Generalization Bounds with Data-dependent Fractal Dimensions

ICML 2023

Benjamin Dupuis[1,2], Georges Deligiannidis[3], Umut Simsekli[1,2]

[1]INRIA Paris, [2]ENS Paris, [3]Oxford University

July 10, 2023

Context

On a data space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ endowed with a probability distribution μ_z, we want to minimize the *population risk*

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

where:

- $h_\cdot(\cdot) : \mathbb{R}^d \times \mathcal{X} \longrightarrow \mathcal{Y}$ is a parametric model,
- $\mathcal{L} : \mathcal{Y} \times \mathcal{Y} \longrightarrow \mathbb{R}$ is a loss function.

We do so by considering the empirical risk over a dataset $S = (z_1, \dots, z_n) \sim \mu_z^{\otimes n}$

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

Our goal is to bound the *worst-case generalization error* over a (potentially random) hypothesis set $\mathcal{W}_{S,U} \subset \mathbb{R}^d$:

$$\mathcal{G}(S) := \sup_{w \in \mathcal{W}_{S,U}} \big(\mathcal{R}(w) - \hat{\mathcal{R}}_S(w) \big). \tag{1}$$

Overview of past results (informal)

Classical theory predicts a generalization in $\sqrt{d/n}$, which has been experimentally challenged by modern neural networks.

Recent results: Under the assumption that the loss is L-Lipschitz and uniformly bounded (by B), we have with probability $1-\zeta$ that [ŞSDE21, CDE$^+$21, BLGŞ21, HŞKM22]:

$$\sup_{w\in\mathcal{W}_{S,U}} |\hat{\mathcal{R}}_S(w) - \mathcal{R}(w)| \lesssim B{\color{red}L}\sqrt{\frac{{\color{blue}\overline{\dim}_B\mathcal{W}_{S,U}} + {\color{green}I_\infty} + \log(1/\zeta)}{n}} \tag{2}$$

- ${\color{red}L}$: Lipschitz constant of the loss ℓ wrt the parameters.
- ${\color{blue}\overline{\dim}_B\mathcal{W}_{S,U}}$: *Upper-box counting dimension, which is a notion of fractal dimension of the trajectory*
- ${\color{green}I_\infty}$: Total mutual information term, measuring the statistical dependence between the data and the hypothesis set.

Goal of this paper

Prove generalization bounds:

- Without Lipschitz assumption.
- Introducing a notion of upper box-counting dimension based on a **data-dependent pseudo-metric** instead of the Euclidean distance
- Prove that we can numerically evaluate it.

$$\sup_{w \in \mathcal{W}_{S,U}} |\hat{\mathcal{R}}_S(w) - \mathcal{R}(w)| \lesssim B\cancel{L}\sqrt{\frac{\cancel{\overline{\dim}_B \mathcal{W}_{S,U}} + I_\infty + \log(1/\zeta)}{n}}$$

Goal of this paper

Prove generalization bounds:

- Without Lipschitz assumption.
- Introducing a notion of upper box-counting dimension based on a **data-dependent pseudo-metric** instead of the Euclidean distance
- Prove that we can numerically evaluate it.

$$\sup_{w \in \mathcal{W}_{S,U}} |\hat{\mathcal{R}}_S(w) - \mathcal{R}(w)| \lesssim B\cancel{L}\sqrt{\frac{\cancel{\overline{\dim}_B \mathcal{W}_{S,U}} + I_\infty + \log(1/\zeta)}{n}}$$

Inspired by classical covering arguments on Rademacher complexity, we use the following random pseudo-metric on $\mathbb{R}^d$:

$$\boxed{\rho_S(w, w') = \frac{1}{n}\sum_{i=1}^{n} |\ell(w, z_i) - \ell(w', z_i)|,} \qquad (3)$$

and the corresponding upper box-counting dimension $\overline{\dim}_B^{\rho_S}(\mathcal{W}_{S,U})$.

Rademacher complexity

Warm-up: Result for fixed hypothesis space Proof

Assumptions: The loss is bounded by $B > 0$ and the learning algorithm is 'measurable' (see paper for details).
Let $\mathcal{W} \subseteq \mathbb{R}^d$ be a fixed closed set and $S \sim \mu_z^{\otimes n}$, we define:

- $\mathcal{G}(S) := \sup_{w \in \mathcal{W}} (\mathcal{R}(w) - \hat{\mathcal{R}}_S(w))$,
- $d(S) := \overline{\dim}_B^{\rho_S}(\mathcal{W})$.

Theorem 1

For all $\epsilon, \gamma, \eta > 0$ we can find $\delta_{n,\gamma,\epsilon} > 0$ such that with probability at least $1 - 2\eta - \gamma$, for all $\delta < \delta_{n,\gamma,\epsilon}$ we have:

$$\mathcal{G}(S) \leq 2\delta + 2B\sqrt{\frac{4(\epsilon + d(S))\log(1/\delta) + 9\log(1/\eta)}{n}}$$

Random hypothesis space Proof

Let us define:

- $\mathcal{G}(S, U) := \sup_{w \in \mathcal{W}_{S,U}} (\mathcal{R}(w) - \hat{\mathcal{R}}_S(w))$
- $I_{n,\delta} := \max_{0 \leq j \leq \lfloor \sqrt{n} \rfloor} I_\infty(S, N_{\delta,j})$ $N_{\delta,j}$ is a covering of R_S^j, which is a set where the empirical risk does not vary much.
- $d(S, U) := \overline{\dim}_B^{\rho_S}(\mathcal{W}_{S,U})$

Theorem 2

For all $\epsilon, \gamma, \eta > 0$ we can find $\delta_{n,\gamma,\epsilon} > 0$ such that with probability at least $1 - \eta - \gamma$ under $\mu_z^{\otimes n} \otimes \mu_u$, for all $\delta < \delta_{n,\gamma,\epsilon}$ we have:

$$\mathcal{G}(S, U) \leq \delta + \frac{B}{\sqrt{n} - 1} + \sqrt{2} B \sqrt{\frac{(\epsilon + d(S, U)) \log(2/\delta) + \log(\sqrt{n}/\eta) + I_{n,\delta}}{n}}$$

Comparison with previous results

- No Lipschitz assumption: ✓
- Data-dependent intrinsic dimension: ✓

Old:

$$\sup_{w \in \mathcal{W}_{S,U}} |\hat{\mathcal{R}}_S(w) - \mathcal{R}(w)| \lesssim B\cancel{L}\sqrt{\frac{\cancel{\overline{\dim}_B(\mathcal{W}_{S,U})} + I_\infty + \log(1/\zeta)}{n}}$$

New:

$$\sup_{w \in \mathcal{W}_{S,U}} |\hat{\mathcal{R}}_S(w) - \mathcal{R}(w)| \lesssim B\sqrt{\frac{\overline{\dim}_B^{\rho_S}(\mathcal{W}_{S,U}) + I_{n,\delta} + \log(1/\zeta)}{n}}$$

Comparison with previous results

- No Lipschitz assumption: ✓
- Data-dependent intrinsic dimension: ✓

Old: $$\sup_{w \in \mathcal{W}_{S,U}} |\hat{\mathcal{R}}_S(w) - \mathcal{R}(w)| \lesssim B\cancel{L}\sqrt{\frac{\cancel{\overline{\dim}_B(\mathcal{W}_{S,U})} + I_\infty + \log(1/\zeta)}{n}}$$

New: $$\sup_{w \in \mathcal{W}_{S,U}} |\hat{\mathcal{R}}_S(w) - \mathcal{R}(w)| \lesssim B\sqrt{\frac{\overline{\dim}_B^{\rho_S}(\mathcal{W}_{S,U}) + I_{n,\delta} + \log(1/\zeta)}{n}}$$

Main drawback:

- More complex mutual information term than in [HŞKM22, ŞSDE21].

$\Rightarrow$ Geometric stability.

Geometric stability to avoid complex mutual information

Proof

Theorem 3

We define $d(S,U)$ and $\mathcal{G}(S,U)$ as in theorem 2 and further define $I := I_\infty(S, \mathcal{W}_{S,U})$. We assume local coverings stability with parameters (α, β), then we have: Then there exists a constant $n_\alpha > 0$ such that for all $n \geq n_\alpha$, with probability $1-\gamma-\eta$, for all δ smaller than some $\delta_{\gamma,\epsilon,n} > 0$ we have:

$$\mathcal{G}(S,U) \leq \delta + \frac{3B+2\beta}{n^{\alpha/3}} + B\sqrt{\frac{\log(n/\eta) + I + \big(\epsilon + d(S,U)\big)\log(4/\delta)}{2n^{\frac{2\alpha}{3}}}}$$

We even show that $n_\alpha = \max\{2^{\frac{3}{2\alpha}}, 2^{1+\frac{3}{3-2\alpha}}\}$.

The mutual information term in this theorem is the same than the one appearing in previous results [HŞKM22]. ✓

Experimental setup

- We extend results from [AAF+20, BLGŞ21] to prove that $\overline{\dim}_B^{\rho_S}(\mathcal{W}_{S,U})$ can be numerically approximated using Topological Data Analysis (TDA) tools.
- Namely, we prove that our proposed dimension can be related to the corresponding 'persistent homology dimension',

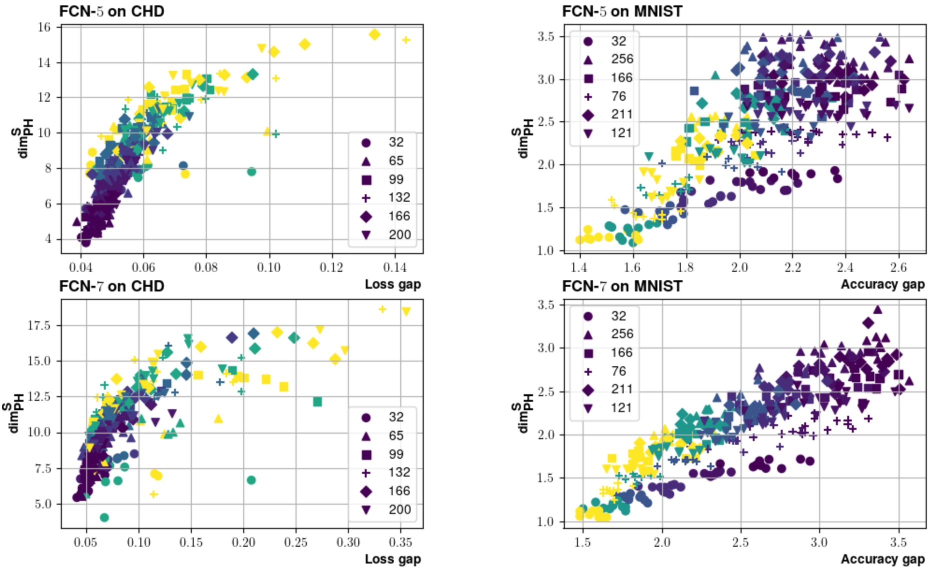
$$\overline{\dim}_B^{\rho_S}(\mathcal{W}_{S,U}) = \dim_{\mathrm{PH}^0}^{\rho_S}(\mathcal{W}_{S,U}),$$

Proof

that we can estimate with some Python libraries [Bau21, PHL+21].

- We (approximately) evaluate it on SGD trajectories.
- Experiments on various datasets (MNIST, CIFAR-10, CIFAR-100 California Housing Dataset) and models (FCN, AlexNet, LeNet, Resnet-18).

Proposed intrinsic dimension VS Generalization gap



We compare with [BLGŞ21] using correlation statistics: Kendall's τ, Spearman's ρ and Average Granulated Kendall's coefficient introduced in [JNM+19].

Table: Correlation coefficients on MNIST

MODEL	DIM.	ρ	$\boldsymbol{\Psi}$	τ
FCN-5	$\dim_{\mathrm{PH}^0}^{\mathrm{EUCL}}$	$0.62_{\pm 0.10}$	$0.78_{\pm 0.08}$	$0.47_{\pm 0.07}$
FCN-5	$\dim_{\mathrm{PH}^0}^{\rho_S}$	$\mathbf{0.73}_{\pm 0.07}$	$\mathbf{0.81}_{\pm 0.07}$	$\mathbf{0.56}_{\pm 0.06}$
FCN-7	$\dim_{\mathrm{PH}^0}^{\mathrm{EUCL}}$	$0.80_{\pm 0.04}$	$0.88_{\pm 0.04}$	$0.62_{\pm 0.04}$
FCN-7	$\dim_{\mathrm{PH}^0}^{\rho_S}$	$\mathbf{0.89}_{\pm 0.02}$	$\mathbf{0.90}_{\pm 0.04}$	$\mathbf{0.73}_{\pm 0.03}$

Table: Correlation coefficients with AlexNet on CIFAR-10

MODEL	DIM.	ρ	$\boldsymbol{\Psi}$	τ
ALEXNET	$\dim_{\mathrm{PH}^0}^{\mathrm{EUCL}}$	0.86	0.81	0.68
ALEXNET	$\dim_{\mathrm{PH}^0}^{\rho_S}$	**0.93**	**0.84**	**0.78**

Thank you!

Correspondence: *benjamin.dupuis@inria.fr, umut.simsekli@inria.fr*

Appendix

Sketch of proof, euclidean case, fixed $\mathcal{W}$ [ŞSDE21]

Back to presentation

- It all starts with the following decomposition: if $\|w - w'\| \leq \delta$ then:

$$
\begin{aligned}
|\mathcal{R}(w) - \hat{\mathcal{R}}_S(w)| \leq & |\mathcal{R}(w') - \hat{\mathcal{R}}_S(w')| \\
& + |\mathcal{R}(w) - \mathcal{R}(w')| + |\hat{\mathcal{R}}_S(w) - \hat{\mathcal{R}}_S(w')| \\
\leq & 2L\delta + |\mathcal{R}(w') - \hat{\mathcal{R}}_S(w')|.
\end{aligned}
$$

- Therefore, introducing a minimal covering for the Euclidean distance:

$$
|\mathcal{R}(w) - \hat{\mathcal{R}}_S(w)| \leq \max_{w' \in N_\delta} |\mathcal{R}(w') - \hat{\mathcal{R}}_S(w')|.
$$

- Using the union bound and Hoeffding's inequality:

$$
\mathbb{P}\left(\max_{w' \in N_\delta} |\mathcal{R}(w') - \hat{\mathcal{R}}_S(w')| \geq \epsilon \right) \leq 2|N_\delta^{\mathrm{Eucl}}(\mathcal{W})| \exp\left\{ -\frac{n\epsilon^2}{2B^2} \right\}
$$

Sketch of proof, euclidean case, fixed $\mathcal{W}$ [ŞSDE21] (2)

- Rearranging terms, with probability $1 - \zeta$:

$$\max_{w' \in N_\delta} |\mathcal{R}(w') - \hat{\mathcal{R}}_S(w')| \leq \sqrt{\frac{2B^2}{n}\left(\log\left(2|N_\delta^{\mathrm{Eucl}}(\mathcal{W})|\right) + \log(1/\zeta)\right)}$$

The general case (random hypothesis set) follows by:

- Decoupling techniques based on (total) mutual information
- Use of Egoroff's theorem

Rademacher complexity

Let $(\sigma_1, \dots, \sigma_n)$ i.i.d. random variables with Bernoulli distribution in $\{-1, 1\}$.

$$\mathbf{Rad}(A) := \frac{1}{n}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{x \in A} \sum_{i=1}^{n} \sigma_i x_i\right]$$

Proposition 4.1

Assume that the loss is uniformly bounded by B. For all $\eta > 0$, we have with probability $1 - 2\eta$ that:

$$\sup_{w \in \mathcal{W}} \left(\mathcal{R}(w) - \hat{\mathcal{R}}_S(w)\right) \leq 2\boldsymbol{Rad}(\ell(\mathcal{W}, S)) + 3\sqrt{\frac{2B^2}{n}\log(1/\eta)}$$

It opens the door to a classical **covering argument** [Reb20], to bound the generalization error.

Total mutual information

Definition 4

For two random variables X and Y:

$$I_\infty(X, Y) := \log \left(\sup_B \frac{\mathbb{P}_{X,Y}(B)}{\mathbb{P}_X \otimes \mathbb{P}_Y(B)} \right)$$

It can equivalently been defined as limit of α-mutual information, for $\alpha \to \infty$, see [vH14]

For all borelian set B:

$$\mathbb{P}_{X,Y}(B) \leq e^{I_\infty(X,Y)} \mathbb{P}_X \otimes \mathbb{P}_Y(B) \tag{4}$$

Proof of Theorem 1 (Sketch)

Back to presentation

Using a classical covering technique for Rademacher complexity [Reb20], with pseudo-metric ρ_S:

$$\mathbf{Rad}(\ell(\mathcal{W}, S)) \leq \delta + B\sqrt{\frac{2\log(|N_\delta^{\rho_S}|)}{n}}.$$

Then we use Egoroff's theorem to bound $\log(|N_\delta^{\rho_S}|)$ by $(\overline{\dim}_B^{\rho_S}(\mathcal{W}) + \epsilon)\log(1/\delta)$, for δ small enough, uniformly on $S \in \mathcal{Z}^n$.

Then the result follows from Proposition 4.1.

Can we control $\delta_{n,\gamma,\epsilon}$?

- The result could be expressed only in terms of covering numbers.
- By applying Egoroff's theorem, we can make their convergence uniform in S, but not in n.
- It controls the dependence on n of the convergence of the upper-box-counting dimension.
- For instance, the assumption that for all $S \in \mathcal{Z}^\infty$:

$$\sup_n \left| \frac{\log |N_\delta^{\rho_{S_n}}|}{\log(1/\delta)} - \overline{\dim}_B^{\rho_{S_n}}(\mathcal{W}_{S_n,U}) \right| \underset{\delta \to 0}{\longrightarrow} 0,$$

where S_n is the natural projection $S \in \mathcal{Z}^\infty \longrightarrow \mathcal{Z}^n$.
- This removes the dependence on n, so that we can set $\delta = 1/\sqrt{n}$ and have as a result that for n big enough:

$$\mathcal{G}(S) \leq \frac{1}{\sqrt{n}} + 2B\sqrt{\frac{(\epsilon + d(S))\log(n) + 9\log(1/\eta)}{n}}$$

Can we control $\delta_{n,\gamma,\epsilon}$? (2)

- Indeed, $\mathcal{Z}^\infty$ can be endowed with the cylindrical σ-algebra $\mathcal{F}^{\otimes\infty}$ and the associated probability measure $\mu_z^{\otimes\infty}$.
- Let $\epsilon > 0$, we can apply Egoroff's theorem in this probability space to get that there exists $\Omega_\gamma \subset \mathcal{Z}^\infty$ and $\delta_{\gamma,\epsilon}$, such that $\mu_z^{\otimes\infty}(\Omega_\gamma) \geq 1-\gamma$ and:

$$\forall \delta \leq \delta_{\gamma,\epsilon},\ \forall n,\ \log|N_\delta^{\rho_{S_n}}| \leq \log(1/\delta)(\overline{\dim}_B^{\rho_{S_n}}(\mathcal{W}_{S_n,U}) + \epsilon).$$

- We conclude by noting that, if $\pi_n : \mathcal{Z}^\infty \longrightarrow \mathcal{Z}^n$ is the natural projection, we have:

$$\mu_z^{\otimes n}(\pi_n(\Omega_\gamma)) \geq \mu_z^{\otimes\infty}(\Omega_\gamma) \geq 1-\gamma$$

Proof of Theorem 2 (Sketch) Back to presentation

Recall slide **??**:

$$\sup_{w \in \mathcal{W}_{S,U}} |\mathcal{R}(w) - \hat{\mathcal{R}}_S(w)| \leq \max_j \max_{w' \in N_{\delta,j}} |\mathcal{R}(w') - \hat{\mathcal{R}}_S(w')| + \delta + \frac{B}{\sqrt{n} - 1}$$

Applying Hoeffding's inequality, along with union bounds and decoupling inequality (4), with probability $1 - \zeta$:

$$\sup_{w \in \mathcal{W}_S} \left(\mathcal{R}(w) - \hat{\mathcal{R}}_S(w)\right) \leq \delta + \frac{B}{K} + \sqrt{\frac{2B^2}{n}\left(\log(K/\eta) + \log\left(\max_j |N_{\delta,j}|\right) + \max_j I_\infty(S, N_{\delta,j})\right)}$$

Then we remark that $\max_j |N_{\delta,j}| \leq |N_{\delta/2}(\mathcal{W}_{S,U})|$ and use Egoroff's theorem and the definition of upper box-counting dimension as before.

Proof of Theorem 3 (Sketch) Back to presentation

The main idea is to divide the dataset $S \in \mathcal{Z}^n$ into H groups $J_1, \dots, J_H$ of size J with $J, H \in \mathbb{N}_+$ and $JH = n$.

$$\sup_{w \in \mathcal{W}_{s,U}} |\mathcal{R}(w) - \hat{\mathcal{R}}_S(w)| \leq \delta + \frac{B}{K} + \max_{0 \leq j \leq K-1} \underbrace{\max_{w \in N_{\delta,j}(S,S,U)} |\mathcal{R}(w) - \hat{\mathcal{R}}_S(w)|}_{:= E_j}$$

Then:

$$E_j \leq \frac{2J\beta}{n^\alpha} + \frac{1}{n} \sum_{k=1}^{H} \max_{w \in N_{\delta,j}(S, S^{\setminus J_k}, U)} \left| \sum_{i \in J_k} \left(\ell(w, z_i) - \mathcal{R}(w) \right) \right|$$

Then the proof proceeds by applying Hoeffding's inequality and decoupling as before.

Proof of Theorem 3 (Sketch) (2)

Back to presentation

The key point is that the nested Markov chains:

$$\begin{tikzcd}
S_{J_k} \arrow[rr] & & \mathcal{W}_{S,U} \arrow[r] \arrow[dl] & N_{\delta,j}(S, S^{\setminus J_k}, U) \\
& S \arrow[ul] & &
\end{tikzcd}$$

allow to write: $I_\infty(N_{\delta,j}(S, S^{\setminus J_k}, U), S_{J_k}) \leq I_\infty(S, \mathcal{W}_{S,U})$.

The covering numbers $|N_{\delta,j}(S, S^{\setminus J_k}, U), S_{J_k}|$ are then controlled in terms of $|N_\delta(\mathcal{W}_{S,U})|$.

A **trade-off** appears, requiring that:

$$\boxed{J = n^{\frac{2\alpha}{3}}}$$

Persistent homology in pseudo-metric spaces

Back to presentation

- (X, ρ) a pseudo-metric space,
- $\pi : X \longrightarrow X/\sim$ its metric identification,
- $P \subset X$ a finite subset and $\tilde{P} = \pi(P)$.

Given a Vietoris-Rips filtration of P:

$$\emptyset \to K^{\delta_0,1} \to \cdots \to K^{\delta_0,\alpha_0} \to K^{\delta_1,1} \to \cdots \to K^{\delta_c,\alpha_C} = K,$$

For 'death time' of connected components greater than 0, the homology groups of degree 0 are the same.

$$\begin{tikzcd}
\dots \arrow[r] \arrow[d] & H_0(K^{0,\alpha_0-1}) \arrow[r] \arrow[d] & H_0(K^{0,\alpha_0}) \arrow[r] \arrow[d, equal, "\wr"] & \dots \arrow[r] \arrow[d, equal, "\wr"] & H_0(K^{\delta_C,\alpha_C}) \arrow[d, equal, "\wr"] \\
\dots \arrow[r] & H_0(\tilde{K}^{0,\alpha_0-1}) \arrow[r] & H_0(\tilde{K}^{0,\alpha_0}) \arrow[r] & \dots \arrow[r] & H_0(\tilde{K}^{\delta_C,\alpha_C})
\end{tikzcd}$$

Hence we get the same dimensions.

References I

Henry Adams, Manuchehr Aminian, Elin Farnell, Michael Kirby, Chris Peterson, Joshua Mirth, Rachel Neville, Patrick Shipman, and Clayton Shonkwiler.
A fractal dimension for measures via persistent homology.
Topological Data Analysis, Abel Symposia, vol. 15, pages 1–31, 2020.

Ulrich Bauer.
Ripser: Efficient computation of Vietoris-Rips persistence barcodes.
Journal of Applied and Computational Topology, 5(3):391–423, September 2021.

Tolga Birdal, Aaron Lou, Leonidas Guibas, and Umut Şimşekli.
Intrinsic Dimension, Persistent Homology and Generalization in Neural Networks.
Advances in Neural Information Processing Systems 34 (NeurIPS 2021), November 2021.

Alexander Camuto, George Deligiannidis, Murat A. Erdogdu, Mert Gürbüzbalaban, Umut Şimşekli, and Lingjiong Zhu.
Fractal Structure and Generalization Properties of Stochastic Optimization Algorithms.
Advances in Neural Information Processing Systems 34 (NeurIPS 2021), June 2021.

References II

Liam Hodgkinson, Umut Şimşekli, Rajiv Khanna, and Michael W. Mahoney.
Generalization Bounds using Lower Tail Exponents in Stochastic Optimizers.
Proceedings of the 39th International Conference on Machine Learning, July 2022.

Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio.
Fantastic Generalization Measures and Where to Find Them.
ICLR 2020, December 2019.

Julián Burella Pérez, Sydney Hauke, Umberto Lupo, Matteo Caorsi, and Alberto Dassatti.
Giotto-ph: A Python Library for High-Performance Computation of Persistent Homology of Vietoris-Rips Filtrations, August 2021.

Patrick Rebeschini.
Algorithmic fundations of learning, 2020.

References III

Umut Şimşekli, Ozan Sener, George Deligiannidis, and Murat A. Erdogdu.
Hausdorff Dimension, Heavy Tails, and Generalization in Neural Networks.
Journal of Statistical Mechanics: Theory and Experiment, 2021(12):124014, December 2021.

Tim van Erven and Peter Harremoës.
Renyi Divergence and Kullback-Leibler Divergence.
IEEE Transactions on Information Theory, 60(7):3797–3820, July 2014.

Notations

$\ell : \mathbb{R}^d \times \mathcal{Z} \longrightarrow \mathbb{R}$ Loss function

$\hat{\mathcal{R}}_S(\bullet)$ Empirical risk

$\mathcal{R}(\bullet)$ Risk

$\mathcal{W}_S$ Sample path generated by the learning algorithm

$\tilde{X}$ Independent copy of the random element X

$\overline{\dim}_B^d(\bullet)$ Upper box counting dimension computed with (pseudo-)metric d

N_δ^d Centers of a covering by closed δ-balls for (pseudo-)metric d

$S = (z_1, \dots, z_n) \in \mathcal{Z}^n$ Dataset of i.i.d. random variables