt{\frac{2\log|\mathcal{N}(\mathcal{W}_{S,U},\delta)|}{n}}+\frac{I_\infty(\mathcal{W}_{S,U},S)+\log\frac{1}{\zeta}}{\lambda}+C\lambda\frac{B^2}{n}$$

By combining Cor. A.6 and Cor. A.8 and applying a union bound, we can now state the following theorem, which serves as a key foundation for our subsequent analysis.

Corollary A.9. *Assume that the loss $\ell(w,z)$ is L-Lipschitz in w, bounded by B, and that the random set $\mathcal{W}_{S,U}$ is almost surely of finite diameter. Then there exists a constant $C>0$ such that, for any $\lambda>0$, with probability at least $1-\zeta-\gamma$ under the joint law of (S,U), there exists $\delta_{n,\gamma}>0$ such that for all $0<\delta<\delta_{n,\gamma}$,*

$$\sup_{w\in\mathcal{W}_{S,U}}\big(\mathcal{R}(w)-\widehat{\mathcal{R}}_S(w)\big)\le 2L\delta+2B\sqrt{\frac{4\,\overline{\dim}_{\mathrm{B}}(\mathcal{W}_{S,U})\,\log(1/\delta)}{n}}+\frac{I_\infty(\mathcal{W}_{S,U},S)+\log(1/\zeta)}{\lambda}+\frac{C\lambda B^2}{n}.\tag{13}$$

Remark A.10. The parameter δ_n appearing in Theorem A.13 deserves a brief comment. As can be seen from the proof, it quantifies the uniformity in n of the limit defining the upper box-counting dimension: if this convergence is uniform in n, then δ becomes independent of n. A similar parameter already arises in [16], and [60] shows that δ can indeed be taken independent of n under a suitable convergence assumption on the distributions of the random sets. We refer the reader to these works for further details.

Generalization Bounds via Random Set Stability. A different perspective was taken by Tuci et al. [18], who showed that various 'topological generalization' bounds can be recovered within a novel stability framework. In our case, we make use of both frameworks. Before stating our main bound in terms of the stability parameter, we recall the following assumption on the random set, denoted by $\mathcal{W}_{S,U}$, which, in our setting, is the random attractor.

Definition A.11. A data-dependent selection of $\mathcal{W}_{S,U}$ is a deterministic mapping $\omega:\mathrm{CL}(\mathbb{R}^d)\times\mathcal{Z}^n\to\mathbb{R}^d$ such that $\omega(\mathcal{W}_{S,U},S')\in\mathcal{W}_{S,U}$, almost surely. In particular, we assume the existence of a random variable $\omega_0(\mathcal{W}_{S,U},S')$ such that, almost surely, $\omega_0(\mathcal{W}_{S,U},S')\in\arg\max_{w\in\mathcal{W}_{S,U}}G_{S'}(w)$.

Assumption A.12 (Random set stability by Tuci et al. [18])**.** $\mathcal{W}_{S,U}$ is β_n-random set stable if for any $J\in\mathbb{N}^\star$ and any data-dependent selection ω of $\mathcal{W}_{S,U}$, there exists a map $\omega':\mathrm{CL}(\mathbb{R}^d)\times\mathbb{R}^d\to\mathbb{R}^d$ such that:

- For any S, U and $w\in\mathbb{R}^d$, $\omega'(\mathcal{W}_{S,U},w)\in\mathcal{W}_{S,U}$.
- For all $z\in\mathcal{Z}$ and two datasets $S,S'\in\mathcal{Z}^n$ differing by J elements we have:

$$\mathbb{E}_U[|\ell(\omega(\mathcal{W}_{S,U},S),z)-\ell(\omega'(\mathcal{W}_{S',U},\omega(\mathcal{W}_{S,U},S)),z)|]\le\beta_nJ.$$

[8] We use the notation $\lesssim$ in informal statements where absolute constants have been omitted.

As noted by Tuci et al. [18], in the absence of algorithmic randomness (that is, when U is constant), Assump. A.12 reduces to a special case of the celebrated random set stability notion introduced by Foster et al. [78]. Stability assumptions are typically formulated for neighboring datasets; here, we instead consider a variant in which datasets differ by J elements. The two formulations are equivalent, and we adopt this version to streamline subsequent proofs and simplify notation.

Theorem A.13 (Tuci et al. [18] Theorem 4.3.)**.** *Suppose that the loss ℓ is bounded by B, L-Lipschitz and Assump. A.12 hold, and that $\mathcal{W}_{S,U}$ is a.s. of finite diameter. Without loss of generality, assume that $\beta_n^{-2/3}$ is an integer divisor of n. There exists $\delta_n > 0$ such that for all $\delta < \delta_n$*

$$\mathbb{E}\Bigg[\sup_{w\in\mathcal{W}_{S,U}}\Big(\mathcal{R}(w)-\widehat{\mathcal{R}}_S(w)\Big)\Bigg] \le 2\mathbb{E}\Bigg[\frac{B}{n}+\delta L+\beta_n^{1/3}\left(1+B\sqrt{4\overline{\dim}_{\mathrm{B}}(\mathcal{W}_{S,U})\log\frac{1}{\delta}}\right)\Bigg],$$

where $\overline{\dim}_{\mathrm{B}}(\mathcal{W}_{S,U})$ is the upper box-counting dimension (see Dfn. A.4)

Indeed, we present a simplified version of the theorem here. Since we assume that the loss ℓ is globally L-Lipschitz, this constant appears directly in the bound. It would in fact suffice to assume Lipschitz continuity only on the random set itself.

We now provide formal statements for the main assumptions in our paper.

B Assumptions

Assumption B.1 (Boundedness of Loss)**.** We assume the loss function $\ell : \mathbb{R}^d \times \mathcal{Z} \to \mathbb{R}$ to be bounded. That is, there exists a constant $B > 0$ such that for all weights $w \in \mathbb{R}^d$ and $z \in \mathcal{Z}$ holds: $0 \le \ell(w, \omega) \le B$.

Assumption B.2 (Lipschitz Continuity of Loss)**.** We assume that the loss function $\ell : \mathbb{R}^d \times \mathcal{Z} \to \mathbb{R}$ satisfies the following properties for all $z \in \mathcal{Z}$: The function $w \mapsto \ell(w, z)$ is L-Lipschitz continuous for some $L > 0$. That is, for all $w_1, w_2 \in \mathbb{R}^d$: $|\ell(w_1, z) - \ell(w_2, z)| \le L\|w_1 - w_2\|$.

Assumption B.3 (Regular random dynamics)**.** For each $S \in \mathcal{Z}^n$, let $(\Omega, \mathcal{F}, \mathbb{P}, \theta, \phi_S)$ be a C^2 discrete-time RDS with unique compact random pullback attractor $\mathcal{A}_S(\omega)$. Assume:

1. $\inf_{x\in\mathcal{A}_S(\omega)} \sigma_d(D\phi_S(1, \omega, x)) > 0$ for $\mathbb{P}$-a.e. ω;
2. $\mathbb{E}[\sup_{x\in\mathcal{A}_S(\omega)} \log \|D\phi_S(1, \omega, x)\|] < \infty$ and $\mathbb{E}[\sup_{x\in\mathcal{A}_S(\omega)} \log \|D^2\phi_S(1, \omega, x)\|] < \infty$;
3. There exists $j^* \in \{1, \dots, d-1\}$ such that $\sum_{i=1}^{j^*} \lambda_i \ge 0 > \sum_{i=1}^{j^*+1} \lambda_i$;
4. Define for $A \in \mathbb{R}^{d\times d}$: $\|A\|_j := \prod_{i=1}^{j} \sigma_i(A)$ and assume

$$\lim_{m\to\infty} \frac{1}{m}\mathbb{E}\left[\sup_{x\in\mathcal{A}_S(\omega)} \log \|D\phi_S(m, \omega, x)\|_j - \inf_{x\in\mathcal{A}_S(\omega)} \log \|D\phi_S(m, \omega, x)\|_j\right] = 0$$

   for every $j \in \{1, \dots, d\}$.

Intuitively, the j-volume growth rates along different orbits in $\mathcal{A}(\omega)$ may differ at sub-exponential order, but coincide to leading exponential order as $m \to \infty$. Conditions of this type, requiring the spatial variation of the cocycle to be subexponential, are commonly imposed in the smooth ergodic theory of random dynamical systems see, e.g., Arnold [55] for related formulations.

Discussion about the assumptions. Proposition 3.2 guarantees a compact random attractor whenever a bounded pullback-absorbing set exists. This is weaker than assuming that the dynamics are globally contractive. Several standard mechanisms provide a bounded absorbing set. Bounded activations are one sufficient condition, as shown in [45].Global dissipativity is another sufficient condition and is assumed in prior work [19, 77]. Although several sufficient conditions for existence have been established in prior work, verifying them for the large-scale architectures used in our experiments remains theoretically challenging.

Regarding Assumption B.3.3, contraction along the stable directions dominates expansion along the unstable directions, causing d-dimensional volume elements to contract under the training dynamics.

We verify this condition empirically: the spectra reported in Appendix Figure 8 have strictly negative sums, consistent with the corresponding SD values being strictly smaller than d. We conjecture that this condition is characteristic of training regimes in which the loss decreases over time. Our intuition is that sustained expansion across all principal directions would cause nearby trajectories to diverge, making the observed decrease in loss unlikely. This argument is only heuristic, however, and we do not currently provide a theoretical result establishing the assumption.

About the role of the bounded-distortion assumption (Assump. B.3.4). The assumption requires the ratio between the maximal and minimal j-dimensional volume expansions across the random attractor to grow subexponentially in time, in the sense that its expected logarithm is $o(m)$. It does not require this ratio to remain uniformly bounded.

Moreover, the assumption is not needed to get a less 'clean' result at the cost of an error term in Thm.C.4. To see this, we refer the reader to the argument leading to Eq. (60). More precisely, the argument would change as follows: The compactness of the random attractor and the nonsingularity of the derivative ensure that the maximum one-step volume expansion is finite and the minimum one is positive. The chain rule and the invariance of the random attractor allow us to control the accumulated distortion over several steps by the product of the corresponding one-step distortions. Submultiplicativity of the maximal volume expansion, supermultiplicativity of the minimal volume expansion, and Birkhoff's theorem then yield a finite asymptotic residual distortion rate.

At the transition index, this residual weakens the asymptotic contraction gap by the same amount. If the residual distortion is strictly smaller than the original contraction gap, then the first contracting direction remains exponentially contracting, and the covering radii converge to zero. Therefore, the remainder of the recursive covering argument remains unchanged; however, the resulting dimension bound contains an additional error, which is controlled by dividing the residual by the remaining contraction gap. Exact bounded distortion corresponds to a vanishing residual, which recovers the original bound.

We will next provide the proofs omitted from the main paper.

C Omitted Proofs

Before we present our main theorem. We recall some definition and basic facts, which will be used in the proof and improve readability.

C.1 Preliminaries

Important Sets

Definition C.1 (Minkowski Sum)**.** Let U and V be two non-empty subsets of a vector space (in this context, $\mathbb{R}^d$). The Minkowski sum, denoted by $U \oplus V$, is the set formed by adding every element of U to every element of V:

$$U \oplus V := \{u + v \mid u \in U, v \in V\}$$

In the context of the proof, the **Minkowski sum** provides a rigorous way to define the "thickening" of a transformed set to account for approximation errors. Let $U \subset \mathbb{R}^d$ be a compact set and let $B(0, \epsilon)$ be the closed ball of radius $\epsilon > 0$ centered at the origin. The Minkowski sum $U \oplus B(0, \epsilon)$ corresponds exactly to the **closed ϵ-neighborhood** of U:

$$U \oplus B(0, \epsilon) = \{y \in \mathbb{R}^d \mid \exists u \in U, \|y - u\| \leq \epsilon\} \tag{14}$$

Equivalently, this can be expressed using the distance function $\text{dist}(y, U) = \inf_{u \in U} \|y - u\|$:

$$U \oplus B(0, \epsilon) = \{y \in \mathbb{R}^d \mid \text{dist}(y, U) \leq \epsilon\} \tag{15}$$

Geometric Interpretation: Ellipsoids and Linear Images of Balls

Let $L \in \mathbb{R}^{d \times d}$ be a real matrix of rank

$$k := \text{rank}(L) \leq d.$$

We study the geometry of the image of the unit ball

$$B(0,1) := \{\xi \in \mathbb{R}^d : \|\xi\|_2 \le 1\}$$

under the linear map L.

Define

$$E := L(B(0,1)) = \{L\xi \in \mathbb{R}^d : \|\xi\|_2 \le 1\}.$$

Then E is an ellipsoid if L is invertible, and a *degenerate ellipsoid* (i.e. a flattened ellipsoid lying in a lower-dimensional subspace) if $\mathrm{rank}(L) < d$.

Canonical Form of the Ellipsoid. By the Singular Value Decomposition, there exist orthonormal bases

$$\{u_1, \dots, u_d\} \subset \mathbb{R}^d, \qquad \{v_1, \dots, v_d\} \subset \mathbb{R}^d$$

and singular values

$$\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_k > 0, \qquad \sigma_{k+1} = \cdots = \sigma_d = 0$$

such that

$$L = \sum_{i=1}^{k} \sigma_i \, v_i u_i^T.$$

Equivalently,

$$Lu_i = \sigma_i v_i \quad \text{for } i = 1, \dots, k, \qquad Lu_i = 0 \quad \text{for } i = k+1, \dots, d.$$

The singular values satisfy

$$\sigma_i = \sqrt{\lambda_i(L^T L)},$$

where $\lambda_i(L^T L)$ are the eigenvalues of $L^T L$ ordered in decreasing order. Since $\{u_1, \dots u_d\}$ an orthonormal basis we rewrite $\xi \in \mathbb{R}^d$ with $\|\xi\| \le 1$ as

$$\xi = \sum_{i=1}^{d} a_i u_i, \qquad \sum_{i=1}^{d} a_i^2 \le 1.$$

Then

$$L\xi = \sum_{i=1}^{k} \sigma_i a_i v_i.$$

Therefore for $t_i := \sigma_i a_i$,

$$E = \left\{ \sum_{i=1}^{k} t_i v_i \;\middle|\; \sum_{i=1}^{k} \frac{t_i^2}{\sigma_i^2} \le 1 \right\}.$$

Geometrically, E lies in the k-dimensional subspace

$$\mathrm{Im}(L) = \mathrm{span}\{v_1, \dots, v_k\},$$

its principal axes are the directions $v_1, \dots, v_k$, and the length of the i-th semi-axis is σ_i.

Image of a Ball of Radius ρ. Let $\rho > 0$. Since

$$B(0,\rho) = \rho \, B(0,1),$$

by linearity we have

$$L(B(0,\rho)) = \rho \, L(B(0,1)) = \rho \, E.$$

Thus the scaled ellipsoid is

$$\rho E = \left\{ \sum_{i=1}^{k} t_i v_i \;\middle|\; \sum_{i=1}^{k} \frac{t_i^2}{(\rho\sigma_i)^2} \le 1 \right\},$$

with principal semi-axis lengths

$$\rho\sigma_1, \ \rho\sigma_2, \ \dots, \ \rho\sigma_k.$$

Image of a Shifted Ball.. For any $x \in \mathbb{R}^d$,

$$B(x,\rho) = \{x\} \oplus B(0,\rho),$$

and therefore

$$L(B(x,\rho)) = \{Lx\} \oplus \rho E.$$

Covering Estimates for Ellipsoid.

Lemma C.2 (Ellipsoid covering)**.** *Let $E \subset \mathbb{R}^d$ be an ellipsoid with semi-axes*

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_d > 0.$$

Let $\rho > 0$, and let $j \in \{0, 1, \ldots, d\}$ be such that

$$\sigma_{j+1} \leq \rho \quad (\textit{with } \sigma_{d+1} := 0).$$

Then the minimal number of Euclidean balls of radius ρ needed to cover E satisfies

$$\mathcal{N}(E, \rho) \;\leq\; 3^d \prod_{i=1}^{j} \frac{\sigma_i}{\rho}.$$

Proof. Up to translation and rotation, we may assume that

$$E = \left\{ x \in \mathbb{R}^d : \sum_{i=1}^{d} \frac{x_i^2}{\sigma_i^2} \leq 1 \right\}.$$

Then E is contained in the axis-aligned box

$$Q := \prod_{i=1}^{d} [-\sigma_i, \sigma_i].$$

Hence

$$\mathcal{N}(E, \rho) \leq \mathcal{N}(Q, \rho).$$

We cover Q by a grid of cubes of side length ρ. Along the i-th coordinate direction, the interval $[-\sigma_i, \sigma_i]$ can be covered by at most

$$\left\lceil \frac{2\sigma_i}{\rho} \right\rceil \;\leq\; 1 + \frac{2\sigma_i}{\rho}$$

intervals of length ρ. Therefore,

$$\mathcal{N}(Q, \rho) \leq \prod_{i=1}^{d} \left(1 + \frac{2\sigma_i}{\rho} \right).$$

Now assume that $\sigma_{j+1} \leq \rho$.

Step 1: Small axes. For $i \geq j + 1$, since $\sigma_i \leq \rho$, we have

$$1 + \frac{2\sigma_i}{\rho} \leq 1 + 2 = 3.$$

Step 2: Large axes. For $i \leq j$, since $\sigma_i \geq \rho$, we have

$$1 + \frac{2\sigma_i}{\rho} \;\leq\; \frac{3\sigma_i}{\rho}.$$

Indeed, this is equivalent to $1 \leq \sigma_i/\rho$, which holds.

Step 3: Combine. Therefore,

$$\mathcal{N}(E, \rho) \;\leq\; \prod_{i=1}^{j} \frac{3\sigma_i}{\rho} \;\cdot\; \prod_{i=j+1}^{d} 3 \;=\; 3^d \prod_{i=1}^{j} \frac{\sigma_i}{\rho}.$$

This completes the proof. □

Covering for Minkowski Sum.

Lemma C.3 (Covering of Minkowski sums)**.** *Let $X, Y \subset \mathbb{R}^d$ be bounded sets and let $\varepsilon, \delta > 0$. Assume that*

$$X \subset \bigcup_{i=1}^{N} B(x_i, \varepsilon), \qquad Y \subset B(0, \delta),$$

where $N \in \mathbb{N}$ and $x_i \in \mathbb{R}^d$ for all $i \in \{1, \dots, N\}$. Then the Minkowski sum $A \oplus B$ satisfies

$$X \oplus Y \subset \bigcup_{i=1}^{N} B(x_i, \varepsilon + \delta).$$

In particular, the covering numbers satisfy

$$\mathcal{N}(X \oplus Y, \varepsilon + \delta) \leq \mathcal{N}(X, \varepsilon).$$

Proof. Let $a \in X \oplus Y$. Then $a = x + y$ with $x \in X$ and $y \in Y$. By assumption, there exists $i \in \{1, \dots, N\}$ such that $x \in B(x_i, \varepsilon)$, hence $\|x - x_i\| \leq \varepsilon$. Moreover, since $y \in B(0, \delta)$, we have $\|y\| \leq \delta$. Therefore,

$$\|a - x_i\| = \|x + y - x_i\| \leq \|x - x_i\| + \|y\| \leq \varepsilon + \delta,$$

which shows that $a \in B(x_i, \varepsilon + \delta)$. This proves the claim. □

We will now prove the main result of our work. We begin by bounding the Minkowski dimension (see Dfn. A.4)

Theorem C.4 (Minkowski Dimension Bound)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P}, \theta, \phi)$ be a discrete-time random dynamical system according to Dfn. 2.1. Suppose the assumption B.1, B.2 and B.3hold:*

Then, for $\mathbb{P}$-almost every $\omega \in \Omega$, the upper Minkowski dimension *of the random pullback attractor $\mathcal{A}(\omega)$ is bounded by the Sharpness Dimension* $\dim_{\mathrm{S}} \mathcal{A}$*:*

$$\overline{\dim}_M(\mathcal{A}(\omega)) \leq \dim_{\mathrm{S}} \mathcal{A}.$$

Proof. The main idea behind the proof is the fact that the sets $\mathcal{A}(\omega)$ and $\mathcal{A}(\theta^K \omega)$ have the same distribution for any K. Hence, we will estimate the covering number of $\mathcal{A}(\theta^K \omega)$ which will enable us to link the covering number to the singular values of $D\phi$.

Step 1: General Covering. Let $(\Omega, \mathcal{F}, \mathbb{P}, \phi)$ be a C^2 random dynamical system. We begin by covering the set $\mathcal{A}(\omega)$, which is almost surely compact and, therefore, bounded. Hence, for any $R > 0$ we have a finite integer $N_1(\omega, R)$ such that

$$\mathcal{A}(\omega) \subset \bigcup_{i=1}^{N_1(\omega, R)} B(x_i(\omega), R), \tag{16}$$

where $\{x_1(\omega), \dots, x_{N_1}(\omega)\}$ denote the centers of a finite covering of $\mathcal{A}(\omega)$ by balls of radius R. These centers can be chosen in a measurable way [79]. We denote the corresponding covering number $N_1(\omega, R)$ simply by N_1, or, to make the dependence on ω and R explicit, by $N_1(\omega, R)$.

Since by assumption $\mathcal{A}(\omega)$ is ϕ-invariant, we have $\mathcal{A}(\theta\omega) = \phi(1, \omega, \mathcal{A}(\omega))$ and thus by (16)

$$\mathcal{A}(\theta\omega) \subset \bigcup_{i=1}^{N_1(\omega, R)} \phi(1, \omega, B(x_i, R)). \tag{17}$$

Step 2: Local Approximation. In the second step, we approximate ϕ via a second-order Taylor approximation.

By assumption, for every $\omega \in \Omega$ the map $x \mapsto \phi(1, \omega, x)$ is twice continuously differentiable. We define the random variable

$$C(\omega, R) := \frac{1}{2} \sup_{x \in \mathcal{A}(\omega)_R} \left\| D^2 \phi(1, \omega, x) \right\|, \tag{18}$$

where $\mathcal{A}(\omega)_R$ denotes the closed R-neighborhood of $\mathcal{A}(\omega)$, cf. (14).

Since $\mathcal{A}(\omega)$ is bounded almost surely, its closed R-neighborhood $\mathcal{A}(\omega)_R$ is also compact almost surely. Because $D^2\phi(1,\omega,\cdot)$ is continuous, the supremum in the definition of $C(\omega,R)$ is finite for almost every ω. Hence,

$$C(\omega,R) < \infty \quad \text{for almost every } \omega \in \Omega.$$

We now study the image of the random ball $B(x_i(\omega),R)$ under the map $\phi(1,\omega,\cdot)$. As argued above, the random constant $C(\omega)$ is finite only almost surely and not uniformly in ω. Therefore, we fix a set $\Omega_0 \subset \Omega$ with full measure, i.e.

$$\mathbb{P}(\Omega_0) = 1,$$

such that the constant $C(\omega,R)$ is finite for all $\omega \in \Omega_0$. In what follows, all arguments are carried out pathwise for $\omega \in \Omega_0$. Fix $i \in \{1,\dots,N_1\}$ and we investigate the image of $\phi(1,\omega,B(x_i(\omega),R))$.

Fix $\omega \in \Omega_0$. Let $i := i(\omega) \in \{1,\dots,N_1(\omega)\}$ (for notational simplicity we suppress the dependence on ω in the index). We consider the image of the ball $B(x_i(\omega),R)$ under $\phi(1,\omega,\cdot)$. Let $y(\omega) \in B(x_i(\omega),R)$ be arbitrary. Then there exists $\xi(\omega)$ with $\|\xi(\omega)\| \le R$ such that

$$y(\omega) = x_i(\omega) + \xi(\omega).$$

Since $\phi(1,\omega,\cdot)$ is C^2, Taylor's theorem yields

$$\phi(1,\omega,x_i(\omega)+\xi(\omega)) = \phi(1,\omega,x_i(\omega)) + D\phi(1,\omega,x_i(\omega))\,\xi(\omega) + R(\omega,x_i(\omega),\xi(\omega)),$$

where the remainder term satisfies the uniform bound

$$\|R(\omega,x_i(\omega),\xi)\| \le C(\omega,R)\,\|\xi\|^2 \le C(\omega,R)\,R^2.$$

Since the point $y(\omega) \in B(x_i(\omega),R)$ was arbitrary, the above estimate holds uniformly for all $y(\omega) \in B(x_i(\omega),R)$. Therefore, the image of the ball $B(x_i(\omega),R)$ under $\phi(1,\omega,\cdot)$ satisfies the set inclusion

$$\phi(1,\omega,B(x_i(\omega),R)) \subset \{\phi(1,\omega,x_i(\omega))\} \;\oplus\; D\phi(1,\omega,x_i(\omega))\,B(0,R) \;\oplus\; B\big(0,C(\omega,R)\,R^2\big), \tag{19}$$

where $\oplus$ denotes the Minkowski sum of sets. Since i was arbitrary, this holds for each $i \in \{1,\dots N_1\}$. In other words, we showed the set inclusion for almost all ω.

Step 3: Covering Numbers. We recall we have now by (17) and (19)

$$\mathcal{A}(\theta\omega) \subset \bigcup_{i=1}^{N_1} \{\phi(1,\omega,x_i(\omega))\} \;\oplus\; D\phi(1,\omega,x_i(\omega))\,B(0,R) \;\oplus\; B\big(0,C(\omega,R)\,R^2\big). \tag{20}$$

Now, we will obtain an esimate on the covering number of $\mathcal{A}(\theta\omega)$ by using the fact that each element in the union

$$\{\phi(1,\omega,x_i(\omega))\} \;\oplus\; D\phi(1,\omega,x_i(\omega))\,B(0,R) \;\oplus\; B\big(0,C(\omega,R)\,R^2\big)$$

is a dilated ellipsoid, so that we can use our result for estimating covering numbers for ellipsoids.

More precisely, we have that for any $\rho > 0$

$$\mathcal{N}(\mathcal{A}(\theta\omega),\rho) \le \sum_{i=1}^{N_1} \mathcal{N}\left(\{\phi(1,\omega,x_i(\omega))\} \oplus D\phi(1,\omega,x_i(\omega))B(0,R) \oplus B(0,C(\omega,R)R^2),\rho\right) \tag{21}$$

We now apply Lemma C.3 with

$$X := D\phi(1,\omega,x_i(\omega))\,B(0,R), \qquad Y := B\big(0,C(\omega,R)\,R^2\big).$$

The translation by $\phi(1,\omega,x_i(\omega))$ does not affect the covering number and can be ignored in the following estimates.

We further define as covering radius

$$\rho_1(\omega) := R \sup_{w \in \mathcal{A}(\omega)} \sigma_{j^*+1}(D\phi(1,\omega,w)).$$

By Lemma C.3 we obtain

$$\mathcal{N}(X \oplus Y, \rho_1(\omega) + C(\omega, R)R^2) \leq \mathcal{N}(X, \rho_1(\omega)). \tag{22}$$

We recall that j^* is the transition index and $\sigma_i(D\phi(1, \omega, x))$ denotes the i-th singular value of $D\phi(1, \omega, x)$ for $x \in \mathbb{R}^d$. Now we recall that X is an ellipsoid with semi axis lengths

$$R\sigma_1(D\phi(1, \omega, x_i(\omega))), \ldots, R\sigma_d(D\phi(1, \omega, x_i(\omega)))$$

since

$$X = D\phi(1, \omega, x_i(\omega))\, B(0, R) = R \cdot D\phi(1, \omega, x_i(\omega))\, B(0, 1).$$

Therefore, we are now ready to apply Lemma C.2. We note that

$$R\sigma_{j^*+1}(D\phi(1, \omega, x_i(\omega))) \leq R \sup_{w \in \mathcal{A}(\omega)} \sigma_{j^*+1}(D\phi(1, \omega, w) = \rho_1(\omega).$$

Hence, by Lemma C.2 we obtain

$$\mathcal{N}(X, \rho_1(\omega)) \leq 3^d \prod_{k=1}^{j^*} \frac{R\sigma_k(D\phi(1, \omega, x_i(\omega)))}{\rho_1(\omega)} \leq 3^d \prod_{k=1}^{j^*} \frac{R \sup_{x \in \mathcal{A}(\omega)} \sigma_k(D\phi(1, \omega, x)}{\rho_1(\omega)} =: 3^d \mu(\omega). \tag{23}$$

Notice that even though X depends on the particular center $x_i(\omega)$, $\mu(\omega)$ is uniform over all the centers and does not depend on i.

Now, let us define

$$\rho(\omega) := \rho_1(\omega) + C(\omega, R)R^2$$

and observe that by combining (21) and (22), we have that

$$\begin{aligned}
\mathcal{N}(\mathcal{A}(\theta\omega), \rho(\omega)) \leq & \sum_{i=1}^{N_1} \mathcal{N}\left(\{\phi(1, \omega, x_i(\omega))\} \oplus D\phi(1, \omega, x_i(\omega))B(0, R) \oplus B(0, C(\omega, R)R^2), \rho(\omega)\right) \\
= & \sum_{i=1}^{N_1} \mathcal{N}\left(D\phi(1, \omega, x_i(\omega))B(0, R) \oplus B(0, C(\omega, R)R^2), \rho(\omega)\right) \\
\leq & \sum_{i=1}^{N_1} \mathcal{N}(D\phi(1, \omega, x_i(\omega))B(0, R), \rho_1(\omega) \qquad (24) \\
\leq & N_1(\omega, R)3^d \mu(\omega), \qquad (25)
\end{aligned}$$

where (24) follows from (22) and (25) follows from (23).

Now, we choose R sufficiently small such that

$$C(\omega, R)R^2 \leq \rho_1(\omega). \tag{26}$$

This is possible since when R tends to 0, $C(\omega, R)$ converges to a constant by definition (cf. (18)) hence $C(\omega, R)R^2$ tends to 0 as R tends to 0. On the other hand, since by assumption $\inf_{x \in \mathcal{A}(\omega)} \sigma_d(D\phi(1, \omega, x)) > 0$, that makes $\rho_1(\omega) > 0$, which yields (26).

Therefore, for sufficiently small R, we have that

$$\rho(\omega) \leq 2\rho_1(\omega) =: \epsilon_1(\omega).$$

By the monotonicity of covering numbers, a larger radius requires fewer balls. Therefore, we have the following chain of inequalities:

$$\mathcal{N}(\mathcal{A}(\theta\omega), \epsilon_1(\omega)) \leq \mathcal{N}(\mathcal{A}(\theta\omega), \rho(\omega)) \leq N_1(\omega, R)3^d \mu(\omega), \tag{27}$$

where the last step follows from (25).

At this stage we have estimated the covering number of $\mathcal{A}(\theta^1\omega)$. In the next step, we will iterate this idea to get an estimate on the covering number of $\mathcal{A}(\theta^K\omega)$.

Step 4: Global Iteration and Recursive Covering. We now extend the one-step covering estimate to multiple time steps by induction over the discrete time index K. Fix a sufficiently small initial

radius $R > 0$ and suppose that $\mathcal{A}(\omega)$ is covered by $N_1(\omega, R)$ balls of radius R. By the invariance of the random attractor,

$$\mathcal{A}(\theta^{K+1}\omega) = \phi(1, \theta^K\omega, \mathcal{A}(\theta^K\omega)) \quad \text{for all } K \in \mathbb{N}.$$

Define the sequence of radii $\{\epsilon_K\}_{K\geq 0}$ by

$$\epsilon_{K+1}(\omega) := 2 \cdot \epsilon_K(\omega) \sup_{x\in\mathcal{A}(\theta^K\omega)} \sigma_{j^*+1}(D\phi(1, \theta^K\omega, x)), \qquad \epsilon_0 := R. \tag{28}$$

Moreover, the one-step covering argument derived above applies to any sufficiently small covering radius. Consequently, if $\mathcal{A}(\theta^K\omega)$ is covered by balls of radius $\epsilon_K(\omega)$, then applying the same argument with ω replaced by $\theta^K\omega$ yields a cover of $\mathcal{A}(\theta^{K+1}\omega)$ by balls of radius $\epsilon_{K+1}(\omega)$. In other words, all previous estimates remain valid after the replacements $\mathcal{A}(\omega) \leftarrow \mathcal{A}(\theta^K\omega)$ and $\mathcal{A}(\theta\omega) \leftarrow \mathcal{A}(\theta^{K+1}\omega)$.

By the one-step covering estimate from (27), each ball of radius $\epsilon_K(\omega)$ is mapped into a set that can be covered by at most $\mu(\theta^K\omega)$ balls of radius $\epsilon_{K+1}(\omega)$. Consequently,

$$\mathcal{N}(\mathcal{A}(\theta^K\omega), \epsilon_K(\omega)) \;\leq\; N_1(\omega, R) \prod_{k=0}^{K-1} 3^d \cdot \mu(\theta^k\omega). \tag{29}$$

Iterating the radius recursion yields the explicit expression

$$\epsilon_K(\omega) = R \prod_{k=0}^{K-1} \left(2 \cdot \sup_{x\in\mathcal{A}(\theta^k\omega)} \sigma_{j^*+1}(D\phi(1, \theta^k\omega, x)) \right). \tag{30}$$

Step 5: Ergodic Limits and Dimension Bound. To determine the asymptotic growth rate of the covering number, we begin by recalling the definition of the i-th sharpness exponent:

$$\lambda_i := \mathbb{E}\left[\sup_{x\in\mathcal{A}(\omega)} \ln \sigma_i(D\phi(1, \omega, x)) \right]. \tag{31}$$

Taking logarithms in the covering estimate from (29) and normalizing by K, we obtain

$$\frac{1}{K} \ln \mathcal{N}(\mathcal{A}(\theta^K\omega), \epsilon_K(\omega)) \leq \frac{1}{K} \ln N_1(\omega, R) + \frac{1}{K} \sum_{k=0}^{K-1} \ln \mu(\theta^k\omega) + d \ln 3. \tag{32}$$

Since $N_1(\omega, R)$ is finite almost surely and independent of K, we have $\frac{1}{K} \ln N_1(\omega, R) \to 0$ as $K \to \infty$. By the Birkhoff Ergodic Theorem (see Yunis [80, Theorem 3.10]) applied to the ergodic process $\omega \mapsto \ln \mu(\omega)$[9], we obtain

$$\limsup_{K\to\infty} \frac{1}{n} \ln \mathcal{N}(\mathcal{A}(\theta^K\omega), \epsilon_K(\omega)) \leq \mathbb{E}\left[\ln \mu(\omega)\right] + d \ln 3 \tag{33}$$

$$= \mathbb{E}\left[\sum_{i=1}^{j^*} \sup_{x\in\mathcal{A}(\omega)} \ln \sigma_i(D\phi(1, \omega, x)) - j^* \sup_{x\in\mathcal{A}(\omega)} \ln \sigma_{j^*+1}(D\phi(1, \omega, x)) \right] + d \ln 3. \tag{34}$$

Step 5: Asymptotic Scale Decay.

By using (30), taking logarithms and normalizing by K, we compute

$$\frac{1}{K} \ln \epsilon_K(\omega) = \frac{1}{K} \sum_{k=0}^{K-1} \left(\sup_{x\in\mathcal{A}(\theta^k\omega)} \ln \sigma_{j^*+1}(D\phi(1, \theta^k\omega, x)) \right) + \ln 2 + \frac{1}{K} \ln R. \tag{35}$$

Applying the Birkhoff Ergodic Theorem once more, we obtain

$$\lim_{K\to\infty} \frac{1}{K} \ln \epsilon_K(\omega) = \mathbb{E}\left[\sup_{x\in\mathcal{A}(\omega)} \ln \sigma_{j^*+1}(D\phi(1, \omega, x)) \right] + \ln 2 = \lambda_{j^*+1} + \ln 2. \tag{36}$$

[9] By our integrability assumption μ is also integrable, therefore the theorem can be applied.

Our goal is to study the asymptotic growth of the covering numbers $\mathcal{N}(\mathcal{A}(\theta^K\omega), \epsilon_K(\omega))$ by dividing their logarithm by K and passing to the limit $K \to \infty$. However, the current bound involve several scale-independent multiplicative constants (such as 3^d) that obscure the leading exponential behavior and would not disappear when taking this limit. To resolve this, we refine the analysis by rewriting the recursive scales $\epsilon_K(\omega)$ in a form that makes their multiplicative structure explicit.

Step 6: Computation of the Dimension Bound. We consider the m-th iterate of the cocycle, $\phi(m, \omega, x)$, $x \in \mathbb{R}^d$. By the cocycle property we have for $m, k \in \mathbb{N}$ and $x \in \mathbb{R}^d$

$$\phi(m + k, \omega, x) = \phi(k, \theta^m\omega, \phi(m, \omega, x)), \tag{37}$$

and by applying the chain rule:

$$D\phi(m + k, \omega, x) = D\phi(k, \theta^m\omega, \phi(m, \omega, x)) \cdot D\phi(m, \omega, x). \tag{38}$$

For a linear map $A \in \mathbb{R}^{d\times d}$ and $j \in \{1, \dots, d\}$, define the *j-th exterior-power norm*

$$\|A\|_j := \sigma_1(A)\, \sigma_2(A) \cdots \sigma_j(A), \tag{39}$$

where $\sigma_1(A) \geq \sigma_2(A) \geq \cdots \geq \sigma_d(A) \geq 0$ are the singular values of A. We also set $\|A\|_0 := 1$ for convenience, so that $\sigma_j(A) = \|A\|_j/\|A\|_{j-1}$ for all $j \geq 1$.

By the functoriality of the exterior power, $\bigwedge^j(AB) = (\bigwedge^j A)(\bigwedge^j B)$, and the submultiplicativity of the operator norm, we have the fundamental inequality

$$\|AB\|_j \leq \|A\|_j\, \|B\|_j \qquad \text{for all } A, B \in \mathbb{R}^{d\times d}. \tag{40}$$

6.1. Subadditivity and Fekete limits. Define the expected log-growth of ω_j along the cocycle:

$$\Omega_j^{(m)} := \mathbb{E}\left[\ln \sup_{w\in\mathcal{A}(\omega)} \|\big(D\phi(m, \omega, w)\big)\|_j\right], \qquad m \in \mathbb{N}^*. \tag{41}$$

Combining the cocycle identity (38) with the submultiplicativity (40) gives

$$\|\big(D\phi(m + k, \omega, x)\big)\|_j \leq \|\big(D\phi(k, \theta^m\omega, \phi(m, \omega, x))\big)\|_j \cdot \|\big(D\phi(m, \omega, x)\big)\|_j.$$

Taking the supremum over $x \in \mathcal{A}(\omega)$ and using the invariance $\phi(m, \omega, \mathcal{A}(\omega)) = \mathcal{A}(\theta^m\omega)$:

$$\sup_{x\in\mathcal{A}(\omega)} \|\big(D\phi(m + k, \omega, x)\big)\|_j \leq \sup_{y\in\mathcal{A}(\theta^m\omega)} \|\big(D\phi(k, \theta^m\omega, y)\big)\|_j \cdot \sup_{x\in\mathcal{A}(\omega)} \|\big(D\phi(m, \omega, x)\big)\|_j.$$

Taking logarithms, then expectations, and using the θ^m-invariance of $\mathbb{P}$, we obtain the subadditivity

$$\Omega_j^{(m+k)} \leq \Omega_j^{(m)} + \Omega_j^{(k)} \qquad \text{for all } m, k \in \mathbb{N}^*. \tag{42}$$

By the integrability assumption and Fekete's lemma, the following limit exists for each $j \in \{1, \dots, d\}$:

$$\Lambda_j := \lim_{m\to\infty} \frac{1}{m}\, \Omega_j^{(m)} = \inf_{m\geq 1} \frac{1}{m}\, \Omega_j^{(m)}. \tag{43}$$

We set $\Lambda_0 := 0$. By Fekete's lemma, $\Lambda_j \leq \Omega_j^{(1)}$. Moreover, since $\ln$ is monotone increasing and $\prod_{i=1}^j \sigma_i(D\phi(1, \omega, x)) > 0$ by the non-singularity assumption,

$$\begin{aligned}\Omega_j^{(1)} &= \mathbb{E}\left[\ln \sup_{x\in\mathcal{A}(\omega)} \prod_{i=1}^j \sigma_i(D\phi(1, \omega, x))\right] = \mathbb{E}\left[\sup_{x\in\mathcal{A}(\omega)} \sum_{i=1}^j \ln \sigma_i(D\phi(1, \omega, x))\right] \\ &\leq \sum_{i=1}^j \mathbb{E}\left[\sup_{x\in\mathcal{A}(\omega)} \ln \sigma_i(D\phi(1, \omega, x))\right] = \sum_{i=1}^j \lambda_i,\end{aligned} \tag{44}$$

where the inequality uses $\sup_x(f_1(x) + \cdots + f_j(x)) \leq \sup_x f_1(x) + \cdots + \sup_x f_j(x)$. Hence $\Lambda_j \leq \sum_{i=1}^j \lambda_i$. In particular, if $\Lambda_j \geq 0$ then $\sum_{i=1}^j \lambda_i \geq 0$, so the transition index $j^*(\Lambda)$ (the largest $j \in \{0, \dots, d - 1\}$ with $\Lambda_j \geq 0$) satisfies $j_\Lambda^* \leq j^*$ the transition index defined in terms of the one-step exponents λ_i.

Now define the m-step sharpness exponents

$$\lambda_i^{(m)} := \mathbb{E}\left[\sup_{x\in\mathcal{A}(\omega)} \ln \sigma_i\big(D\phi(m,\omega,x)\big)\right], \qquad i = 1,\dots,d. \tag{45}$$

6.3. Covering bound for the m-step map. We apply the covering argument from Steps 1–3 to the m-step map $\phi(m,\omega,\cdot)$ in place of $\phi(1,\omega,\cdot)$, with the replacement $\theta\omega \leftarrow \theta^m\omega$. Steps 1 and 2 (general covering and Taylor approximation) carry over without modification. In Step 3, the ellipsoid covering lemma (Lemma C.2) involves a free index j at which the singular values are split into expanding directions $\sigma_1,\dots,\sigma_j$ and a contracting direction σ_{j+1} that determines the covering radius. We choose $j = j_\Lambda^*$, where j_Λ^* is the largest integer in $\{0,\dots,d-1\}$ such that $\Lambda_{j_\Lambda^*} \geq 0$. We now verify that this choice is well-defined and that the resulting covering argument is valid.

First, j_Λ^* is well-defined: by Assumption 4, $\sum_{i=1}^{j^*+1}\lambda_i < 0$. Since $\Lambda_{j^*+1} \leq \sum_{i=1}^{j^*+1}\lambda_i < 0$, and $\Lambda_0 = 0 \geq 0$, there exists at least one j with $\Lambda_j \geq 0$ and at least one with $\Lambda_j < 0$, so j_Λ^* is well-defined and satisfies $0 \leq j_\Lambda^* \leq j^*$.

Second, $\Lambda_{j_\Lambda^*+1} < 0$ by definition of j_Λ^*. By Fekete's lemma, $\frac{1}{m}\,\Omega_{j_\Lambda^*+1}^{(m)} \to \Lambda_{j_\Lambda^*+1} < 0$. Since for any $A \in \mathbb{R}^{d\times d}$ it holds $\sigma_1(A) \geq \sigma_2(A) \geq \cdots \geq \sigma_d(A)$, we have $\sigma_k(A) \geq \sigma_{j_\Lambda^*+1}(A)$ for all $k \leq j_\Lambda^*+1$, and therefore

$$\|A\|_{j_\Lambda^*+1} = \prod_{k=1}^{j_\Lambda^*+1}\sigma_k(A) \geq \sigma_{j_\Lambda^*+1}(A)^{j_\Lambda^*+1}.$$

Taking logarithms and dividing by j_Λ^*+1:

$$\ln \sigma_{j_\Lambda^*+1}(A) \leq \frac{1}{j_\Lambda^*+1}\,\ln\|A\|_{j_\Lambda^*+1}.$$

Applying this with $A = D\phi(m,\omega,x)$ and taking the supremum over $x \in \mathcal{A}(\omega)$:

$$\sup_{x\in\mathcal{A}(\omega)} \ln \sigma_{j_\Lambda^*+1}(D\phi(m,\omega,x)) \;\leq\; \frac{1}{j_\Lambda^*+1}\ln \sup_{x\in\mathcal{A}(\omega)} \|(D\phi(m,\omega,x))\|_{j_\Lambda^*+1},$$

where we used the monotonicity of $\ln$ and the fact that $t \mapsto t^{1/(j_\Lambda^*+1)}$ is monotone increasing, so the supremum passes inside. Taking expectations and dividing by m:

$$\frac{1}{m}\,\lambda_{j_\Lambda^*+1}^{(m)} \;\leq\; \frac{1}{j_\Lambda^*+1}\cdot\frac{1}{m}\,\Omega_{j_\Lambda^*+1}^{(m)} \;\xrightarrow{m\to\infty}\; \frac{\Lambda_{j_\Lambda^*+1}}{j_\Lambda^*+1} < 0,$$

by Fekete's lemma. In particular, $\lambda_{j_\Lambda^*+1}^{(m)} \to -\infty$, which shows that $\sup_{x\in\mathcal{A}(\omega)}\sigma_{j_\Lambda^*+1}(D\phi(m,\omega,x))$ decays exponentially in m. Therefore, the covering radius $\rho_1(\omega,m) = R\,\sup_x \sigma_{j_\Lambda^*+1}(D\phi(m,\omega,x))$ shrinks with each iteration of the m-step map, and the iterative covering argument of Steps 4–5 produces a finite dimension bound.

With this choice, the covering number of each image ellipsoid satisfies

$$\mathcal{N}\big(D\phi(m,\omega,x_i)\,B(0,R),\ \rho_1(\omega,m)\big) \;\leq\; 3^d\,\frac{\prod_{k=1}^{j_\Lambda^*}\sigma_k\big(D\phi(m,\omega,x_i)\big)}{\big(\sup_{x\in\mathcal{A}(\omega)}\sigma_{j_\Lambda^*+1}(D\phi(m,\omega,x))\big)^{j_\Lambda^*}}, \tag{46}$$

where $\rho_1(\omega,m) := R\,\sup_{x\in\mathcal{A}(\omega)}\sigma_{j_\Lambda^*+1}(D\phi(m,\omega,x))$. To obtain a bound that is uniform over all centers $x_i \in \mathcal{A}(\omega)$, we use

$$\prod_{k=1}^{j_\Lambda^*}\sigma_k\big(D\phi(m,\omega,x_i)\big) = \big\|\big(D\phi(m,\omega,x_i)\big)\big\|_{j_\Lambda^*} \leq \sup_{x\in\mathcal{A}(\omega)}\big\|\big(D\phi(m,\omega,x)\big)\big\|_{j_\Lambda^*}.$$

Define the multiplier

$$\mu^{(m)}(\omega) := \frac{\sup_{x\in\mathcal{A}(\omega)}\big\|\big(D\phi(m,\omega,x)\big)\big\|_{j_\Lambda^*}}{\Big(\sup_{x\in\mathcal{A}(\omega)}\sigma_{j_\Lambda^*+1}\big(D\phi(m,\omega,x)\big)\Big)^{j_\Lambda^*}}. \tag{47}$$

Then the one-step covering estimate (applied to the m-step map) gives

$$\mathcal{N}\big(\mathcal{A}(\theta^m\omega),\, \epsilon_1(\omega,m)\big) \le N_1(\omega,R)\, 3^d\, \mu^{(m)}(\omega), \tag{48}$$

where $\epsilon_1(\omega,m) = 2\rho_1(\omega,m) = 2R\, \sup_{x\in\mathcal{A}(\omega)} \sigma_{j_\Lambda^*+1}(D\phi(m,\omega,x))$.

6.4. Iterating the m-step map. Define the sequence of radii for the K-fold iteration of the m-step map:

$$\epsilon_{K+1}(\omega,m) := 2\,\epsilon_K(\omega,m) \sup_{x\in\mathcal{A}(\theta^{mK}\omega)} \sigma_{j_\Lambda^*+1}\big(D\phi(m,\theta^{mK}\omega,x)\big), \qquad \epsilon_0 := R. \tag{49}$$

By the inductive argument of Step 5 (with θ replaced by θ^m):

$$\mathcal{N}\big(\mathcal{A}(\theta^{mK}\omega),\, \epsilon_K(\omega,m)\big) \le N_1(\omega,R) \prod_{k=0}^{K-1} 3^d\, \mu^{(m)}(\theta^{mk}\omega). \tag{50}$$

The radii iterate to

$$\epsilon_K(\omega,m) = R \prod_{k=0}^{K-1} 2 \sup_{x\in\mathcal{A}(\theta^{mk}\omega)} \sigma_{j_\Lambda^*+1}\big(D\phi(m,\theta^{mk}\omega,x)\big). \tag{51}$$

6.5. Ergodic limits. Taking logarithms in (50), dividing by K, and applying the Birkhoff Ergodic Theorem to $\omega \mapsto \ln \mu^{(m)}(\omega)$ (with respect to θ^m), we obtain

$$\limsup_{K\to\infty} \frac{1}{K} \ln \mathcal{N}\big(\mathcal{A}(\theta^{mK}\omega), \epsilon_K(\omega,m)\big) \le \mathbb{E}\big[\ln \mu^{(m)}(\omega)\big] + d\ln 3. \tag{52}$$

From the definition of $\mu^{(m)}$:

$$\mathbb{E}\big[\ln \mu^{(m)}(\omega)\big] = \Omega^{(m)}_{j_\Lambda^*} - j_\Lambda^*\, \lambda^{(m)}_{j_\Lambda^*+1}. \tag{53}$$

Similarly, from (51) and the Birkhoff Ergodic Theorem:

$$\lim_{K\to\infty} \frac{1}{K} \ln \epsilon_K(\omega,m) = \lambda^{(m)}_{j_\Lambda^*+1} + \ln 2. \tag{54}$$

6.6. Sign condition and passage to the limit. We need $\lambda^{(m)}_{j_\Lambda^*+1} + \ln 2 < 0$, i.e.,

$$\lambda^{(m)}_{j_\Lambda^*+1} < -\ln 2. \tag{55}$$

Since $\sigma_{j_\Lambda^*+1}(A)^{j_\Lambda^*+1} \le \omega_{j_\Lambda^*+1}(A)$ for any $A \in \mathbb{R}^{d\times d}$, we have the upper bound

$$\lambda^{(m)}_{j_\Lambda^*+1} \le \frac{1}{j_\Lambda^*+1}\, \Omega^{(m)}_{j_\Lambda^*+1}. \tag{56}$$

By Fekete's lemma, $\frac{1}{m}\,\Omega^{(m)}_{j_\Lambda^*+1} \to \Lambda_{j_\Lambda^*+1} < 0$, so $\Omega^{(m)}_{j_\Lambda^*+1} \to -\infty$ as $m\to\infty$. Combined with (56), this gives $\lambda^{(m)}_{j_\Lambda^*+1} \to -\infty$, and in particular (55) holds for all m sufficiently large.

For such m, the box-counting dimension satisfies

$$\overline{\dim}_M\, \mathcal{A}(\omega) \le \frac{\Omega^{(m)}_{j_\Lambda^*} - j_\Lambda^*\, \lambda^{(m)}_{j_\Lambda^*+1} + d\ln 3}{-\lambda^{(m)}_{j_\Lambda^*+1} - \ln 2}. \tag{57}$$

Dividing numerator and denominator by m:

$$\overline{\dim}_M\, \mathcal{A}(\omega) \le \frac{\frac{1}{m}\,\Omega^{(m)}_{j_\Lambda^*} - j_\Lambda^*\,\frac{1}{m}\,\lambda^{(m)}_{j_\Lambda^*+1} + \frac{d\ln 3}{m}}{-\frac{1}{m}\,\lambda^{(m)}_{j_\Lambda^*+1} - \frac{\ln 2}{m}}. \tag{58}$$

We now pass to the limit $m \to \infty$. By Fekete's lemma, $\frac{1}{m}\,\Omega^{(m)}_{j^*_\Lambda} \to \Lambda_{j^*_\Lambda}$. The terms $\frac{d\ln 3}{m}$ and $\frac{\ln 2}{m}$ vanish. It remains to identify the limit of $\frac{1}{m}\,\lambda^{(m)}_{j^*_\Lambda+1}$.

Since $\sigma_{j^*_\Lambda+1}(A) = \omega_{j^*_\Lambda+1}(A)/\omega_{j^*_\Lambda}(A)$, we have the two-sided bounds

$$\Omega^{(m)}_{j^*_\Lambda+1} - \Omega^{(m)}_{j^*_\Lambda} \;\le\; \lambda^{(m)}_{j^*_\Lambda+1} \;\le\; \Omega^{(m)}_{j^*_\Lambda+1} - \hat{\Omega}^{(m)}_{j^*_\Lambda}, \tag{59}$$

where the lower bound uses $\sup(f-g) \ge \sup f - \sup g$ and the upper bound uses $\sup(f-g) \le \sup f - \inf g$. By Fekete's lemma, $\frac{1}{m}(\Omega^{(m)}_{j^*_\Lambda+1} - \Omega^{(m)}_{j^*_\Lambda}) \to \Lambda_{j^*_\Lambda+1} - \Lambda_{j^*_\Lambda}$. By the Bounded Distortion assumption, $\frac{1}{m}(\Omega^{(m)}_{j^*_\Lambda} - \hat{\Omega}^{(m)}_{j^*_\Lambda}) \to 0$, and hence $\frac{1}{m}\,\hat{\Omega}^{(m)}_{j^*_\Lambda} \to \Lambda_{j^*_\Lambda}$, so that $\frac{1}{m}(\Omega^{(m)}_{j^*_\Lambda+1} - \hat{\Omega}^{(m)}_{j^*_\Lambda}) \to \Lambda_{j^*_\Lambda+1} - \Lambda_{j^*_\Lambda}$. By the squeeze theorem,

$$\lim_{m\to\infty} \frac{1}{m}\,\lambda^{(m)}_{j^*_\Lambda+1} = \Lambda_{j^*_\Lambda+1} - \Lambda_{j^*_\Lambda} =: \tilde{\lambda}_{j^*_\Lambda+1}. \tag{60}$$

Substituting all limits into (58), we obtain

$$\overline{\dim}_M\,\mathcal{A}(\omega) \;\le\; \frac{\Lambda_{j^*_\Lambda} - j^*_\Lambda\,\tilde{\lambda}_{j^*_\Lambda+1}}{-\tilde{\lambda}_{j^*_\Lambda+1}} = j^*_\Lambda + \frac{\Lambda_{j^*_\Lambda}}{\Lambda_{j^*_\Lambda} - \Lambda_{j^*_\Lambda+1}} =: D_s. \tag{61}$$

We now show that $D_s \le \dim_{\mathrm{S}} \mathcal{A}_S$, where $\dim_{\mathrm{S}} \mathcal{A}_S := j^* + \frac{\sum_{i=1}^{j^*}\lambda_i}{|\lambda_{j^*+1}|}$ denotes the bound with one-step exponents. Define $g(a,b) := \frac{a}{a-b}$ for $a \ge 0 > b$, so that $D_s = j^*_\Lambda + g(\Lambda_{j^*_\Lambda}, \Lambda_{j^*_\Lambda+1})$. Since $b < 0$ and $a \ge 0$:

$$\frac{\partial g}{\partial a} = \frac{-b}{(a-b)^2} > 0, \qquad \frac{\partial g}{\partial b} = \frac{a}{(a-b)^2} \ge 0,$$

so g is monotonically increasing in both arguments, and $g(a,b) < 1$ since $a < a - b$ (because $b < 0$).

We distinguish two cases.

Case 1: $j^*_\Lambda = j^*$. Since $\Lambda_j \le \sum_{i=1}^{j}\lambda_i$ for all j, the monotonicity of g in the first argument gives

$$g(\Lambda_{j^*},\,\Lambda_{j^*+1}) \;\le\; g\left(\sum_{i=1}^{j^*}\lambda_i,\;\Lambda_{j^*+1}\right).$$

By Assumption 4, $\sum_{i=1}^{j^*+1}\lambda_i < 0$, and since $\Lambda_{j^*+1} \le \sum_{i=1}^{j^*+1}\lambda_i < 0$, the monotonicity of g in the second argument yields

$$g\left(\sum_{i=1}^{j^*}\lambda_i,\;\Lambda_{j^*+1}\right) \;\le\; g\left(\sum_{i=1}^{j^*}\lambda_i,\;\sum_{i=1}^{j^*+1}\lambda_i\right) = \frac{\sum_{i=1}^{j^*}\lambda_i}{|\lambda_{j^*+1}|}.$$

Therefore $D_s = j^* + g(\Lambda_{j^*}, \Lambda_{j^*+1}) \le j^* + \frac{\sum_{i=1}^{j^*}\lambda_i}{|\lambda_{j^*+1}|} = \dim_{\mathrm{S}}\mathcal{A}$.

Case 2: $j^*_\Lambda < j^*$. Since $g(\Lambda_{j^*_\Lambda}, \Lambda_{j^*_\Lambda+1}) < 1$, we have $D_s < j^*_\Lambda + 1 \le j^*$. On the other hand, $\dim_{\mathrm{S}}\mathcal{A} = j^* + \frac{\sum_{i=1}^{j^*}\lambda_i}{|\lambda_{j^*+1}|} \ge j^*$, since $\sum_{i=1}^{j^*}\lambda_i \ge 0$ by Assumption. Therefore $D_s < j^* \le \dim_{\mathrm{S}}\mathcal{A}$.

Combining both cases, we conclude that for $\mathbb{P}$-almost every ω:

$$\overline{\dim}_M(\mathcal{A}(\omega)) \;\le\; D_s \;\le\; \dim_{\mathrm{S}}\mathcal{A} \;=\; j^* + \frac{\sum_{i=1}^{j^*}\lambda_i}{|\lambda_{j^*+1}|}. \tag{62}$$

Note: The bound D_s is strictly sharper than $\dim_{\mathrm{S}}\mathcal{A}_S$ whenever $\Lambda_j < \sum_{i=1}^{j}\lambda_i$ for some $j \in \{j^*_\Lambda, j^*_\Lambda + 1\}$, which occurs when different points on the attractor maximize different singular values σ_i. This completes Step 6.

Step 7: Transition to Arbitrary Radii. The covering estimates obtained in Step 6 are formulated along the discrete sequence of radii $\{\epsilon_K(\omega,m)\}_{K\in\mathbb{N}}$. In this step, we show that the dimension

bound extends to arbitrary radii $\epsilon \to 0$, thereby establishing an upper bound on the upper Minkowski dimension.

Fix $m \in \mathbb{N}$ sufficiently large such that $\lambda^{(m)}_{j^*_\Lambda+1} + \ln 2 < 0$ (which is possible by Step 6.6). Let $\{\epsilon_K(\omega, m)\}_{K\in\mathbb{N}}$ be the sequence of radii defined in (49). By (54), the Birkhoff Ergodic Theorem gives

$$\lim_{K\to\infty} \frac{1}{K} \ln \epsilon_K(\omega, m) = \lambda^{(m)}_{j^*_\Lambda+1} + \ln 2 =: \ell(m) < 0 \qquad \mathbb{P}\text{-a.s.} \tag{63}$$

In particular, $\epsilon_K(\omega, m) \to 0$ exponentially as $K \to \infty$, and for $\mathbb{P}$-a.e. ω the sequence $\{\epsilon_K(\omega, m)\}_K$ is eventually strictly decreasing.

We now work with a fixed realization ω in the full-measure set where (63) and the covering estimate (50) both hold. The estimates in Step 6 bound $\mathcal{N}(\mathcal{A}(\theta^{mK}\omega), \epsilon_K(\omega, m))$. Since θ^{mK} preserves $\mathbb{P}$ and the entire argument from Steps 1–6 applies with ω replaced by any ω' in the underlying full-measure set, the same estimates hold with $\theta^{mK}\omega$ replaced by ω. That is, for $\mathbb{P}$-a.e. ω there exists a sequence of radii $\epsilon_K(\omega, m) \to 0$ such that

$$\limsup_{K\to\infty} \frac{1}{K} \ln \mathcal{N}(\mathcal{A}(\omega), \epsilon_K(\omega, m)) \;\le\; \mathbb{E}[\ln \mu^{(m)}(\omega)] + d \ln 3. \tag{64}$$

For any $\epsilon > 0$ sufficiently small, there exists $K = K(\epsilon)$ such that $\epsilon \in [\epsilon_{K+1}(\omega, m),\, \epsilon_K(\omega, m))$. Since covering numbers are monotonically non-increasing in the radius:

$$\mathcal{N}(\mathcal{A}(\omega), \epsilon_K(\omega, m)) \;\le\; \mathcal{N}(\mathcal{A}(\omega), \epsilon) \;\le\; \mathcal{N}(\mathcal{A}(\omega), \epsilon_{K+1}(\omega, m)). \tag{65}$$

Taking logarithms and dividing by $-\ln \epsilon > 0$:

$$\frac{\ln \mathcal{N}(\mathcal{A}(\omega), \epsilon_K(\omega, m))}{-\ln \epsilon_{K+1}(\omega, m)} \;\le\; \frac{\ln \mathcal{N}(\mathcal{A}(\omega), \epsilon)}{-\ln \epsilon} \;\le\; \frac{\ln \mathcal{N}(\mathcal{A}(\omega), \epsilon_{K+1}(\omega, m))}{-\ln \epsilon_K(\omega, m)}, \tag{66}$$

where we used $\epsilon_{K+1}(\omega, m) \le \epsilon < \epsilon_K(\omega, m)$ to bound $-\ln \epsilon$ from below and above in the denominators.

It remains to show that the lower and upper bounds in (66) have the same $\limsup$. By (63), $\frac{1}{K} \ln \epsilon_K(\omega, m) \to \ell(m) < 0$, and therefore

$$\frac{\ln \epsilon_{K+1}(\omega, m)}{\ln \epsilon_K(\omega, m)} = \frac{\frac{1}{K+1} \ln \epsilon_{K+1}(\omega, m)}{\frac{1}{K} \ln \epsilon_K(\omega, m)} \cdot \frac{K+1}{K} \;\xrightarrow{K\to\infty}\; \frac{\ell(m)}{\ell(m)} \cdot 1 = 1. \tag{67}$$

Consequently, replacing $-\ln \epsilon_{K+1}$ by $-\ln \epsilon_K$ (or vice versa) in the denominators of (66) does not affect the $\limsup$. Both the lower and upper bounds converge to the same value, and we conclude

$$\limsup_{\epsilon\to 0} \frac{\ln \mathcal{N}(\mathcal{A}(\omega), \epsilon)}{-\ln \epsilon} \;=\; \limsup_{K\to\infty} \frac{\ln \mathcal{N}(\mathcal{A}(\omega), \epsilon_K(\omega, m))}{-\ln \epsilon_K(\omega, m)}. \tag{68}$$

Combining (68) with the covering estimate (64) and the radii asymptotics (63), we obtain for each fixed m sufficiently large:

$$\overline{\dim}_M(\mathcal{A}(\omega)) \;=\; \limsup_{\epsilon\to 0} \frac{\ln \mathcal{N}(\mathcal{A}(\omega), \epsilon)}{-\ln \epsilon} \;\le\; \frac{\mathbb{E}[\ln \mu^{(m)}(\omega)] + d \ln 3}{-\lambda^{(m)}_{j^*_\Lambda+1} - \ln 2}. \tag{69}$$

Passing to the limit $m \to \infty$ as in Step 6.6 yields

$$\overline{\dim}_M(\mathcal{A}(\omega)) \;\le\; \dim_{\mathrm{S}} \mathcal{A}_S$$

for $\mathbb{P}$-almost every ω. This completes the proof. □

Theorem C.5. *Generalization via Sharpness Dimension*

Let $S = \{z_1, \ldots, z_n\} \sim \mu_z^{\otimes n}$ be a dataset of size n. Let $(\Omega, \mathcal{F}, \mathbb{P}, \theta, \phi)$ be a discrete-time RDS according to Dfn 2.1 such that Assump. B.3 holds. Under Assumps. B.1 and B.2, there exists a constant $C > 0$ s.t. with probability at least $1 - \zeta - \gamma$ over the joint draw $(S, \omega) \sim \mu_z^{\otimes n} \otimes \mathbb{P}$, there exists $\delta_{n,\gamma} > 0$ such that for all $0 < \delta < \delta_{n,\gamma}$,

$$\mathcal{G}_S(\mathcal{A}(\omega)) \le 2L\delta + 2B\sqrt{\frac{4 \dim_{\mathrm{S}} \mathcal{A}_S \, \log(1/\delta)}{n}} + \frac{I_\infty(\mathcal{A}_S(\omega), S) + \log(1/\zeta)}{\sqrt{n}} + \frac{CB^2}{\sqrt{n}}.$$

We recall that $\mathcal{G}_S(\mathcal{A}(\omega))$ denotes the worst-case generalization gap (see (1)*) and $I_\infty(\mathcal{A}_S(\omega), S)$ (see Dfn. A.7) denotes the total mutual information between the random pullback attractor $\mathcal{A}_S(\omega)$ and S.*

Proof. The result is an immediate consequence of Cor. A.9 and Thm. C.4. □

Theorem C.6 (Stability Bound)**.** *Under the same assumptions as in Thm. C.5, we assume for simplicity that $\beta_n^{-2/3}$ is an integer divisor of n. Then, there exists $\delta_n > 0$ such that for all $0 < \delta < \delta_n$:*

$$\mathbb{E}\left[\sup_{w \in \mathcal{W}_{S,U}} \left(\mathcal{R}(w) - \widehat{\mathcal{R}}_S(w)\right)\right] \leq 2\mathbb{E}\left[\frac{B}{n} + \delta L + \beta_n^{1/3}\left(1 + B\sqrt{4 \dim_{\mathrm{S}} \mathcal{A}_S \log \frac{1}{\delta}}\right)\right].$$

Proof. The proof follows from Thm. A.13 and Thm. C.4. □

D Implementation Details and Computational Complexity

Experimental Setup Details. For the 3-layer MNIST experiments, we use a width of 16, ReLU activation, and no bias, for a total of 12,960 parameters. Networks are trained with SGD, without momentum or weight decay, for 250 epochs using cross-entropy loss. For the $\eta \times b$ hyperparameter grid, we use the learning rates $0.005, 0.01, 0.025, 0.05, 0.1$ and batch sizes $25, 50, 100, 250, 500$. For the SLQ-based approximation SD-SLQ, we use 500 Lanczos runs of 100 steps each with random Rademacher initializations and full reorthogonalization; the expectation is estimated using a separate minibatch per Lanczos run, followed by Gaussian quadrature and kernel smoothing, and SD is then computed by integrating the estimated spectral density and the eigenvalue-weighted spectral density using the trapezoidal rule.

For the 5-layer MNIST experiments, we use width 200, corresponding to 278,800 parameters. Here, the number of Lanczos runs varies with minibatch size to correspond to two training epochs. We report both the SLQ-based histogram and the Gaussian kernel-smoothed estimate, denoted SD-SLQ and SD-KDE, respectively, and do not report exact SD because exact SVD is no longer feasible to compute in this regime. Both 3-layer and 5-layer MNIST experiments are run on an internal cluster using NVIDIA L40S GPUs. Training, including baseline topological complexity calculations, takes 35 minutes per model on average. On 3-layer MNIST, model analysis takes 30 minutes on average (25 minutes for the full Hessian, 5 minutes per SLQ). On 5-layer MNIST, model analysis with SLQ takes 1 hour on average. The total GPU time estimate for MNIST experiments, including additional, baseline, and preliminary experiments, is 600 GPU hours.

For Grokking experiments, the relevant hyperparameters are reported in Fig. 4 and Fig. 5, and in the discussion. In the 2-layer MLP grokking experiments, models are trained for 20000 epochs, and our quantities SD and RDS Sharpness and baseline complexities are evaluated every 500 epochs. In the 3-layer MLP grokking experiments with no momentum, models are trained for 30000 epochs, and all quantities are evaluated every 300 epochs. The grokking experiments are performed on NVIDIA L40S GPUs on an internal cluster. Training, including baseline topological complexity calculations, takes an average of 30 minutes per model. SD and RDS-Sharpness calculations take 1.5 minutes per checkpoint on average using 20 minibatches. The total GPU time estimate for Grokking experiments, including additional, baseline, and preliminary experiments, is 350 GPU hours.

For the GPT-2 [11] experiments, we evaluate SGD and SGD with momentum on the 5 by 5 $\eta \times b$ hyperparameter grid with the learning rates $1 \times 10^{-4}, 2 \times 10^{-4}, 5 \times 10^{-4}, 1 \times 10^{-3}, 2 \times 10^{-3}$ and effective batch sizes (batch size multiplied by number of gradient accumulation steps) of $8, 16, 32, 64, 128$. We evaluate AdamW on the $4 \times 3 \times 3$ hyperparameter grid over learning rates $1 \times 10^{-4}, 3 \times 10^{-4}, 1 \times 10^{-3}, 3 \times 10^{-3}$, effective batch sizes $16, 32, 64$ and weight decays $0, 1 \times 10^{-3}, 1 \times 10^{-2}$. In all settings, we use a block size of 256 and a batch size of 8 for 50 epochs, with learning rate warmup for the first 500 steps and maximum gradient norm clipping to 1. SLQ was performed with 4096 training samples and 64 SLQ probes, with 100 Lanczos steps per probe 64 and full reorthogonalization. Sharpness was calculated using the maximum from 128-step Lanczos runs with 8 restarts and full reorthogonalization. For the highest effective batch size 128 with SGD and SGD with momentum, 32 SLQ probes and 4 sharpness restarts were used to due to the computational cost and to have similar analysis time.

We reconstruct the spectral densities using Gaussian smoothing, with separate bandwidths for the Hessian eigenvalues and their transformed log singular values. Table 1 lists these bandwidths, the histogram and integration grid sizes, and the number of equal-mass quantiles used by SD-PS. The GPT-2 experiments, including both training and Hessian evaluation, are performed on a cluster using NVIDIA GH200 systems with Grace CPUs and H200 GPUs, totaling 800 GPU-hours across all GPT-2-related experiments, baselines, and preliminary experiments. On average, 1 hour per model is spent on training and 3 hours on analysis for SD, RDS Sharpness, and Sharpness estimation. See Table 2 for the per-Lanczos iteration runtime for the stochastic Lanczos quadrature (SLQ) estimation in GPT-2 models per batch size.

Table 1: **GPT-2 Density estimation settings.** Density estimation parameters used for the GPT-2 experiments. SGD and SGD with momentum use the same settings. Additional details are explained in text.

Setting	SGD	AdamW
Hessian bandwidth	0.03	0.03
Transformed bandwidth	7.5×10^{-4}	10^{-4}
Histogram bins	512	512
Grid points	1024	1024
SD-PS quantiles	50,000	50,000

Table 2: **SLQ Iteration Time.** Mean and standard deviation of the per 100-step Lanczos iteration runtime, measured in seconds, for the stochastic Lanczos quadrature (SLQ) estimation in GPT-2 models, with 100. Batch size refers to the effective training batch size. Results for batch sizes 8 and 128 include both SGD and SGD with momentum runs. Results for the remaining batch sizes include all experiments.

Batch Size	SLQ Iteration Time (s)
8	38.99 ± 0.08
16	72.58 ± 0.17
32	139.89 ± 0.34
64	274.13 ± 0.77
128	550.25 ± 1.37

Lanczos and SLQ For all SLQ-based estimators, each Lanczos run uses one Rademacher probe vector and one minibatch Hessian operator. The operator is fixed throughout the Lanczos iterations of that run, so the Krylov subspace and Gaussian quadrature interpretation are well defined. We resample only between independent runs, average the resulting quadrature measures or their histogram/smoothed density estimates, and then compute SD-SLQ, SD-KDE, or SD-PS from this averaged spectral measure. This procedure estimates the expectation over minibatch Hessian operators and probe vectors without mixing Hessian-vector products from different operators inside a single Lanczos run.

More explicitly, let N be the number of parameters and let R_{SLQ} be the number of independent SLQ runs. In run i, we sample a minibatch $\mathcal{B}_i$ and define the corresponding Hessian operator

$$H_i := \nabla_w^2 \ell(w^\star; \mathcal{B}_i). \tag{70}$$

Given an independent Rademacher probe $v_i \in \{\pm 1\}^N$, we set $q_{i,1} = v_i/\|v_i\|$ and run m Lanczos steps on the fixed operator H_i. This produces a basis $Q_i = [q_{i,1}, \ldots, q_{i,m}]$ and a symmetric tridiagonal matrix T_i, with diagonal entries $\alpha_{i,1}, \ldots, \alpha_{i,m}$ and off-diagonal entries $\beta_{i,1}, \ldots, \beta_{i,m-1}$, such that

$$H_i Q_i = Q_i T_i + \beta_{i,m} q_{i,m+1} e_m^\top. \tag{71}$$

Here e_m denotes the m-th standard basis vector in $\mathbb{R}^m$). Let

$$T_i = U_i \operatorname{diag}(\widetilde{\alpha}_{i,1}, \ldots, \widetilde{\alpha}_{i,m}) U_i^\top \tag{72}$$

be the eigendecomposition of this tridiagonal matrix. The eigenvalues $\widetilde{\alpha}_{i,\ell}$ of T_i are the Ritz values, i.e. Lanczos approximations to the eigenvalues of H_i. The associated columns of U_i are the Ritz vectors, and the corresponding Gaussian quadrature weights are given by their squared first components:

$$\omega_{i,\ell} = (U_i)_{1\ell}^2, \qquad \ell = 1, \ldots, m. \tag{73}$$

Thus, for a test function f,

$$q_{i,1}^\top f(H_i) q_{i,1} \approx e_1^\top f(T_i) e_1 = \sum_{\ell=1}^{m} \omega_{i,\ell} f(\widetilde{\alpha}_{i,\ell}). \tag{74}$$

Because $q_{i,1}$ is a normalized Rademacher probe, multiplying by N gives an unbiased trace-scale estimate in expectation over the probe. Averaging across the independently sampled minibatch Hessians and probes yields the empirical spectral measure

$$\widehat{\nu}_H = \frac{N}{R_{\mathrm{SLQ}}} \sum_{i=1}^{R_{\mathrm{SLQ}}} \sum_{\ell=1}^{m} \omega_{i,\ell}\, \delta_{\widetilde{\alpha}_{i,\ell}}. \tag{75}$$

Here δ_x denotes the Dirac measure at x. This is the object averaged by our SLQ procedure: not full spectra, but the quadrature measures induced by Ritz values and weights from independent fixed-operator Lanczos runs. The histogram estimator used for SD-SLQ is obtained by binning this measure. For a bin $I_b = [a_b, a_{b+1})$ of width $\Delta_b = a_{b+1} - a_b$, we set

$$\widehat{\rho}_{\mathrm{hist}}(\alpha) = \frac{N}{R_{\mathrm{SLQ}}\,\Delta_b} \sum_{i=1}^{R_{\mathrm{SLQ}}} \sum_{\ell=1}^{m} \omega_{i,\ell} \mathbf{1}\{\widetilde{\alpha}_{i,\ell} \in I_b\}, \qquad \alpha \in I_b. \tag{76}$$

The smoothed estimator used for SD-KDE replaces the bin indicator by a kernel K_h with bandwidth h:

$$\widehat{\rho}_{\mathrm{kde}}(\alpha) = \frac{N}{R_{\mathrm{SLQ}}} \sum_{i=1}^{R_{\mathrm{SLQ}}} \sum_{\ell=1}^{m} \omega_{i,\ell} K_h(\alpha - \widetilde{\alpha}_{i,\ell}). \tag{77}$$

To compute the Sharpness Dimension estimators, we apply the one-step Jacobian map

$$g_\eta(\alpha) = \log(\max\{|1 - \eta\alpha|, \varepsilon\}) \tag{78}$$

to the Ritz values and estimate the corresponding push-forward density. SD-SLQ and SD-KDE apply the Sharpness Dimension formula to the histogram or smoothed density of these transformed values, respectively. More explicitly, let $\widehat{\rho}_z$ denote the estimated density of the transformed values, with spectral mass scaled by the parameter dimension N. We compute

$$D(t) = \int_t^\infty \widehat{\rho}_z(z)\, dz, \qquad S(t) = \int_t^\infty z\, \widehat{\rho}_z(z)\, dz, \tag{79}$$

which are the continuous counterparts of the spectral count and partial sum in Dfn. 3.3. Starting from the largest transformed values, we accumulate spectral mass through D and its weighted sum through S. The estimated dimension is the accumulated mass where S crosses zero, computed by interpolation between adjacent integration points. If the total weighted integral is nonnegative, we return N. In this formulation, the density enters through the cumulative integrals used for interpolation, rather than through division by its value at a point. Binning, bandwidth, and probe count can affect these integrals and consequently the estimated dimension.

SD-PS, described next, uses the same smoothed-density viewpoint but first converts the density into equal-mass pseudo-eigenvalues before applying the finite-dimensional weighted sharpness dimension formula.

Pseudo-spectrum SD (SD-PS). We also report a pseudo-spectrum version of the SLQ estimator, denoted SD-PS, which discretizes the smoothed spectral density before applying the same sharpness-dimension formula as in Dfn. 3.3. Let $\widehat{\rho}(\alpha)$ be the smoothed SLQ estimate of the Hessian spectral density for a model with N parameters, normalized so that $\int \widehat{\rho}(\alpha)\, d\alpha = N$. We form its cumulative mass function $\widehat{F}(\alpha) = \int_{-\infty}^{\alpha} \widehat{\rho}(s)\, ds$ and choose equal-mass quantiles

$$\widehat{\alpha}_r := \widehat{F}^{-1}\left(\frac{r + 1/2}{M} N\right), \qquad r = 0, \ldots, M-1, \tag{80}$$

where $M \le N$ is the number of pseudo-eigenvalues used in the discretization. Each pseudo-eigenvalue carries weight N/M. We then map these values through the one-step GD Jacobian spectrum,

$$\widehat{z}_r = \log(|1 - \eta\widehat{\alpha}_r| + \varepsilon), \tag{81}$$

sort the values and reindex them so that $\widehat{z}_0 \ge \cdots \ge \widehat{z}_{M-1}$, and compute the weighted Sharpness Dimension: if j is the largest index for which $\sum_{r=0}^{j} (N/M)\widehat{z}_r \ge 0$, then

$$\widehat{\dim}_{\mathrm{S}}^{\mathrm{PS}} = \sum_{r=0}^{j} \frac{N}{M} + \frac{\sum_{r=0}^{j} (N/M)\widehat{z}_r}{|\widehat{z}_{j+1}|}, \tag{82}$$

with the usual truncation to $[0, N]$ and the boundary conventions from Dfn. 3.3. Here, the density determines the equal-mass quantiles, while the denominator in the fractional correction is the magnitude of the next transformed spectral value, $|\widehat{z}_{j+1}|$ (as opposed to the continuous interpolation in SD-SLQ and SD-KDE cases). Bandwidth and probe count can also affect the resulting dimension through the quantiles.

Table 3: **GPT-2 SLQ Sensitivity Analysis.** Sensitivity of Kendall's τ with the loss gap to SLQ configuration. Entries are mean $\pm$ sample standard deviation pooled over 3 seeds.

Varied setting	Estimator	SGD	SGD + momentum	AdamW
Bin count	SD-SLQ	-0.039 ± 0.154	0.160 ± 0.136	0.295 ± 0.104
Bandwidth	SD-KDE	0.266 ± 0.114	0.320 ± 0.122	0.128 ± 0.083
Bandwidth	SD-PS	0.149 ± 0.096	0.491 ± 0.044	0.683 ± 0.035
Probe count	SD-SLQ	-0.056 ± 0.116	0.196 ± 0.081	0.328 ± 0.089
Probe count	SD-KDE	0.347 ± 0.023	0.336 ± 0.037	0.166 ± 0.075
Probe count	SD-PS	0.204 ± 0.031	0.507 ± 0.000	0.693 ± 0.030

In our GPT-2 experiments we use this estimator as an alternative discretization of the same smoothed SLQ measure used for SD-KDE. Thus SD-PS differs from SD-KDE only in how the density is converted into the ordered log-singular-value spectrum: SD-KDE integrates the density directly, whereas SD-PS first constructs equal-mass pseudo-eigenvalues and then applies the finite-dimensional weighted Sharpness Dimension formula.

Sensitivity to SLQ configuration.. Using the same trained checkpoints as in the reported experiments, we vary the histogram bin count over $256, 384, 768, 1024$, the smoothing bandwidth over $1/3, 2/3, 4/3, 2$ times its reference value, and the probe count over $8, 16, 32$ for all three optimizers. For each checkpoint, we repeat the analysis with three probe-selection seeds. Table 3 reports the mean and standard deviation. SD-KDE and SD-PS are independent of histogram bin count, while SD-SLQ is independent of the smoothing bandwidth. Across the bandwidth and probe-count sweeps, the smoothed estimators retain positive average correlations. SD-PS is particularly consistent for AdamW and SGD with momentum, whereas SD-KDE shows weaker correlations for AdamW. The histogram-based SD-SLQ is more sensitive to binning and probe selection, with near-zero or negative average correlations for SGD. These results support the findings for the smoothed estimators while showing that histogram-based estimates are affected more by the numerical settings.

Complexity analysis. Regarding the SD computation, for a network with N parameters and k mini-batches, assume the cost of one Hessian–vector product (HVP) is C_{hvp}. Explicit Hessian computation has memory complexity $\Theta(N^2)$ and worst-case runtime complexity $\mathcal{O}(N \cdot C_{\text{hvp}})$. Singular value or eigenvalue decompositions have memory complexity $\Theta(N^2)$ and runtime complexity $\Theta(N^3)$. In this case, since SD computation requires $\Theta(N)$ memory and runtime, the overall algorithm has $\mathcal{O}(kNC_{\text{hvp}} + kN^3)$ runtime complexity and $\Theta(N^2)$ memory complexity. In the approximate SLQ variant, assuming m iterations per Lanczos run, each run has runtime complexity $\Theta(m \cdot C_{\text{hvp}})$ for HVPs and $\Theta(m^2N)$ for reorthogonalization, with $\Theta(mN)$ memory complexity. Hence, the SLQ-based algorithm has total runtime complexity $\Theta(kmC_{\text{hvp}} + km^2N)$ and memory complexity $\Theta(mN)$. For typical settings where $m \ll N$, both runtime and memory complexity are significantly reduced.

E Additional Results on Hessian Spectra & RDS Sharpness Spectrum

Fig. 8 illustrates our spectral estimators for Hessian spectral density and the RDS sharpness spectrum on GPT-2 across SGD, SGD with momentum, and AdamW. The left image in each pair visualizes the expectation of minibatch Hessian spectral density over eigenvalues α. The corresponding right images show the push-forward of this density through the transformation $\alpha \mapsto \log|1 - \eta\alpha|$, i.e., the transformed spectral density underlying the RDS Sharpnesses of Order k, λ_k. We observe that, across all three optimizers, a large fraction of the Hessian spectrum is concentrated near $\alpha = 0$. This mass near $\alpha = 0$ produces a peak near 0 in the RDS sharpness spectrum and corresponds to neutral or nearly neutral directions (i.e., expanding and contracting with almost 0 log-singular values). In contrast, the isolated spikes, positive tails, and negative curvature components of the Hessian spectrum produce the contractive and expansive directions of the RDS Sharpness spectrum. Positive eigenvalues less than $2/\eta$ yield negative RDS sharpness, corresponding to the contractive directions. Positive eigenvalues that are larger than $2/\eta$ and all negative eigenvalues yield positive RDS sharpness, corresponding to expansive directions instead. Hence, this visualization clarifies why the Sharpness Dimension depends on the full Hessian spectrum rather than only the top Hessian eigenvalue or a small part of it: SD is determined by the balance between the positive RDS sharpness directions, the near-neutral bulk, and the negative tail. Moreover, the estimated Hessian densities are consistent with prior observations on neural-network and transformer Hessian spectra [65, 81], providing additional evidence that our SLQ-based procedure captures the relevant spectral structure.

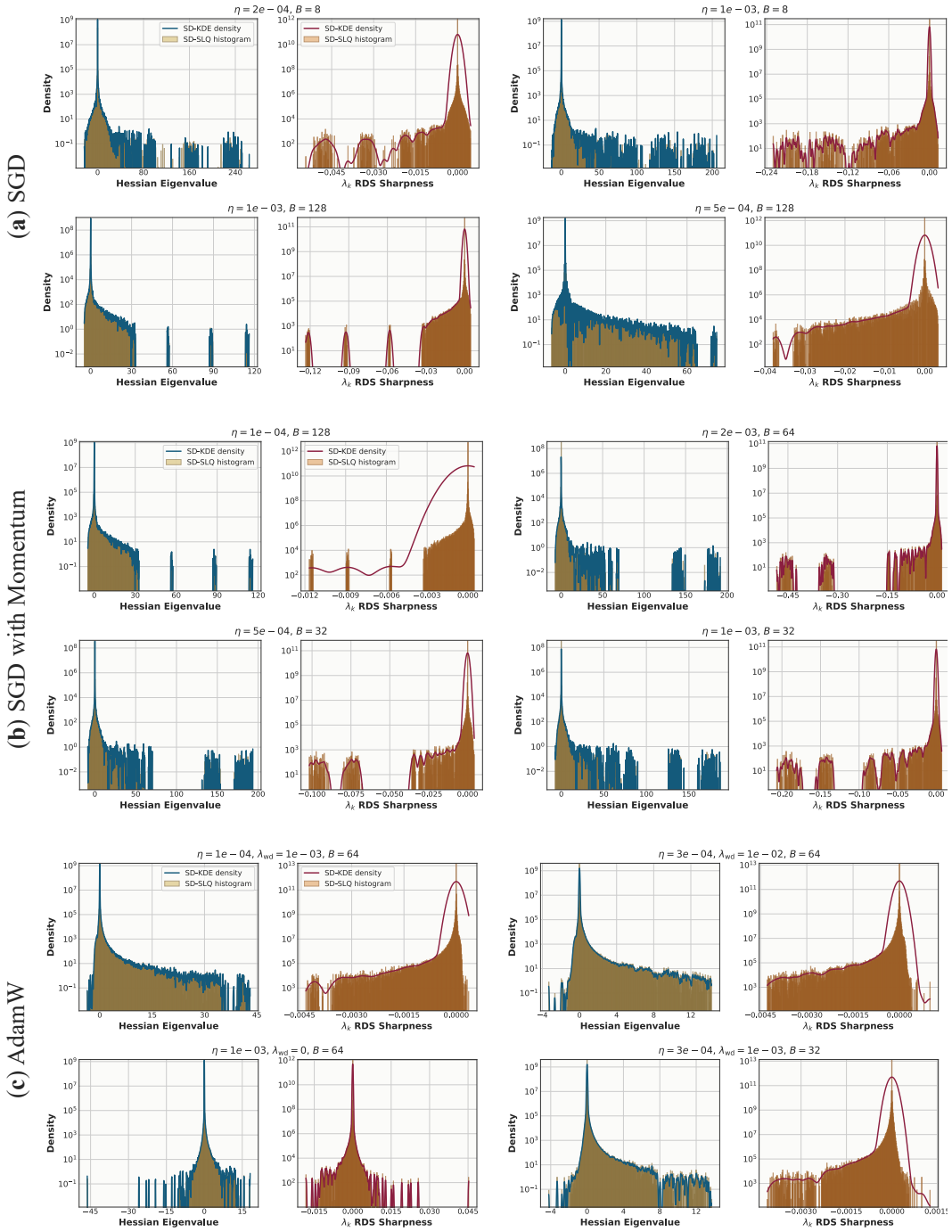

Figure 8: **Hessian Spectra & RDS Sharpness Spectrum.** For selected GPT-2 runs trained with (**a**) SGD, (**b**) SGD with momentum, and (**c**) AdamW, we show the SLQ-based histogram estimate (SD-SLQ) and the kernel-smoothed estimate (SD-KDE) for both the raw Hessian spectrum (**left** in each pair) and the transformed RDS sharpness spectrum $\log|1-\eta\alpha|$ (**right** in each pair). Together, the panels show how optimizer and hyperparameter choices affect both the Hessian spectrum and its induced RDS sharpness spectrum. These examples show that the SLQ histogram and the corresponding kernel-smoothed estimate provide consistent views of the underlying spectrum across a range of training configurations.

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 claims are supported by theory and experiments.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

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

   Answer: [Yes]

   Justification: We discuss the limitations in a dedicated section (see 5).

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

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: We provide concrete assumptions and proofs to our results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

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: For every experimental result, we provide the concrete procedure, including the hyperparameters, datasets, and code required to reproduce it.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying 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: All experiments are conducted on publicly available datasets. We provide the code necessary to reproduce the results with the submission (link in the introduction).

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

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: All the experimental settings are provided in Sec. 4 with additional details in App. D and our code.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

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: [No]

   Justification: Error bars are not reported because it would be too computationally expensive with the training and analysis over hyperparameter grids.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

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: In the **Experimental Setup Details** paragraph of App. D, we detail our computational resources, training and analysis times and total GPU usage throughout the full research project.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

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]

   Yes, we fully adhere to the NeurIPS ethics code and do not have human trial or data-related concerns.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

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

    Answer: [N/A]

    Justification: We do not anticipate any specific positive or negative societal impacts due to the theoretical nature of this work.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

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: Our work is primarily theoretical and does not release high-risk data or models. The experiments are used only to validate the theoretical results on publicly available datasets, so responsible release safeguards of this kind are not applicable.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

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 assets are open source and have been properly cited.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [N/A]

    Justification: Our work is theoretical and tests theory on public available datasets. We do not provide new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

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: Our work proved theoretical results. No crowdsourcing or experiments on human subjects are involved.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

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: We provide new theory and test it empirically.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

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: Our work does not rely on LLMs in any important, original, or non standard compents.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.