On the Limitations of Fractal Dimension
as a Measure of Generalization

Charlie B. Tan[*]
University of Oxford

Inés García-Redondo[*]
Imperial College London

Qiquan Wang[*]
Imperial College London

Michael M. Bronstein
University of Oxford / Aithyra

Anthea Monod
Imperial College London

Abstract

Bounding and predicting the generalization gap of overparameterized neural networks remains a central open problem in theoretical machine learning. There is a recent and growing body of literature that proposes the framework of fractals to model optimization trajectories of neural networks, motivating generalization bounds and measures based on the fractal dimension of the trajectory. Notably, the persistent homology dimension has been proposed to correlate with the generalization gap. This paper performs an empirical evaluation of these persistent homology-based generalization measures, with an in-depth statistical analysis. Our study reveals confounding effects in the observed correlation between generalization and topological measures due to the variation of hyperparameters. We also observe that fractal dimension fails to predict generalization of models trained from poor initializations. We lastly reveal the intriguing manifestation of model-wise double descent in these topological generalization measures. Our work forms a basis for a deeper investigation of the causal relationships between fractal geometry, topological data analysis, and neural network optimization.

1 Introduction

Deep learning enjoys widespread empirical success despite limited theoretical support. Measures from statistical learning theory, such as Rademacher complexity [Bartlett and Mendelson, 2002] and VC-Dimension [Hastie et al., 2009], indicate that without explicit regularization, over-parameterized models will generalize poorly. In contrast, neural networks are able to generalize strongly despite having sufficient capacity to simply memorize their training data [Liu et al., 2020, Zhang et al., 2021]. Remarkably, neural networks often exhibit improved generalization for increases in capacity [Nakkiran et al., 2021]. Describing the generalization behavior of neural networks therefore requires the development of novel learning theory [Zhang et al., 2021]. Ultimately, deep learning theory seeks to define generalization bounds for given experimental configurations [Valle-Pérez and Louis, 2020], such that the generalization error of a given model can be bounded and predicted [Jiang et al., 2020].

Generalization is typically attributed to the implicit bias of gradient-based optimization. A number of works have considered the geometry of generalizing solutions within parameter space [Dinh et al., 2017, Garipov et al., 2018], and the bias of optimization methods towards such solutions [He et al., 2019, Izmailov et al., 2018]. Şimşekli et al. [2020] propose *random fractal* structure for neural network optimization trajectories and compute generalization bounds based on *fractal dimensions*. However, this work requires rigid topological and statistical conditions on the optimization trajectory as well as the learning algorithm. Subsequent work by Birdal et al. [2021] proposes the use of
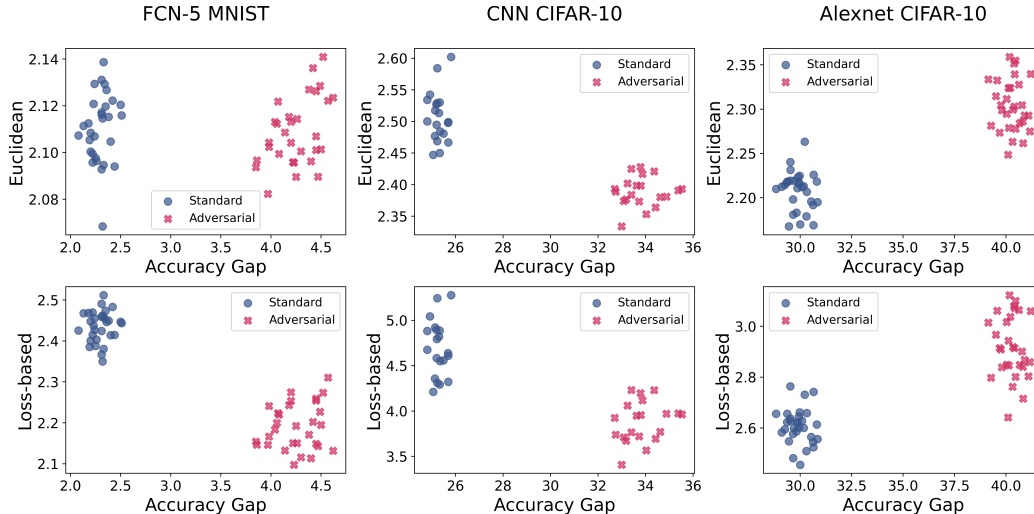

Figure 1: **Adversarial initialization is a failure mode for PH dimension-based generalization measures**. Training models from an adversarial initialization leads to higher accuracy gap than for models trained from random initialization. Both PH dimensions fail to correctly attribute higher values to the poorly generalizing models on FCN-5 MNIST and CNN CIFAR-10.

persistent homology (PH) dimension [Adams et al., 2020]—a measure of fractal dimension deriving from topological data analysis (TDA)—to relax these assumptions. They propose an efficient procedure for estimating PH dimension, and apply this both as a measure of generalization and as a scheme for explicit regularization. Dupuis et al. [2023] extend this approach using a data-dependent pseudometric to further relax continuity assumptions on the network loss.

Our paper constitutes an extended empirical evaluation of the performance and viability of these proposed topological measures of generalization; in particular, robustness and failure modes are explored in a wider range of experiments than those considered by Birdal et al. [2021] and Dupuis et al. [2023]. Our main contributions are as follows:

- We reproduce the learning rate/batch size grid experiments of Dupuis et al. [2023], observing comparable correlation coefficients, aside from the loss-based PH dimension on AlexNet CIFAR-10 where models trained with high learning rate are attributed high dimension values despite having small generalization gap.
- We extend the statistical analysis of Dupuis et al. [2023] to include partial correlations. We observe that in some cases learning rate has a significant influence on observed correlation between PH dimensions and generalization gap for fixed batch sizes.
- We further conduct a conditional independence test using the conditional mutual information [Jiang et al., 2020], observing that both Euclidean and loss-based PH dimension are conditionally independent of generalization gap on MNIST.
- We train models of varying generalization gap using adversarial initialization [Liu et al., 2020]. As presented in Figure 1, we observe that both dimensions fail to correctly attribute high values to poorly generalizing models for some architectures and datasets.
- We train a CNN architecture at a range of width multipliers to reproduce the model-wise double descent of Nakkiran et al. [2021]. Neither PH dimension correctly correlates with generalization gap in this setting. Interestingly, by correlating with test accuracy, double descent manifests in Euclidean PH dimension.

2 Background

Following Dupuis et al. [2023], let $(\mathcal{Z}, \mathcal{F}_{\mathcal{Z}}, \mu_{\mathcal{Z}})$ be the data space, where $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, and $\mathcal{X}, \mathcal{Y}$ represent the feature and label spaces respectively. We aim to learn a parametric approximation $h_w : \mathcal{X} \times \mathcal{W} \to \mathcal{Y}$ of the unknown data generating distribution $\mu_{\mathcal{Z}}$ from a finite set of i.i.d. training points $S := \{z_1, \ldots, z_n\} \sim \mu_{\mathcal{Z}}^{\otimes n}$. The quality of our parametric approximation is measured using a loss function $\mathcal{L} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ composed with the parametric approximation $\ell(\omega, z) := \mathcal{L}(h_\omega(x), y)$.

The learning task then amounts to solving an optimization problem over parameters $w \in \mathbb{R}^d$, where we seek to minimize the *empirical risk* $\hat{\mathcal{R}}(w, S) := \frac{1}{n} \sum_{i=1}^{n} \ell(w, z_i)$ over a finite set of training points. To measure performance on unseen data samples, we consider the *population risk* $\mathcal{R}(w) := \mathbb{E}_z[\ell(w, z)]$ and define the *generalization gap* to be the difference of the population and empirical risks $\mathcal{G}(S, w) := |\mathcal{R}(w) - \hat{\mathcal{R}}(S, w)|$. For a given training dataset and some initial value for the weights $w_0 \in \mathbb{R}^d$ we refer to the optimization trajectory as $\mathcal{W}_S$.

2.1 Persistent Homology and Fractal Dimension

Fractals arise in recursive processes [Mandelbrot, 1983, Prähofer and Spohn, 2000]; chaotic dynamical systems [Briggs, 1992, Mandelbrot et al., 2004]; and real-world data [Mandelbrot, 1967, Coleman and Pietronero, 1992, Falconer, 2007, Pietronero and Tosatti, 2012]. A key characteristic is their *fractal dimension*, first introduced as the *Hausdorff dimension* [Hausdorff, 1918]. Due to its computational complexity, more efficient measures such as the *box-counting dimension* Sarkar and Chaudhuri [1994] were later developed. An alternative fractal dimension can also be defined in terms of *minimal spanning trees* of finite metric spaces [Kozma et al., 2006]. A recent line of work by Adams et al. [2020] and Schweinhart [2021, 2020] extended and reinterpreted fractal dimension using PH. Originating from algebraic topology, PH provides a systematic framework for capturing and quantifying the multi-scale topological features in complex datasets through a topological summary called the *persistence diagram* or *persistence barcode*; further details on PH are given in Appendix A and a more comprehensive exposition of fractal dimension can be found in Appendix B.

We follow the approach by Schweinhart [2021] to define a PH-based fractal dimension. Let $\mathbf{x} = \{x_1, \ldots, x_n\}$ be a finite subset of a metric space (X, ρ). Let $\mathrm{PH}_i(\mathbf{x})$ be the persistence diagram obtained from the PH of dimension i computed from the Vietoris–Rips filtration and define

$$E_\alpha^i(\mathbf{x}) := \sum_{(b,d) \in \mathrm{PH}_i(\mathbf{x})} (d - b)^\alpha. \tag{1}$$

Definition 2.1 ([Schweinhart, 2020]). Let S be a bounded subset of a metric space (X, ρ). The *ith PH dimension* (PH_i-dim) for the Vietoris–Rips complex of S is

$$\dim_{\mathrm{PH}}{}^i(S) := \inf \left\{ \alpha : \exists\, C > 0 \text{ s.t. } E_\alpha^i(\mathbf{x}) < C, \, \forall\, \mathbf{x} \subset S \text{ finite subset} \right\}.$$

Fractal dimensions need not be well-defined for all subsets of a metric space. However, under a certain regularity condition (*Ahlfors regularity*), the Hausdorff and box-counting dimensions are well defined and coincide [Falconer, 2007]. Additionally, for any metric space, the minimal spanning tree is equal to the upper box dimension [Kozma et al., 2006]. The relevance of PH appears when considering the minimal spanning tree in fractal dimensions. Specifically, there is a bijection between the edges of the Euclidean minimal spanning tree of a finite metric space $\mathbf{x} = \{x_1, \ldots, x_n\}$ and the points in the persistence diagram $\mathrm{PH}_0(\mathbf{x})$ obtained from the Vietoris–Rips filtration. This automatically gives the equivalence $\dim_{\mathrm{PH}}{}^0(S) = \dim_{\mathrm{MST}}(S) = \dim_{\mathrm{box}}(S)$.

2.2 Fractal Dimension-Based Generalization Bounds

Şimşekli et al. [2020] empirically observe that the gradient noise exhibits heavy-tailed behavior, which they use to model stochastic gradient descent (SGD) as a discretization of a *decomposable Feller process*. They also impose initialization with zeros; that ℓ is bounded and Lipschitz continuous in w; and that $\mathcal{W}_S$ is bounded and Ahlfors regular. In this setting, they compute two bounds for the worst-case generalization error, $\max_{w \in \mathcal{W}_S} \mathcal{G}(S, w)$, in terms of the Hausdorff dimension of $\mathcal{W}_S$. They first prove bounds related to covering numbers (used to define the upper-box counting dimension) and then use Ahlfors regularity to link the bounds to the Hausdorff dimension.

Subsequently, Birdal et al. [2021] further develop the bounds in Şimşekli et al. [2020] by reformulating them in terms of the 0-dimensional PH dimension [Schweinhart, 2021] of $\mathcal{W}_S$. The link between the upper-box dimension and the 0-dimensional PH dimension, which is the cornerstone of their proof, only requires boundedness of $\mathcal{W}_S$ (which is also one of the assumptions by Şimşekli et al. [2020]), thus eliminating the Ahlfors regularity condition. In order to estimate the PH dimension, they prove (see Proposition 2, [Birdal et al., 2021]) that for all $\epsilon > 0$ there exists a constant $D_{\alpha,\epsilon}$ such that

$$E_\alpha^0(W_n) \leq D_{\alpha,\epsilon}\, n^\beta, \tag{2}$$

3

where $\beta := \frac{\dim_{\mathrm{PH}}^0(\mathcal{W}_S)+\epsilon-\alpha}{\dim_{\mathrm{PH}}^0(\mathcal{W}_S)+\epsilon}$ for all $n \geq 0$, all i.i.d. samples W_n with n points on the optimization trajectory $\mathcal{W}_s$, and $E_\alpha^0(W_n)$ as defined in (7). Using this result, they estimate and bound the PH-dimension by fitting a power law to the pairs $(\log(n), \log(E_1^0(W_n)))$. They then use (2) to estimate $\dim_{\mathrm{PH}}^0(\mathcal{W}_s) \approx \frac{\alpha}{1-m}$, where m is the slope of the regression line. Concurrently, Camuto et al. [2021] take a different, non-topological approach by studying stationary distributions of Markov chains and describing the optimization algorithm as an iterated function system (IFS). They establish generalization bounds with respect to the upper Hausdorff dimension of a limiting measure. Most recently, Dupuis et al. [2023] further develop the topological approach by Birdal et al. [2021] to circumvent the Lipschitz condition on the loss function that was required in all previous works by obtaining a bound depending on a data-driven pseudometric ρ_S instead of the $\mathbb{R}^d$ Euclidean metric,

$$\rho_S(w,w') := \frac{1}{n} \sum_{i=1}^n |\ell(w,z_i) - \ell(w',z_i)|, \quad \forall\, w,\, w' \in \mathbb{R}^d. \tag{3}$$

They derive bounds for the worst-case generalization gap, where the only assumption is that the loss $\ell : \mathbb{R}^d \times \mathcal{Z} \to \mathbb{R}$ is continuous in both variables and uniformly bounded by some $B > 0$. These bounds are established with respect to the upper-box dimension of the set $\mathcal{W}_S$ using ρ_S (see Theorem 3.9. [Dupuis et al., 2023]). They additionally prove that for pseudometric-bounded spaces, the corresponding upper-box counting dimension coincides with the 0-dimensional PH-dimension, which they estimate as in Birdal et al. [2021].

3 Experimental Setup

Our experiments closely follow the setting of Dupuis et al. [2023]. We train with SGD until convergence, then continue for 5000 additional iterations to obtain a sample optimization trajectory $\{w_k : 0 < k \leq 5000\}$ about the local minimum attained. We then compute the 0-PH dimension using both the Euclidean metric and the loss-based pseudometric (3) to obtain the generalization measures of Birdal et al. [2021] and Dupuis et al. [2023]. In keeping with the assumptions of this theory, we omit explicit regularization such as dropout or weight decay, and maintain constant learning rates in all experiments. Following Dupuis et al. [2023] we define the generalization gap as the the absolute accuracy gap for classification tasks, and the absolute loss gap in regression tasks. Further details on the experimental setup, expanding on this section, are provided in Appendix C and a note on the stability of PH dimension estimates post-convergence can be found in Appendix D.

Datasets and Architectures. We employ the same datasets and architectures as Dupuis et al. [2023]: (i) fully-connected network of 5 (FCN-5) and 7 (FCN-7) layers on the California housing dataset (CHD) [Kelley Pace and Barry, 1997]; (ii) FCN-5 and FCN-7 on the MNIST dataset [Lecun et al., 1998]; and (iii) AlexNet [Krizhevsky et al., 2017] on the CIFAR-10 dataset [Krizhevsky, 2009]. We additionally conduct experiments on a 5-layer convolutional neural network, defined by Nakkiran et al. [2021] as *standard CNN*, trained on CIFAR-10 and CIFAR-100, with batch normalization removed to align with the bound assumptions.

Overview of Experiments. We divide our experiments into three categories. In the first, we replicate the 6×6 grid of learning rates and batch sizes considered in Dupuis et al. [2023]. We use these results to perform a statistical analysis of the correlation between PH dimension (both Euclidean and loss-based) and generalization error, particularly exploring the influence of the hyperparameters. Second, we use *adversarial pre-training* as a method to generate poorly generalizing models [Liu et al., 2020]. Finally, we consider model-wise double descent, the phenomenon in which test accuracy is non-monotonic with respect to increasing model parameters [Nakkiran et al., 2021].

Computational Resources. All experiments were run on high performance computing clusters using GPU nodes with Quadro RTX 6000 (128 CPU cores) or NVIDIA H100 (192 CPU cores). The runtime for training and computing the PH dimension vary for different architectures, datasets, and hardware used, with the longest experiments taking several hours.

4 Grid Experiments and Correlations

We first reproduce the experiments of Dupuis et al. [2023], wherein learning rate and batch size are defined in a 6×6 grid as defined in Appendix C. In Figure 2 we present the results of these
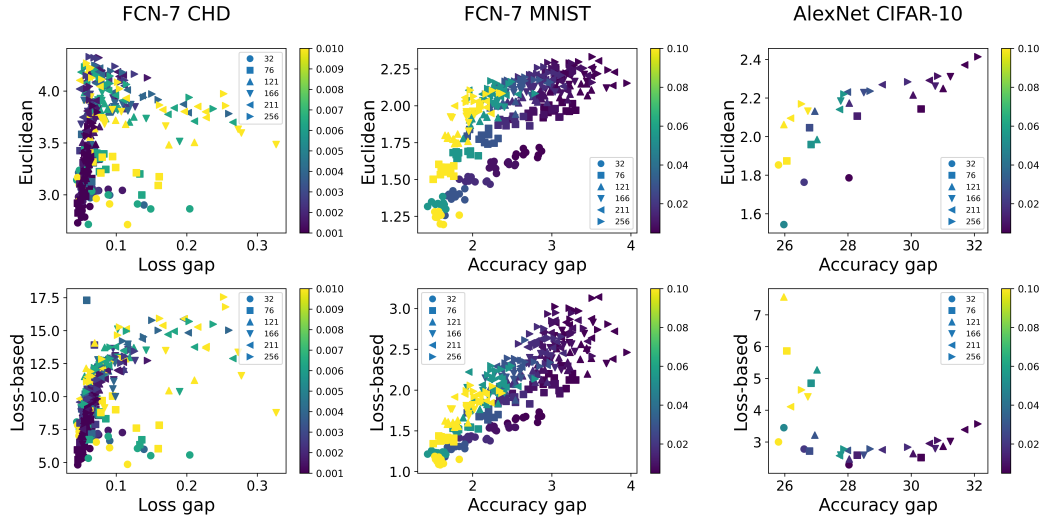
Figure 2: Learning rate/batch size grid results. Euclidean (top) and loss-based (bottom) PH dimension plotted against generalization gap for range of learning rates and batch sizes.

experiments for the PH dimensions in the FCN-7 and AlexNet experiments, additional results for the FCN-5 models and other generalization measures are presented in Appendix E.

4.1 Correlation Analysis

In Table 1, we present correlation coefficients between the PH dimensions (Euclidean and loss-based) and generalization gap. As in Dupuis et al. [2023], we present Spearman's rank correlation coefficient ρ; the mean granulated Kendall rank correlation coefficient Ψ [Jiang et al., 2020]; and the standard Kendall rank correlation coefficient τ. We also compute correlations with the ℓ^2 norm of the final parameter vector $\|w_{5000}\|_2$, and the learning rate/batch size ratio. We emphasize the learning rate/batch size ratio is not a measure of generalization as it is not a measurable experimental output.

The results on CHD and MNIST align with those reported by Dupuis et al. [2023]: both the Euclidean and loss-based measures have positive correlation with generalization gap, with the correlation of the loss-based being slightly stronger than that of the Euclidean. However, for the AlexNet CIFAR-10 experiment, we obtain negative correlations for the loss-based measure, in contrast to the theory and results of Dupuis et al. [2023]. Observing the results in Figure 2, we see this is due to several points with high learning rate achieving very high PH dimension. We are unable to determine why these points appear in our experiments but not prior studies. However, we assert that all points considered attain 100% training accuracy hence meet the convergence assumption of Dupuis et al. [2023].

The ℓ^2 norm has the strongest absolute correlations for all experiments, but this correlation is positive for regression and negative for classification. The positive correlation on regression experiments is unexpected, although similar behavior has been observed by Jiang et al. [2020]. The learning rate/batch size ratio has strong negative correlation on classification experiments and weak positive correlation on the regression experiments. The strong correlation of learning rate/batch size ratio on classification experiments aligns with trends observable in Figure 2, indicating potential confounding effects of these variables in the observed correlations. Dupuis et al. [2023] included the mean granulated Kendall rank correlation coefficient Ψ in their analysis to mitigate the influence of hyperparameters when computing rank correlations. This coefficient is computed by taking the average over Kendall coefficients at fixed values of the hyperparameters. However, by averaging over all hyperparameter ranges, significant correlations for different fixed values of the hyperparameters might be masked by lower correlations, resulting in inconclusive findings.

4.2 Partial Correlation

Given the correlation between learning rate/batch size ratio and generalization gap, we study partial correlations to isolate the influence of these hyperparameters on the observed correlation between generalization and PH dimensions. Suppose X and Y are our variables of interest and Z is a

Table 1: Spearman's ρ, mean granulated Kendall Ψ and Kendall τ rank coefficients of the correlation with generalization gap. For CHD and MNIST mean of 10 seeds presented with standard deviation.

	Measure	ρ	Ψ	τ
FCN-5 CHD	Euclidean	$0.71_{\pm 0.07}$	$0.48_{\pm 0.07}$	$0.54_{\pm 0.07}$
	Loss-based	$0.78_{\pm 0.06}$	$0.64_{\pm 0.06}$	$0.64_{\pm 0.06}$
	Norm	$0.92_{\pm 0.04}$	$0.84_{\pm 0.05}$	$0.81_{\pm 0.06}$
	LR / BS	$0.29_{\pm 0.11}$	$0.11_{\pm 0.08}$	$0.21_{\pm 0.07}$
FCN-7 CHD	Euclidean	$0.45_{\pm 0.07}$	$0.19_{\pm 0.07}$	$0.32_{\pm 0.06}$
	Loss-based	$0.67_{\pm 0.11}$	$0.51_{\pm 0.09}$	$0.54_{\pm 0.08}$
	Norm	$0.87_{\pm 0.05}$	$0.78_{\pm 0.05}$	$0.72_{\pm 0.06}$
	LR / BS	$0.15_{\pm 0.01}$	$0.04_{\pm 0.05}$	$0.10_{\pm 0.09}$
FCN-5 MNIST	Euclidean	$0.67_{\pm 0.05}$	$0.73_{\pm 0.05}$	$0.50_{\pm 0.04}$
	Loss-based	$0.77_{\pm 0.05}$	$0.79_{\pm 0.05}$	$0.60_{\pm 0.04}$
	Norm	$-0.93_{\pm 0.03}$	$-0.79_{\pm 0.04}$	$-0.81_{\pm 0.04}$
	LB / BS	$-0.95_{\pm 0.02}$	$-0.83_{\pm 0.05}$	$-0.84_{\pm 0.04}$
FCN-7 MNIST	Euclidean	$0.78_{\pm 0.04}$	$0.88_{\pm 0.04}$	$0.61_{\pm 0.04}$
	Loss-based	$0.88_{\pm 0.03}$	$0.90_{\pm 0.03}$	$0.71_{\pm 0.04}$
	Norm	$-0.97_{\pm 0.01}$	$-0.83_{\pm 0.05}$	$-0.88_{\pm 0.04}$
	LR / BS	$-0.98_{\pm 0.00}$	$-0.91_{\pm 0.03}$	$-0.90_{\pm 0.02}$
AlexNet CIFAR-10	Euclidean	0.85	0.85	0.69
	Loss-based	-0.31	-0.07	-0.14
	Norm	-0.98	-0.94	-0.91
	LR / BS	-0.98	-0.94	-0.91

multivariate variable. The partial correlation between X and Y given Z is the correlation between the residuals of the regressions of X with Z and of Y with Z. If the correlation between X and Y can be fully explained by Z, then the partial correlation should yield a low coefficient. To report statistical significance, we conduct a non-parametric permutation-type hypothesis test for the assumption that the partial correlation is equal to zero. In our setting, the null hypothesis implies the correlation observed between generalization and PH dimensions is explained by the influence of other hyperparameters.

In Table 2, we report the partial Spearman's and (standard) Kendall rank correlation between generalization gap and PH dimensions, conditioned on learning rate for fixed batch sizes. We provide the corresponding p-values for the stated hypothesis test in parentheses. Recall that a p-value lower than 0.05 implies the rejection of the null hypothesis, or equivalently, that there is a correlation between PH dimension and generalization gap that cannot be explained by the influence of the hyperparameters; a p-value greater than 0.05 implies a significant influence of the corresponding hyperparameter in the apparent correlation. We observe for most batch sizes, the correlation present between Euclidean dimension and generalization gap can be explained by the influence of learning rate, particularly for larger batch sizes. There is no consistent trend for the loss-based dimension, and the influence of the learning rate is found to be significant in fewer cases, indicating it may be a better-suited measure.

4.3 Conditional Independence

We have established that for some batch sizes, the correlation between PH dimension and generalization is significantly influenced by learning rate. We now seek to determine the existence of a causal relationship between the PH dimension and the generalization gap, basing our study on that of Jiang et al. [2020]. If a causal relationship does exist, then a low PH dimension caused by a variation of hyperparameters would consequently cause the generalization gap to be small. If a causal relationship does not exist, then the variation of the hyperparameter would cause both the PH dimension and generalization gap to be low, without any meaningful effect between the PH dimension and generalization gap. An illustrative example of these scenarios is provided in Figure 3.

Table 2: Partial Spearman's ρ and Kendall τ correlation computed between PH dimensions and generalization error for fixed batch sizes given learning rate. p-values in parentheses; bolded entries have p-value ≥ 0.05 signaling a significant influence of learning rate.

	Batch size	Euclidean		Loss-based	
		ρ	τ	ρ	τ
FCN-5 CHD	32	0.10 (**0.43**)	0.06 (**0.48**)	0.06 (**0.64**)	0.04 (**0.66**)
	65	-0.03 (**0.85**)	-0.01 (**0.90**)	-0.10 (**0.47**)	-0.08 (**0.39**)
	99	-0.41 (0.00)	-0.29 (0.00)	-0.67 (0.00)	-0.49 (0.00)
	132	-0.31 (0.02)	-0.21 (0.02)	-0.65 (0.00)	-0.47 (0.00)
	166	-0.04 (**0.76**)	-0.02 (**0.79**)	-0.49 (0.00)	-0.33 (0.00)
	200	-0.05 (**0.70**)	-0.03 (**0.75**)	-0.65 (0.00)	-0.48 (0.00)
FCN-7 CHD	32	0.48 (0.00)	0.32 (0.00)	0.37 (0.00)	0.24 (0.01)
	65	0.10 (**0.46**)	0.07 (**0.42**)	-0.02 (**0.88**)	-0.02 (**0.86**)
	99	-0.35 (0.01)	-0.24 (0.01)	-0.73 (0.00)	-0.55 (0.00)
	132	0.04 (**0.74**)	0.02 (**0.87**)	-0.18 (**0.19**)	-0.14 (**0.13**)
	166	0.08 (**0.56**)	0.03 (**0.76**)	-0.70 (0.00)	-0.51 (0.00)
	200	0.12 (**0.39**)	0.08 (**0.37**)	-0.82 (0.00)	-0.66 (0.00)
FCN-5 MNIST	32	0.63 (0.00)	0.42 (0.00)	0.46 (0.00)	0.32 (0.00)
	76	-0.08 (**0.54**)	-0.06 (**0.51**)	0.43 (0.00)	0.29 (0.00)
	121	0.17 (**0.21**)	0.13 (**0.14**)	0.37 (0.00)	0.26 (0.00)
	166	0.00 (**0.99**)	0.01 (**0.95**)	0.16 (**0.22**)	0.12 (**0.18**)
	211	0.22 (**0.10**)	0.15 (**0.09**)	0.17 (**0.20**)	0.12 (**0.18**)
	256	0.08 (**0.55**)	0.07 (**0.48**)	0.10 (**0.45**)	0.09 (**0.34**)
FCN-7 MNIST	32	0.81 (0.00)	0.61 (0.00)	0.82 (0.00)	0.62 (0.00)
	76	0.68 (0.00)	0.46 (0.00)	0.79 (0.00)	0.58 (0.00)
	121	0.29 (0.03)	0.21 (0.02)	0.69 (0.00)	0.50 (0.00)
	166	0.26 (**0.05**)	0.17 (**0.05**)	0.50 (0.00)	0.34 (0.00)
	211	0.26 (**0.46**)	0.20 (0.03)	0.45 (0.00)	0.31 (0.00)
	256	0.19 (**0.15**)	0.16 (**0.07**)	0.30 (0.02)	0.21 (0.02)

To distinguish between these two scenarios, we conduct a conditional independence test by computing the conditional mutual information (CMI) [Jiang et al., 2020] as defined in Appendix F. The CMI vanishes to zero if and only if $X \perp Y \mid Z$, i.e., X (PH dimension) and Y (generalization gap) are conditionally independent given Z (hyperparameter). Given the discrete nature of our selected hyperparameters, we empirically determine the probability density functions. To assess the significance of the computed CMI, we generate a null distribution for the CMI under local permutations of X or Y for fixed hyperparameter values, where "local" here refers to the group of realizations of X and Y generated under the same hyperparameters Z [Kim et al., 2022]. The null hypothesis implies that X and Y are conditionally independent, in which case the CMI is invariant to permutations. We reject the assumption of conditional independence if the CMI lies in the extremes of the null distribution.

Table 3 contains the results of the conditional independence test between PH dimensions and generalization conditioned on learning rate for fixed batch sizes. Within the table, a p-value > 0.05 implies the acceptance of the null hypothesis of conditional independence (H_0 in Figure 3), whereas a p-value ≤ 0.05 indicates the existence of conditional dependence (H_1 in Figure 3). Hence, we observe that for the models trained on MNIST data and for most batch sizes, the PH dimensions and generalization can be considered to be conditionally independent. For the models trained on the CHD, for most batch sizes, the PH dimensions and generalization are seen to be conditionally dependent.

5 Adversarial Initialization

The theory of Birdal et al. [2021] and Dupuis et al. [2023] proposes a positive correlation between generalization and the respective PH dimensions. However, neither work makes an assumption on the initialization scheme applied a the start of training. To investigate the sensitivity of the measures to initialization, we employ the *adversarial initialization* technique proposed by Liu et al. [2020].
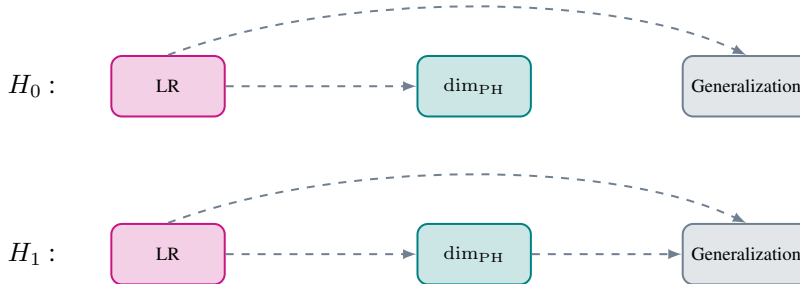
Figure 3: **Diagram of causal relationships under investigation in the conditional independence test.** In H_0 the PH dimension is conditionally independent of PH dimension given learning rate and there is no direct causal relationship between these variables. In H_1 generalization gap is conditionally dependent of the PH dimension indicating a causal relationship may exist.

Table 3: Table of p-values from conditional independence tests between PH dimensions and generalization gap conditioned on learning rate using conditional mutual information (CMI) as test statistic with local permutations for given batch sizes. Bolded p-values indicate conditional independence between PH dimension and generalization.

	PH dimension	Batch size					
		32	65	99	132	166	200
FCN-5 CHD	Euclidean	0.01	**0.27**	0.02	0.01	0.00	**0.06**
	Loss-based	0.00	0.02	0.00	0.00	0.00	0.00
FCN-7 CHD	Euclidean	0.00	**0.28**	0.00	0.00	0.00	0.00
	Loss-based	0.00	**0.33**	0.00	0.00	0.00	0.00
		Batch size					
		32	76	121	166	211	256
FCN-5 MNIST	Euclidean	**0.18**	**0.57**	**0.35**	**0.11**	**0.18**	**0.40**
	Loss-based	**0.23**	**0.28**	**0.07**	**0.09**	**0.11**	0.01
FCN-7 MNIST	Euclidean	**0.15**	0.04	**0.41**	**0.25**	**0.92**	**0.75**
	Loss-based	0.02	0.00	**0.30**	**0.71**	**0.88**	**0.38**

This entails a pre-training phase on training data with fixed, random labels until the network has successfully interpolated this random data. The resulting parameters are then used as initialization for training on the true dataset, leading to a poorly generalizing model.

In Figure 1, we present the results of this experiment, where the adversarial initialization models are contrasted against models trained from a standard (random) initialization. 30 seeds are evaluated for CHD and MNIST, and 20 seeds for AlexNet. We find that for the CNN CIFAR-10 and the FCN-5 MNIST the adversarial initialization models present lower PH dimensions despite having higher generalization gap, in contrast to the proposed theory. On AlexNet CIFAR-10 the PH dimensions both successfully identify the poorly generalization models, prescribing high values in this case.

6 Model-Wise Double Descent

We lastly explore model-wise double descent through the PH dimensions [Nakkiran et al., 2021]. In model-wise double descent the test accuracy of a classifier is non-monotonic with respect to the number of parameters—a surprising result in contrast with classical learning theory.

In Figure 4, we present results for the CNN trained on CIFAR-100 at a variety of width multipliers. We follow the "noiseless" configuration of Nakkiran et al. [2021], although we do not use batch normalization or learning rate decay to align with the assumptions of Dupuis et al. [2023]. The
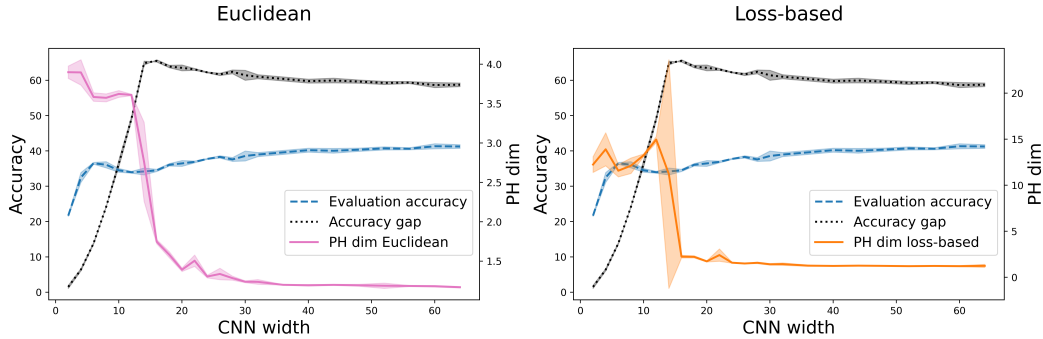
Figure 4: **Model-wise double descent manifests in Euclidean PH dimension, whilst neither PH dimension correlates with generalization gap in this setting.** Test accuracy, generalization gap, and PH dimensions for range of CNN widths. The double descent behavior is clearly visible in test accuracy and Euclidean PH dimension, but the generalization gap is monotonic in this critical region. Mean of three seeds with standard deviation shaded.

mean of 3 seeds is presented with standard deviation shaded. In evaluation accuracy, we observe the classical double descent behavior, but note that the generalization gap is monotonic in the range up to width multiplier of 16. We additionally observe a double descent behavior in Euclidean PH dimension. The behavior of loss-based PH dimension does not follow the double descent pattern, with notable instability and variance in the region of the evaluation accuracy double descent critical region. These results suggest a connection between the convergence properties of the model in the critical region of double descent with respect to the model width, itself a poorly understood phenomenon, and the Euclidean PH dimension. They also show the failure of either PH dimension to correlate with generalization gap for varying model width, which is monotonic in this width region.

7   Discussion

Our results demonstrate two modes of failure for PH dimensions as measures of generalization: adversarial initialization and model width variation in the critical region of double descent. Furthermore, we show that in some cases the correlation observed between PH dimension and generalization gap is significantly influenced by hyperparameter values. An explanation for the influence of hyperparameters—in particular learning rate—in the value of PH dimension is that the underlying space used in the PH computations is determined by samples from the optimization trajectory; and the influence of the learning rate on the geometry of these samples is notable. We further demonstrate that for some architectures and datasets the generalization measures and generalization gap are conditionally independent given the confounding hyperparameters, implying that the observed relationship between the two variables is not directly causal in these settings.

Evidently, the PH dimensions are not universally successful in correlating with generalization gap, in contrast to the general bounds proven by Birdal et al. [2021] and Dupuis et al. [2023]. We have two main conjectures for this disconnect between theory and practice. Firstly, the technical assumptions required by Dupuis et al. [2023] to prove generalization bounds, and their implications on architecture and hyperparameters valid for variation, are not clear. Preceding works [Şimşekli et al., 2020, Birdal et al., 2021] involve various technical assumptions about optimization trajectories and loss functions that may not be met in practice. Secondly, the term in the generalization bounds involving mutual information between the training data and the optimization trajectory may dominate, leading to vacuous bounds with respect to fractal dimension. These mutual information terms are complex and less studied than the fractal dimensions; it is unclear if they can be empirically estimated. The assumption that the mutual information does not dominate the bounds has not been proven and has been scarcely explored. We believe some experiments may enter a regime where this term dominates, disrupting the expected correlation.

7.1   Limitations

Models and settings. The goal of our study was to better understand the conclusions drawn by Birdal et al. [2021] and Dupuis et al. [2023]. We are thus subject to the following restrictions arising

from the assumptions in the theoretical results of Birdal et al. [2021] and Dupuis et al. [2023]: (i) we use only "vanilla" SGD; (ii) we only work with a constant learning rate; (iii) we do not use batch normalization; (iv) we do not study the addition of explicit regularization such as dropout or weight decay. We address this limitation by extending the study to include adversarial initialization scenarios [Liu et al., 2020], and studying the connection between double descent [Nakkiran et al., 2021] and PH dimension whilst still remaining within the theoretical assumptions. Future research directions include alternative optimization algorithms and common neural network architecture choices, such as batch normalization or annealing learning rates, prevented by the current setting.

Choice of hyperparameters. Our study exhibits a limited range of batch sizes and learning rates, along with unconventional grid values that varied between different architectures. These choices were made to align with, and ensure a fair comparison with, the experiments of [Dupuis et al., 2023]. We believe that these design choices were made by Dupuis et al. [2023] to ensure convergence of the models within reasonable numbers of iterations, due to the computational cost of repeatedly training different models with various seeds, and may have contributed to the statistically significant results reported in their work. For an extended analysis, we would explore a wider range of hyperparameters.

Computational limitations. Most of the runtime in our experiments was devoted to computing the loss-based PH dimension. Though efficient computation of PH is an active field of research in TDA Chen and Kerber [2011], Bauer et al. [2014a,b], Guillou et al. [2023], Bauer [2021], PH remains a computationally intensive methodology, limiting the number of experiments it was possible to run.

Lack of identifiable patterns in the correlation failure. An important limitation of our work is our failure to identify any clear patterns offering explanations as to when and why the PH dimension can fail to correlate with the generalization gap when conditioning on the network hyperparameters. We further cannot identify a pattern to explain the success and failure of PH dimension measures to correlate with generalization when using adversarial initialization.

Theoretical limitations. Despite providing extended experimental analyses of relationship between PH dimension and generalization gap, we do not make any theoretical contributions to explain the disconnect observed between theory and practice.

8 Conclusion

In this work, we extend previous evaluations of PH dimension-based generalization measures. Although theoretical results on the fractal and topological measures of generalization gaps were provided by Şimşekli et al. [2020], Birdal et al. [2021], Camuto et al. [2021] and Dupuis et al. [2023], experimentally, our study shows that there is in some cases a disparity between theory and practice. We suggest two directions for further investigation: (i) considering probabilistic definitions of fractal dimensions [Adams et al., 2020, Schweinhart, 2020] may offer a more natural interpretation for generalization compared to metric-based approaches; (ii) exploring multifractal models for optimization trajectories could better capture the complex interplay of network architecture and hyperparameters in understanding generalization. Overall, our work demonstrates that there is still much to understand concerning the complex interplay between generalization gap and TDA-based fractal dimension of optimization trajectories.

References

Henry Adams, Manuchehr Aminian, Elin Farnell, Michael Kirby, Joshua Mirth, Rachel Neville, Chris Peterson, and Clayton Shonkwiler. A Fractal Dimension for Measures via Persistent Homology. In Nils A. Baas, Gunnar E. Carlsson, Gereon Quick, Markus Szymik, and Marius Thaule, editors, *Topological Data Analysis*, Abel Symposia, pages 1–31, Cham, 2020. Springer International Publishing. ISBN 978-3-030-43408-3. doi: 10.1007/978-3-030-43408-3_1.

Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Ulrich Bauer. Ripser: Efficient computation of Vietoris–Rips persistence barcodes. *Journal of Applied and Computational Topology*, 5(3):391–423, September 2021. ISSN 2367-1734. doi: 10.1007/s41468-021-00071-5.

Ulrich Bauer, Michael Kerber, and Jan Reininghaus. Clear and compress: Computing persistent homology in chunks. In *Topological Methods in Data Analysis and Visualization III: Theory, Algorithms, and Applications*, pages 103–117. Springer, 2014a.

Ulrich Bauer, Michael Kerber, and Jan Reininghaus. Distributed computation of persistent homology. In *2014 proceedings of the sixteenth workshop on algorithm engineering and experiments (ALENEX)*, pages 31–38. SIAM, 2014b.

Tolga Birdal, Aaron Lou, Leonidas J Guibas, and Umut Şimşekli. Intrinsic Dimension, Persistent Homology and Generalization in Neural Networks. In *Advances in Neural Information Processing Systems*, volume 34, pages 6776–6789. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/hash/35a12c43227f217207d4e06ffefe39d3-Abstract.html.

John Briggs. *Fractals: The patterns of chaos: A new aesthetic of art, science, and nature*. Simon and Schuster, 1992.

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.

Chao Chen and Michael Kerber. Persistent homology computation with a twist. In *Proceedings 27th European workshop on computational geometry*, volume 11, pages 197–200, 2011.

Paul H Coleman and Luciano Pietronero. The fractal structure of the universe. *Physics Reports*, 213 (6):311–389, 1992.

Umut Şimşekli, Ozan Sener, George Deligiannidis, and Murat A Erdogdu. Hausdorff Dimension, Heavy Tails, and Generalization in Neural Networks. In *Advances in Neural Information Processing Systems*, volume 33, pages 5138–5151. Curran Associates, Inc., 2020. URL https://papers.nips.cc/paper/2020/hash/37693cfc748049e45d87b8c7d8b9aacd-Abstract.html.

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.

Benjamin Dupuis, George Deligiannidis, and Umut Şimşekli. Generalization Bounds using Data-Dependent Fractal Dimensions. In *Proceedings of the 40th International Conference on Machine Learning*, pages 8922–8968. PMLR, July 2023. URL https://proceedings.mlr.press/v202/dupuis23a.html.

Kenneth Falconer. *Fractal geometry: mathematical foundations and applications*. John Wiley & Sons, 2007.

Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry P Vetrov, and Andrew G Wilson. Loss surfaces, mode connectivity, and fast ensembling of dnns. *Advances in neural information processing systems*, 31, 2018.

Pierre Guillou, Jules Vidal, and Julien Tierny. Discrete morse sandwich: Fast computation of persistence diagrams for scalar data–an algorithm and a benchmark. *IEEE Transactions on Visualization and Computer Graphics*, 2023.

Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

Felix Hausdorff. Dimension und äußeres Maß. *Math. Ann.*, 79(1-2):157–179, 1918. ISSN 0025-5831,1432-1807. doi: 10.1007/BF01457179. URL `https://doi.org/10.1007/BF01457179`.

Haowei He, Gao Huang, and Yang Yuan. Asymmetric Valleys: Beyond Sharp and Flat Local Minima. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL `https://proceedings.neurips.cc/paper/2019/hash/01d8bae291b1e4724443375634ccfa0e-Abstract.html`.

Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization, 2018. Publisher Copyright: © 34th Conference on Uncertainty in Artificial Intelligence 2018. All rights reserved.; 34th Conference on Uncertainty in Artificial Intelligence 2018, UAI 2018 ; Conference date: 06-08-2018 Through 10-08-2018.

Jonathan Jaquette and Benjamin Schweinhart. Fractal dimension estimation with persistent homology: a comparative study. *Communications in Nonlinear Science and Numerical Simulation*, 84:105163, 2020.

Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=SJgIPJBFvH`.

R. Kelley Pace and Ronald Barry. Sparse spatial autoregressions. *Statistics & Probability Letters*, 33 (3):291–297, May 1997. ISSN 0167-7152. doi: 10.1016/S0167-7152(96)00140-X.

Ilmun Kim, Matey Neykov, Sivaraman Balakrishnan, and Larry Wasserman. Local permutation tests for conditional independence. *The Annals of Statistics*, 50(6):3388–3414, 2022.

Gady Kozma, Zvi Lotker, and Gideon Stupp. The minimal spanning tree and the upper box dimension. *Proceedings of the American Mathematical Society*, 134(4):1183–1187, 2006.

Alex Krizhevsky. Learning Multiple Layers of Features from Tiny Images. 2009.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. volume 60, pages 84–90, May 2017. doi: 10.1145/3065386.

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, November 1998. ISSN 1558-2256. doi: 10.1109/5.726791.

Shengchao Liu, Dimitris Papailiopoulos, and Dimitris Achlioptas. Bad global minima exist and sgd can reach them. *Advances in Neural Information Processing Systems*, 33:8543–8552, 2020.

Benoit Mandelbrot. How long is the coast of britain? statistical self-similarity and fractional dimension. *science*, 156(3775):636–638, 1967.

Benoit B Mandelbrot. The fractal geometry of nature/revised and enlarged edition. *New York*, 1983.

Benoit B Mandelbrot, Carl JG Evertsz, and Martin C Gutzwiller. *Fractals and chaos: the Mandelbrot set and beyond*, volume 3. Springer, 2004.

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt*. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124003, December 2021. ISSN 1742-5468. doi: 10.1088/1742-5468/ac3a74.

Steve Y Oudot. *Persistence theory: from quiver representations to data analysis*, volume 209. American Mathematical Soc., 2017.

Luciano Pietronero and Erio Tosatti. *Fractals in physics*. Elsevier, 2012.

Michael Prähofer and Herbert Spohn. Statistical self-similarity of one-dimensional growth processes. *Physica A: Statistical Mechanics and its Applications*, 279(1-4):342–352, 2000.

Nirupam Sarkar and Bidyut Baran Chaudhuri. An efficient differential box-counting approach to compute fractal dimension of image. *IEEE Transactions on systems, man, and cybernetics*, 24(1): 115–120, 1994.

Benjamin Schweinhart. Fractal dimension and the persistent homology of random geometric complexes. *Advances in Mathematics*, 372:107291, October 2020. ISSN 0001-8708. doi: 10.1016/j.aim.2020.107291. URL `https://www.sciencedirect.com/science/article/pii/S0001870820303170`.

Benjamin Schweinhart. Persistent Homology and the Upper Box Dimension. *Discrete & Computational Geometry*, 65(2):331–364, March 2021. ISSN 1432-0444. doi: 10.1007/s00454-019-00145-3. URL `https://doi.org/10.1007/s00454-019-00145-3`.

Guillaume Tauzin, Umberto Lupo, Lewis Tunstall, Julian Burella PÃ©rez, Matteo Caorsi, Anibal M. Medina-Mardones, Alberto Dassatti, and Kathryn Hess. giotto-tda: : A topological data analysis toolkit for machine learning and data exploration, 2021. URL `http://jmlr.org/papers/v22/20-325.html`.

The GUDHI Project. *GUDHI User and Reference Manual*. GUDHI Editorial Board, 3.1.1 edition, 2020. URL `https://gudhi.inria.fr/doc/3.1.1/`.

Guillermo Valle-Pérez and Ard A. Louis. Generalization bounds for deep learning, December 2020. URL `http://arxiv.org/abs/2012.04115`. arXiv:2012.04115 [cs, stat].

L. Vietoris. Über den höheren Zusammenhang kompakter Räume und eine Klasse von zusammenhangstreuen Abbildungen. *Mathematische Annalen*, 97(1):454–472, December 1927. ISSN 1432-1807. doi: 10.1007/BF01447877. URL `https://doi.org/10.1007/BF01447877`.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, February 2021. ISSN 0001-0782. doi: 10.1145/3446776.

Afra Zomorodian and Gunnar Carlsson. Computing Persistent Homology. *Discrete & Computational Geometry*, 33(2):249–274, February 2005. ISSN 1432-0444. doi: 10.1007/s00454-004-1146-y.

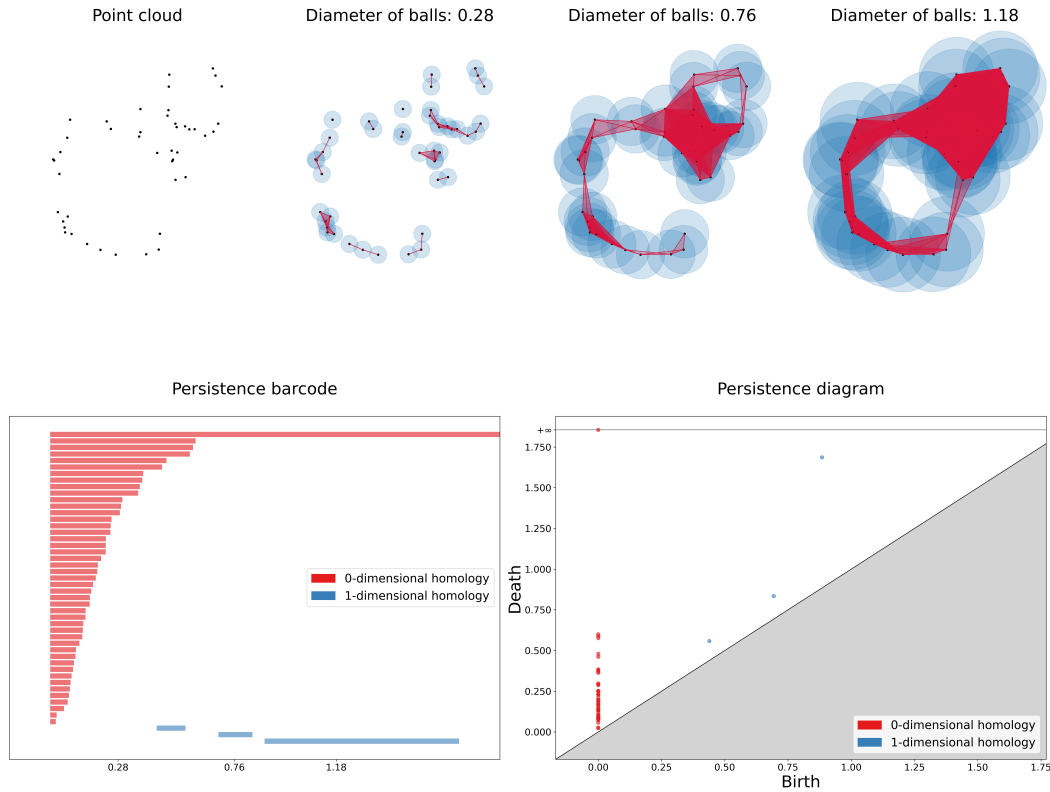# A   Persistent Homology and Vietoris–Rips Filtrations

Figure 5: Vietoris–Rips filtration over two noisy circles (with 30 and 15 points each) at 4 different filtration values; and corresponding persistence barcode and diagram (0-dimensional PH in red, 1-dimensional PH in blue). Images produced using GUDHI [The GUDHI Project, 2020].

PH is a methodology for computing topological representations of data. It achieves this using a *filtration*, producing a compact topological summary of the topological features over this filtration often presented in the form of a *persistence barcode* or *persistence diagram*. Here, we briefly overview these key concepts. For a more complete introduction to PH see Zomorodian and Carlsson [2005], Oudot [2017].

A *simplicial complex* is a combinatorial object built over a finite set and defined as a family of subsets of such finite set, called simplices, which is closed under inclusion. Geometrically, this can be understood as a set of vertices, edges, triangles, tetrahedra, and higher-order geometric objects—i.e. higher dimensional simplices. Being closed under inclusion means, for instance, that if a triangle is present in the family, all the edges and vertices in the boundary of the triangle also belong to the complex. For finite subset $S \subset X$ of a metric space (X, ρ), an example of such an object is the Vietoris–Rips [Vietoris, 1927] simplicial complex at scale $t \in [0, +\infty)$, defined as the family of all simplices of diameter less or equal than t that can be formed with the finite set S as set of vertices.

A *filtration* is defined as a family of nested simplicial complexes, that is, a parameterized set $\{K_t : t \in T\}$ with totally ordered indexing set T, such that if $s \leq t$ then $K_s \subset K_t$. The *Vietoris–Rips filtration* is then the family of Vietoris–Rips complexes at all scales $t \in [0, +\infty)$. More generally, a filtration is a family $\{F_t : t \in \mathbb{R}\}$ of nested simplicial complexes indexed by the real numbers, that is, if $t \leq s$ then we have $F_t \subset F_s$.

The *persistence barcode* provides a compact summary of the lifetime of topological features (components, holes, voids and higher dimensional generalizations) as we allow the filtration parameter to evolve. It is defined precisely as a multiset of bars, each of them spanning the lifetime of a topological

14

feature of the corresponding dimension. An alternative representation of the barcode is given by the *persistence diagram*. This is a scatterplot of points in the first quadrant of the two-dimensional plane, in the region above the diagonal, where each point is in correspondence with a bar in the barcode and has as the first coordinate its starting point and as the second coordinate the ending point of the bar. An example of a Vietoris–Rips filtration and the corresponding persistence barcode and diagram can be found in Figure 5. The Vietoris–Rips filtration is often utilized due to its fast computation [Bauer, 2021] and is also the method of choice for Dupuis et al. [2023] in computing the PH dimensions.

B Fractal Dimensions

Fractal dimensions describe the geometry of fractals—spaces that are rough and irregular on a finer scale. Informally, we want to say that such an object has dimension d when its "local geometry" at scale ϵ scales as ϵ^d or ϵ^{-d}, for some positive real number d which need not be integral. Fractal shapes arise in a number of situations, such as spaces built from self-similar recursions, chaotic dynamical systems, and even real data.

We now review several notions of fractal dimension that are interesting for the purposes of this work, following the presentation in Adams et al. [2020], where many of the original references can be found. We begin by providing the definitions of *metric* fractal dimensions, defined in terms of subsets S of a metric space (X, ρ). The first fractal dimension of this kind to appear in the literature was the following.

Definition B.1. Let $d \in [0, \infty)$. The *d-Hausdorff measure* of S is

$$H_d(S) := \inf_{\delta > 0} \left(\inf \left\{ \sum_{j=1}^{\infty} \text{diam}(B_j)^d : S \subseteq \bigcup_{j=1}^{\infty} B_j \text{ and } \text{diam}(B_j) \leq \delta \right\} \right)$$

where the inner infimum is taken over all coverings of S by balls B_j of diameter at most δ.

Definition B.2. The *Hausdorff dimension* of S is

$$\dim_{\text{H}}(S) := \inf_d \{ H_d(S) = 0 \}. \tag{4}$$

In practice, it is difficult to compute the Hausdorff dimension, which lead to the introduction of more computable notions of fractal dimension. Let N_ϵ be the infimum over the number of balls of radius $\epsilon > 0$ required to cover S.

Definition B.3. The *box-counting dimension* of S is

$$\dim_{\text{box}}(S) = \lim_{\epsilon \to 0} \frac{\log(N_\epsilon)}{\log(1/\epsilon)} \tag{5}$$

provided this limit exists. Replacing the limit with a lim sup yields the *upper* box-counting dimension, and a lim inf gives the *lower* box-counting dimension.

There is also a notion of metric fractal dimension defined in terms of minimal spanning trees. Let $T(\mathbf{x})$ denote the minimal spanning tree of a finite subset $\mathbf{x} := \{x_1, \ldots, x_n\} \subset X$ and define

$$E_\alpha^0(\mathbf{x}) = \frac{1}{2} \sum_{e \in T(\mathbf{x})} |e|^\alpha.$$

Definition B.4. Let S be a bounded subset of a metric space (X, ρ). The *minimal spanning tree dimension* of S is

$$\dim_{\text{MST}}(S) := \inf \left\{ \alpha : \exists\, C > 0 \text{ s.t. } E_\alpha^0(\mathbf{x}) < C, \, \forall\, \mathbf{x} \subset S \text{ finite subset } \right\} \tag{6}$$

Many authors [Adams et al., 2020, Schweinhart, 2021, 2020] extended these ideas using persistent homology. We first present the approach followed in Schweinhart [2021]. Let $\mathbf{x} = \{x_1, \ldots, x_n\}$ be a finite metric space. Call $\text{PH}_i(\mathbf{x})$ the persistence diagram obtained from the PH of dimension i computed from the Čech complex of $\mathbf{x}$ and $\widetilde{\text{PH}}_i(\mathbf{x})$ the one obtained from the Vietoris–Rips filtration. For any of these persistence diagrams we can define

$$E_\alpha^i(\mathbf{x}) := \sum_{(b,d) \in \text{PH}_i(\mathbf{x})} (d - b)^\alpha. \tag{7}$$

Definition B.5. Let S be a bounded subset of a metric space (X, ρ). The *ith persistent homology dimension* (PH_i-dim) for the Čech complex of S is

$$\dim_{\mathrm{PH}}{}^i(S) := \inf \left\{ \alpha : \exists\, C > 0 \text{ s.t. } E_\alpha^i(\mathbf{x}) < C,\, \forall\, \mathbf{x} \subset S \text{ finite subset} \right\}. \tag{8}$$

It is possible to analogously define $\dim_{\widetilde{\mathrm{PH}}}{}^i(S)$ for the Vietoris–Rips PH.

In addition to these metric notions, we also have *probabilistic* fractal dimensions, defined directly in terms of a measure μ supported in a subspace S of a metric space (X, ρ).

Definition B.6. The *lower Hausdorff dimension* of a measure μ with total mass 1 is

$$\underline{\dim_{\mathrm{H}}}(\mu) = \inf\{\dim_{\mathrm{H}}(S) : S \text{ is a Borel subset with } \mu(S) > 0\} \tag{9}$$

while the *upper Hausdorff dimension* is

$$\overline{\dim_{\mathrm{H}}}(\mu) = \inf\{\dim_{\mathrm{H}}(S) : S \text{ is a Borel subset with } \mu(S) = 1\}. \tag{10}$$

Remark 1. *We have $\underline{\dim_{\mathrm{H}}}(\mu) \leq \dim_{\mathrm{H}}(\mathrm{supp}(\mu))$ and this inequality can be strict.*

A probability measure μ defined in a metric space (X, ρ) induces a probability measure ν on the distance set of X, $\mathrm{dist}_X := \{\mathrm{dist}(x, y) : x, y \in X, x \neq y\}$

Definition B.7. The *correlation integral* of X is defined as the cumulative density function of ν

$$C(\epsilon) := \mathbb{P}_\nu \left(\rho(x, y) \leq \epsilon \right).$$

The *correlation dimension* is thus defined as the limit

$$\dim_{\mathrm{corr}}(\mu) = \lim_{\epsilon \to 0} \frac{\log(C(\epsilon))}{\log(\epsilon)} \tag{11}$$

As a final remark, the PH dimension can also be seen as a probabilistic fractal dimension if we consider that the finite metric spaces $\mathbf{x}$ used Definition B.5 are coming from i.i.d. samples drawn from a probability measure μ supported on a subset S of a metric space (X, ρ). This is the approach followed in Schweinhart [2020], Jaquette and Schweinhart [2020], which introduce the following modified definition.

Definition B.8 ([Schweinhart, 2020])**.** The *persistent homology dimension* (PH_i-dimension) of a measure μ, for each $\alpha > 0$ and $i \in \mathbb{N}$, is

$$\dim_{\mathrm{PH}}{}^{i,\alpha}(\mu) := \frac{\alpha}{1 - \beta} \tag{12}$$

where

$$\beta = \limsup_{n \to \infty} \frac{\log(\mathbb{E}(E_\alpha^i(\mathbf{x})))}{\log(n)}$$

where $\mathbf{x} = \{x_1,\ \ldots,\ x_n\}$ is set of i.i.d. samples drawn from μ.

B.1 Relations Between Different Notions of Fractal Dimension

It is worth mentioning here that a notion of fractal dimension does not need to be well-defined for every subset of a metric space or measure supported on it. In fact, there are shapes that present a "multifractal" structure scaling at several values of d in our informal intuition above. However, if we assume the following regularity condition, the Hausdorff, box-counting, and correlation dimension are well-defined and all coincide.

Definition B.9. A probability measure μ supported on a metric space (X, ρ) is *d-Ahlfors regular* if there exist c, $\delta_0 \in \mathbb{R}_+$ such that

$$\frac{1}{c}\delta^d \leq \mu(B_\delta(x)) \leq c\delta^d$$

for all $x \in X$ and $d < \delta_0$, where $B_\delta(x) := \{y \in X : \rho(x, y) < \delta\}$.

If μ is d-Ahlfors regular on X then it is comparable to the d-dimensional Hausdorff measure in X and the Hausdorff measure is also d-regular.

On the other hand, Kozma et al. [2006] prove that for any metric space (with no mention to any regularity conditions) the minimal spanning tree equals the upper box dimension

$$\dim_{\text{box}}(S) = \dim_{\text{MST}}(S). \tag{13}$$

Concerning minimal spanning trees and PH, there is a bijection between the edges of the Euclidean minimal spanning tree of a finite metric space $\mathbf{x} = \{x_1, \ldots, x_n\}$ and the intervals in the PH of $\text{PH}_0(\mathbf{x})$, where we need to halve the length of the intervals in the PH decomposition. For the Vietoris–Rips PH $\widetilde{\text{PH}}_i$ there is no need to halve the length of the intervals. In any case, it is clear from the definitions that

$$\dim_{\text{PH}}{}^0(S) = \dim_{\widetilde{\text{PH}}}{}^0(S) = \dim_{\text{MST}}(S). \tag{14}$$

For higher homological degree, Schweinhart [2021] obtains conditions under which the PH dimension of higher homological degrees agrees with the box-counting dimension, in a similar vein to Kozma et al. [2006].

On the other hand, Schweinhart [2020] additionally presented a thorough study of the asymptotic behavior of the quantities $E_\alpha^i(\mathbf{x})$ for $i \in \mathbb{N}$, $\alpha > 0$, and $\mathbf{x} = \{x_1, \ldots, x_n\}$ a set of random i.i.d. samples drawn from a probability measure μ defined a metric space (X, ρ). For instance, the following result concerning minimal spanning trees is proved.

Theorem B.1 (Theorem 3, [Schweinhart, 2020]). Let μ be a d-Ahlfors regular measure on a metric space and $\mathbf{x} = \{x_1, \ldots, x_n\}$ i.i.d. samples from μ. If $0 < \alpha < d$, then

$$E_\alpha^0(\mathbf{x}) \approx n^{\frac{d-\alpha}{d}}$$

with high probability as $n \to \infty$, where $\approx$ means that the ratio of the two quantities is bounded between positive constants that do not depend on n.

The hypothesis of d-Ahlfors regularity is not strictly needed in the proof of this theorem: it is possible to just assume weaker statements on the measure μ to prove the bounds. However, it is argued that d-Ahlfors regularity is included in the theorem because in an accompanying paper [Jaquette and Schweinhart, 2020], where the authors explore these quantities in applications, they observe that for fractals emerging from chaotic attractors (that do not satisfy Ahlfors regularity), for each $\alpha > 0$ there is a different value of d_α such that $E_\alpha^0(\mathbf{x}) \approx n^{\frac{d_\alpha - \alpha}{d_\alpha}}$. In particular, this means that we cannot replace d in the theorem above with the upper-box or Hausdorff dimension.

Other results regarding the asymptotic behavior for $E_\alpha^i(x_1, \ldots, x_n)$ are also derived in Schweinhart [2020], where d-Ahlfors regularity is also required, in addition to some conditions on the asymptotic behavior of the expectation and variance of the PH_i-dimension. Finally, Schweinhart [2020] establishes a correspondence between the PH_0-dimension and the Hausdorff dimension when a measure is d-Ahlfors regular for $\dim_{\text{PH}}{}^{0,\alpha}$, if $0 < \alpha \leq d$. There is also a connection for higher-order PH_i-dimensions adding some extra conditions: requiring the measure to be defined in an Euclidean space and some asymptotic behavior for the expectation and variance of $\text{PH}_i(x_1, \ldots, x_n)$.

C   Additional Experimental Details

In this appendix we include additional experimental specifications for the experiments.

C.1   Architectures

- As in Dupuis et al. [2023], both FCN-5 and FCN-7 networks have a width of 200 for each hidden layer and use ReLU activation.
- AlexNet follows the construction outlined in Krizhevsky et al. [2017].
- The standard CNN defined in consists of four 3×3 convolutional layers with widths $[c, 2c, 4c, 8c]$, where c is a width (channel) multiplier [Nakkiran et al., 2021]. In all experiments, $c = 64$ unless otherwise stated. Each convolution is followed by a ReLU activation and a MaxPool operation with kernel $=$ stride $= [1, 2, 2, 8]$.

C.2 Training Configuration

We train using mean squared error for the regression experiments and cross-entropy loss for the classification experiments. Similarly to Dupuis et al. [2023] we report accuracy gap instead of loss gap for the classification experiments.

The convergence criteria follow Dupuis et al. [2023], and are as follows:

- For regression, we compute the empirical risk on the full training dataset every 2000 iterations and define convergence when the relative difference between two consecutive evaluations becomes smaller than 0.5%.
- For classification, we define convergence when the model reaches 100% training accuracy, given that the model is evaluated on the full training dataset every 10,000 iterations.

C.3 Computation of PH Dimensions

The computation of the PH dimensions is based on Algorithm 1 by Birdal et al. [2021], using the code of Dupuis et al. [2023]. This codebase relies on the TDA software Giotto-TDA [Tauzin et al., 2021] to compute the PH barcodes in the two (pseudo-)metric spaces under study.

C.4 Grid Experiments

All experiments for the correlation analysis utilize learning rates and batch sizes on a 6×6 grid, defined by Dupuis et al. [2023] and repeated below:

- For CHD, learning rates are logarithmically spaced between 0.001 and 0.01. Batch sizes take values $\{32, 65, 99, 132, 166, 200\}$.
- For MNIST and CIFAR-10, learning rates are logarithmically spaced between 0.005 and 0.1. Batch sizes take values $\{32, 76, 121, 166, 211, 256\}$.

Experiments on MNIST and CHD are repeated with seeds $\{0, \ldots, 9\}$, while experiments on CIFAR-10 use seed 0.

C.5 Adversarial Initialization

A batch size of 128 is used for all experiments, with a constant learning rate of 0.01. We train using seeds $\{0, \ldots, 29\}$ for MNIST and AlexNet CIFAR-10, only seeds $\{0, \ldots, 19\}$ are used for CNN CIFAR-10 due to computational constraints.

C.6 Model-wise Double Descent

We train the standard CNN with CIFAR-100 with a constant learning rate of 0.01, a batch size of 128, and seeds $\{0, 1, 2\}$. Since not all model widths achieve 100% training accuracy on CIFAR-100, we terminate training after 250,000 iterations.

D Stability of PH Dimension Estimates

The theory proposed by Birdal et al. [2021] and Dupuis et al. [2023] establishes PH dimensions as a measure of generalization, but in practice we can only compute an approximation of this quantity. As detailed in Section 3, we compute this estimation using the 5,000 iterations after reaching convergence $\{w_k : 0 < k \leq 5000\}$. To study if this value remains constant after the first 5,000 iterations, we conduct an additional experiment in which we train for 15,000 iterations beyond the convergence criterion $\{w_k : 0 < k \leq 15000\}$. We then compute 3 estimations of the PH dimensions: for the first 5,000 iterations $\dim_{\mathrm{PH}}(w_{0:5000})$, using the samples between 5,000 and 10,000 iterations $\dim_{\mathrm{PH}}(w_{5000:10000})$ and between 10,000 and 15,000 iterations $\dim_{\mathrm{PH}}(w_{10000:15000})$. We then compute rank correlation coefficients between the three estimates, to evaluate their relative order, and their relative difference by computing the following quantity

$$\frac{\dim_{\mathrm{PH}}(w_{j:k}) - \dim_{\mathrm{PH}}(w_{p:q})}{\dim_{\mathrm{PH}}(w_{j:k})} \cdot 100.$$

Table 4 contains the results obtained. We note that the relative differences between values are small, and do not have a consistent sign, indicating a (small) fluctuating pattern that is likely due to randomness. Concerning the rank correlations, they seem to indicate that the relative ordering is consistent between different sets of iterations.

Table 4: Relative difference (%) and Spearman's ρ and Kendall τ rank correlations for the PH dimensions estimates computed using different subsets of trajectory after convergence. Means over 10 seeds of each model are presented with standard deviations as error bars.

Model & Data	Measures	5,000 vs 10,000			10,000 vs 15,000			5,000 vs 15,000		
		% difference	ρ	τ	% difference	ρ	τ	% difference	ρ	τ
FCN-5 CHD	Euclidean	0.15 ± 1.74	0.63	0.43	-0.41 ± 1.05	0.88	0.72	-0.25 ± 1.72	0.68	0.53
	Loss based	-3.83 ± 6.82	0.75	0.56	-3.48 ± 6.62	0.62	0.49	-7.24 ± 7.44	0.72	0.55
FCN-5 MNIST	Euclidean	0.11 ± 0.66	0.65	0.46	0.18 ± 0.60	0.46	0.34	-0.07 ± 0.74	0.45	0.33
	Loss based	1.53 ± 1.35	0.62	0.45	1.01 ± 1.21	0.53	0.38	2.62 ± 1.22	0.65	0.48
AlexNet CIFAR-10	Euclidean	-0.75 ± 0.63	0.86	0.68	0.02 ± 0.40	0.87	0.69	-0.72 ± 0.64	0.92	0.78
	Loss based	-1.05 ± 2.19	0.61	0.46	-1.41 ± 1.67	0.23	0.16	-2.47 ± 2.77	0.14	0.11

E   Additional Experimental Results

In Figure 6, we present the results for the FCN-5 used for the computation of the corresponding correlation coefficients in Table 1, that were not present in Figure 2 in the main text due to space constraints. In Figure 7 we additionally plot $||w_{5000}||_2$ against the generalization gap for the results of Table 1.
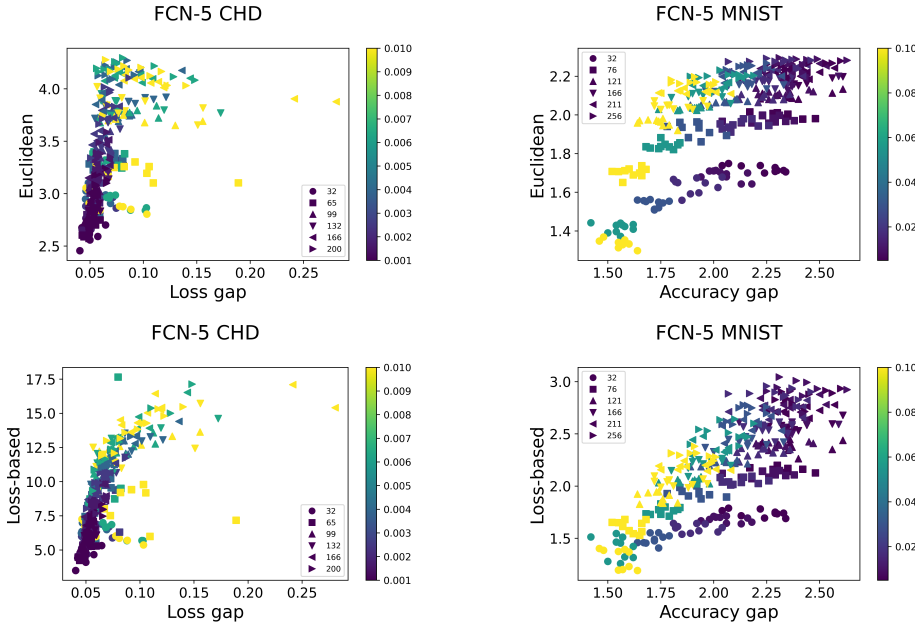

Figure 6: **Learning rate/batch size grid results for FCN-5 architecture.** Euclidean (top) and loss-based (bottom) PH dimension plotted against generalization gap for range of learning rates and batch sizes for FCN-5 architecture.

F   Conditional Mutual Information

The CMI, denoted by I, for discrete random variables is defined as

$$I(X;Y|Z) = \sum_{z \in \mathcal{Z}} p_Z(z) \sum_{y \in \mathcal{Y}} \sum_{x \in \mathcal{X}} p_{X,Y|Z}(x,y|z) \log \frac{p_{X,Y|Z}(x,y|z)}{p_{X,Z}(x,z)p_{Y,Z}(y,z)},$$
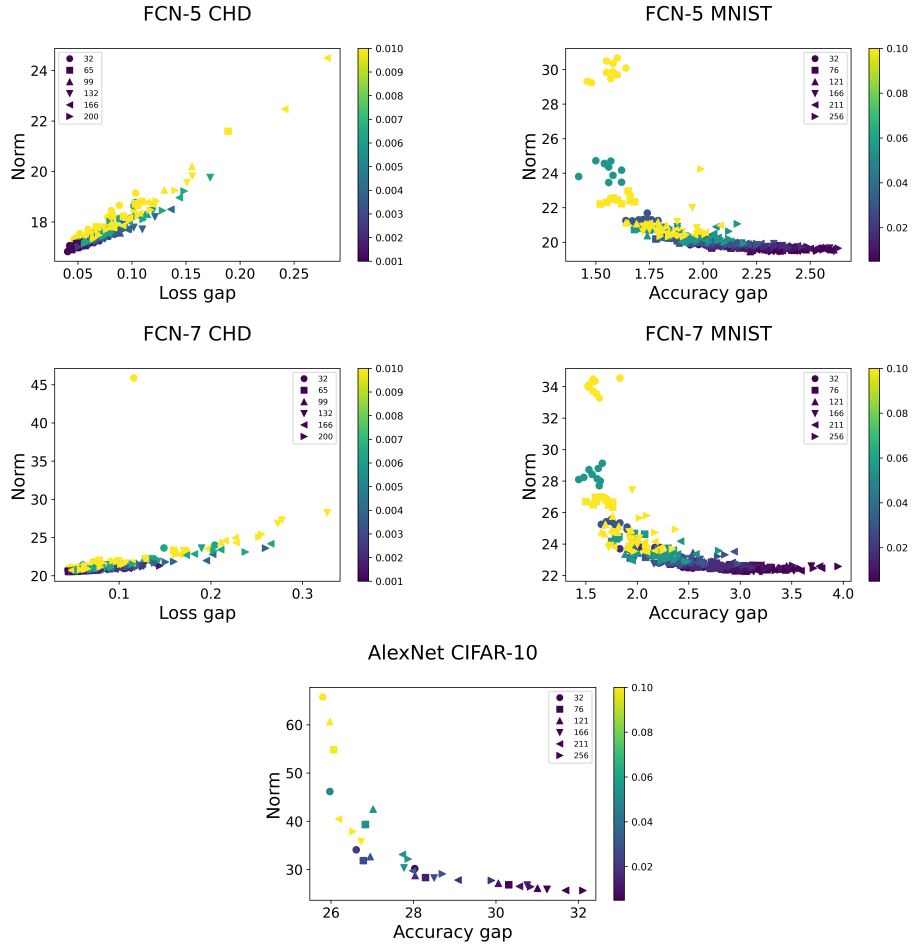
Figure 7: **Learning rate/batch size grid results for parameter norm.** $\|w_{5000}\|_2$ plotted against generalization gap for range of learning rates and batch sizes.

where p is the empirical measure of probability. From this definition, we see that the CMI vanishes if and only if $X \perp Y \mid Z$, i.e., X and Y are conditionally independent given Z. Hence, while changes in X might seem linked to changes in Y, the CMI allows us to isolate the effect of Z and establish whether X and Y are independent when Z is fixed. In Section 4.3, X is the PH dimension, Y the generalization error, and Z the learning rate.

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: *All the contributions are concisely listed in the abstract of the paper and more detailed explanations of these are included at the end of the introduction. These contributions directly relate to experimental evidence provided in sections 4, 5 and 6.*
   Guidelines:
     • The answer NA 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 NA 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: *The limitations from our studies are thoroughly explained in Section 7.1 and Appendix C.*
   Guidelines:
     • The answer NA 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: [NA]
   Justification: *No theoretical results are provided, this is an experimental and statistical work.*
   Guidelines:
     • The answer NA 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: *The experimental setup: architectures, training sets, choice of hyperparameters, optimization algorithms, etc., and references to reproduce the adversarial initialization and double descent experiments are provided in Section 3.*
   Guidelines:
   - The answer NA 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: *Open access code to reproduce our experiments is provided at `https: // github. com/ charliebtan/ fractal_ dimensions`*
   Guidelines:
   - The answer NA means that paper does not include experiments requiring code.

22

- Please see the NeurIPS code and data submission guidelines (`https://nips.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://nips.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, etc.) necessary to understand the results?
   Answer: [Yes]
   Justification: *All details about the experimental settings are included in section 3 and Appendix C.*
   Guidelines:
   - The answer NA 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: [Yes]
   Justification: *Statistics and hypothesis tests are reported with p-values indicating the significance of the finding as can be seen in Section 4. For these kind of statistical tests, these values are more appropriate measures of statistical significand than error bars.*
   Guidelines:
   - The answer NA 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: *Computation resources and details on the runtimes of the experiments are included in Section 3.*
   Guidelines:

   - The answer NA 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]
   Justification: *The compliance with the NeurIPS Code of Ethics is detailed throughout this Checklist.*
   Guidelines:

   - The answer NA 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: [NA]
    Justification: *There is no societal impact of the work performed.*
    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA 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., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: *The paper poses no such risks.*

Guidelines:

- The answer NA 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: *We give credit to the original authors of the code we use in our experiments when required, respecting their licenses and terms of use.*

Guidelines:

- The answer NA 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: [NA]

Justification: *The paper does not release new assets, the experiments are based in pre-existing code.*

Guidelines:

- The answer NA 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: [NA]

Justification: *The paper does not involve crowdsourcing nor research with human subjects.*
Guidelines:

- The answer NA 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: [NA]

Justification: *The paper does not involve crowdsourcing nor research with human subjects.*
Guidelines:

- The answer NA 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.