IMPERIAL

On the Limitations of Fractal Dimension as a Measure of Generalization

Charlie B. Tan, **Inés García-Redondo**, Qiquan Wang, Michael M. Bronstein and Anthea Monod

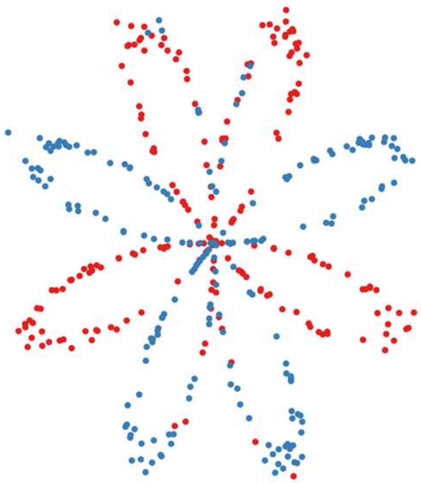
NeurIPS Poster, 2024
arXiv 2406.02234

Erlangen AI Hub Conference
Inés García-Redondo
11/06/2025

Learning Framework

Data Space

$$(\mathcal{Z} = \mathcal{X} \times \mathcal{Y}, \mathcal{F}_{\mathcal{Z}}, \mu_{\mathcal{Z}})$$



$\mathcal{X} = \mathbb{R}^2$

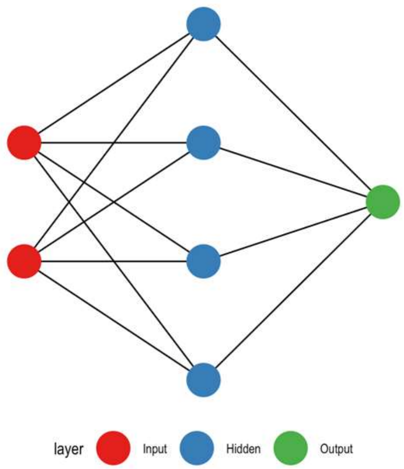
$\mathcal{Y} = \{\text{blue}, \text{red}\}$

$\mu_{\mathcal{Z}}$ data generating distribution (unknown)

$S = \{z_1, \dots, z_n\} \sim \mu_{\mathcal{Z}}^{\otimes n}$

Neural Network

$$h_\omega : \mathcal{X} \to \mathcal{Y}, \qquad \omega \in \mathbb{R}^d$$

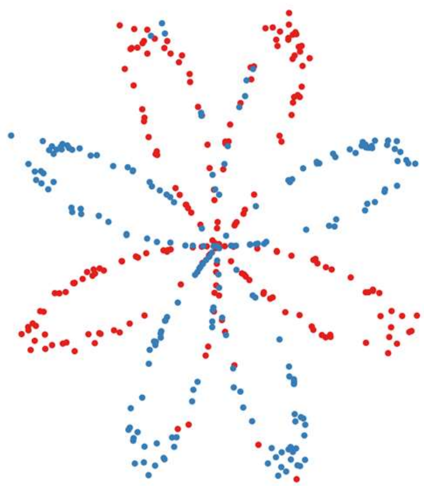


$$\ell: \mathbb{R}^d \times \mathcal{Z} \to \mathbb{R}, \qquad \ell(\omega, z) = \mathcal{L}(h_\omega(x), y)$$

Learning Framework

Data Space

$$(\mathcal{Z} = \mathcal{X} \times \mathcal{Y}, \mathcal{F}_{\mathcal{Z}}, \mu_{\mathcal{Z}})$$



$$\mathcal{X} = \mathbb{R}^2$$

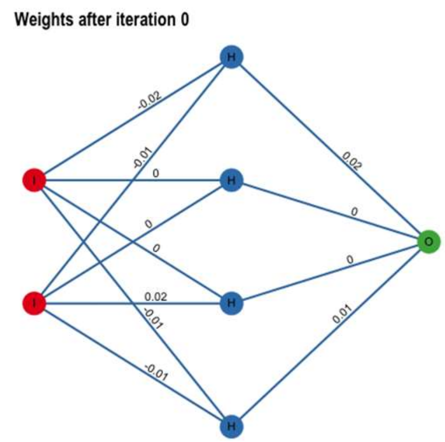
$$\mathcal{Y} = \{\text{blue}, \text{red}\}$$

$\mu_{\mathcal{Z}}$ data generating distribution (unknown)

$$S = \{z_1, \dots, z_n\} \sim \mu_{\mathcal{Z}}^{\otimes n}$$

Neural Network

$$h_\omega : \mathcal{X} \to \mathcal{Y}, \qquad \omega \in \mathbb{R}^d$$



finite sample $\boldsymbol{\omega} = \{\omega_1, \dots, \omega_n\}$ over the <u>optimization trajectory</u> for the weights $\mathcal{W}_S \subset \mathbb{R}^d$

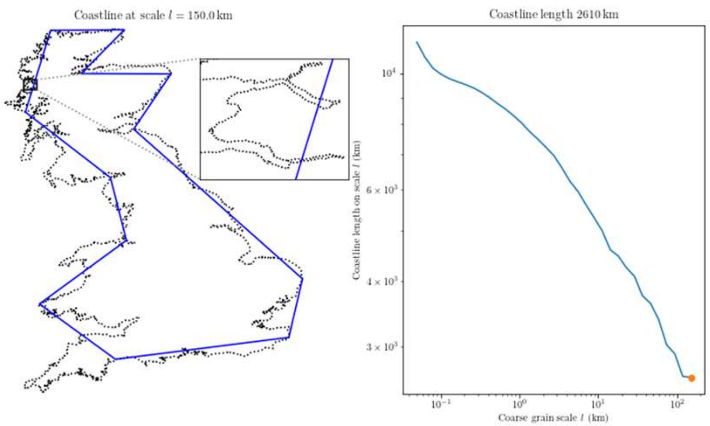
$$\ell : \mathbb{R}^d \times \mathcal{Z} \to \mathbb{R}, \qquad \ell(\omega, z) = \mathcal{L}(h_\omega(x), y)$$

$$\hat{\mathcal{R}}(\omega, S) \coloneqq \frac{1}{n} \sum_{i=1}^{n} \ell(\omega, z_i) \qquad \mathcal{R}(\omega) \coloneqq \mathbb{E}_{z \sim \mu_{\mathcal{Z}}}[\ell(\omega, z)]$$

$$\mathcal{G}(\omega) \coloneqq \left| \mathcal{R}(\omega) - \hat{\mathcal{R}}(\omega, S) \right|$$

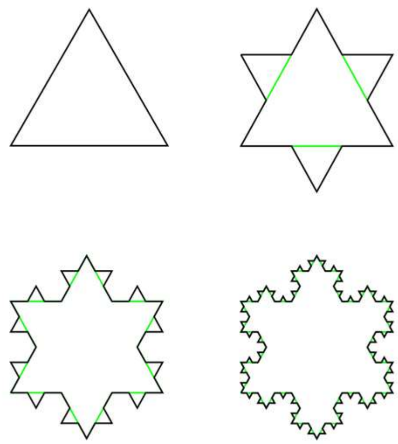
Fractals

How to describe shapes that are *rough* when you zoom in?

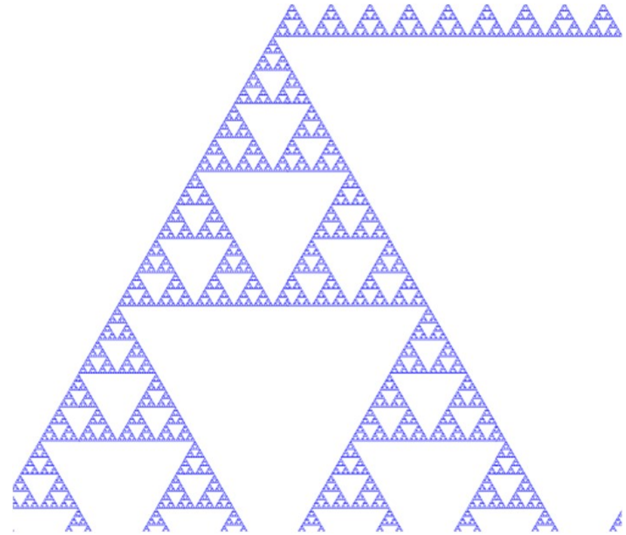


Coast of the UK, from Wikipedia

The coastline paradox



Von Koch snowflake, from Wikipedia

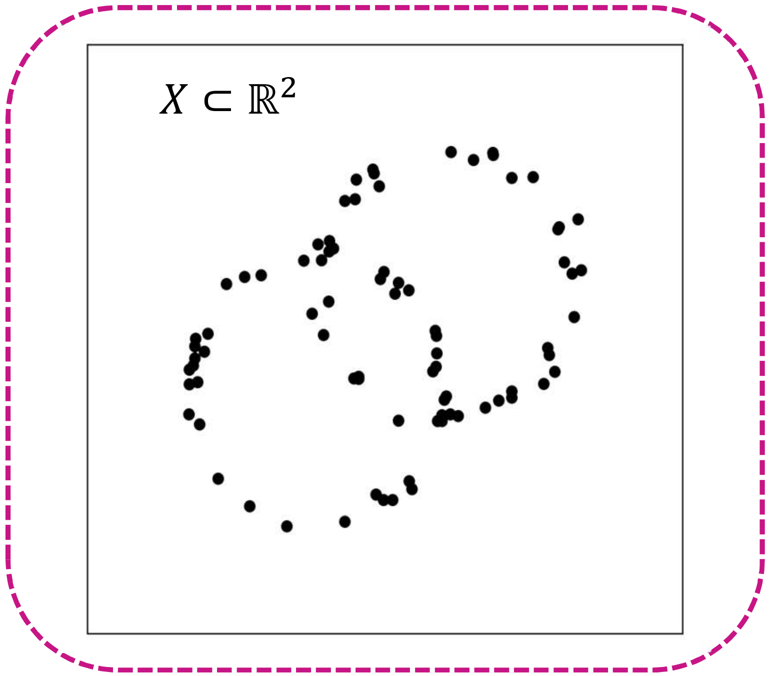


Zooming in the Sierpinski Triangle, from Wikipedia

Define a notion of "dimension" that captures this roughness…

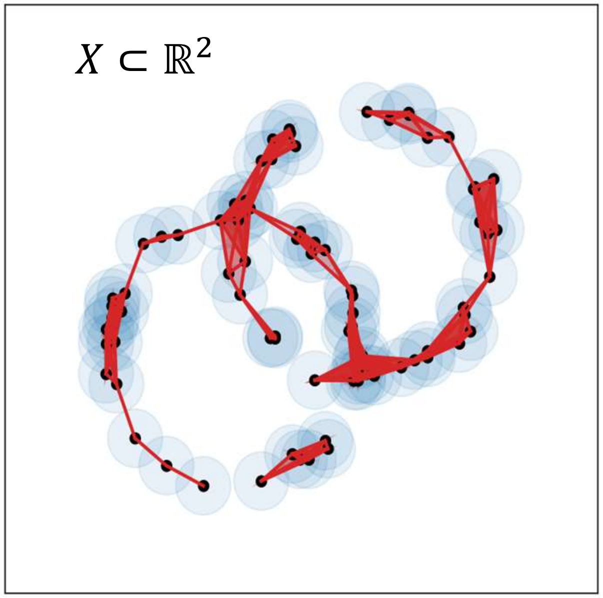
Persistent Homology

Data | Filtration | Invariants | Analysis



Persistent Homology

Data | Filtration | Invariants | Analysis

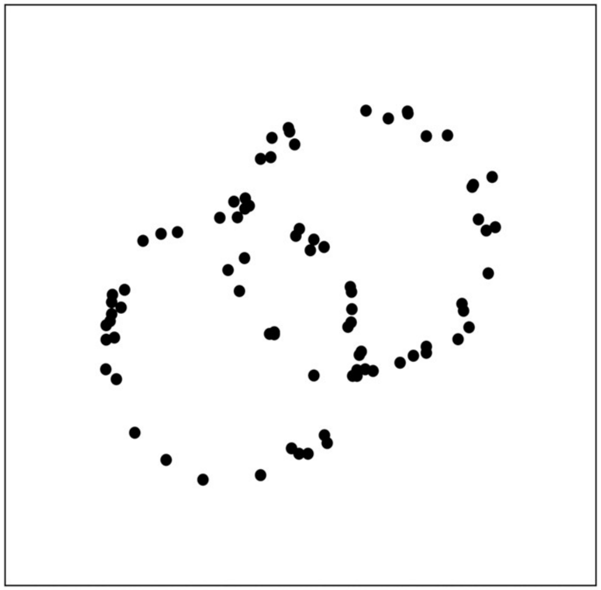


Persistent Homology

Data | Filtration | Invariants | Analysis

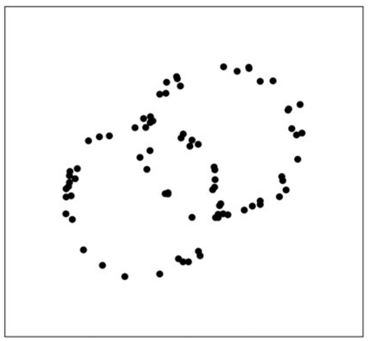
$\{X_t : t \in \mathbb{R}\}$

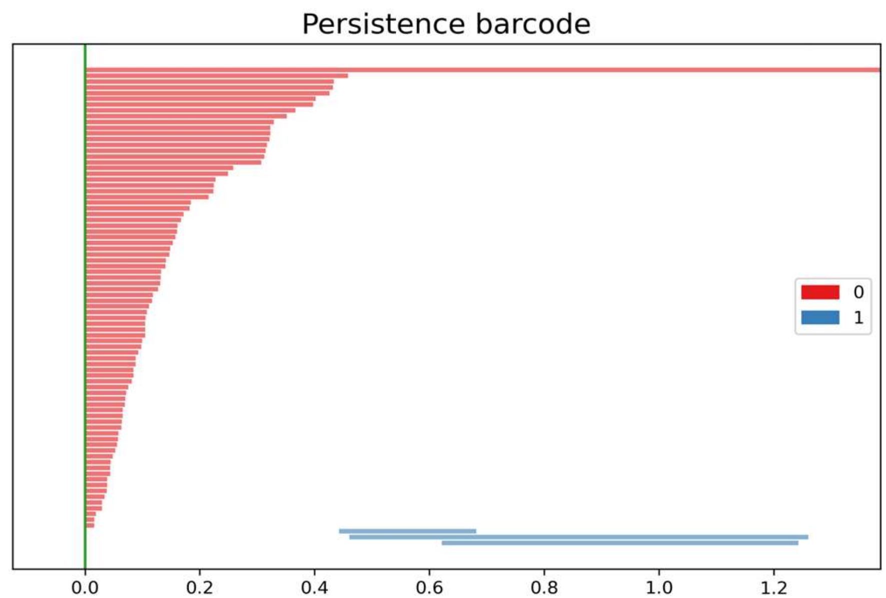
$t \leq s,\ X_t \subset X_s$



Persistent Homology

Data → Filtration → Invariants → Analysis





Persistence Barcode

For homological degree $k \in \mathbb{Z}$:

$$B_k(X) = \{[b_i, d_i) \subset \mathbb{R}: i \in I\}$$

PH dimension

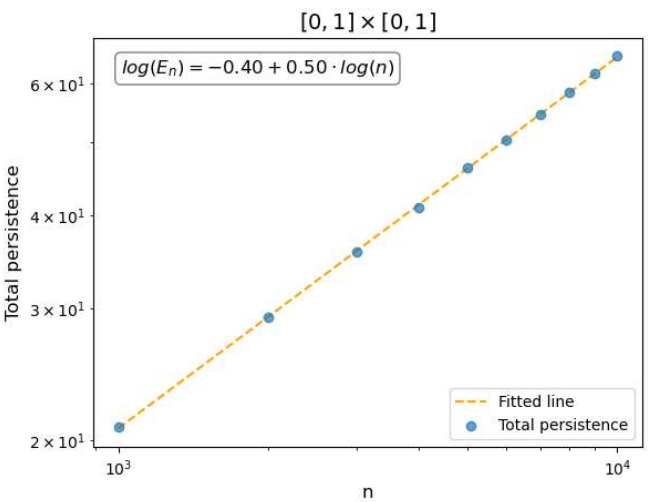
- Let $\boldsymbol{x} = \{x_1, \dots, x_n\} \subset S$ be a sample from some shape
- Compute the sum of the lengths of 0-bars (*total persistence*)

$$E_n(\boldsymbol{x}) = \sum_{(b,d)\in PH_0(\boldsymbol{x})} |d-b|$$

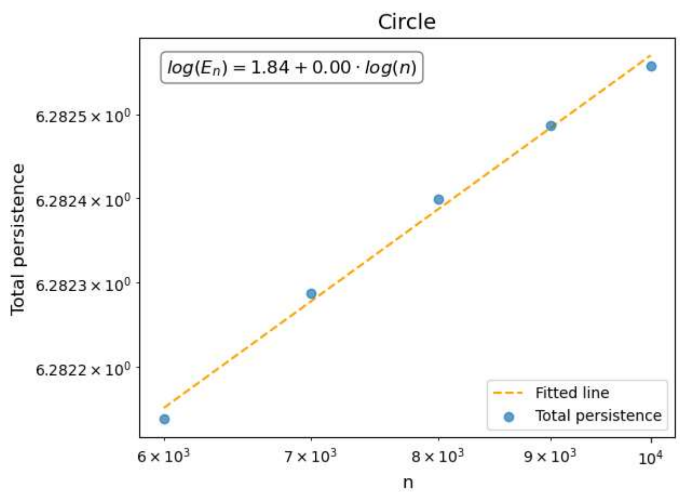
- Repeat for increasing n and fit a line $\log E_n \approx m \cdot \log n + b$

PH dimension ($\dim_{\mathrm{PH}}$)

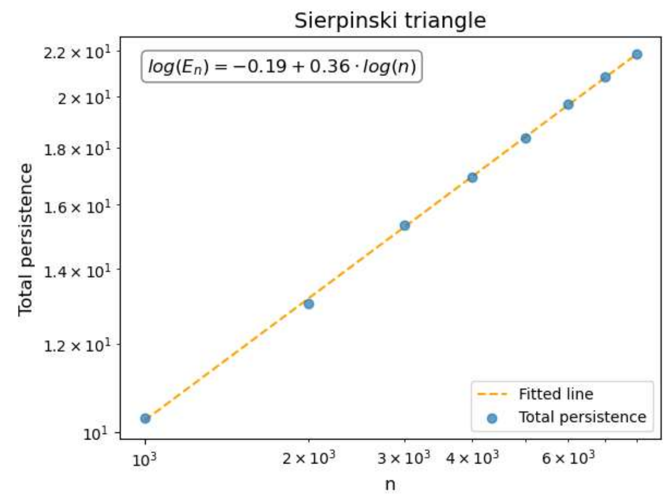
- Thesis of Vanessa Robins
- Adams et al. (2020)
- Schweinhart (2020, 2021)
- Jaquette and Schweinhart (2020)



$$\frac{1}{1-m} = 1.985 \approx 2$$



$$\frac{1}{1-m} \approx 1$$



$$\frac{1}{1-m} = 1.571 \approx \frac{\log 3}{\log 2}$$

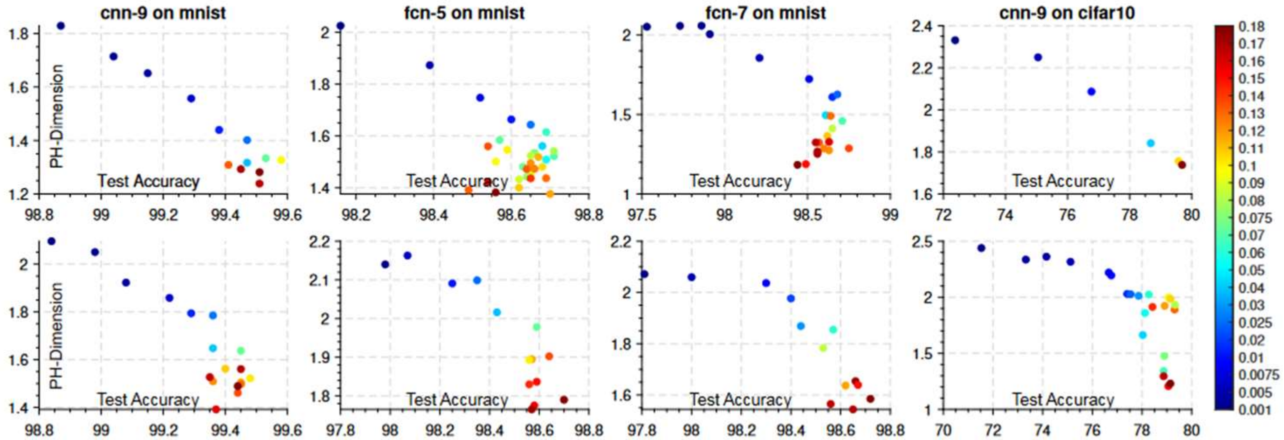
Fractal Dimension and Generalization

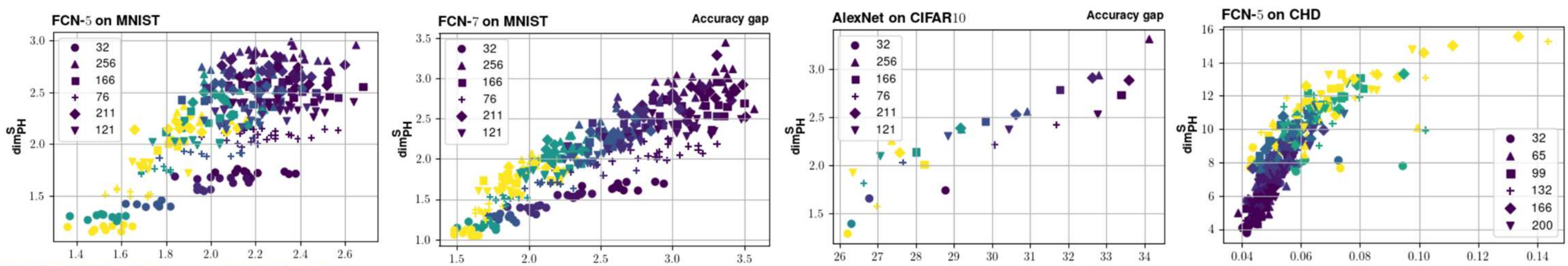
$$\sup_{\omega \in \mathcal{W}_S} \left| \mathcal{R}(\omega) - \hat{\mathcal{R}}(\omega, S) \right| \leq B \sqrt{\frac{\dim_{\mathrm{PH}}(\mathcal{W}_S) + I(\mathcal{W}_S, S) + \log(1/\zeta)}{n}}$$

Birdal et al. (2021) and Dupuis et al. (2023)

They also observed a **positive correlation** between generalization gap and PH dimension supporting this theory.

$$\text{accuracy gap} = \text{train accuracy} - \text{test accuracy}$$





Our experiments and analyses

Experimental design:

- **Networks:** FCN-5, 7 layers, AlexNet and a CNN
- **Datasets**: classification - MNIST, CIFAR-10, CIFAR-100; regression – CHD
- Train using **SGD** (with learning rate and batch sizes in a 6×6 grid) until 100% training accuracy
- Run **5000 additional iterations** to obtain sample of weights near the local minimum
- **Compute 0-dim PH dimension** using
  - Euclidean metric in $\mathbb{R}^d$
  - Loss-based pseudo-metric (Dupuis et al., 2023): $\rho_S(\omega, \omega') = \frac{1}{n}\sum_{i=1}^{n}|\ell(\omega, z_i) - \ell(\omega', z_i)|$
- Compute correlation of PH dimension with **absolute value accuracy/loss gap**

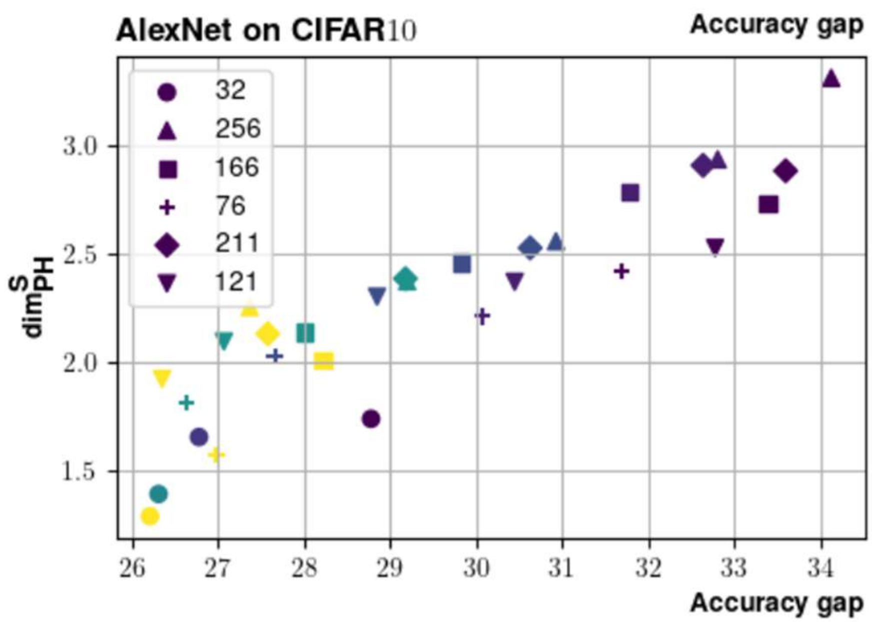
Statistically grounded analysis of the correlation between PH dimension and the generalization error

1. **Grid correlations** + hyperparameters of the network
2. **Partial correlation** analysis
3. **Conditional Independence**

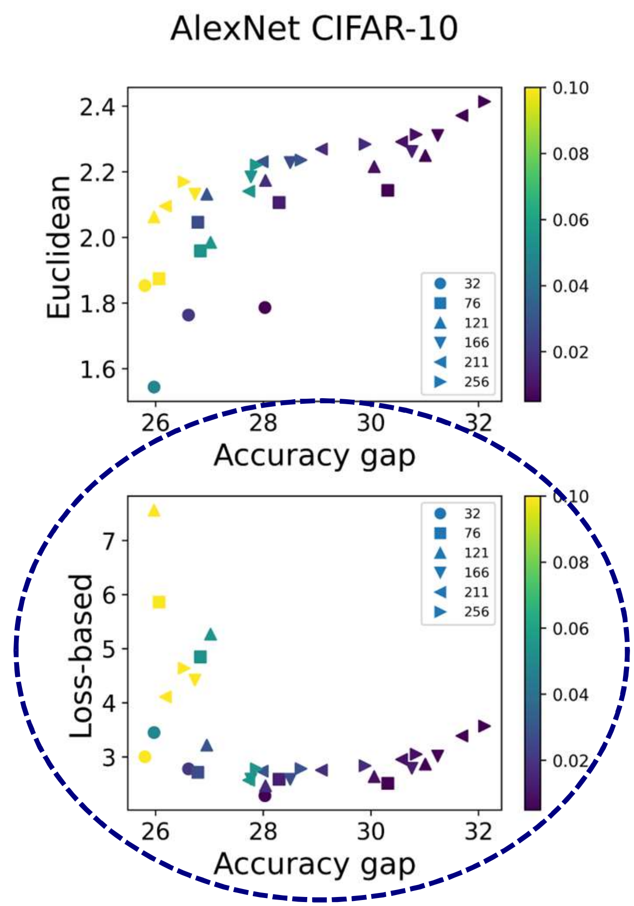
Found two situations where fractal dimension fails to predict the generalization error

1. Adversarial initialization
2. Double-descent model

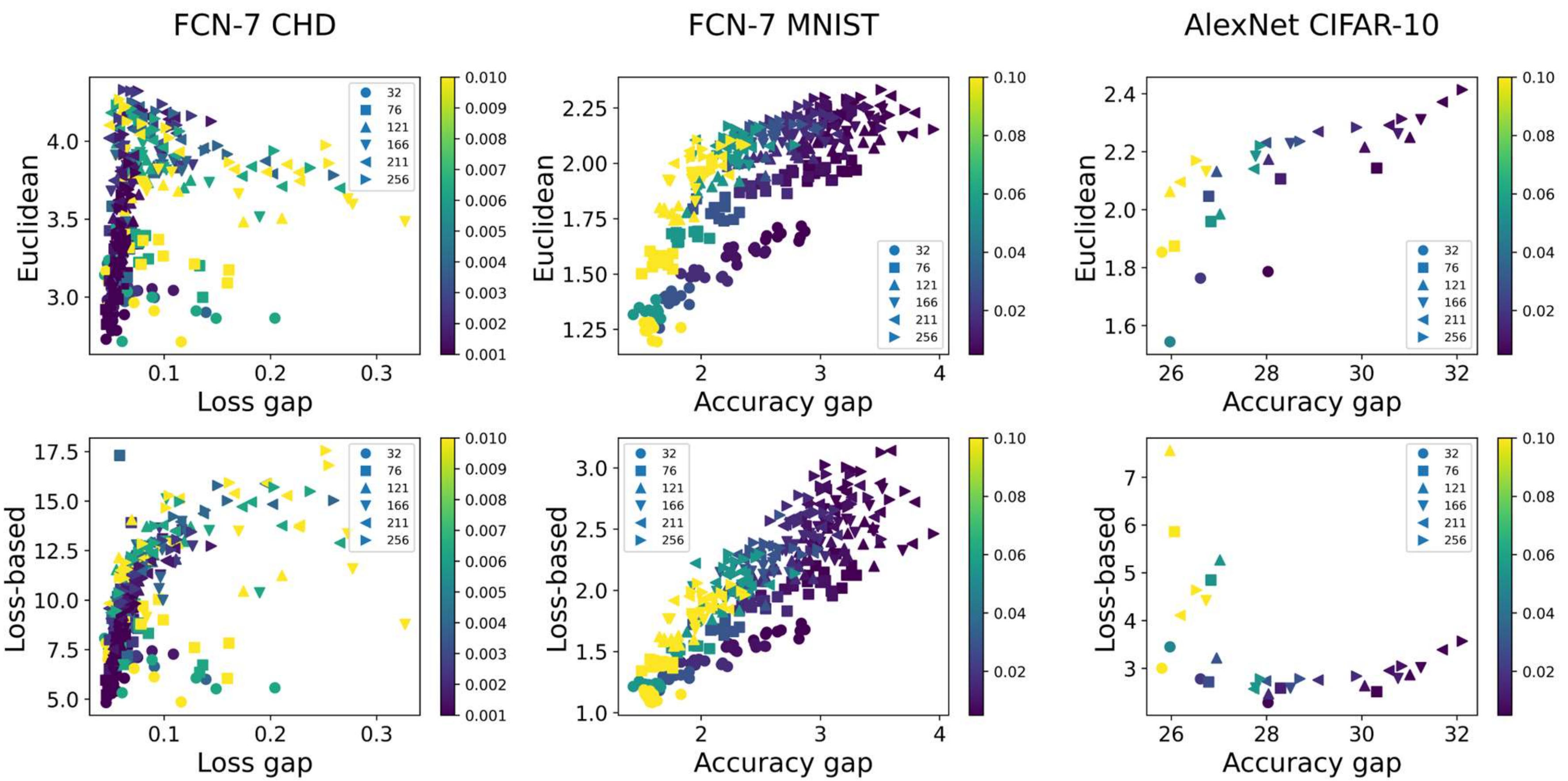
Grid correlations



Dupuis et al., (2023)

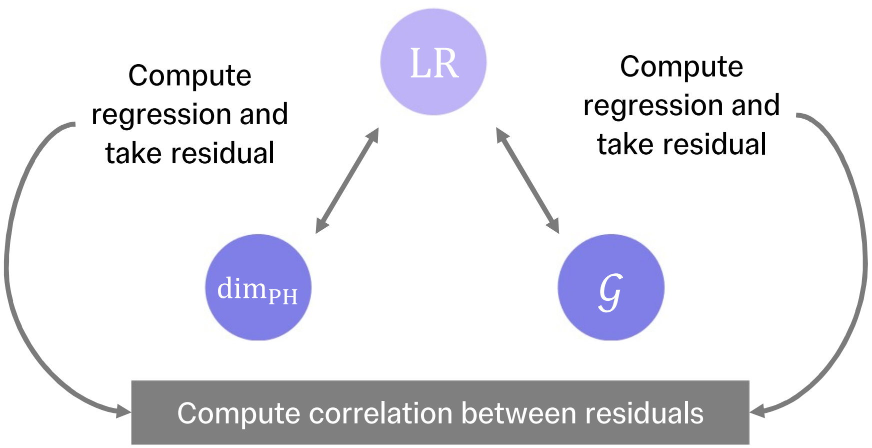


Grid correlations



Partial Correlation Analysis

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



Low coefficient means that the correlation between PH dimension and generalization can be explained by learning rate

+

Non-parametric permutation-type hypothesis test

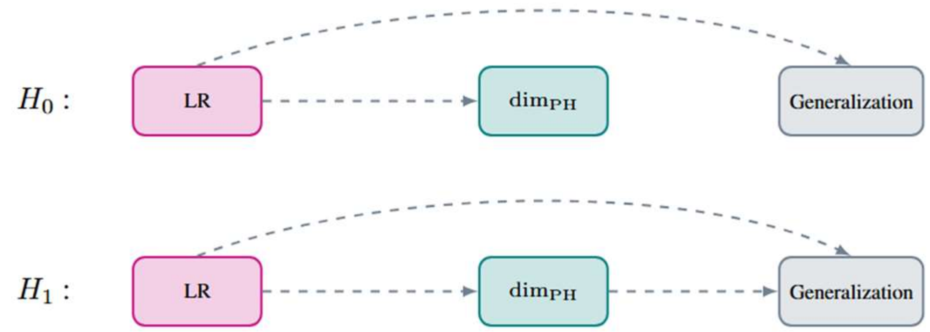
Partial Correlation given Learning Rate is **statistically significant for most cases**

- **Euclidean PH dimension:**
  - FCN-5 with MNIST and CHD shows significant partial correlation for most batch sizes
  - FCN-7 with MNIST and CHD has similar results, except for smaller batch sizes
- **Loss-based PH dimension**
  - FCN-5 with MNIST has significant partial correlation for bigger batch sizes, with CHD with smaller batch sizes
  - FCN-7 shows partial correlation in half of cases, but patterns are not apparent

Conditional independence test

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

- Use **Conditional Mutual Information** (CMI), a statistic that vanishes if and only if
$$\mathrm{dim_{PH}} \perp \mathcal{G} \mid \mathrm{LR}$$
- Generate null distribution for the CMI under **local permutations** of X and Y (Kim et al., 2022).
- **Hypothesis test:** null hypothesis implies that X and Y are conditionally independent



- For all models **trained on MNIST**, for most batch sizes, PH dimensions and Generalization can be considered conditionally independent (H_0)
- For all models **trained on CHD**, for most batch sizes, PH dimensions and Generalization can be considered conditionally independent (H_1)

Main takeaways

Grid correlations

What happens if we study correlation with **other hyperparameters** of our experiments?

Significant correlations with other hyperparameters. Confounding variables?

Partial correlation

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

Significant influence of learning rate, for some batch sizes.

Conditional independence

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

PH dimension and generalization gap conditionally independent on MNIST but not on CHD.

Adversarial Initialization

In the proposed theory there is no mention to how the initialization of the model could affect the proposed correlation. We test this theory on adversarially initialized models.

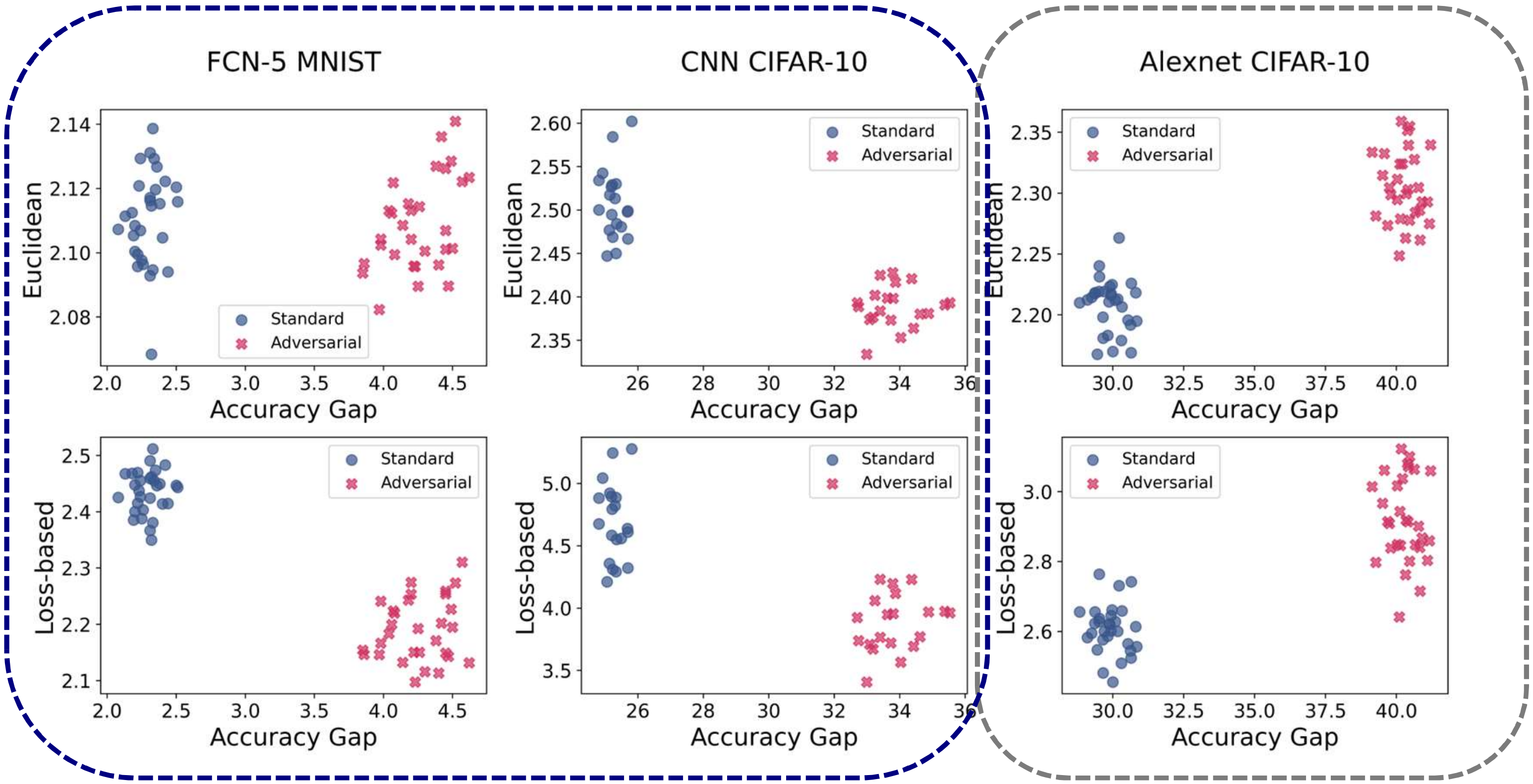
Adversarial initialization (Liu et al., 2020):

- Randomize labels on training data
- Train model in randomized training
- Use optimized model as initialization for a regular training
- The resulting model will have **bad generalization** properties (big generalization gap)
- We expect these models to have **big PH dimension**

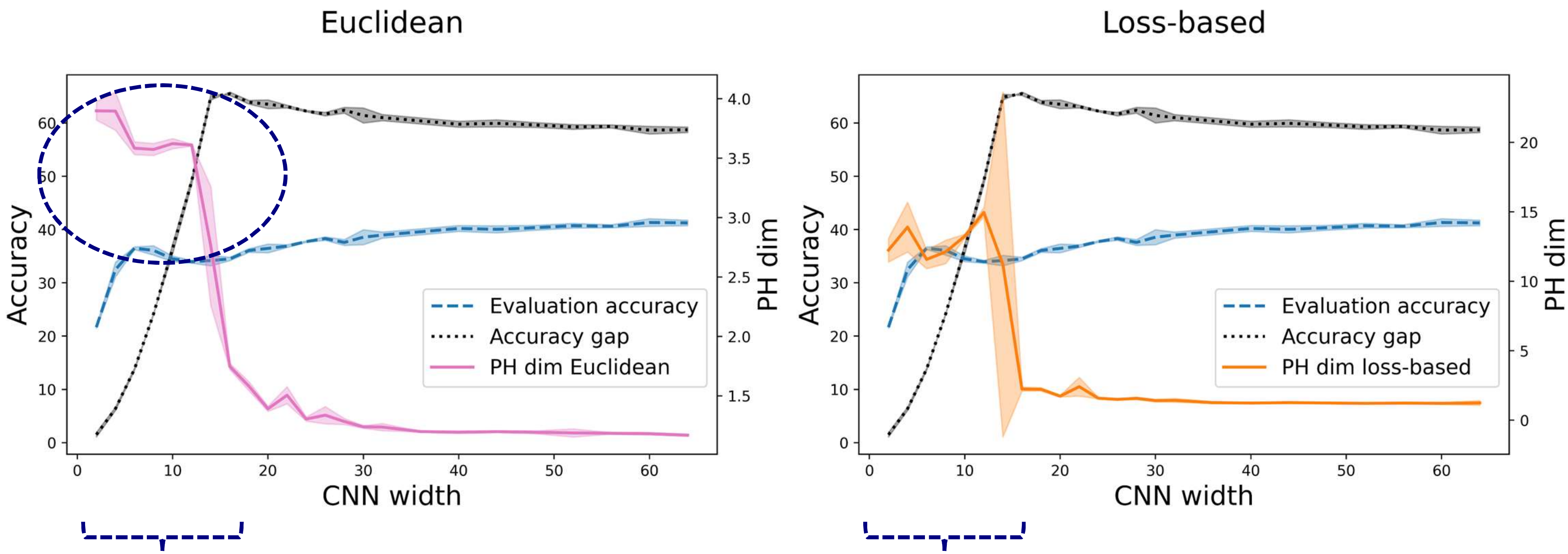
Standard initialization

- Models from standard, random initial points will tend to have **better generalization** properties
- We expect these have **smaller PH dimension**

Failure of PH dimension to predict Generalization



Double Descent (Nakkiran et al., 2021)



Conclusion and future work

The **observed correlations** in previous literature appear to be **influenced by the hyperparameter** choices

PH dimension **fails to positively correlate** with generalization gap for **poorly initialized** models and lower widths of the **double descent** experiment

Future work

- Extend results to larger ranges of hyperparameters
- Extend to other models, more parameters in the networks
- Explore theoretically the bounds
  - Conditional Mutual Information term?
  - Proofs are obscure to us, what are the implications of the assumptions in the choices of the architectures?
- Different topological measures? Different definition of the PH dimension?
  - Andreeva et al. (2024) – Other measures based on magnitude and other topological tools

IMPERIAL

Thank you
Questions?

On the Limitations of Fractal Dimension as a Measure of Generalization
11/06/2025