IMPERIAL

On the Limitations of Fractal Dimension as a Measure of Generalization

arXiv:2406.02234

Charlie Tan[2],
Inés García-Redondo[1],
Qiquan Wang[1],
Michael M. Bronstein[2],
Anthea Monod[1]

Introduction

Deep learning's empirical success contrasts with its limited theoretical foundation, especially regarding why neural networks generalize effectively without explicit regularization, despite predictions from classical statistical learning theory.

Learning Setting

- $(\mathcal{Z}, \mathcal{F}_{\mathcal{Z}}, \mu_{\mathcal{Z}})$ data space:
  - $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, $\mathcal{X}$ feature and $\mathcal{Y}$ label spaces.
  - $\mu_{\mathcal{Z}}$ unknown data-generating distribution.
- Training data: $S = \{z_1, \ldots, z_n\} \sim \mu_{\mathcal{Z}}^{\otimes n}$
- Loss function: $\ell : \mathbb{R}^d \times \mathcal{Z} \to \mathbb{R}_+$, measures quality of our parametric approximation.
- Aim to minimize the empirical risk:

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

- Use optimization algorithms, e.g. Stochastic Gradient Descent (SGD).
- Performance in unseen data measured by population risk: $\mathcal{R}(w) := \mathbb{E}_z[\ell(w, z)]$

Definition 1. Generalization Gap:

$$\mathcal{G}(S, w) := |\mathcal{R}(w) - \hat{\mathcal{R}}(S, w)|.$$

Fractal Dimension

- Fractals are self-similar shapes arising in real world-data.
- Their key defining point is their non-integer fractal dimension, a notion of their roughness.

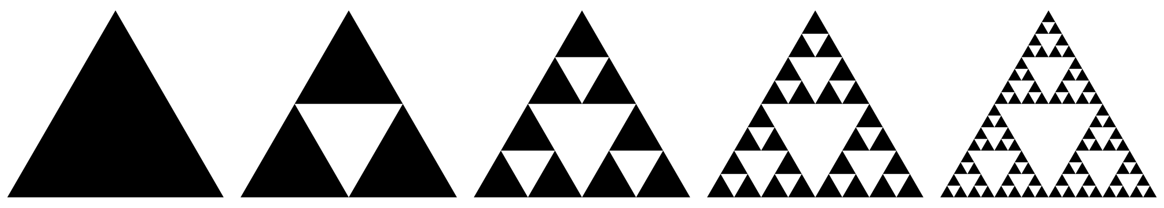

Figure 1: Evolution of the Sierpinski triangle in five iterations. Its fractal dimension is $D = \log_2 3$

Fractal Structure in Optimization Trajectories

- Given the recursive nature of optimization algorithms, several authors have proposed a random fractal structure for neural network optimization trajectories.
- Bounds for the generalization gap have been established with respect to various fractal dimensions.
- Experimental validation is based on an observed correlation between generalization gap and fractal dimension.

Experimental Design

- Train the model using SGD until 100% training accuracy.
- Run 5000 additional iterations to obtain weights near the local minimum and compute two notions of fractal dimension.

Statistically Grounded Correlation Analysis

Correlation Analysis

How does the correlation between the generalization gap and fractal dimension compare to correlations with other common hyperparameters?

- ρ: Spearman's rank correlation coefficient;
- Ψ: mean granulated Kendall rank correlation coefficient;
- τ: standard Kendall rank correlation coefficient.

Table 1: Correlation coefficients with generalization error for different hyperparameters and models.

Model & Data	Coeff.	Measure				
		Dim 1	Dim 2	Norm	Step size	LB ratio
FCN-5	ρ	-0.688	-0.762	-0.910	-0.623	-0.287
&	Ψ	-0.382	-0.559	-0.769	-0.360	-0.106
CHD	τ	-0.501	-0.604	-0.767	-0.460	-0.203
FCN-7	ρ	-0.434	-0.668	-0.866	-0.528	-0.149
&	Ψ	-0.156	-0.500	-0.740	-0.389	-0.032
CHD	τ	-0.304	-0.701	-0.378	-0.378	-0.103
FCN-5	ρ	0.649	0.752	-0.898	0.200	-0.929
&	Ψ	0.601	0.614	-0.579	0.090	-0.690
MNIST	τ	0.473	0.561	-0.725	0.116	-0.779
FCN-7	ρ	0.759	0.850	-0.916	0.491	-0.959
&	Ψ	0.654	0.661	-0.539	0.256	-0.749
MNIST	τ	0.567	0.660	-0.744	0.355	-0.832
AlexNet	ρ	0.851	-0.311	-0.977	0.741	-0.982
&	Ψ	0.850	-0.0722	-0.944	0.450	-0.944
CIFAR-10	τ	0.689	-0.140	-0.906	0.539	-0.910

We observe a stronger correlation with the norm, and significant correlations with other hyperparameters.

Partial Correlation

Is the correlation observed between fractal dimension and generalization gap a product of a correlation with a third variable?

- Compute regressions of generalization error and fractal dimensions with learning rate.
- Calculate correlation between the marginals of both regressions; a low coefficient implies potential influence from shared correlation with learning rate.
- Conduct non-parametric permutation tests for statistical significance.

We find that in most cases the partial correlation with learning rate is statistically significant.

Conditional Independence

Is there a causal relation between changes in the hyperparameter and changes in the generalization and fractal dimension?

Compute Conditional Mutual Information conditioned on the hyperparameters and simulate the null-distribution (conditional independence) using local permutations.

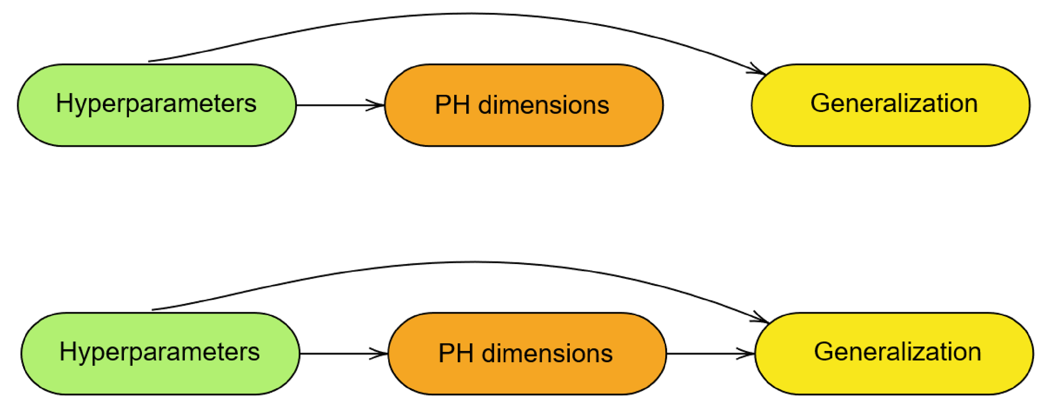

Figure 2: Above (H0): there exists no causal link between fractal dimensions and generalization, both are explained by hyperparameters; Below (H1): causal link between PH dimensions, generalization can be fully explained by fractal dimensions.

We conclude that for MNIST, fractal dimension and generalization are conditionally independent; for CHD, they are conditionally dependent.

Fractal Dimension Fails to Predict Generalization

Adversarial Initialization

Table 2: Spearman's and Kendall rank correlation coefficients between PH dimensions and generalization given standard or adversarial initialization.

	AlexNet & CIFAR-10		CNN & CIFAR-100		CFN-5 & MNIST	
Initialization	Dim 1	Dim 2	Dim 1	Dim 2	Dim 1	Dim 2
Spearman's rank coefficients						
Standard	0.321	0.261	0.237	0.249	0.709	0.455
Adversarial	-0.418	-0.733	-0.212	0.127	0.588	0.552
Kendall rank coefficients						
Standard	0.244	0.200	0.225	0.225	0.467	0.333
Adversarial	-0.289	-0.600	-0.156	0.0667	0.422	0.467

Double Descent

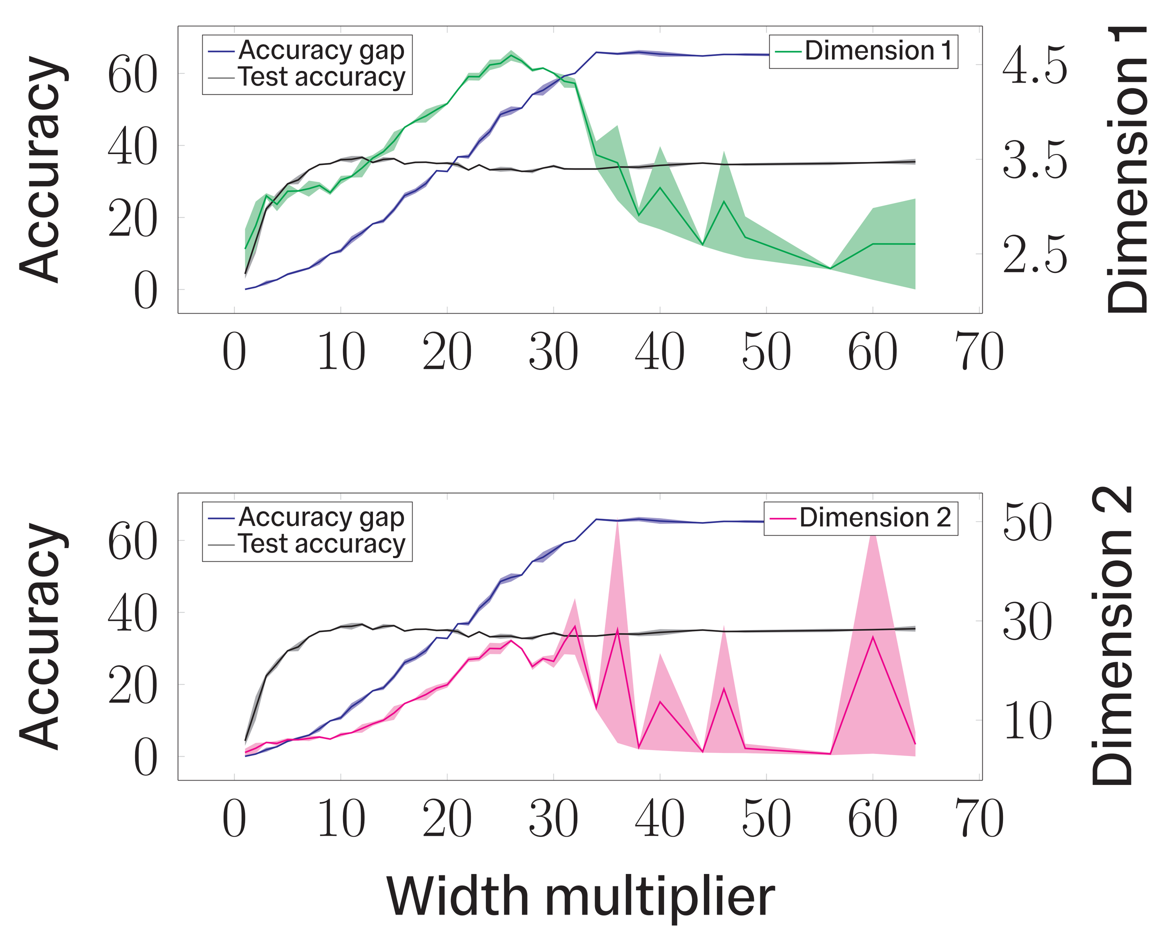

Figure 3: Mean accuracy gap, mean test accuracy and fractal dimensions computed for two seeds of the "standard CNN". The x-axis corresponds to the width multiplier from the architecture design.

Affiliations

[1] Department of Mathematics, Imperial College London

[2] Department of Computer Science, University of Oxford