Fractal Structure and Generalization Bounds with Intrinsic Dimension for Stochastic Optimization Algorithms

12432036 Haiyu Zhang

Department of Mathematics
Southern University of Science and Technology

August 28, 2025

Outlines

1. Background
2. Preliminaries
3. Existing Results
4. Research Directions and Plans
5. References

Towards Understanding Generalization in Deep Learning

Let the data probability space be $(\mathcal{Z}, \mathcal{F}, \mu_z)$, where $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, $x \in \mathcal{X}$ are features and $y \in \mathcal{Y}$ are labels. Let the training dataset be $S := (z_i)_{1 \leq i \leq n} \sim \mu_z^{\otimes n}$, which consists of n independent and identically distributed (i.i.d.) data points. Let $\mathcal{H} := \{h_w : \mathcal{X} \to \mathcal{Y} \mid w \in \mathbb{R}^d\}$ be a hypothesis class parametrized by weights w.

Towards Understanding Generalization in Deep Learning

The main task of deep learning is to solve a population risk minimization problem

$$\min_{w \in \mathbb{R}^d} \left\{ \mathcal{R}(w) := \mathop{\mathbb{E}}_{z \sim \mu_z} [\ell(w, z)] := \mathop{\mathbb{E}}_{(x,y) \sim \mu_z} [\mathcal{L}(h_w(x), y)] \right\},$$

where ℓ is the composition of the loss function $\mathcal{L} : Y \times Y \to \mathbb{R}$ and $\ell(w, z) = \ell(w, (x, y)) = \mathcal{L}(h_w(x), y)$. In practice, since μ_z is unknown, we minimize the empirical risk

$$\widehat{\mathcal{R}}_S(w) := \frac{1}{n} \sum_{i=1}^{n} \ell(w, z_i)$$

through a learning algorithm $\mathcal{A}$. The worst-case generalization error over a weight set $\mathcal{W} \in E$ is defined as

$$\mathcal{G}_S(\mathcal{W}) := \sup_{w \in \mathcal{W}} \left(\mathcal{R}(w) - \widehat{\mathcal{R}}_S(w) \right).$$

Rethinking the Generalization Bounds

VC Generalization bounds:

$$\mathcal{G} \approx O(\sqrt{\frac{d}{n}}),$$

where d is the VC dimension.
Classical tools from statistical learning theory (e.g., Rademacher complexity, VC-dimension) suggest that highly over-parametrized models should suffer from poor generalization in the absence of explicit regularization.

Rethinking the Generalization Bounds

- However, **neural networks frequently generalize well even when they possess enough capacity to memorize training data (the enigma of generalization).**
- Due to factors such as **data dependency** and the **implicit regularization of optimization algorithms**, not all parameters are equally important or fully utilized during the actual learning process.

$\downarrow$

Generalization bounds with **intrinsic dimension**,
which indicates the "effective" geometric complexity of the model.

Fractal Dimension

Definition (Hausdorff dimension)

For any $s \geq 0$ and $\delta > 0$, the *s-dimensional Hausdorff outer measure* is defined as:

$$\mathcal{H}^s_\delta(X) := \inf \left\{ \sum_{i=1}^{\infty} (\mathrm{diam}\, U_i)^s : X \subset \bigcup_{i=1}^{\infty} U_i,\ \mathrm{diam}(U_i) < \delta \right\}$$

Then, the *s-dimensional Hausdorff measure* is given by:

$$\mathcal{H}^s(X) := \lim_{\delta \to 0} \mathcal{H}^s_\delta(X)$$

The Hausdorff dimension of X is defined as:

$$\dim_{\mathrm{H}}(X) := \inf \{ s \geq 0 : \mathcal{H}^s(X) = 0 \} = \sup \{ s \geq 0 : \mathcal{H}^s(X) = \infty \}$$

Fractal Dimension

Definition (Upper-box dimension)

For each $\delta > 0$, let $N_\delta^d(X)$ be the covering number of X, which is the smallest number of sets of diameter at most δ needed to cover X. The upper box dimension is defined as:

$$\dim_{Box}(X) = \limsup_{\delta \to 0} \left(\frac{\log N_\delta^d(X)}{\log(\frac{1}{\delta})} \right).$$

Minimum Spanning Tree Dimension

A tree $\mathcal{T}$ on X is a connected undirected graph. We represent $\mathcal{T}$ by its set of edges, which are denoted $a \to b$ (or equivalently $b \to a$ as the graph is undirected). For an edge e of the form $a \to b$, we define its length by $|e| = d(a, b)$.
Define the cost of a tree by the sum of the length of its edges, *i.e.*,

$$E_1^d(\mathcal{T}) := \sum_{e \in \mathcal{T}} |e|.$$

Given a finite point set $\mathbf{x} \in X$ as vertices, a **minimum spanning tree** $T(\mathbf{x})$ is defined as a tree with minimal cost.

Minimum Spanning Tree Dimension

Definition (Minimum spanning tree dimension)

Let $\mathbf{x} \subset X$ be a finite set, the α-weighted lifetime sum of $\mathbf{x}$ is

$$E_\alpha^d(\mathbf{x}) := \sum_{e \in \mathcal{T}(\mathbf{x})} |e|^\alpha,$$

with $\alpha \geq 0$. Then the minimum spanning tree dimension is defined as

$$\dim_{MST}(X) := \inf \left\{ \alpha : \exists C \text{ so that } E_\alpha^{MST}(\mathbf{x}) < C \ \forall \text{ finite } \mathbf{x} \subset X \right\}.$$

Persistent Homology Dimension

Theorem (Mattila, 1995, Falconer, 2003)

Let $X \subset \mathbb{R}^d$ be a set equipped with a Borel measure μ, and suppose that μ is s-Ahlfors regular, that is, there exist constants $c_1, c_2 > 0$ and $r_0 > 0$ such that

$$c_1 r^s \leq \mu(B(x,r)) \leq c_2 r^s, \quad \forall x \in \mathrm{supp}(\mu), \quad \forall r \in (0, r_0).$$

Then, the following dimensions of W coincide:

$$\dim_{\mathrm{H}}(X) = \dim_{Box}(X) = s.$$

Persistent Homology Dimension

Definition (Pesistent homology)

Given a sequence of filtered simplicial complex $\{\Sigma_{t_i}\}$, we can construct a persistent chain complex $C_k(\Sigma_k)$ by setting $C_k^i(\Sigma_k) = C_k(\Sigma_{t_i})$ and the chain map $x_k^i : C_k^i(\Sigma_k) \hookrightarrow C_k^{i+1}(\Sigma_k)$. For $i < j$, the (i, j)-persistent homology group of C, denoted $H_*^{i \to j}(C)$, is defined to be the image of the induced homomorphism $h_k^{i \to j} : H_k(C_k^i(\Sigma_k)) \to H_k(C_k^j(\Sigma_k))$.

Persistent Homology Dimension

Remark:
Note h_k are linear maps on homology groups and for $\forall i < j$,

$$h_k^{i \to j} = h_k^{j-1} \circ h_k^{j-2} \circ \cdots \circ h_k^{i},$$

we get an $\mathbb{N}$-indexed persistence module which fits into the diagram

$$H_k(C_k^0(\Sigma_k) \xrightarrow{x_k^0} H_k(C_k^1(\Sigma_k) \xrightarrow{x_k^1} \ldots.$$

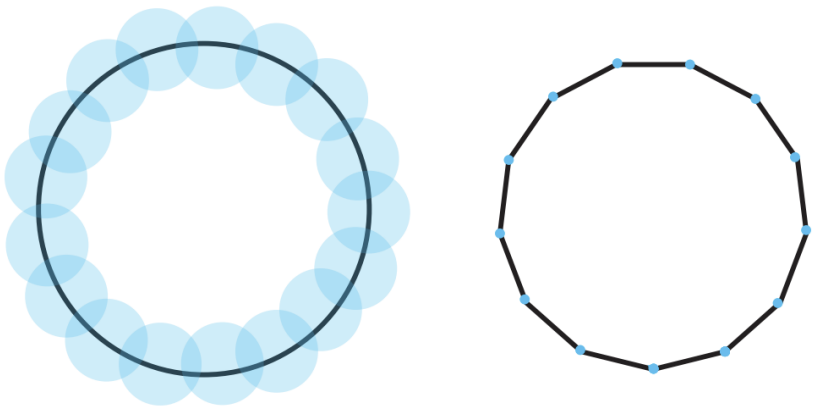
We say that a persistent generator γ in $H_k(C_k^i(\Sigma_k))$ is born at filtration index i if γ does not lie in img h_k^{i-1}; similarly, γ is said to die at filtration index $j > i$ whenever j is the smallest number satisfying $h_k^{j-1}(\gamma) = 0$. By convention, the death index of γ equals $+\infty$ if no such j exists, i.e., if $h_k^{j-1}(\gamma)$ is nonzero for all $j \geq i$. The persistence of γ is defined to be death minus birth, i.e., $(j - i)$.

Persistent Homology Dimension

Definition (Vietoris-Rips complex)

Given a point cloud $Z = \{x_\alpha\} \subseteq \mathbb{R}^n$, the Vietoris-Rips complex R_ϵ is the abstract simplicial complex consisting of k-simplices whose vertices are unordered $(k+1)$-tuples of points $\{x_\alpha\}_0^k$ if and only if

$$d(z_i, z_j) \leq \epsilon.$$



Persistent Homology Dimension

Definition (Persistent homology dimension)

For a finite set $\mathbf{x} \subset X$, the weighted i^{th} homology lifetime sum is defined as follows:

$$E_\alpha^i(\mathbf{x}) = \sum_{\gamma \in \mathrm{PH}_i(\mathcal{VR}(\mathbf{x}))} |I(\gamma)|^\alpha,$$

where $\mathrm{PH}_i(\mathcal{VR}(\mathbf{x}))$ is the i-dimensional persistence module of the Vietoris-Rips complex on $\mathbf{x}$ and $|I(\gamma)|$ is the persistence of some persistent generator γ.
The PH_i-dimension of X is defined as

$$\dim_{PH}^i(X) := \inf \left\{ \alpha \ : \ E_\alpha^i(\mathbf{x}) < C \text{ for some constant } C > 0, \text{ for all finite } \mathbf{x} \subset X \right\}.$$

The Equivalence of Three Definitions of Intrinsic Dimension

Theorem (Kozma et al., 2006)

$$\dim^0_{PH}(X) = \dim_{MST}(X) = \dim_{Box}(X).$$

1 Background

2 Preliminaries

3 Existing Results
- Fractal Structure of Weight Trajectories
- Generalization Bounds with Intrinsic Dimension

4 Research Directions and Plans

5 References

Fractal Structure of Weight Trajectories

We model the SGD learning algorithm as a Feller process which is expressed by the following SDE:

$$\mathrm{dW}_t = -\nabla f(\mathrm{W}_t)\mathrm{d}t + \Sigma_1(\mathrm{W}_t)\mathrm{dB}_t + \Sigma_2(\mathrm{W}_t)\mathrm{dL}_t^{\boldsymbol{\alpha}}(\mathrm{W}_t).$$

where Σ_1, Σ_2 are $d \times d$ matrix-valued functions, B_t denotes the Brownian motion and $\mathrm{L}_t^{\boldsymbol{\alpha}(\cdot)}$ denotes the state-dependent $\boldsymbol{\alpha}$-stable Lévy motion. Let $\{W_t\}_{t\in[0,1]}$ be the solution of SDE, then the weight trajectories $\mathcal{W}_S$ generated by the training set S is defined as

$$\mathcal{W}_S := \left\{ w \in \mathbb{R}^d : \exists t \in [0,1], w = W_t \right\}.$$

Fractal Structure of Weight Trajectories

To illustrate relationship between the heavy-tail property of the learning algorithm (approximated by a decomposable Feller process) and the fractal structure of weight trajectories generated by the algorithm, we have the following theorem.

Theorem (Şimşekli et al., 2020)

Let $\left\{\mathrm{W}^{(S)}\right\}_{S\in\mathcal{Z}^n}$ be a family of Feller processes. Assume that for each S, $\mathrm{W}^{(S)}$ is decomposable at a point w_S with sub-symbol ψ_S. Then for the corresponding weight trajectories $\mathcal{W}_S$, we have

$$\dim_{\mathrm{H}} \mathcal{W}_S \leq \beta_S, \quad \text{where} \quad \beta_S := \inf\left\{\lambda \geq 0 : \lim_{\|\xi\|\to\infty} \frac{|\psi_S(\xi)|}{\|\xi\|^\lambda} = 0\right\}.$$

Fractal Structure of Weight Trajectories

Consider a simple example where the process $W_t^{(S)}$ is selected as a d-dimensional α-stable Lévy process with $d \geq 2$, i.e.

$$\mathrm{d}W_t^{(S)} = \mathrm{d}L_t^{\alpha},$$

then it satisfies the assumptions of theorem with $\beta_S = \alpha$ for all S. In particular, it holds that $\dim_{\mathrm{H}} \mathcal{W}_S = \alpha$. As shown in Figure 1, the fractal structure of weight trajectories tends to be simpler if the stochastic process exhibits a heavier tail.

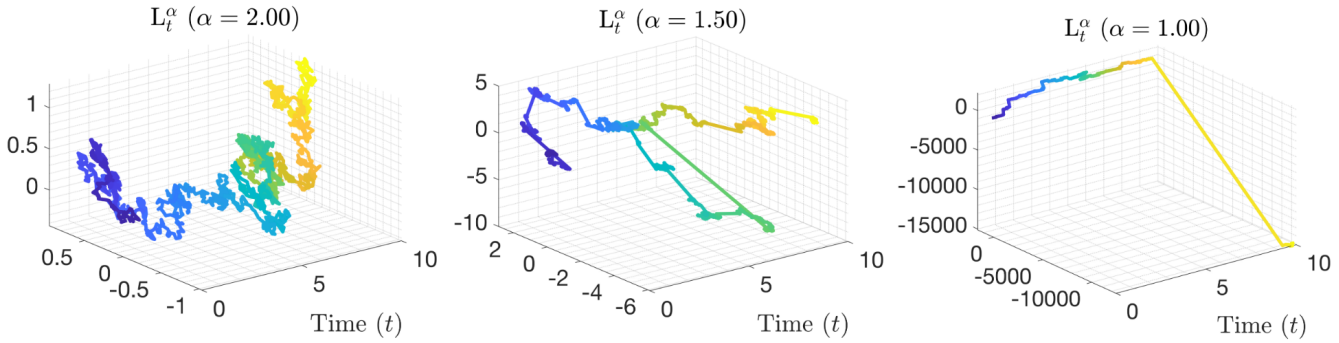


Figure: Weight trajectories of L_t^{α} for $\alpha = 2.0$, 1.5, and 1.0.

1 Background

2 Preliminaries

3 Existing Results
- Fractal Structure of Weight Trajectories
- Generalization Bounds with Intrinsic Dimension

4 Research Directions and Plans

5 References

Mathematical Set-up

Define the parameter space as $\mathbb{R}^d$, equipped with the Borel σ-algebra $\mathcal{B}(\mathbb{R}^d)$. Let E denote the space of closed subsets of $\mathbb{R}^d$, equipped with the Effrös σ-algebra $\mathfrak{E}$. We consider the random weight sets (or weight trajectories) $\mathcal{W} \in E$ and the data-dependent probability distribution on $(E, \mathfrak{E})$. We define that a learning algorithm is a measurable map

$$\mathcal{A} : \mathcal{Z}^n \to \mathcal{P}(E),$$

mapping S to a data-dependent probability measure $\rho_S \in \mathcal{P}(E)$, where $\mathcal{P}(E)$ is the set of all probability measures defined on the measurable space $(E, \mathfrak{E})$. We formalize the **data-dependent probability distribution as a Markov kernel**

$$K : \mathcal{Z}^n \times \mathfrak{E} \to [0, 1]$$

such that:

1. For each $S \in \mathcal{Z}^n$, $A \mapsto K(S, A)$ is a probability measure on $(E, \mathfrak{E})$ (i.e., ρ_S).
2. For each $A \in \mathfrak{E}$, $S \mapsto K(S, A)$ is a measurable function on $(\mathcal{Z}^n, \mathcal{F}^n)$.

PAC-Bayesian Theory on Random Sets

Consider the probability space $(\mathcal{Z}^n \times E, \mathcal{F}^{\otimes n} \otimes \mathfrak{E}, \mathbb{P})$ with probability distribution denoted by $\mu_z^{\otimes n} \otimes \rho_S$. To be more precise, for all $A \in \mathcal{F}^{\otimes n} \otimes \mathfrak{E}$

$$\mathbb{P}_{S,\mathcal{W}\sim\rho_S}(A) = \int_{\mathcal{Z}^n} \rho_S(\{\mathcal{W} : (S, \mathcal{W}) \in A\})\mu_z^{\otimes n}(dS).$$

Definition (Priors and posteriors)

A prior, π, is a data-independent probability distribution on $(E, \mathfrak{E})$. A family of posteriors $(\rho_S)_{S\in\mathcal{Z}^n}$ is defined as a Markov kernel on $E \times \mathcal{Z}^n$. We further require that the posteriors are absolutely continuous with respect to the prior, i.e. $\rho_S \ll \pi$, $\mu_z^{\otimes n}$-almost surely.

PAC-Bayesian Theory on Random Sets

Theorem (PAC-Bayesian bounds for random sets)

Let $\Phi : E \times \mathcal{Z}^n \longrightarrow \mathbb{R}$ be a measurable function with respect to $\mathfrak{E} \otimes \mathcal{F}^{\otimes n}$.Then we have for any $\zeta \in (0,1)$:

$$\mathbb{P}_S \left(\mathbb{E}_{\mathcal{W}\sim\rho_S} \Phi(\mathcal{W}, S) \right.$$
$$\left. \leq \mathbf{KL}\left(\rho_S \| \pi\right) + \log(1/\zeta) + \log \mathbb{E}_S \mathbb{E}_{\mathcal{W}\sim\pi} \left[e^{\Phi(\mathcal{W},S)} \right] \right) \geq 1 - \zeta,$$

as well as the disintegrated bound

$$\mathbb{P}_{S,\mathcal{W}\sim\rho_S} \left(\Phi(\mathcal{W}, S) \right.$$
$$\left. \leq \log \left(\frac{d\rho_S}{d\pi}(\mathcal{W}) \right) + \log(1/\zeta) + \log \mathbb{E}_S \mathbb{E}_{\mathcal{W}\sim\pi} \left[e^{\Phi(\mathcal{W},S)} \right] \right) \geq 1 - \zeta.$$

Generalization Bounds with Intrinsic Dimensions

Assumption 1 [Bounded measurable loss]
The loss function $\ell : \mathbb{R}^d \times \mathcal{Z}$ is measurable and bounded in $[0, B]$, for some constant $B > 0$.

Assumption 2 [Supremum measurability]
Both ℓ and $(E, \mathfrak{E})$ have enough regularity so that, for any coefficients $b, a_1, \ldots, a_n \in \mathbb{R}$, the following is $\mathbb{E} \otimes \mathcal{F}^{\otimes n}$-measurable:

$$(\mathcal{W}, S) \longmapsto \sup_{w \in \mathcal{W}} \sum_{i=1}^{n} \left(a_i \ell\left(w, z_i\right) - b\mathcal{R}(w)\right)$$

Assumption 3 [(q, L, d)-Lipschitz continuity]
ℓ is (q, L, d)-Lipschitz in w on $(\mathbb{R}^d, d)$, i.e.
$\|\mathbf{L}_S(w) - \mathbf{L}_S(w')\|_q \leq L n^{1/q} d(w, w')$ for $\forall w, w' \in \mathbb{R}^d$, where the data-dependent map $L_S : \mathbb{R}^d \longrightarrow \mathbb{R}^n$ is defined by
$L_S(w) = (\ell(w, z_1), \ldots, \ell(w, z_n))$.

Generalization Bounds with Intrinsic Dimensions

Assumption 4 [Measurability of topological complexity]
The covering number, packing number and α-weighted lifetime sum are all measurable with respect to $\mathcal{F}^{\otimes n} \otimes \mathfrak{E}$.

Assumption 5 [Uniformly convergence in n] Given S, then for all $\epsilon > 0$,

$$\sup_{n \in \mathbb{N}^\star} \rho_{S_n} \left(\left| \sup_{0<r<\delta} \frac{\log\left(|N_r(\mathcal{W})|\right)}{\log(1/r)} - \dim_{Box}(\mathcal{W}) \right| \geq \epsilon \right) \xrightarrow[\delta \to 0]{} 0.$$

Generalization Bounds with Intrinsic Dimensions

Let

$$\Phi_\lambda(\mathcal{W}, S) = \lambda G_S(\mathcal{W}) - 2\lambda \operatorname{Rad}_S(\mathcal{W}), \quad \lambda > 0,$$

where $\operatorname{Rad}_S(\mathcal{W})$ is the Rademacher complexity $\mathcal{W}$.

Theorem

Suppose that Assumptions 1, 2, 3, 4 and 5 hold, then for any $\gamma > 0$ and $\epsilon > 0$, there exists $n_{\gamma,\epsilon}$ s.t. for all $n > n_{\gamma,\epsilon}$, we have

$$\mathbb{P}_{S,\mathcal{W}\sim\rho_S}\left(G_S(\mathcal{W}) \leq \frac{2L}{n} + 2B\sqrt{\frac{2\left(\dim_{Box}(\mathcal{W}) + \epsilon\right)\left(\log(n)\right)}{n}} + 3B\sqrt{\frac{\mathrm{I}_\infty\left(S, \mathcal{W}\right) + \log(1/\zeta)}{2n}} \right) \geq 1 - \zeta - \gamma,$$

Generalization Bounds with Intrinsic Dimensions

Theorem

Suppose that Assumptions 1, 2, 3, 4 and 5 hold, then for any $\gamma > 0$ and $\epsilon > 0$, there exists $n_{\gamma,\epsilon}$ s.t. for all $n > n_{\gamma,\epsilon}$, we have

$$\mathbb{P}_{S,\mathcal{W}\sim\rho_S}\left(G_S(\mathcal{W}) \leq \frac{2L}{n} + 2B\sqrt{\frac{2\left(\dim^0_{PH}(\mathcal{W}) + \epsilon\right)(\log(n))}{n}}\right.$$
$$\left. + 3B\sqrt{\frac{\mathrm{I}_\infty(S,\mathcal{W}) + \log(1/\zeta)}{2n}}\right) \geq 1 - \zeta - \gamma.$$

These theorems show that the fractal dimension of the weight trajectories acts as a "capacity metric" and the generalization error is therefore directly linked to this metric.

Generalization Bounds with Intrinsic Dimensions

Remark 1:
We observe that the upper bound of the generalization error can be decomposed as

$$\text{Complexity Term} + \text{Data Dependence Term} + \text{Confidence Term}.$$

Here, the Complexity Term primarily quantifies the fitting capacity of the hypothesis class, while the Data Dependence Term mainly measures the degree to which $\mathcal{W}_S$ depends on S. Together, these two terms assess the overfitting of the hypothesis class to the data from different perspectives.

Generalization Bounds with Intrinsic Dimensions

Remark 2:
We know that if the weight trajectories is regular enough, heavier-tails imply less generalization error. That is

$$\text{heavier-tailed noise (smaller } \alpha) \implies \dim_{\mathrm{Box}}(W_S) \downarrow \implies G_S(W) \downarrow .$$

Intuitively, heavier-tailed stochastic processes in SGD, such as α-stable Lévy processes with small α, promote better generalization by enabling broader exploration of the loss landscape. The occasional large updates help escape sharp minima and bias the optimization toward flatter regions, which are known to generalize better.

Motivation

There are limitations in existing results ...

- These bounds are implicit, in the sense that they cannot be related to algorithm hyperparameters, problem geometry, or data, which causes a disparity between theory and practice and provides only limited insights for practical application.
- The correlation observed between intrinsic dimension and generalization gap is significantly influenced by hyperparameter values, especially the learning rates.
- The term in the generalization bounds involving mutual information between the training data and the optimization trajectory is less explored and the relationship between the complexity term and data dependence term remains unclear.

Fractal Structure of Invariant Measures

We follow the work [Camuto et al., 2021].
Consider the **stochastic gradient descent(SGD)** algotithm as an random iterative function system (IFS):

$$w_k = w_{k-1} - \eta \nabla \tilde{\mathcal{R}}_k \left(w_{k-1}\right),$$
$$\text{where} \quad \nabla \tilde{\mathcal{R}}_k(w) := \nabla \tilde{\mathcal{R}}_{\Omega_k}(w) := (1/b) \sum_{i \in \Omega_k} \nabla \ell \left(w, z_i\right).$$

A simple example:
Consider a 1 -dimensional quadratic problem with loss function $\ell\left(w, z_1\right) = \frac{w^2}{2}$ and $\ell\left(w, z_2\right) = \frac{w^2}{2-w}$. Let $\Omega_k \subset \{1, 2\}$ be uniformly random with batch-size $b = 1$ and the step-size $\eta = \frac{2}{3}$. Then the invariant set under the iteration is the famous "middle-third Cantor set" and the stationary distribution is the Cantor distribution.

Research plan (theoretical part)

- Explore the generalization bounds with **"effective" hypothesis complexity** the persistent homology dimension of invariant measure obtained by an IFS $w_\infty \sim \mu$ defined in [Adams et al, 2019]:

$$\dim_{\mathrm{PH}}^{i}(\mu) = \inf_{d>0} \Big\{ d \mid \exists \text{ constant } C(i,\mu,d) \text{ s.t.} \\ \lim_{n\to\infty} \mathbb{P}\left[L^i\left(X_n\right) \leq C n^{(d-1)/d} \right] = 1 \Big\},$$

where $X_n \subseteq X$ be a random sample of n points from X distributed according to μ, and let $L^i\left(X_n\right)$ be the sum of the lengths of the intervals in the i-dimensional persistent homology for X_n.

Research plan (theoretical part)

- Reconstructing generalization bounds with **hypothesis stability**. My preliminary findings indicate that the positive correlation between generalization gap and intrinsic dimension depends strongly on hyperparameters such as the learning rate, suggesting potential interactions with algorithmic stability. We aim to establish a connection between the information-theoretic quantities between the loss and the data and the algorithmic stability proposed by [Bousquet and Elisseeff, 2002], in order to derive generalization.

Research plan (theoretical part)

- Explicitly establish the relationship between the PH dimension, the hyperparameters of the stochastic optimization algorithm and the generalization gap. By modeling stochastic gradient descent (SGD) as an iterative random function system or a stochastic process, we can investigate how the persistent homology (PH) dimension of the resulting weight trajectory or invariant measure depends on algorithmic hyperparameters such as learning rate or optimizer type. This approach aims to disentangle the interplay between intrinsic dimension, the geometric structure of the loss landscape (e.g., flatness of generalizing minima), and the implicit regularization induced by optimization dynamics. Intuitively, the intrinsic dimension should scale with the learning rate, but whether this reflects an implicit regularization mechanism that enhances generalization remains an open problem.

Research plan (experimental part)

1. Empirically support the theoretical results.
2. Other questions:
   - What is the appropriate ambient dimension for a given intrinsic dimension?
   - How to design a regularizer (with respect to PH dimension) to the optimization problem?

Research plan (experimental part)

We investigate the **Grokking phenomenon** by training a 3-layer MLP on a subset of the MNIST dataset consisting of 1000 samples and test on the whole dataset (10000 samples). The model architecture includes two hidden layers with ReLU activations and an output layer with 10 classes. The network is trained using the AdamW optimizer with a mean squared error (MSE) loss over 500,000 steps.

Research plan (experimental part)

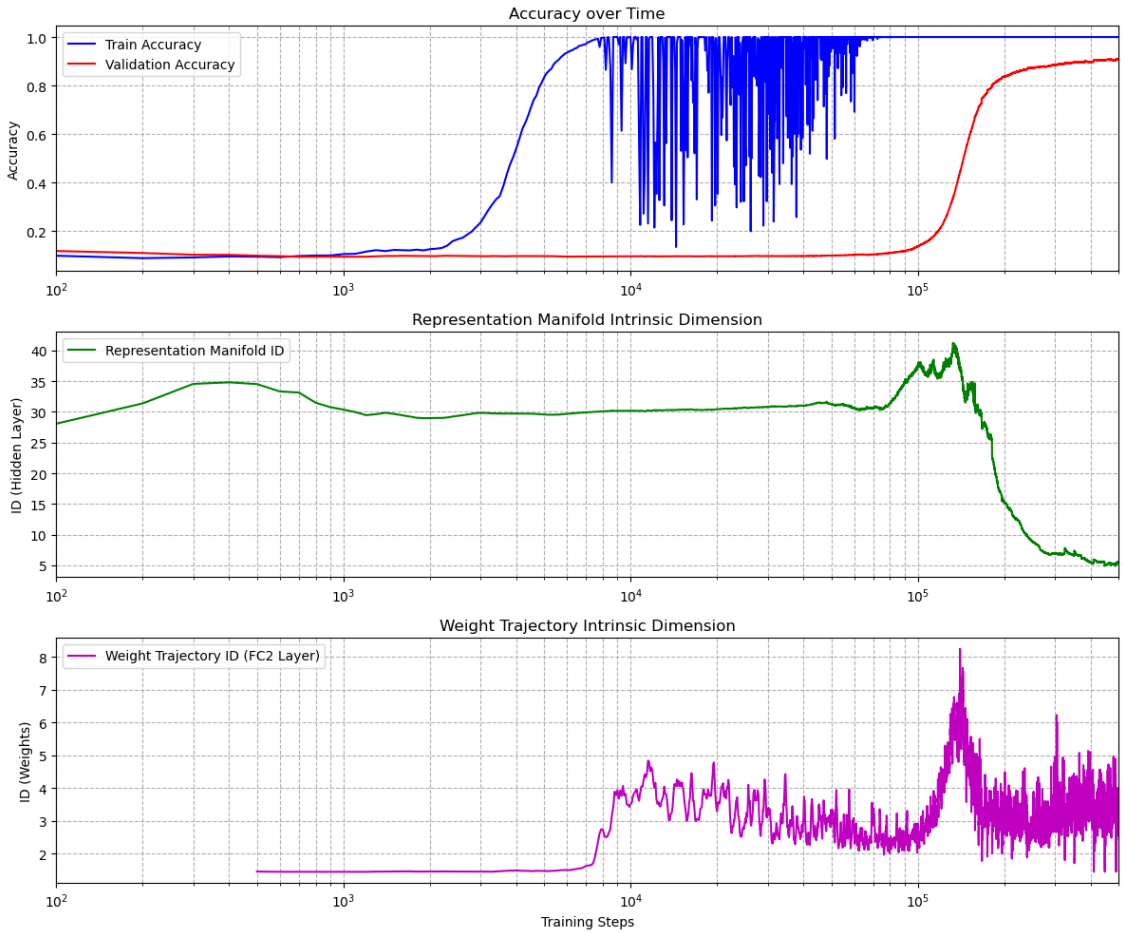


Figure: The Evolution of Intrinsic Dimension during the Grokking Phenomenon

Main References

- Adams, H., Aminian, M., Farnell, E., Kirby, M., Peterson, C., Mirth, J., Neville, R., Shipman, P., Shonkwiler, C. (2020). A fractal dimension for measures via persistent homology.
- Andreeva, R., Dupuis, B., Sarkar, R., Birdal, T., Şimşekli, U. (2024). Topological Generalization Bounds for Discrete-Time Stochastic Optimization Algorithms.
- Ansuini, A., Laio, A., Macke, J. H., Zoccolan, D. (2019). Intrinsic dimension of data representations in deep neural networks.
- Birdal, T., Lou, A., Guibas, L. Intrinsic Dimension, Persistent Homology and Generalization in Neural Networks.
- Camuto, A., Deligiannidis, G., Erdogdu, M. A., Gürbüzbalaban, M. . Fractal Structure and Generalization Properties of Stochastic Optimization Algorithms.
- Dupuis, B., Deligiannidis, G., Şimşekli, U. Generalization Bounds using Data-Dependent Fractal Dimensions.

- Dupuis, B., Viallard, P., Deligiannidis, G., Simsekli, U. Uniform Generalization Bounds on Data-Dependent Hypothesis Sets via PAC-Bayesian Theory on Random Sets.
- Kozma, G., Lotker, Z., Stupp, G. (2005). The minimal spanning tree and the upper box dimension.
- Rivasplata, O., Kuzborskij, I., Szepesvari, C., Shawe-Taylor, J. (2020). PAC-Bayes Analysis Beyond the Usual Bounds.
- Schilling, R. L. (1998). Feller Processes Generated by Pseudo-Differential Operators: On the Hausdorff Dimension of Their Sample Paths.
- Schweinhart, B. (2020). Fractal Dimension and the Persistent Homology of Random Geometric Complexes.
- Schweinhart, B. (2021). Persistent Homology and the Upper Box Dimension.
- Şimşekli, U., Sener, O., Deligiannidis, G., Erdogdu, M. A. (2021). Hausdorff dimension, heavy tails, and generalization in neural networks.
- Tan, C. B., GarcÃa-Redondo, I., Wang, Q., Bronstein, M. M., Monod, A. (2024). On the Limitations of Fractal Dimension as a Measure of Generalization.