AIDOS LAB
AI FOR DATA-ORIENTED SCIENCE

UNI FR

Using topology and geometry to understand learning: generalization

Session 4 — Topological and Geometric Deep Learning: Theory, Methods and Applications

Universidad Complutense de Madrid

Inés García Redondo — May 13, 2026

Session 4 — Using topology and geometry to understand learning: generalization

Outline

- The problem of generalization in overparameterized neural networks
- Neural persistence: a complexity measure based on PH
- Predicting generalization from the PH of activations
- Fractal dimension
  - Computing fractal dimension with PH
- Bounding the generalization gap with the fractal dimension of the optimization trajectory
- Bounding the generalization gap with the fractal dimension of the optimization trajectory: behavior in practice

Motivation

Session 4 — Using topology and geometry to understand learning: generalization

Inés García Redondo — May 13, 2026

Training NNs: revisited

Generalization

Data space: Probability space: $(Z, \mathcal{F}, \mu)$ $Z = X \times Y$

Training data: $\mathcal{Z} = \{(x_i, y_i) \in X \times Y : 1 \leq i \leq n\}$

Task: Learn: $\hat{y} = f_\theta(x),\ \theta \in \mathbb{R}^m$ (Model)

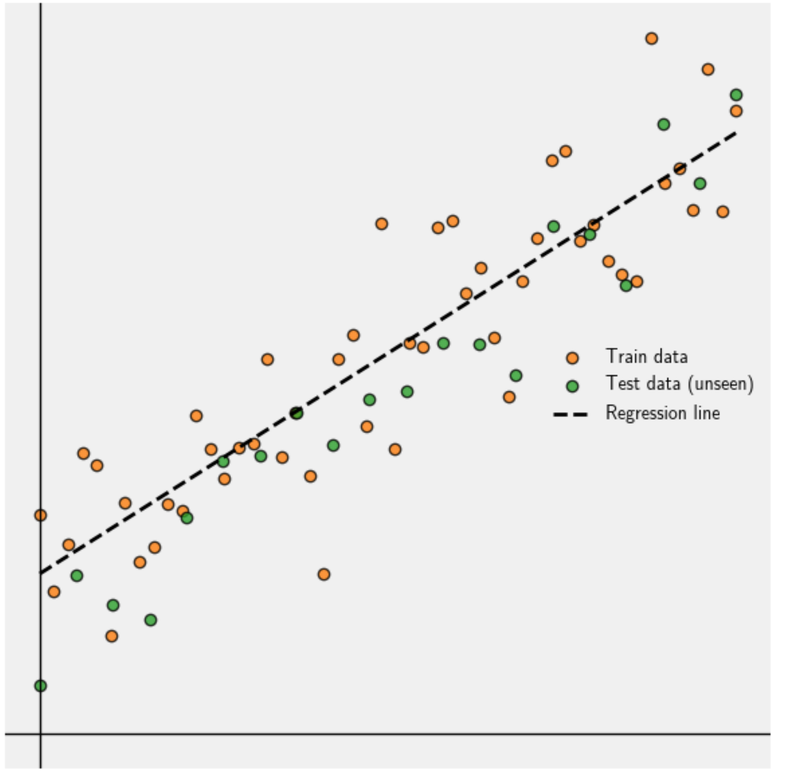
s.t. $\widehat{\mathcal{R}}(\mathcal{Z}, \theta) = \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}\left(f_\theta\left(x_i\right), y_i\right)$ is minimized

(Empirical risk; Loss)

Population risk: $\mathcal{R}(\mu, \theta) = \mathbb{E}_{z \sim \mu}[\mathcal{L}(f_\theta(x), y)]$

Generalization gap: $\mathcal{G}(\mathcal{Z}, \mu, \theta) = |\mathcal{R}(\mu, \theta) - \widehat{\mathcal{R}}(\mathcal{Z}, \theta)|$

Typically approximated as the difference between the risk over the training and the test data



$X = \mathbb{R}, \quad Y = \mathbb{R}$

Linear regression $f_{m,b}(x) = mx + b$

Mean Squared Error $\widehat{\mathcal{R}} = MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

How to estimate the generalization error

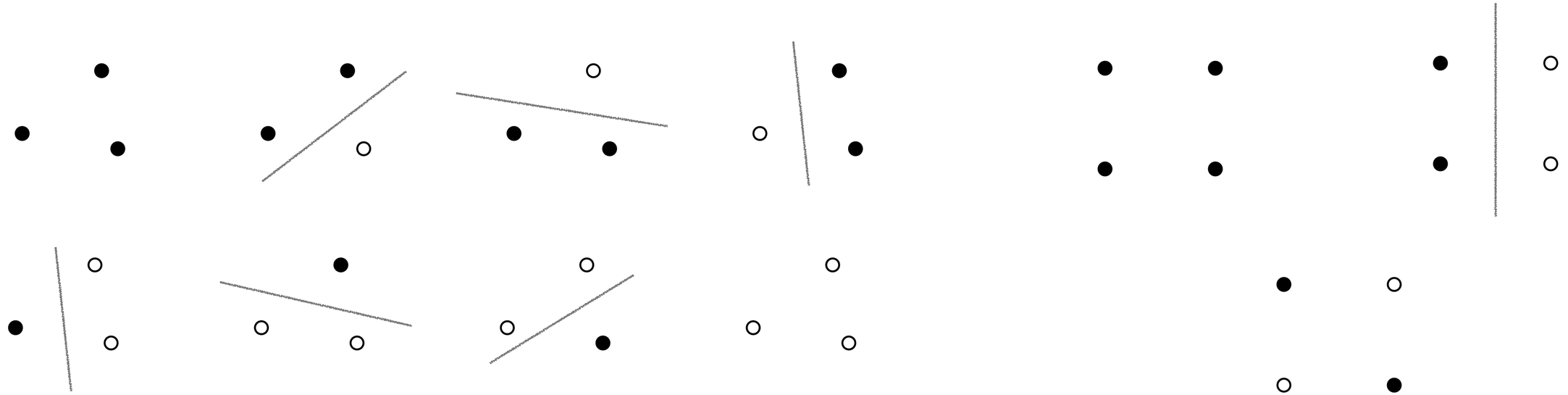
Classical statistical learning theory

Vapnik-Chervonenkis (VC) dimension

Size of the largest set of points that the model can *shatter,* i.e. classify correctly under all possible labels

$\mathrm{VC}(\mathcal{H}) = n \iff$ a model $h \in \mathcal{H}$ can shatter all sets of n points, but no set of $n+1$ points

Example: linear classifier in $\mathbb{R}^d$ has VC dimension of $d+1$



Vapnik, V. N., and A. Ya. Chervonenkis. "On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities." In *Theory of Probability and Its Applications,* Volume XVI, Number 2, pages 264–280, 1971.

How to estimate the generalization error

Classical statistical learning theory

Vapnik-Chervonenkis (VC) dimension — Size of the largest set of points that the model can *shatter,* i.e. classify correctly under all possible labels

$\mathrm{VC}(\mathcal{H}) = n \iff$ a model $h \in \mathcal{H}$ can shatter all sets of n points, but no set of $n+1$ points

Let $d = \mathrm{VC}(\mathcal{H})$

with probability at least $1 - \delta$

$$\mathcal{R}(h) \leq \widehat{\mathcal{R}}(\mathcal{Z}, h) + O\left(\sqrt{\frac{d \log(n/d) + \log 1/\delta}{n}}\right)$$

For a NN with M parameters and linear activations $\mathrm{VC}(\mathcal{H}_{\mathrm{NN}}) = O(M \log M)$

Typically $M \approx 10^6$, so you need millions of samples for reliable generalization guarantees

Vapnik, V. N., and A. Ya. Chervonenkis. "On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities." In *Theory of Probability and Its Applications,* Volume XVI, Number 2, pages 264–280, 1971.

How to estimate the generalization error

Classical statistical learning theory

Rademacher complexity *given some specific training points*, how well can a model $h \in \mathcal{H}$ correlate with pure random noise?

$$\mathfrak{R}_{\mathcal{Z}}(\mathcal{H}) = \mathbb{E}_{\sigma}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i h(x_i)\right]$$

$\sigma_i \in \{-1, +1\}$ random noise labels

With probability at least $1-\delta$

$$\mathcal{R}(h) \leq \widehat{\mathcal{R}}(\mathcal{Z}, h) + 2\mathfrak{R}_{\mathcal{Z}}(\mathcal{H}) + O\left(\sqrt{\frac{\log(1/\delta)}{n}}\right)$$

RC gets worse with number of neurons in the layers and the values of the weights, and improves with the number of samples

Bartlett, Peter L., and Shahar Mendelson. "Rademacher and Gaussian Complexities: Risk Bounds and Structural Results." *Journal of Machine Learning Research* 3, no. Nov (2002): 463–82.

Generalization in NNs

The paradox

Takeaways:

- VC dimension grows with the number of parameters, so networks should overfit
- Rademacher complexity increases with the richness of the model, so networks should fit noise

However, in practice, neural networks generalize well. **Why?**

We haven't yet mentioned an important ingredient in our construction: the *optimization algorithm*

These measures assume that all models can be reached, but maybe this is not what we have in practice

We need to develop *new complexity measures* that are able to capture this to bound the generalization error more tightly

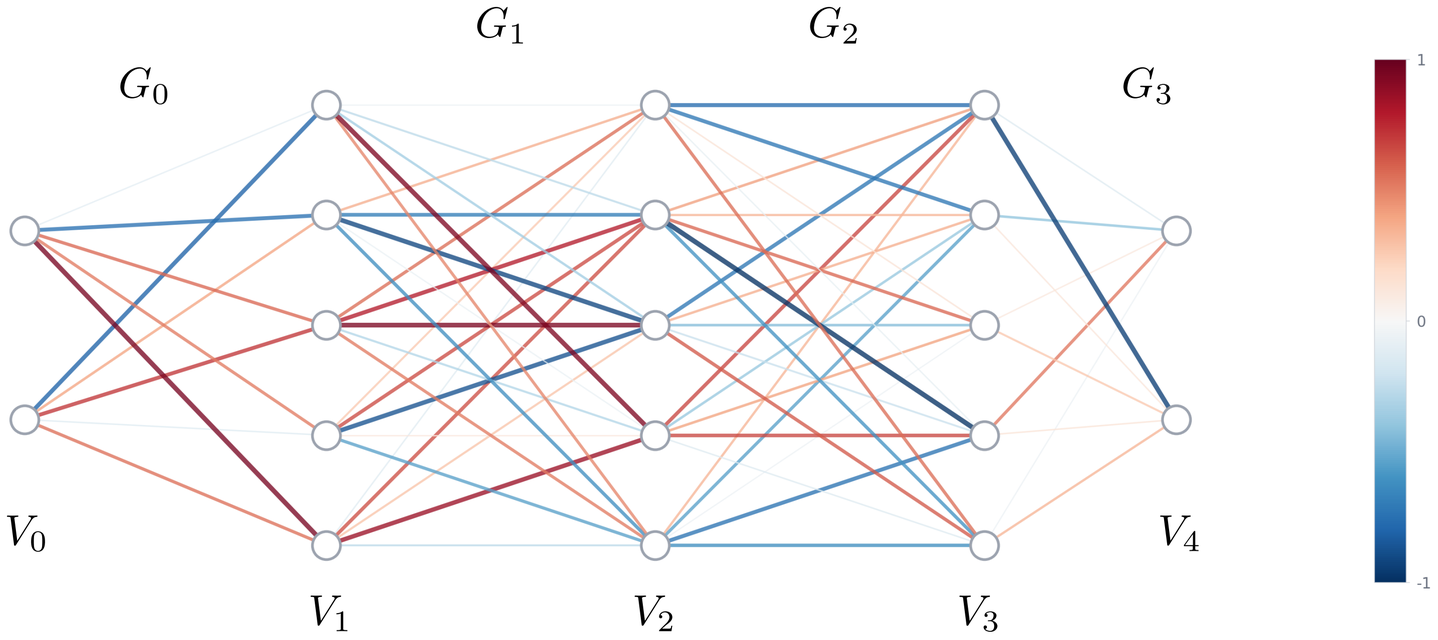
Zhang, Chiyuan, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. "Understanding Deep Learning Requires Rethinking Generalization." arXiv:1611.03530. Preprint, arXiv, February 26, 2017.

Neural persistence

Session 4 — Using topology and geometry to understand learning: generalization

Inés García Redondo — May 13, 2026

Seeing an MLP as a stratified graph



We can see an MLP as a *stratified graph*: $G = (V, E)$

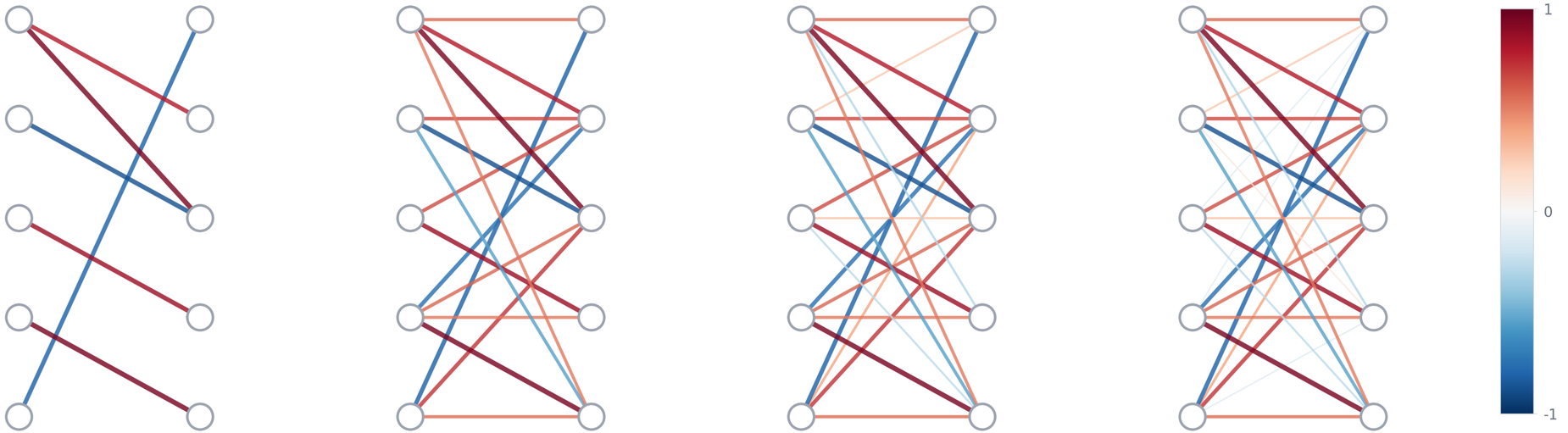
$$V = V_0 \sqcup V_1 \sqcup \cdots \sqcup V_L \qquad (u, v) \in E \iff \exists k \in \{0, \ldots, L\},\ u \in V_k \wedge v \in V_{k+1}$$

With a weight function on the edges: $\varphi : E \to \mathcal{W}$

$$G_k = (V_k \sqcup V_{k+1}, E \cap \{(u, v) \in V_k \times V_{k+1}\})$$

Rieck, Bastian, Matteo Togninalli, Christian Bock, et al. Neural Persistence: A Complexity Measure for Deep Neural Networks Using Algebraic Topology. May 2023, 25 p.

Obtaining a filtration



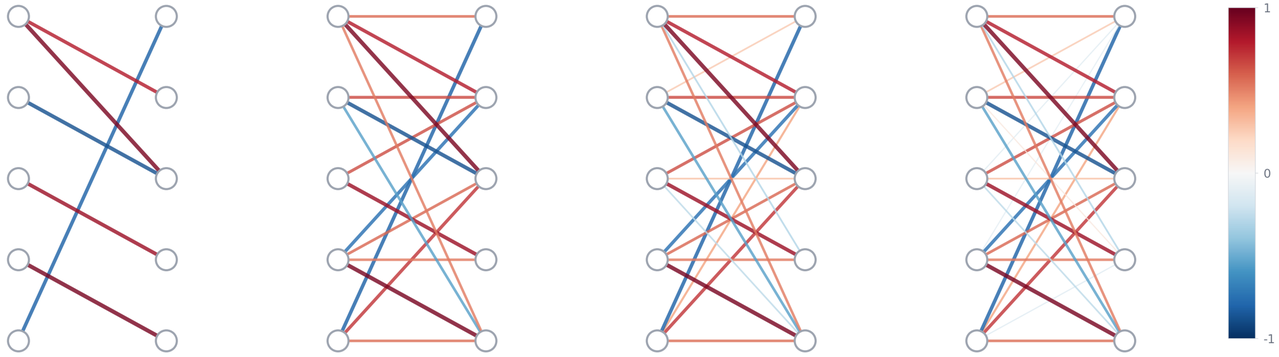
$$G_1^{(1)} \subseteq G_1^{(2)} \subseteq G_1^{(3)} \subseteq G_1^{(4)}$$

$$|w| > 0.75 \qquad |w| > 0.5 \qquad |w| > 0.25 \qquad |w| > 0$$

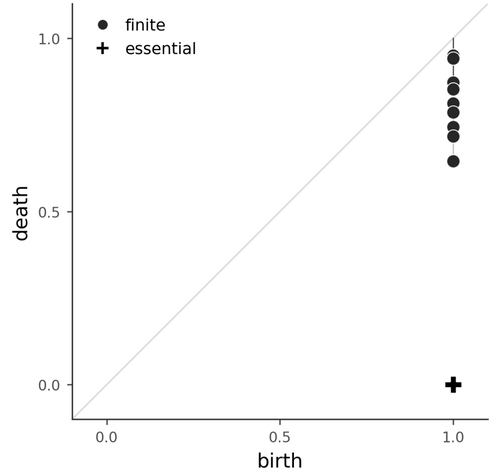
Rieck, Bastian, Matteo Togninalli, Christian Bock, et al. Neural Persistence: A Complexity Measure for Deep Neural Networks Using Algebraic Topology. May 2023, 25 p.

Compute PH

Filtration for G_k



Persistence diagram $\mathcal{D}_k$



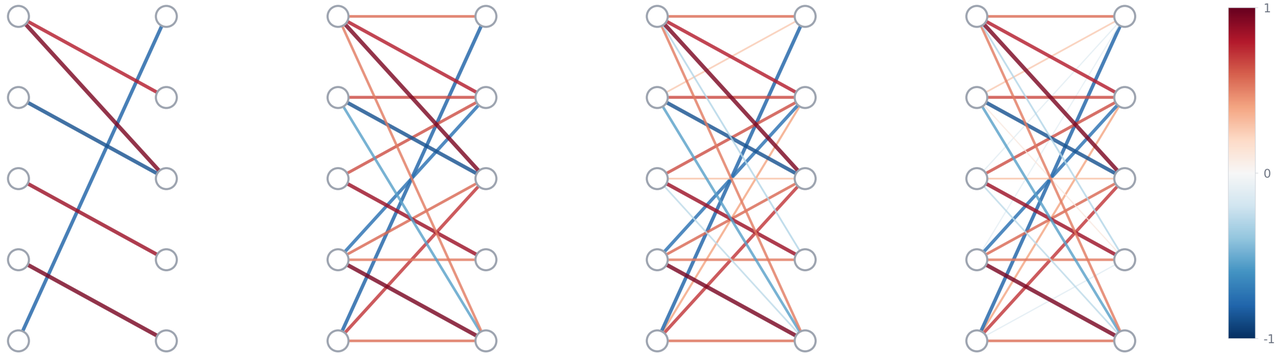
- Compute 0-dimensional PH
- All points are born at $w = 1$, and die for $w < 1$ as we add edges,
- Hence, the points in the diagram appear below the diagonal
- There is one essential component that stays alive until the end of the filtration

Rieck, Bastian, Matteo Togninalli, Christian Bock, et al. Neural Persistence: A Complexity Measure for Deep Neural Networks Using Algebraic Topology. May 2023, 25 p.

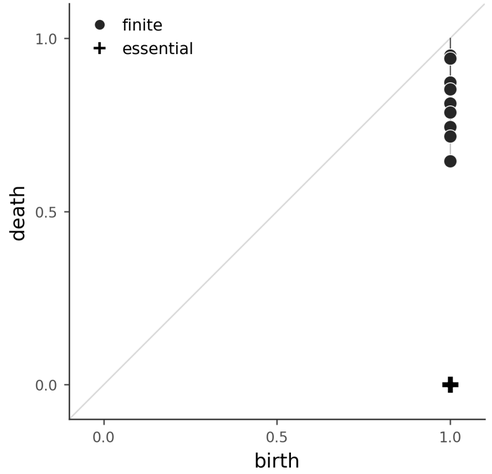
Neural persistence

For one layer

Filtration for G_k



Persistence diagram $\mathcal{D}_k$



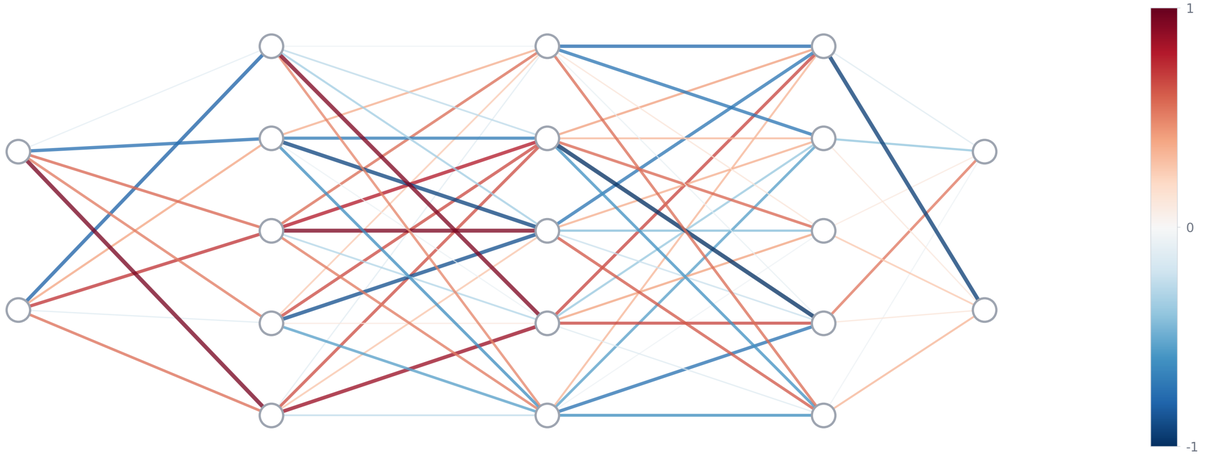
$$\mathrm{NP}(G_k) := \|\mathcal{D}_k\| = \left(\sum_{(c,d)\in\mathcal{D}_k} |d-c|^p \right)^{1/p}$$

$$0 \leq \mathrm{NP}(G_k) \leq \left(\max_{e\in G_k} w_e - \min_{e\in G_k} w_e \right) \left(|V_k \times v_{k+1}| - 1 \right)^{1/p}$$

Rieck, Bastian, Matteo Togninalli, Christian Bock, et al. Neural Persistence: A Complexity Measure for Deep Neural Networks Using Algebraic Topology. May 2023, 25 p.

Neural persistence

For the whole network



$$\mathrm{NP}(G_k) \coloneqq \|\mathcal{D}_k\| = \left(\sum_{(c,d)\in\mathcal{D}_k} |d-c|^p \right)^{1/p} \qquad 0 \leq \mathrm{NP}(G_k) \leq \mathrm{NP}(G_k)^+$$

$$\mathrm{NP}(G) \coloneqq \frac{1}{L} \sum_{l=1}^{L} \frac{\mathrm{NP}(G_k)}{\mathrm{NP}(G_k)^+}$$

Rieck, Bastian, Matteo Togninalli, Christian Bock, et al. Neural Persistence: A Complexity Measure for Deep Neural Networks Using Algebraic Topology. May 2023, 25 p.

Neural persistence in practice

Neural persistence captures complexity

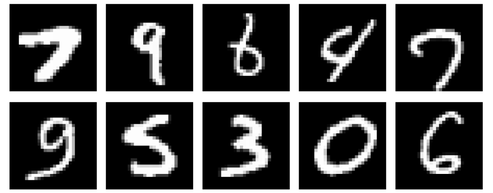
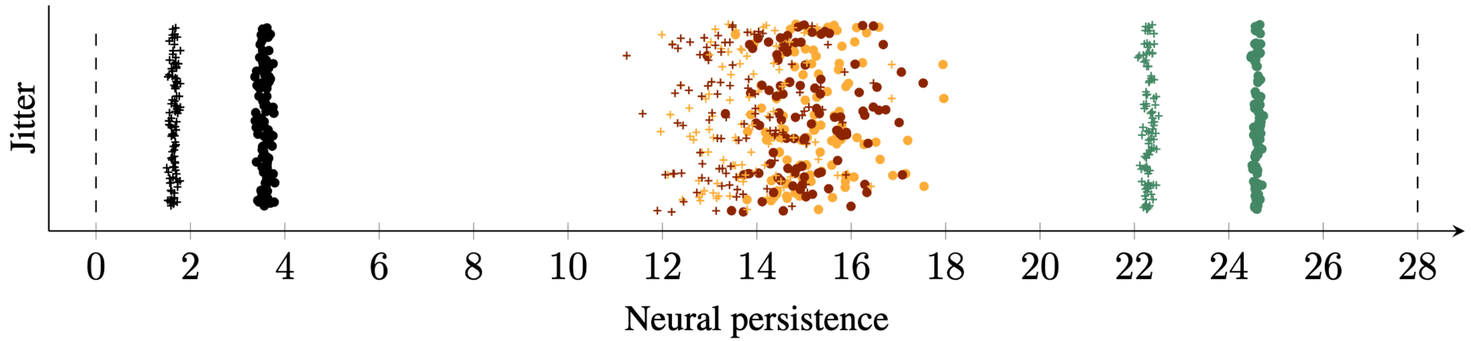


Figure from: Rieck et *al.* (2019)



$$\mathrm{NP}(G) := \frac{1}{L}\sum_{l=1}^{L}\frac{\mathrm{NP}(G_k)}{\mathrm{NP}(G_k)^+}$$

- Trained, converging models
- Trained, diverging models
- Gaussian weights
- Uniform weights

Rieck, Bastian, Matteo Togninalli, Christian Bock, et al. Neural Persistence: A Complexity Measure for Deep Neural Networks Using Algebraic Topology. May 2023, 25 p.

Neural persistence in relation to training

Good training practices increase neural persistence

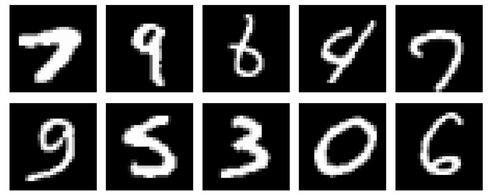
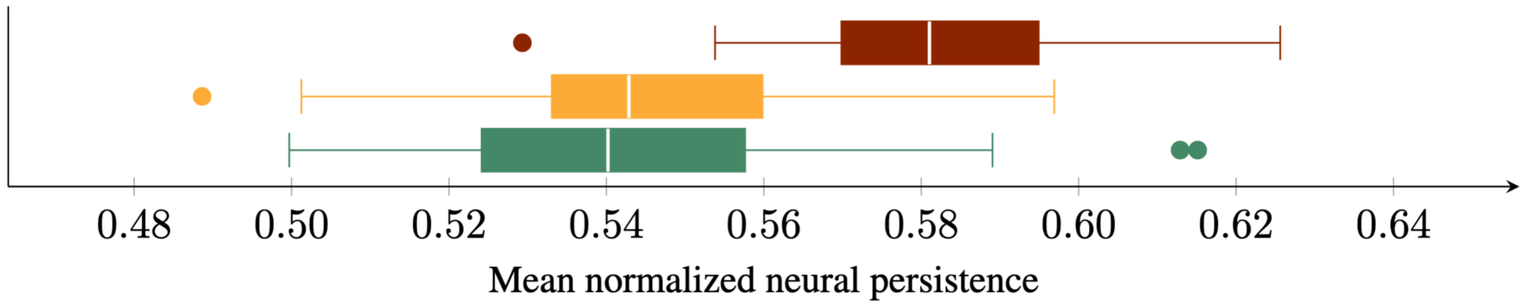


Figure from: Rieck et *al.* (2019)



$$\mathrm{NP}(G) := \frac{1}{L}\sum_{l=1}^{L}\frac{\mathrm{NP}(G_k)}{\mathrm{NP}(G_k)^+}$$

- No modification
- Batch normalization
- 50% dropout

Rieck, Bastian, Matteo Togninalli, Christian Bock, et al. Neural Persistence: A Complexity Measure for Deep Neural Networks Using Algebraic Topology. May 2023, 25 p.

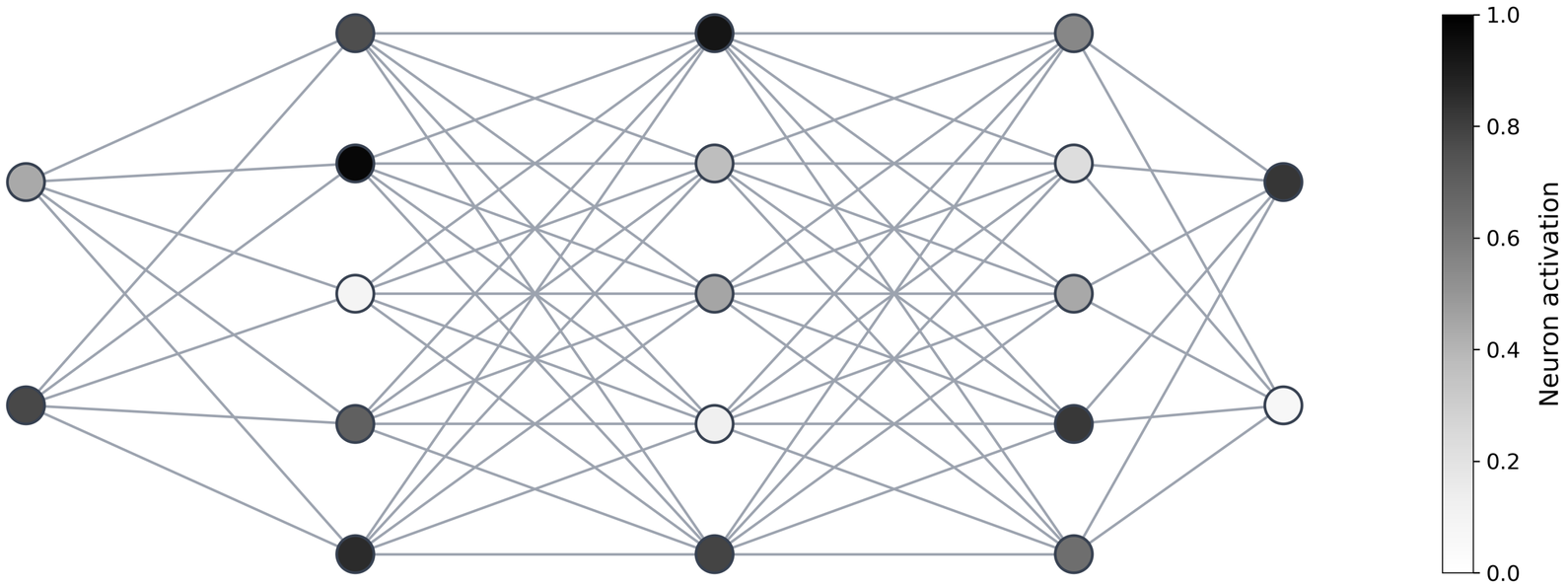
PH to predict the generalization gap

Session 4 — Using topology and geometry to understand learning: generalization

Inés García Redondo — May 13, 2026

A PH construction to predict generalization

Set up



Training data

$$\mathcal{Z} = \{(x_i, y_i) : 1 \leq i \leq n\}$$

Activation value at v of x

$$N_v(x)$$

Activation vectors

$$A_v(\mathcal{Z}) = (N_v(x))_{(x,y)\in\mathcal{Z}}$$

$$A_N(\mathcal{Z}) = \{A_v(\mathcal{Z}) : v \in N\}$$

Correlation distance

$$d(v_i, v_j) = 1 - |\mathrm{corr}(A_{v_i}(\mathcal{Z}), A_{v_j}(\mathcal{Z}))|$$

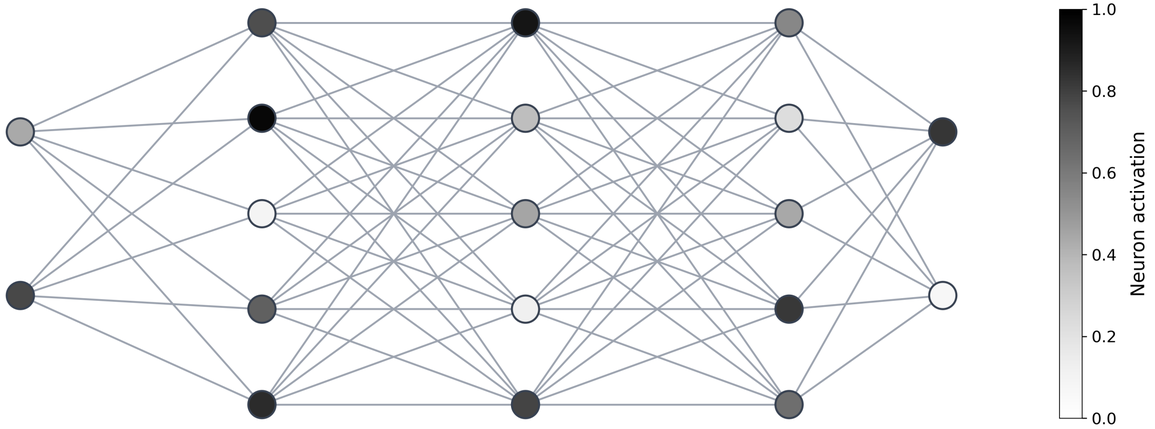
Pearson correlation coefficient

Ballester, Rubén, Xavier Arnal Clemente, Carles Casacuberta, Meysam Madadi, Ciprian A. Corneanu, and Sergio Escalera. "Predicting the Generalization Gap in Neural Networks Using Topological Data Analysis." *Neurocomputing* 596 (September 2024): 127787.

A PH construction to predict generalization

Persistence computations

Networks coming from the Predicting Generalization In Deep Learning (PGDL) competition



Activation vectors

$$A_v(\mathcal{Z}) = (N_v(x))_{(x,y)\in\mathcal{Z}}$$

$$A_N(\mathcal{Z}) = \{A_v(\mathcal{Z}) : v \in N\}$$

$$|A_N(\mathcal{Z})| = |\mathcal{Z}| \times |V|$$

Correlation distance

$$d(v_i, v_j) = 1 - |\mathrm{corr}(A_{v_i}(\mathcal{Z}), A_{v_j}(\mathcal{Z}))|$$

Pearson correlation coefficient

Computing PH for the whole point clouds becomes unfeasible

Instead consider subsamples:

$\mathcal{Z}' \subset \mathcal{Z}, \quad |\mathcal{Z}'| = 2000$ Randomly

$N' \subset N, \quad |N'| = 3000$ According to an *importance score*

Repeat 20 times, compute statistical representations of the diagrams, and combine

Jiang, Yiding, Pierre Foret, Scott Yak, et al. "NeurIPS 2020 Competition: Predicting Generalization in Deep Learning." *CoRR* abs/2012.07976 (2020).

Ballester, Rubén, Xavier Arnal Clemente, Carles Casacuberta, Meysam Madadi, Ciprian A. Corneanu, and Sergio Escalera. "Predicting the Generalization Gap in Neural Networks Using Topological Data Analysis." *Neurocomputing* 596 (September 2024): 127787.

A PH construction to predict generalization

Results

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y}_i)^2}$$

Table 1
Top three combinations of persistence summaries per task according to their respective mean of R^2 test values in the 10 experiments of the 5 × 2-fold cross-validation statistical test. **ASD:** Average and standard deviation of births and deaths. **ASDSQ:** Average and standard deviation of births and deaths, concatenated with the corresponding squared values; see Section 4.1.

Task 1		
Top TDA summaries	Best dim	R^2 score
ASDSQ	0 and 1	0.5601 ± 0.13
ASDSQ	1	0.4321 ± 0.12
ASD	1	0.3720 ± 0.14
Task 2		
Top TDA summaries	Best dim	R^2 score
ASD	1	0.9337 ± 0.01
ASD	0 and 1	0.9198 ± 0.02
ASDSQ	1	0.9166 ± 0.03

Ballester, Rubén, Xavier Arnal Clemente, Carles Casacuberta, Meysam Madadi, Ciprian A. Corneanu, and Sergio Escalera. "Predicting the Generalization Gap in Neural Networks Using Topological Data Analysis." *Neurocomputing* 596 (September 2024): 127787.

A PH construction to predict generalization

Results

Table 2
Comparison of our best performing summaries with state of the art: Average and standard deviation of R^2 scores for Task 1 and Task 2 computed from linear models trained in the ten cases of the 5 × 2-fold cross-validation.

	Task 1	Task 2
Interpex	-0.0518 ± 0.06	0.9500 ± 0.01
Always Generalize	0.9715 ± 0.01	0.8893 ± 0.02
BrAIn	0.4520 ± 0.08	0.7180 ± 0.04
Ours	0.5601 ± 0.13	0.9337 ± 0.01

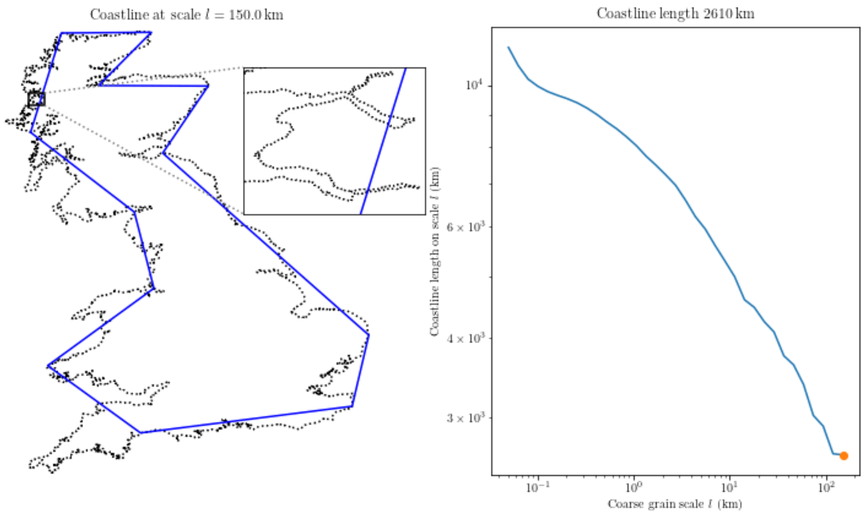
Ballester, Rubén, Xavier Arnal Clemente, Carles Casacuberta, Meysam Madadi, Ciprian A. Corneanu, and Sergio Escalera. "Predicting the Generalization Gap in Neural Networks Using Topological Data Analysis." *Neurocomputing* 596 (September 2024): 127787.

Fractal dimension

Session 4 — Using topology and geometry to understand learning: generalization

Inés García Redondo — May 13, 2026

Fractals

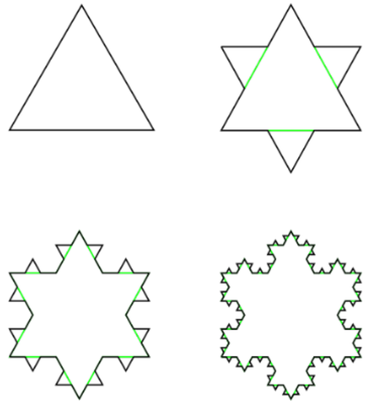


By Tveness - Own work, CC0, https://commons.wikimedia.org/w/index.php?curid=113571720

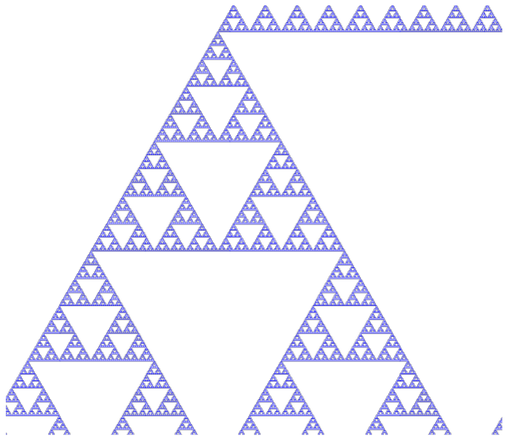
The coastline paradox

Self-similar shapes

Fractals can be seen as shapes that are *rough* when we zoom in



Von Koch snowflake, from Wikipedia



Zooming in the Sierpinski triangle, from Wikipedia

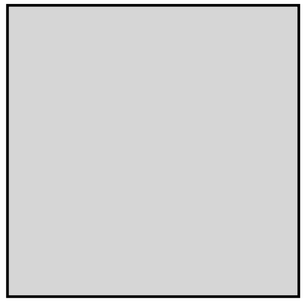
How can we define a notion of dimension that captures this *roughness?*

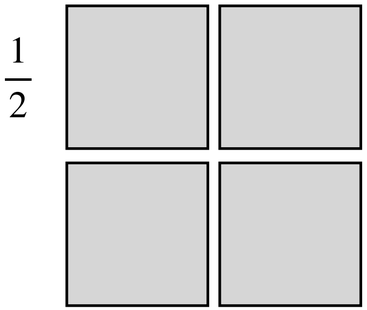
Intuition behind fractal dimension

	Line	**Square**	**Cube**	**Sierpinski**
Scale factor	$s = \frac{1}{2}$			
Measure of the scaled-down shape	$\mu(L_s) = \frac{1}{2}$			

Intuition behind fractal dimension

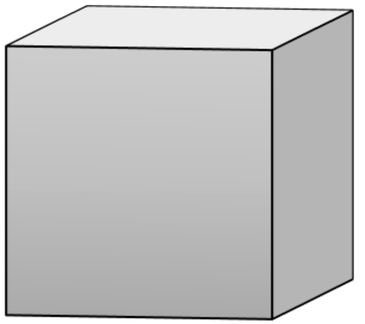
	Line	**Square**	**Cube**	**Sierpinski**
Scale factor	$s = \frac{1}{2}$	$s = \frac{1}{2}$		
Measure of the scaled-down shape	$\mu(L_s) = \frac{1}{2}$	$\mu(L_s) = \frac{1}{4}$		

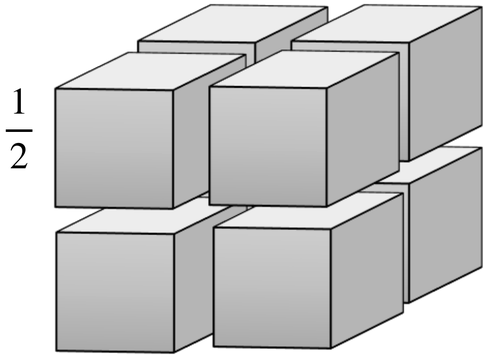




Intuition behind fractal dimension

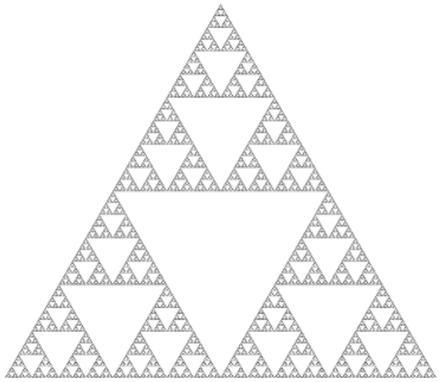
	Line	**Square**	**Cube**	**Sierpinski**
Scale factor	$s = \frac{1}{2}$	$s = \frac{1}{2}$	$s = \frac{1}{2}$	
Measure of the scaled-down shape	$\mu(L_s) = \frac{1}{2}$	$\mu(L_s) = \frac{1}{4}$	$\mu(L_s) = \frac{1}{8}$	

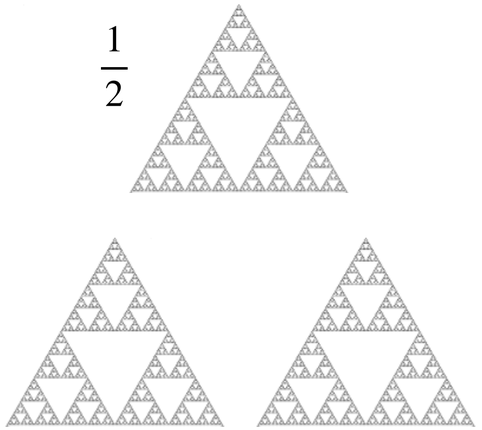




Intuition behind fractal dimension

	Line	**Square**	**Cube**	**Sierpinski**
Scale factor	$s = \frac{1}{2}$	$s = \frac{1}{2}$	$s = \frac{1}{2}$	$s = \frac{1}{2}$
Measure of the scaled-down shape	$\mu(L_s) = \frac{1}{2}$	$\mu(L_s) = \frac{1}{4}$	$\mu(L_s) = \frac{1}{8}$	$\mu(L_s) = \frac{1}{3}??$





Intuition behind fractal dimension

	Line	**Square**	**Cube**	**Sierpinski**
Scale factor	$s = \frac{1}{2}$	$s = \frac{1}{2}$	$s = \frac{1}{2}$	$s = \frac{1}{2}$
Measure of the scaled-down shape	$\mu(L_s) = \frac{1}{2}$	$\mu(L_s) = \left(\frac{1}{2}\right)^2$	$\mu(L_s) = \left(\frac{1}{2}\right)^3$	$\mu(L_s) = \left(\frac{1}{2}\right)^D$

Fractal dimension: D such that, if L is D-dimensional and we scale it down by s

$$\mu(L_s) = s^D$$

$$\left(\frac{1}{2}\right)^D = \frac{1}{3} \implies D = \log_2 3$$

Fractals have non-integer dimension

Box-counting dimension

A first definition

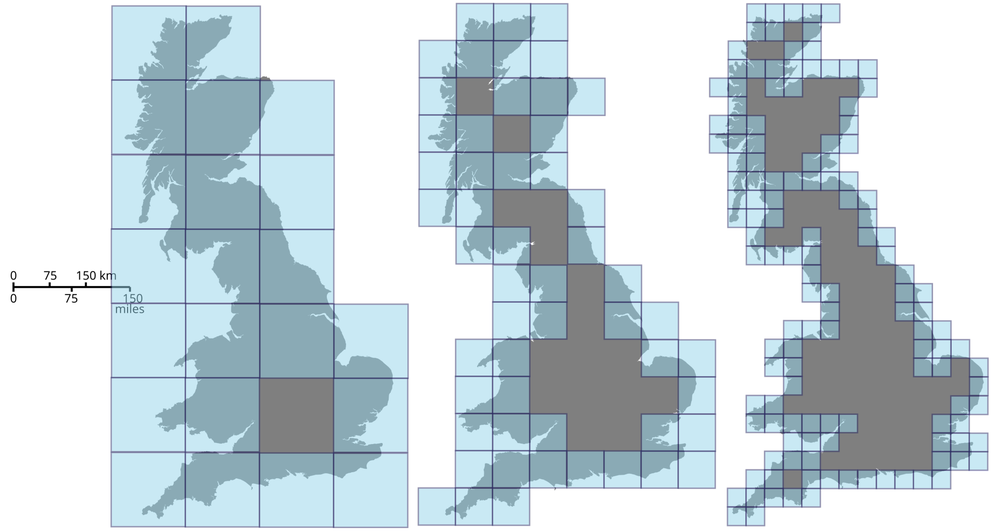
Metric space

balls to cover S

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

Radius

The limit might not exist



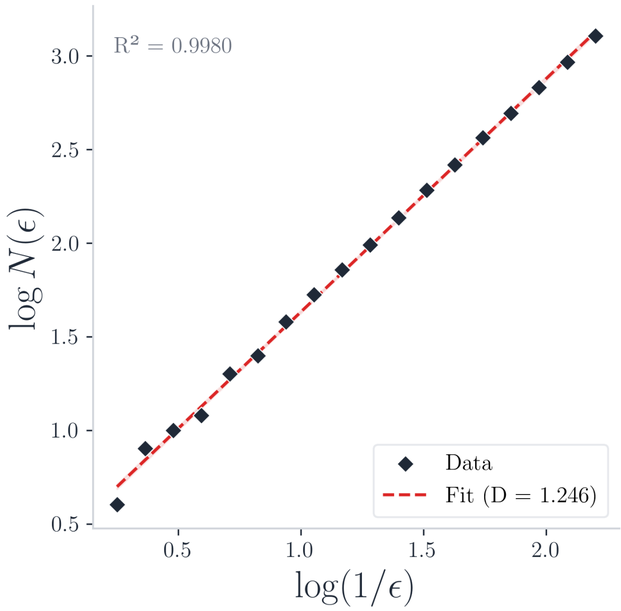
By Prokofiev - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=12042116

In practice, this means

$$N_\epsilon \approx C\epsilon^{-D}$$

$$\log N_\epsilon \approx \log C - d \log \epsilon$$

So we can estimate the box counting dimension interpolating a log-log plot



PH dimension

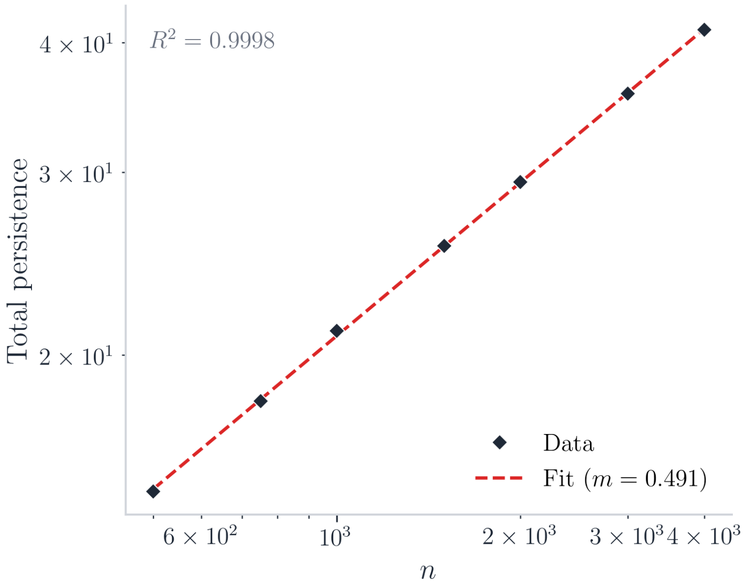
Computing fractal dimension using PH

$$\mathbf{x} = \{x_1, \ldots, x_n\} \subset S \longrightarrow$$

Total persistence

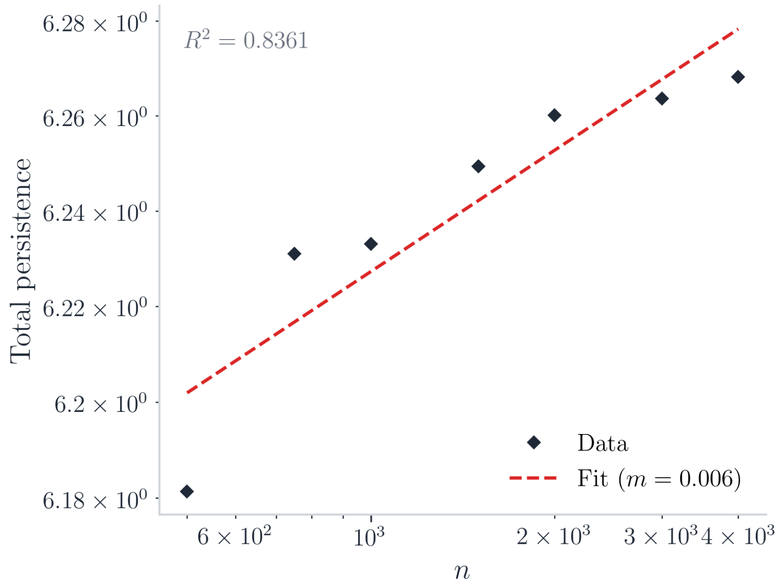
$$E_n(\mathbf{x}) = \sum_{(b,d) \in \mathrm{PD}_0(\mathbf{x})} |d - b|$$

$[0,1]^2$



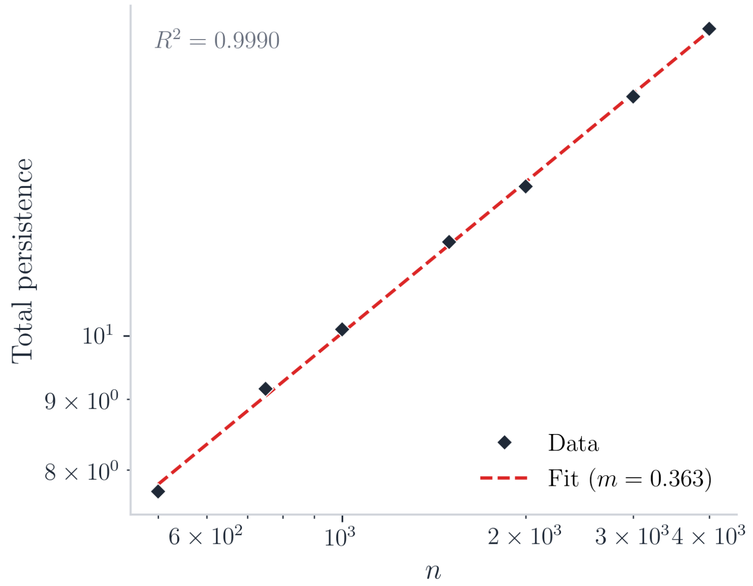
$$\frac{1}{1-m} \approx 1.996$$

Circle



$$\frac{1}{1-m} \approx 1.006$$

Sierpiński triangle



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

Schweinhart, Benjamin. "Fractal Dimension and the Persistent Homology of Random Geometric Complexes." *Advances in Mathematics* 372 (October 2020): 107291

Jaquette, Jonathan, and Benjamin Schweinhart. "Fractal Dimension Estimation with Persistent Homology: A Comparative Study." *Communications in Nonlinear Science and Numerical Simulation* 84 (May 2020): 105163.

Adams, Henry, Manuchehr Aminian, Elin Farnell, et al. *A Fractal Dimension for Measures via Persistent Homology*. Vol. 15. 2020.

PH dimension and box-counting dimension

You can formalize this to define a notion of *PH fractal dimension* and prove that is equivalent to the box-counting

$$\dim_{\mathrm{box}}(S) = \dim_{\mathrm{PH}}(S)$$

Kozma, Gady, Zvi Lotker, and Gideon Stupp. "The Minimal Spanning Tree and the Upper Box Dimension." *Proceedings of the American Mathematical Society* 134, no. 4 (2006): 1183–87.

Schweinhart, Benjamin. "Persistent Homology and the Upper Box Dimension." *Discrete & Computational Geometry* 65, no. 2 (2021): 331–64.

Generalization bounds using fractal dimension

Session 4 — Using topology and geometry to understand learning: generalization

Inés García Redondo — May 13, 2026

Bounding generalization with fractal dimension

Approximating gradient descent with SDEs

Solve $$\min_{\theta\in\mathbb{R}^m}\left\{\widehat{\mathcal{R}}(\theta,\mathcal{Z})=\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(f_\theta(x_i),y_i)\right\}$$

Using $$\theta_{k+1}=\theta_k-\eta\nabla\widetilde{\mathcal{R}}_k(\theta_k)\quad\text{with}\quad\nabla\widetilde{\mathcal{R}}_k(\theta):=\frac{1}{B}\sum_{i\in\widetilde{B}_k}\nabla\widehat{\mathcal{L}}(f_\theta(x_i),y_i)$$

Gradient noise: $\nabla\widetilde{\mathcal{R}}_k(\theta)-\nabla\widehat{\mathcal{R}}(\theta)\sim$ Gaussian

Euler-Mayorama discretization

$$\mathrm{d}\Theta_t=-\nabla\widehat{\mathcal{R}}(\Theta_t)\mathrm{d}t+\Sigma(\Theta_t)\mathrm{d}B_t$$

Diffusion coefficient — Brownian motion

$\nabla\widetilde{\mathcal{R}}_k(\theta)-\nabla\widehat{\mathcal{R}}(\theta)$ has heavy tails in practice!

$$\mathrm{d}\Theta_t=-\nabla\widehat{\mathcal{R}}(\Theta_t)\mathrm{d}t+\Sigma_1(\Theta_t)\mathrm{d}B_t+\Sigma_2(\Theta_t)\mathrm{d}\mathrm{L}_t^{\alpha(\Theta_t)}$$

State-dependent α-stable Lévy motion

Şimşekli, Umut, Mert Gürbüzbalaban, Thanh Huy Nguyen, Gaël Richard, and Levent Sagun. "On the Heavy-Tailed Theory of Stochastic Gradient Descent for Deep Neural Networks." arXiv:1912.00018. Preprint, arXiv, November 29, 2019.

Bounding generalization with fractal dimension

Fractal nature of the optimization trajectory

$$\nabla\widetilde{\mathcal{R}}_k(\theta) - \nabla\widehat{\mathcal{R}}(\theta)$$ has heavy tails in practice!

$$\mathrm{d}\Theta_t = -\nabla\widehat{\mathcal{R}}(\Theta_t)\mathrm{d}t + \Sigma_1(\Theta_t)\mathrm{d}B_t + \Sigma_2(\Theta_t)\mathrm{d}\mathrm{L}_t^{\alpha(\Theta_t)}$$

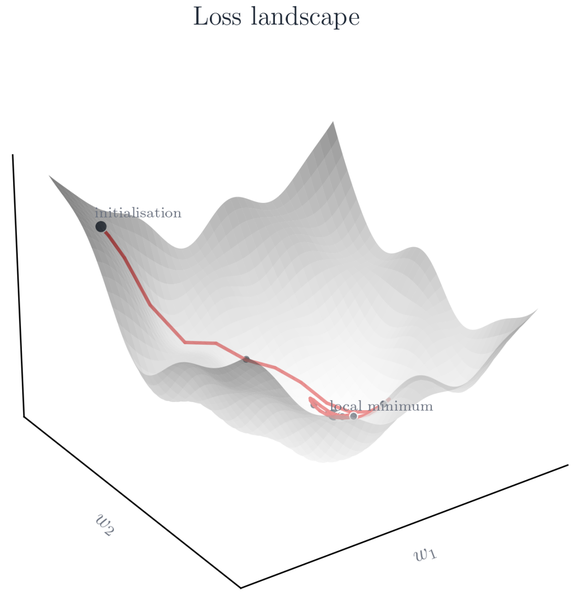
State-dependent α-stable Lévy motion

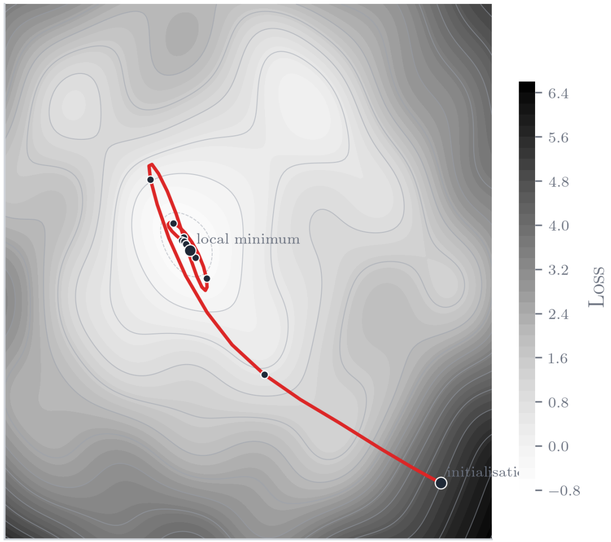
Sample paths (realizations) of a Markov process often exhibit *fractal-like* structure

A realization is just a collection of checkpoints (weights) as we train our model

Optimization trajectory $\mathcal{W}_{\mathcal{Z}} = \{\theta_1, \ldots, \theta_T\} \subset \mathbb{R}^m$

Worst-case generalization error over the optimization trajectory $\max_{\theta \in \mathcal{W}_{\mathcal{Z}}} |\mathcal{R}(\theta) - \widehat{\mathcal{R}}(\theta, \mathcal{Z})|$





Şimşekli, Umut, Ozan Sener, George Deligiannidis, and Murat A. Erdogdu. "Hausdorff Dimension, Heavy Tails, and Generalization in Neural Networks." *Advances in Neural Information Processing Systems* 33 (2020): 5138–51.

Bounding generalization with fractal dimension

Bounds for the worst-case generalization using PH dimension

With probability $1 - \delta$

$$\max_{\theta \in \mathcal{W}_{\mathcal{Z}}} |\mathcal{R}(\theta) - \widehat{\mathcal{R}}(\theta, \mathcal{Z})| \leq C \sqrt{\frac{\dim_{\mathrm{PH}}(\mathcal{W}_{\mathcal{Z}}) + I(\mathcal{W}_{\mathcal{Z}}, \mathcal{Z}) + \log(1/\delta)}{n}}$$

- *Euclidean* distance in $\mathbb{R}^m$
- Data-dependent, *loss-based* pseudo metric:

$$\rho_{\mathcal{Z}}(\theta, \theta') = \frac{1}{n} \sum_{i=1}^{n} |\mathcal{L}(f_\theta(x_i), y_i) - \mathcal{L}(f_{\theta'}(x_i), y_i)|$$

Birdal, Tolga, Aaron Lou, Leonidas J. Guibas, and Umut Simsekli. "Intrinsic Dimension, Persistent Homology and Generalization in Neural Networks." *Advances in Neural Information Processing Systems* 34 (2021): 6776–89.

Dupuis, Benjamin, George Deligiannidis, and Umut Simsekli. "Generalization Bounds Using Data-Dependent Fractal Dimensions." *Proceedings of the 40th International Conference on Machine Learning*, July 3, 2023, 8922–68.

Generalization bounds using fractal dimension in practice

Session 4 — Using topology and geometry to understand learning: generalization

Inés García Redondo — May 13, 2026

Bounding generalization with PH dimension

Experimental setup

- Train MLPs and CNNs with MNIST, CIFAR-10 (classification) and CHD (regression)
- Train with stochastic gradient descent, taking learning rates and batch sizes in a 6×6 grid

$$\theta_{k+1} = \theta_k - \eta \nabla \widetilde{\mathcal{R}}_k(\theta_k) \quad \text{with} \quad \nabla \widetilde{\mathcal{R}}_k(\theta) := \frac{1}{B} \sum_{i \in \widetilde{B}_k} \nabla \widehat{\mathcal{R}}(\theta, \mathcal{Z})$$

- Train until convergence and run 5,000 iterations more to compute PH dimension
- Plot against generalization gap of the trained model

Train accuracy - test accuracy

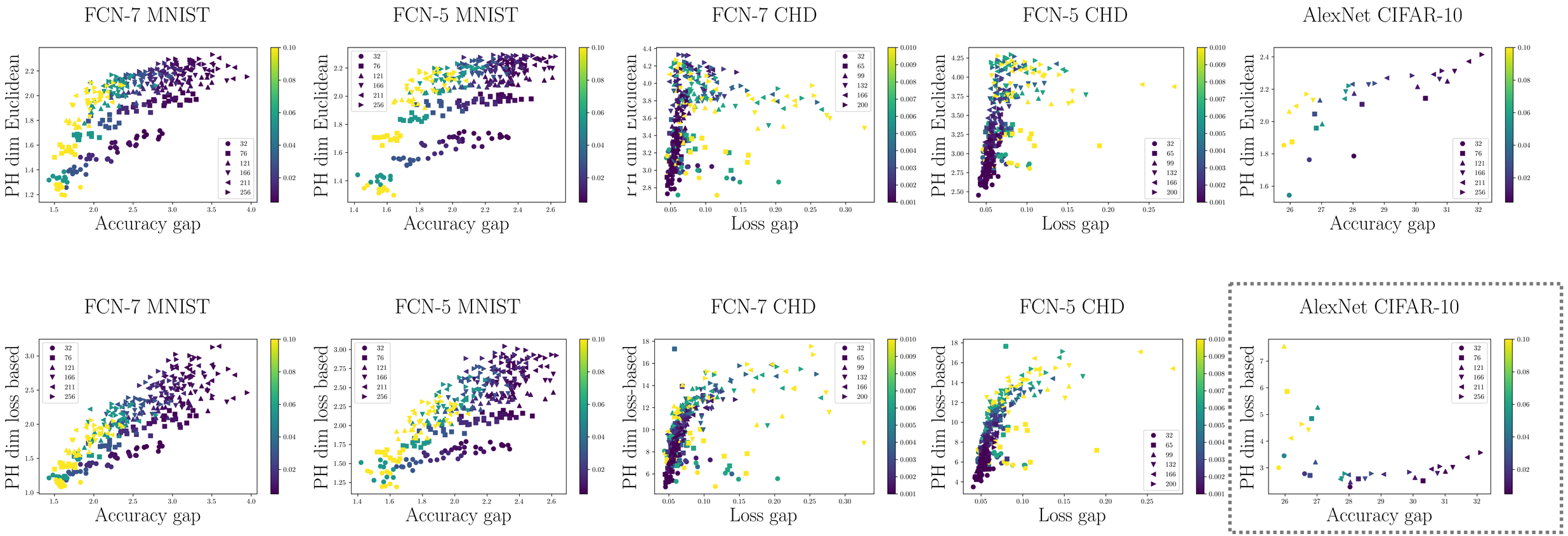
Want to observe a **positive correlation** between PH dimension and generalization

Dupuis, Benjamin, George Deligiannidis, and Umut Simsekli. "Generalization Bounds Using Data-Dependent Fractal Dimensions." *Proceedings of the 40th International Conference on Machine Learning*, July 3, 2023, 8922–68.

Tan, Charlie B., Inés García-Redondo, Qiquan Wang, Michael M. Bronstein, and Anthea Monod. "On the Limitations of Fractal Dimension as a Measure of Generalization." In *Advances in Neural Information Processing Systems*, vol. 37, edited by A. Globerson, L. Mackey, D. Belgrave, et al. Curran Associates, Inc., 2024.

Bounding generalization with PH dimension

Results of the experiments

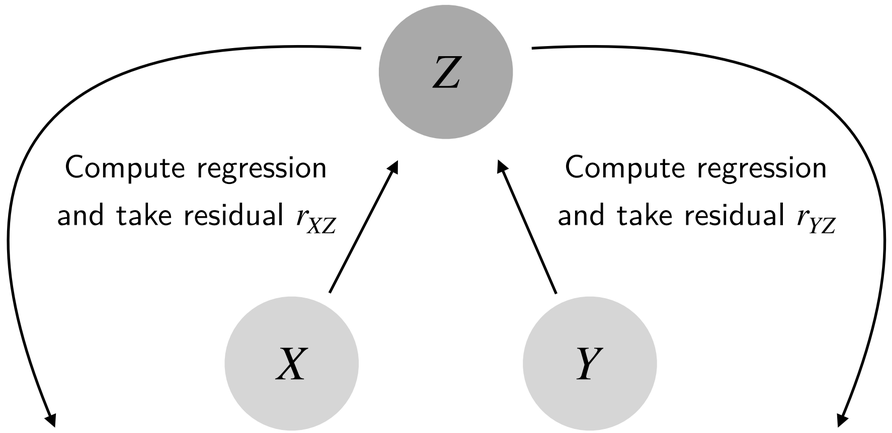


Tan, Charlie B., Inés García-Redondo, Qiquan Wang, Michael M. Bronstein, and Anthea Monod. "On the Limitations of Fractal Dimension as a Measure of Generalization." In *Advances in Neural Information Processing Systems*, vol. 37, edited by A. Globerson, L. Mackey, D. Belgrave, et al. Curran Associates, Inc., 2024.

Partial correlation analysis

Bounds for the worst-case generalization using PH dimension

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



Compute correlation between r_{XZ} and r_{YZ}

Low coefficient $\Longrightarrow$ correlation between X and Y can be explained by a common correlation with Z

+ non-parametric permutation-type hypothesis test

Statistically significant partial correlation in many cases

	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)

Tan, Charlie B., Inés García-Redondo, Qiquan Wang, Michael M. Bronstein, and Anthea Monod. "On the Limitations of Fractal Dimension as a Measure of Generalization." In *Advances in Neural Information Processing Systems*, vol. 37, edited by A. Globerson, L. Mackey, D. Belgrave, et al. Curran Associates, Inc., 2024.

Conditional independence analysis

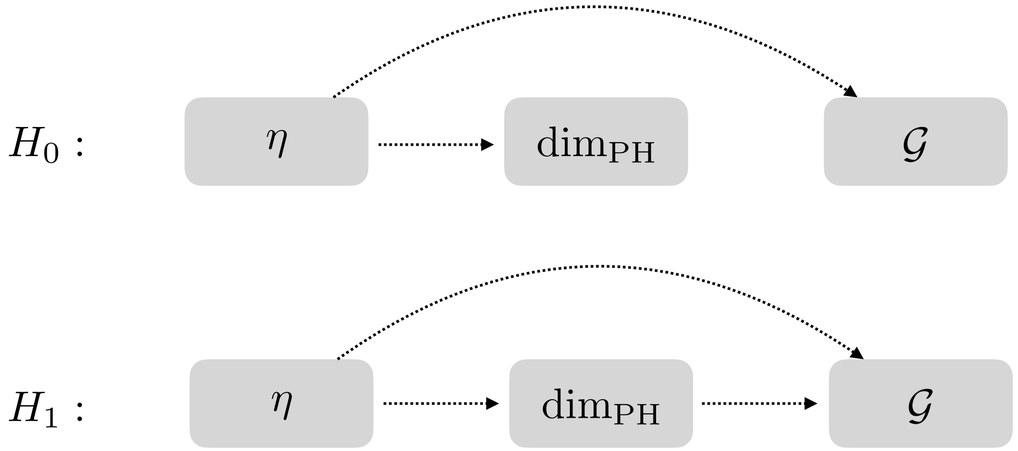
Bounds for the worst-case generalization using PH dimension

Is there a *causal connection* between changes in the hyperparameters and changes in the generalization and PH dimension?

- Use conditional metal information (CMI), a statistical test that vanishes iff

$$\mathcal{G} \perp \dim_{\mathrm{PH}} | \eta$$

- Generate null distribution under local permutations
- Hypothesis test: null hypothesis means that generalization and PH dimension are conditionally independent



Depends on the task: for classification (MNIST) there is conditional independence, for regression (CHD), there is not

Tan, Charlie B., Inés García-Redondo, Qiquan Wang, Michael M. Bronstein, and Anthea Monod. "On the Limitations of Fractal Dimension as a Measure of Generalization." In *Advances in Neural Information Processing Systems*, vol. 37, edited by A. Globerson, L. Mackey, D. Belgrave, et al. Curran Associates, Inc., 2024.

Failure modes

Adversarial initialization

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**

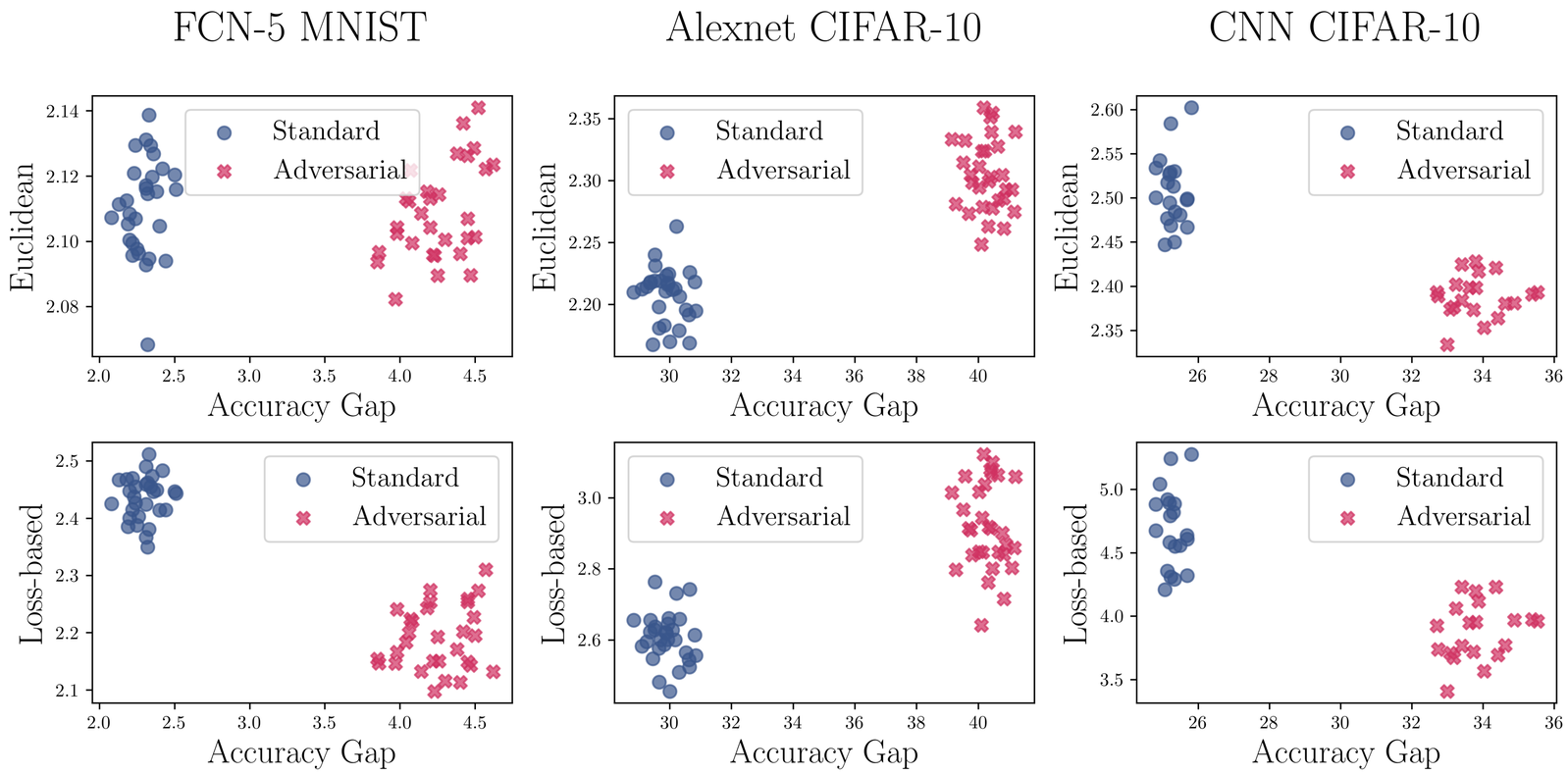
Standard initialization

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

Tan, Charlie B., Inés García-Redondo, Qiquan Wang, Michael M. Bronstein, and Anthea Monod. "On the Limitations of Fractal Dimension as a Measure of Generalization." In *Advances in Neural Information Processing Systems*, vol. 37, edited by A. Globerson, L. Mackey, D. Belgrave, et al. Curran Associates, Inc., 2024.

Failure modes

Adversarial initialization



Tan, Charlie B., Inés García-Redondo, Qiquan Wang, Michael M. Bronstein, and Anthea Monod. "On the Limitations of Fractal Dimension as a Measure of Generalization." In *Advances in Neural Information Processing Systems*, vol. 37, edited by A. Globerson, L. Mackey, D. Belgrave, et al. Curran Associates, Inc., 2024.

Takeaway message

In practice, NNs are very complex systems and it is very hard to isolate effects in correlations and to attribute causal relations between variables

There are cases where the bound seems to be vacuous, why?

$$\max_{\theta \in \mathcal{W}_{\mathcal{Z}}} |\mathcal{R}(\theta) - \widehat{\mathcal{R}}(\theta, \mathcal{Z})| \leq C \sqrt{\frac{\dim_{\mathrm{PH}}(\mathcal{W}_{\mathcal{Z}}) + I(\mathcal{W}_{\mathcal{Z}}, \mathcal{Z}) + \log(1/\delta)}{n}}$$

New updated measures using *magnitude*

Andreeva, Rayna, Benjamin Dupuis, Rik Sarkar, Tolga Birdal, and Umut Şimşekli. "Topological Generalization Bounds for Discrete-Time Stochastic Optimization Algorithms." *Advances in Neural Information Processing Systems* 37 (December 2024): 4765–818.

Thank you for your attention!

AIDOS LAB
AI FOR DATA-ORIENTED SCIENCE

UNI FR

Session 4 — Using topology and geometry to understand learning: generalization

Inés García Redondo — May 13, 2026