École polytechnique fédérale de Lausanne

Master thesis

Master of engineering Mathematics

Generalization Bounds with Data-dependent Fractal Dimensions

New theoretical and empirical results

Author:
Benjamin Dupuis

Supervisors:
Clément Hongler
Umut Şimşekli
George Deligiannidis

Summary

Providing generalization guarantees for modern neural networks has been a crucial task in statistical learning. Recently, several studies have attempted to analyze the generalization error in such settings by using tools from fractal geometry. While these works have successfully introduced new mathematical tools to apprehend generalization, they heavily rely on a Lipschitz continuity assumption, which in general does not hold for neural networks and might make the bounds vacuous.

In this Master Thesis, we address this issue and prove fractal geometry-based generalization bounds *without* requiring any Lipschitz assumption. To achieve this goal, we build up on a classical covering argument in learning theory and introduce a *data-dependent fractal dimension*. Despite introducing a significant amount of technical complications, this new notion lets us control the generalization error (over either fixed or random hypothesis spaces) along with certain mutual information (MI) terms. To provide a clearer interpretation to the newly introduced MI terms, as a next step, we introduce a notion of 'geometric stability' and link our bounds to the prior art.

Finally, we make a rigorous connection between the proposed data-dependent dimension and topological data analysis tools, which then enables us to compute the dimension in a numerically efficient way. We support our theory with experiments conducted on various settings.

P.S.: This Master Thesis, conducted at INRIA Paris, under the supervision of Umut Şimşekli and George Deligiannidis (from Oxford University) resulted, among other things, in the submission of a paper at the International Conference on Machine Learning (ICML 2023). This submission can be found online (Dupuis *et al.*, 2023), which I have written.

The subject of the aforementioned paper being precisely the one of this thesis, some parts of it are reproduced or adapted directly in this work. Moreover, this thesis makes a lot more detailed introduction to the technical background on which the results and experiments rely. Moreover, several additional results are presented, such as technical measure-theoretic statements (in particular Theorems 4.1, 5.1 and 5.2) which imply that some assumptions in the paper are not necessary. Other supplementary results, such as Theorem B.1, present interesting applications of the proof techniques used in the paper. For consistency with the published paper, 'we' is used everywhere in this report.

Contents

1 Introduction

Understanding the generalization properties of modern neural networks has been one of the major challenges in statistical learning theory over the last decade. In a classical supervised learning setting, this task boils down to understanding the so-called *generalization error*, which arises from the classical population risk minimization problem, given as by

$$\min_{w\in\mathbb{R}^d}\Big\{\mathcal{R}(w) := \mathbb{E}_{(x,y)\sim\mu_z}[\mathcal{L}(h_w(x), y)]\Big\}, \tag{1}$$

where $x \in \mathcal{X}$ denotes the features, $y \in \mathcal{Y}$ denotes the labels, μ_z denotes a probability distribution on the *data space* $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, referred to as the *data distribution*, $h_w : \mathcal{X} \longrightarrow \mathcal{Y}$ denotes a parametric predictor with $w \in \mathbb{R}^d$ being its parameter vector, $\mathcal{L} : \mathcal{Y} \times \mathcal{Y} \longrightarrow \mathbb{R}$ denotes the loss function.

In all the following we will write the elements of $\mathcal{Z}$ as $z = (x, y) \in \mathcal{X} \times \mathcal{Y}$ and ℓ for the composition of the loss function $\mathcal{L}$ and the parametric predictor h_w, i.e.

$$\ell : \mathbb{R}^d \times \mathcal{Z} \ni (w, z) \longmapsto \ell(w, z) = \ell(w, (x, y)) = \mathcal{L}(h_w(x), y).$$

With a slight abuse of notation, we will refer to ℓ as the 'loss' in all the following.

As μ_z is unknown, in practice, to approximate the optimization problem described by (1), one resorts to the minimization of the empirical risk, given by

$$\hat{\mathcal{R}}_S(w) := \frac{1}{n}\sum_{i=1}^{n} \ell(w, z_i), \tag{2}$$

where $S := (z_i)_{1\leq i\leq n} \sim \mu_z^{\otimes n}$ is a set of independent and identically distributed (i.i.d.) data points. Then, our goal is to bound the worst-case generalization error that is defined as the gap between the population and empirical risk over a (potentially random) hypothesis set $\mathcal{W} \subset \mathbb{R}^d$:

$$\mathcal{G}(S) := \sup_{w\in\mathcal{W}} \big(\mathcal{R}(w) - \hat{\mathcal{R}}_S(w)\big). \tag{3}$$

This will be formalized more rigorously in Section 1.1.

In the context of neural networks, one peculiar observation has been that, even when a network contains millions of parameters (i.e., $d \gg 1$), it might still generalize well (Zhang *et al.*, 2017), despite accepted wisdom in classical statistical learning theory suggesting that typically $\mathcal{G} \approx \sqrt{d/n}$ (Anthony and Barlett, 1999).

To provide a theoretical understanding for this behavior, several directions have been explored, such as compression-based approaches (Arora *et al.*, 2018; Suzuki *et al.*, 2020; Barsbey *et al.*, 2021) and the approaches focusing on the double-descent phenomenon (Belkin *et al.*, 2019; Nakkiran *et al.*, 2019). Recently, there has been an increasing interest in examining the role of 'algorithm dynamics' on this phenomenon. In particular, it has been illustrated that, in the case where a stochastic optimization algorithm is used for minimizing (2), the optimization trajectories or their invariant distribution can exhibit a fractal structure (Şimşekli *et al.*, 2021; Camuto *et al.*, 2021; Birdal *et al.*, 2021; Hodgkinson *et al.*, 2022). This master thesis aims at studying and improving this fractal geometric approach. Let us first informally describe their results.

Under the assumption that ℓ is uniformly bounded by some B and uniformly L-Lipschitz with respect to w, the aforementioned can be *informally* summarized as follows: with probability $1 - \zeta$, we have that

$$\mathcal{G} \lesssim LB\sqrt{\frac{\bar{d}(\mathcal{W}) + I_\infty(\mathcal{W}, S) + \log(1/\zeta)}{n}}, \tag{4}$$

where the symbol $\lesssim$ means that we didn't write absolute constants and/or small logarithmic terms appearing in those results. Equation 4 informally sums-up various results[1]. In Equation (4), the following notations are used: $\mathcal{W}$ is a *data-dependent hypothesis set*, which is provided by the learning algorithm, $\bar{d}(\mathcal{W})$ is a notion of *fractal dimension* of $\mathcal{W}$, and $I_\infty(\mathcal{W}, S)$ denotes the *total mutual information* between

[1] In (Şimşekli *et al.*, 2021; Camuto *et al.*, 2021) the bound is logarithmic in L. (Şimşekli *et al.*, 2021) only requires sub-Gaussian losses while (Camuto *et al.*, 2021) requires sub-exponential losses. Their common points is to require a Lipschitz assumption.

the data S and the hypothesis set $\mathcal{W}$. These two last notions will be formally defined in Sections 2.3 and 3 respectively. In the case where the intrinsic dimension $\bar{d}(\mathcal{W})$ is significantly smaller than the ambient dimension d (which has been empirically illustrated in (Şimşekli *et al.*, 2021; Birdal *et al.*, 2021)), the bound in (4) provides an explanation on why overparameterized networks might not overfit in practice. Moreover, existing results provide examples of stochastic processes having low fractal dimensions compared to the ambient space dimension (Xiao, 2004).

While these bounds have brought a new perspective on understanding generalization, they also possess an important drawback, that is they all rely on a *uniform Lipschitz continuity* assumption on ℓ (with respect to the parameters), which is too strict to hold for deep learning models. While it is clear that we cannot expect Lipschitz continuity of a neural network when the parameter space is unbounded, Herrera *et al.* (2020) showed that, even for the bounded domains, the Lipschitz constants of fully connected networks are typically polynomial in the width, exponential in depth and linear in the input data, which may be excessively large in practical settings; hence might make the bounds vacuous.

The Lipschitz assumption is required in (Şimşekli *et al.*, 2021; Birdal *et al.*, 2021; Camuto *et al.*, 2021) as it enables the use of a fractal dimension defined through *the Euclidean distance* on the hypothesis set $\mathcal{W}$ (which is independent of the data). Hence, another downside of the Lipschitz assumption is that, the Euclidean distance-based dimension unfortunately ignores certain important components of the learning problem, such as how the loss ℓ behaves over the hypothesis set or trajectory $\mathcal{W}$. As shown in (Jiang *et al.*, 2019) in the case sharpness measures (Keskar *et al.*, 2017), which measure the sensitivity of the empirical risk around local minima and correlate well with generalization, the data-dependence may improve the ability of a complexity measure to explain generalization.

In this study, our main goal is to address the aforementioned issues by proving fractal geometric generalization bounds without requiring any Lipschitz assumptions. Inspired by a classical approach for bounding the Rademacher complexity (defined formally in Section 2.2), we achieve this goal by making use of a *data-dependent* pseudo-metric on the hypothesis set $\mathcal{W}$.

1.1 Mathematical framework

Before describing the structure of this thesis, let us clearly state the mathematical settings in which our results will be proven.

We formalize the learning algorithm as follows. The probability data-space is denoted by $(\mathcal{Z}, \mathcal{F}, \mu_z)$. For measure theoretic reasons, we will assume that $\mathcal{Z}$ may be endowed with a structure of separable metric space such that $\mathcal{F}$ is the corresponding Borel σ-algebra. A *learning algorithm* $\mathcal{A}$ is seen as a map generating a random closed set $\mathcal{W}_{S,U}$ (see (Molchanov, 2017, Definition 1.1.1)) from the data S and an external random variable U accounting for the randomness of the learning algorithm. The external randomness U takes values in some probability space $(\Omega_U, \mathcal{F}_U, \mu_u)$ and has distribution μ_u. Moreover, we assume that U is independent of S.

Therefore if we write $\mathbf{CL}(\mathbb{R}^d)$ for the set of closed sets of $\mathbb{R}^d$ endowed with the Effrös σ-algebra, as in (Molchanov, 2017), the algorithm will be written as a measurable map:

$$\mathcal{A} : (S, U) \in \bigcup_{n=0}^{\infty} \mathcal{Z}^n \times \Omega_U \longrightarrow \mathbf{CL}(\mathbb{R}^d) \ni \mathcal{W}_{S,U}. \tag{5}$$

The technical details regarding this setting, especially the notion of random sets, will be given in Section 5. This formulation encompasses several settings, such as the following examples.

Example 1. Given a continuous time process of the form $\mathrm{d}W_t = -\nabla f(W_t)\mathrm{d}t + \Sigma(W_t)\mathrm{d}X_t$ where X_t is typically a Brownian motion or a Lévy process, as considered in various studies (Mandt *et al.*, 2016; Chaudhari and Soatto, 2018; Hu *et al.*, 2018; Jastrzebski *et al.*, 2018; Şimşekli *et al.*, 2021), we can view $\mathcal{W}_{S,U}$ as the set of points of the trajectory $\{W_t,\ t \in [0,T]\}$, where U accounts for randomness coming from quantities defining the model like X_t.

Example 2. Consider a neural network $h_w(\cdot)$ and denote the output of the stochastic gradient descent (SGD) iterates by $A(x_0, S, U)$, where U accounts for random batch indices and x_0 is the initialization. This induces a learning algorithm $\mathcal{W}_{S,U} = \bigcup_{x_0 \in X_0} \{A(x_0, S, U)\}$, which is closed if X_0 is compact under a continuity assumption on A.

Example 3. As described in Section 6, our experiments will approximate $\mathcal{W}_{S,U}$ as the set of parameters computed through the iterations of a stochastic optimization algorithm used to train a neural network, as in prior works (Birdal *et al.*, 2021; Şimşekli *et al.*, 2021). In this case U accounts for randomness in the batch indices.

Under specific assumptions on the learning algorithm (5) and the loss $\ell : \mathbb{R}^d \times \mathcal{Z} \longrightarrow \mathbb{R}$ (continuity, boundedness), which will be described later on, we aim to bound the worst-case generalization error over $\mathcal{W}_{S,U}$, defined by Equation (3).

1.2 Contributions and plan of this work

The first two sections, Sections 2 and 3, will depict some classic mathematical background used throughout the project. A few classical results will be prove for the sake of completeness. Then, the main theoretical contributions are presented In Sections 4 and 5. Finally, in Section 6, we will describe our experimental results, after introducing some tools from topological data analysis.

More precisely, in Section 2 we discuss some probability theory results, namely concentration inequalities (Section 2.1) and classical statistical learning theory (Section 2.2), especially the notion of *Rademacher complexity* on which some of our results are built. Section 2.3 will also introduce various information theoretic tools that we need in our proofs and results because of their decoupling properties, as described for example in (Xu and Raginsky, 2017; Hodgkinson *et al.*, 2022).

In Section 3, we expose some basic notions of fractal geometry in metric spaces and describe their link with generalization as it appears in aforementioned previous results (Şimşekli *et al.*, 2021; Camuto *et al.*, 2021).

Section 4.3 is devoted to the introduction of a data-dependent pseudo-metric from which we will define a notion of data-dependent fractal dimension. Based on this notion, we will extend known covering arguments (Rebeschini, 2020) to prove our first result, which is a generalization bounds without Lipschitz assumption, in the case of a fixed hypothesis space.

We will then focus on the case of a random hypothesis set, as described in Section 1.1, in Section 5. In particular we prove in Section 5.2 a result of the following form, with similar notations than in Equation (4):

$$\mathcal{G} \lesssim B\sqrt{\frac{\bar{d}_S(\mathcal{W}_{S,U}) + I + \log(1/\zeta)}{n}}, \tag{6}$$

where $\bar{d}_S$ denotes our introduced notion of *data-dependent* fractal dimension and I is a (total) mutual information term (see Section 2.3). As opposed to prior work, this bound does not require any Lipschitz assumption and therefore applies to more general settings. However, this improvement comes at the expense of having a more complicated mutual information term compared to the one in (4). We will try to overcome this issue in Section 5.3 by introducing a notion of uniform geometric stability to prove another result involving mutual information terms which were already used in the literature (Hodgkinson *et al.*, 2022). To give a stronger theoretical basis to our analysis, we give some basic notions from the theory of random sets (Molchanov, 2017) and prove a few technical results in Section 5.1, hence solving potential measure-theoretic issues of this work.

Finally, in Section 6 we extend known results on 'persistent homology' (Adams *et al.*, 2020; Birdal *et al.*, 2021) to prove that our data-dependent intrinsic dimension may be estimated in a numerically efficient way, using a Python library presented in (Pérez *et al.*, 2021). We are therefore able to confirm the theory with experiments conducted on various settings.

2 Probability theory background

In this first section we review a few probabilistic tools which we will use throughout this work.

Sections 2.1 and 2.3 examine some classical notions related to high dimensional probability and information theory, including proofs of well-known results. Some old and recent applications to generalization theory are also presented to highlight the link with the problems presented in Section 1, especially the notion of Rademacher complexity, in Section 2.2, on which we build some of our results.

2.1 Concentration inequalities

One of the key quantity appearing in the generalization defined in Equation (3) is the pointwise generalization $\mathcal{R}(w) - \hat{\mathcal{R}}_S(w) = (1/n)\sum_{i=1}^n (\mathbb{E}_z[\ell(w,z)] - \ell(w,z))$, which is known to converge to 0 for a fixed w, as stated by the strong law of large numbers. In order to get more precise and non-asymptotic improvements of this result, one can refer to *concentration inequalities*, from which we will introduce some basic notion. More details can be found in (Boucheron *et al.*, 2013; Vershynin, 2020).

2.1.1 Hoeffding's inequalities

Inspired by concentration property of Bernoulli and normal random variables, 'sub-Gaussian' distributions are introduced as random variables with sub-Gaussian tails. This includes many classical distribution, like bounded random variables, and is widely used in statistical learning theory. We will define it as follows (Vershynin, 2020, Definition 2.5.6):

Definition 1 (Sub-Gaussian random variables)**.** *Let X be a random variable on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$, then X is said to be sub Gaussian if there exists some $\sigma > 0$ such that:*

$$\forall \epsilon > 0, \ \mathbb{P}(|X| \geq \epsilon) \leq 2e^{-\epsilon^2/\sigma^2}.$$

In that case, we may say that X is σ^2-sub-Gaussian.

Note that as $\mathbb{E}[|X|] = \int_0^\infty \mathbb{P}(|X| \geq \epsilon) d\epsilon$ this implies that X is integrable. The following proposition gives some of the equivalent definitions of the sub-Gaussian property:

Proposition 1 (Proposition 2.5.2 in (Vershynin, 2020))**.** *The variable X is sub-Gaussian if and only if either one of the following is verified (in each statement, Σ denotes a quantity equal to $C\sigma$, where C is an absolute constant, not necessarily always the same one):*

i There exists $\Sigma > 0$ such that $\forall p \geq 1, \ \mathbb{E}[|X|^p] \leq \left(\Sigma\sqrt{p}\right)^p$.

ii There exists $\Sigma > 0$ such that $\mathbb{E}[e^{X^2/\Sigma^2}] \leq 2$.

iii If X has zero mean, then $\forall \lambda \in \mathbb{R}, \ \mathbb{E}[e^{\lambda X}] \leq e^{\Sigma^2 \lambda^2}$.

Our main interest in sub-Gaussian distributions resides in their concentration properties given by the following inequality from Hoeffding:

Theorem 2.1. Hoeffding's inequality (Hoeffding, 1963)

Let $X_1, \dots, X_n$ be independent, mean-zero, sub-Gaussian random variables with parameters $\sigma_1^2, \dots, \sigma_n^2$. Let $S_n := X_1 + \cdots + X_n$, then for any $\epsilon > 0$ we have

$$\mathbb{P}(|S_n| \geq \epsilon) \leq 2\exp\left(- \frac{a\epsilon^2}{\sum_{i=1}^n \sigma_i^2} \right),$$

where a is an absolute constant.

Proof. Let us fix some $\lambda > 0$ and consider the constant $\Sigma_1, \dots, \Sigma_n > 0$ appearing in the third point of Proposition 1, associated to the variables $X_1, \dots, X_n$, we can write:

$$\begin{aligned}
\mathbb{P}(S_n \geq \epsilon) &\leq \mathbb{P}(e^{\lambda S_n} \geq e^{\lambda \epsilon}) \\
&\leq e^{-\lambda\epsilon} \mathbb{E}[e^{\lambda S_n}] \quad \text{(Markov's inequality)} \\
&\leq e^{-\lambda\epsilon} \prod_{i=1}^n \mathbb{E}[e^{\lambda X_i}] \quad \text{(Independence)} \\
&\leq e^{-\lambda\epsilon} \prod_{i=1}^n \mathbb{E}[e^{\lambda^2 \Sigma_i^2}] \quad \text{(Sub-Gaussian property)} \\
&= \exp\left\{ -\lambda\epsilon + \lambda^2 \sum_{i=1}^n \Sigma_i^2 \right\}.
\end{aligned}$$

A second degree term appears inside the exponential, we can reach its minimum by setting $\lambda = \epsilon/(2\sum_i \Sigma_i^2)$, which gives us

$$\mathbb{P}(S_n \geq \epsilon) \leq \exp\left(- \frac{-\epsilon^2}{4\sum_{i=1}^n \Sigma_i^2} \right). \tag{7}$$

Doing the same reasoning with $-S_n$ in place of S_n and reminding that $\Sigma_i = C\sigma_i$ gives us the result. □

An important class of sub-Gaussian random variables is that of bounded random variables. In most of our work, we will need a bounded loss assumption in order to get finite fractal dimensions. However, as in prior works (Şimşekli *et al.*, 2021; Birdal *et al.*, 2021), some results may hold by replacing the boundedness assumption by a weaker sub-Gaussian assumption. This is made formal by the well-known *Hoeffding's lemma*:

Lemma 1 (Hoeffding's lemma)**.** *Let $a, b \in \mathbb{R}$, if X is a random variable such that $a \leq X \leq b$ almost surely, then*

$$\mathbb{E}[e^{\lambda(X - \mathbb{E}[X])}] \leq \exp\left(\lambda^2 \frac{(b-a)^2}{8} \right).$$

Proof. One classical proof of this result is based on the convexity of the exponential and uses a second order Taylor expansion of a well-chosen function. Here, we present a proof inspired from (Boucheron *et al.*, 2013), which may give more intuition.

Without loss of generality, up to use $X' = X - \mathbb{E}[X]$, we can assume that $\mathbb{E}[X] = 0$. Let us fix some $\lambda \in \mathbb{R}$.

Let Y be any random variable with values in $[a, b]$ almost surely. Then one has that $|Y - (a+b)/2| \leq (b-a)/2$, so that

$$\mathbb{V}[Y] = \mathbb{V}\left[Y - \frac{a+b}{2} \right] \leq \frac{(b-a)^2}{4}. \tag{8}$$

Inspired by this, let us introduce a random variable Y with values in $[a, b]$ such that its law has the following Radon-Nykodym's derivative with respect to the law of X:

$$\frac{\mathrm{d}Y_\star \mathbb{P}}{\mathrm{d}X_\star \mathbb{P}} = \frac{e^{\lambda X}}{\mathbb{E}[e^{\lambda X}]}.$$

By Equation (8) we get that

$$\frac{\mathbb{E}[X^2 e^{\lambda X}]}{\mathbb{E}[e^{\lambda X}]} - \frac{\mathbb{E}[X e^{\lambda X}]^2}{\mathbb{E}[e^{\lambda X}]^2} \leq \frac{(b-a)^2}{4}.$$

Now let us introduce the function $\psi : \lambda \mapsto \log\left(\mathbb{E}[e^{\lambda X}]\right)$. Using the boundedness of X and the smoothness of the function $\lambda \mapsto e^{\lambda x}$, it is obvious that we can differentiate under the expectation and that ψ is of class $\mathcal{C}^2$. Therefore

$$\psi'(\lambda) = \frac{\mathbb{E}[X e^{\lambda X}]}{\mathbb{E}[e^{\lambda X}]}, \quad \psi''(\lambda) = \frac{\mathbb{E}[X^2 e^{\lambda X}]}{\mathbb{E}[e^{\lambda X}]} - \frac{\mathbb{E}[X e^{\lambda X}]^2}{\mathbb{E}[e^{\lambda X}]^2}.$$

Therefore, by using Taylor-Lagrange's formula around $\lambda = 0$, we get that

$$\log\left(\mathbb{E}[e^{\lambda X}]\right) \leq 0 + 0 + \frac{\lambda^2}{2}.\frac{(b-a)^2}{4},$$

hence the lemma. □

As a corollary, we may state Hoeffding's inequality for bounded random variables, which we will use in several proofs.

Corollary 1 (Hoeffding's inequality for bounded random variables)**.** *Let $X_1, \dots, X_n$ be random variables such that for each i we have $a_i \leq X_i \leq b_i$ almost surely. We define $S_n := X_1 + \dots + X_n$, then the two following inequalities hold:*

$$\mathbb{P}(S_n - \mathbb{E}[S_n] \geq \epsilon) \leq \exp\left(- \frac{-2\epsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right), \quad \textit{and}$$
$$\mathbb{P}(|S_n - \mathbb{E}[S_n]| \geq \epsilon) \leq 2\exp\left(- \frac{-2\epsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

Proof. We follow the exact same steps as in the proof of Theorem 2.1 but we use Hoeffding's lemma 1 instead of the sub-Gaussian property, which gives us immediately that

$$\mathbb{P}(S_n - \mathbb{E}[S_n] \geq \epsilon) \leq e^{-\lambda\epsilon} \prod_{i=1}^n \mathbb{E}[e^{\lambda(X_i - \mathbb{E}[X_i])}] \leq \exp\left\{ -\lambda\epsilon + \frac{\lambda^2}{8}\sum_{i=1}^n (b_i - a_i)^2\right\}.$$

Setting $\lambda = 4\epsilon / \sum_{i=1}^n (b_i - a_i)^2$ gives us:

$$\mathbb{P}(S_n - \mathbb{E}[S_n] \geq \epsilon) \leq \exp\left(- \frac{-2\epsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

The second inequality is immediately deduced from a symmetry argument. □

Example 4. Let us consider the worst-case generalization setting described in the introduction by Equation 3, where $\mathcal{W}$ is assumed to be a non-random (fixed) finite set of cardinal $|\mathcal{W}|$. Let us also assume that the loss ℓ is uniformly bounded by some B (i.e. $|\ell(w,z)| \leq B$), then applying Hoeffding's inequality along with a union bound we get, for every $\epsilon > 0$ that

$$\begin{aligned}
\mathbb{P}\left(\sup_{w \in \mathcal{W}} \left(\mathcal{R}(w) - \hat{\mathcal{R}}_S(w)\right) \geq \epsilon\right) &\leq \mathbb{P}\left(\bigcup_{w \in \mathcal{W}} \left(\mathcal{R}(w) - \hat{\mathcal{R}}_S(w)\right) \geq \epsilon\right) \\
&\leq \sum_{w \in \mathcal{W}} \mathbb{P}\left(\mathcal{R}(w) - \hat{\mathcal{R}}_S(w) \geq \epsilon\right)\Big) \\
&\leq \sum_{w \in \mathcal{W}} \exp\left\{ - \frac{n\epsilon^2}{2B^2}\right\} \\
&= |\mathcal{W}| \exp\left\{ - \frac{n\epsilon^2}{2B^2}\right\}.
\end{aligned}$$

Rearranging the above equation by setting its right hand term equal to some $\zeta \in (0,1)$, we get that with probability at least $1 - \zeta$ over $S \sim \mu_z^{\otimes n}$ one has the well known result (Rebeschini, 2020):

$$\sup_{w \in \mathcal{W}} \left(\mathcal{R}(w) - \hat{\mathcal{R}}_S(w)\right) \leq \sqrt{2}B\sqrt{\frac{\log|\mathcal{W}| + \log(1/\zeta)}{n}}. \tag{9}$$

2.1.2 Mc-Diarmid's inequality

Let us now introduce another fundamental concentration inequality. Mc-Diarmid inequality is also a fundamental building block of statistical learning theory, used for instance to prove high probability bounds on the worst case generalization error (Rebeschini, 2020) with fixed hypothesis set or to leverage stability conditions, as described by Bousquet (2002). Rademacher complexity is mainly used for fixed hypothesis sets, however, more recently, some authors have tried to extend this notion to random hypothesis sets (Foster *et al.*, 2020). In the context of this project, we were be able to build upon Rademacher complexity to get our result for a fixed $\mathcal{W}$, while another technique has to be used in the general case.

This result is based on the so-called *bounded difference property*, from which we give the following definition.

Definition 2 (Bounded difference property)**.** *Let $f : \mathcal{X}_1 \times \cdots \times \mathcal{X}_n \to \mathbb{R}$ be some arbitrary measurable function such that there exists some $a_1, \dots, a_n \geq 0$ satisfying, for each i and all $(x_1, \dots, x_n) \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$:*

$$\forall x_i' \in \mathcal{X}_i, \ |f(x_1, \dots, x_{i-1}, x_i', x_{i+1}, \dots, x_n) - f(x_1, \dots, x_{i-1}, x_i, x_{i+1}, \dots, x_n)| \leq a_i$$

Let us now state Mc-Diarmid's inequality.

Theorem 2.2. Mc-Diarmid's inequality

If f has the bounded difference property and $X_1, \dots, X_n$ are *independent* random variables in the sets $\mathcal{X}_1, \dots, \mathcal{X}_n$, then the two following inequalities hold:

$$\mathbb{P}(f(X_1, \dots, X_n) - \mathbb{E}[f(X_1, \dots, X_n)] \geq \epsilon) \leq \exp\left(-\frac{-2\epsilon^2}{\sum_{i=1}^n a_i^2}\right), \quad \text{and}$$
$$\mathbb{P}(|f(X_1, \dots, X_n) - \mathbb{E}[f(X_1, \dots, X_n)]| \geq \epsilon) \leq 2\exp\left(-\frac{-2\epsilon^2}{\sum_{i=1}^n a_i^2}\right).$$

2.2 Rademacher complexity

We give in this subsection an important example of the concentration inequalities developed in Section 2.1. Some approaches to explore the generalization property of learning algorithm attempt to relate generalization to notions of intrinsic dimension of the hypothesis set $\mathcal{W}$, such as the well-known VC dimension (Vapnik and Chervonenkis, 2015), or notions of complexity of $\mathcal{W}$, such that the celebrated *Rademacher complexity*, which we shall now describe. We will then go over the arguments allowing us to relate this complexity to the worst-case generalization error (3), among other things based on Theorem 2.2, and the concentration properties it exhibits.

As explained in (Rebeschini, 2020), Rademacher complexity measures the capacity of a set to replicate some random, Bernoulli like, signals. More precisely, we call *Rademacher random variables* a tuple $(\sigma_1 \dots, \sigma_n)$ of mutually independent Bernoulli distributions with values in the set $\{-1, 1\}$, which means that for each i we have $\mathbb{P}(X_i = 1) = \mathbb{P}(X_i = -1) = 1/2$.

Using Rademacher complexity in our setting becomes interesting under the following assumption, which we make in all this Subsection and further sections when specified:

Assumption 1. *The loss $\ell : \mathbb{R}^d \times \mathcal{Z}$ is continuous and uniformly bounded by some $B > 0$.*

While the continuity is merely used for measure-theoretic reasons, boundedness is essential to apply and concentration properties related to Rademacher complexity. The following definition introduces the Rademacher complexity of any fixed set, we will then make an informed choice of such a set.

Definition 3 (Rademacher complexity of a set)**.** *Let us consider a fixed set $A \subset \mathbb{R}^n$ and $\boldsymbol{\sigma} := (\sigma_1 \dots, \sigma_n$ some Rademacher random variables, the Rademacher complexity of A is defined as*

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

Let us consider a fixed *closed* hypothesis set $\mathcal{W}$ and some dataset $S = (z_1, \dots, z_n) \sim \mu_z^{\otimes n}$, we will use the following notation:

$$\ell(\mathcal{W}, S) := \{(\ell(w, z_i)_{1 \leq i \leq n} \in \mathbb{R}^n, \ w \in \mathcal{W}\} \subseteq \mathbb{R}^n. \tag{10}$$

Intuitively, $\ell(\mathcal{W}, S)$ is the image of the hypothesis set $\mathcal{W}$ by an embedding from the parameter space $\mathbb{R}^d$ to the 'output space' $\mathbb{R}^n$. The interest of Rademacher complexity in statistical learning theory comes from the consideration of its evaluation on the set (10), hence denoted $\mathbf{Rad}(\ell(\mathcal{W}, S))$.

Remark 1. *One could legitimately inquire about the measurability of $\boldsymbol{Rad}(\ell(\mathcal{W}, S))$ with respect to $\mathcal{F}^{\otimes n}$ (recall that the data space is denoted $(\mathcal{Z}, \mathcal{F}, \mu_z)$). Thanks to the closeness of $\mathcal{W} \subseteq \mathbb{R}^d$ we can introduce a dense countable subset $\mathcal{C}$ of $\mathcal{W}$ and write that, thanks to the continuity of ℓ,*

$$R(\boldsymbol{\sigma}, S) := \frac{1}{n} \sup_{w \in \mathcal{W}} \sum_{i=1}^{n} \sigma_i \ell(w, z_i) = \frac{1}{n} \sup_{w \in \mathcal{C}} \sum_{i=1}^{n} \sigma_i \ell(w, z_i),$$

which is measurable as a countable supremum of random variables. As ℓ is bounded, so is $R(\boldsymbol{\sigma}, S)$; it is therefore integrable with respect to $(\boldsymbol{\sigma}, S)$. Thus $\boldsymbol{Rad}(\ell(\mathcal{W}, S))$ is integrable (and measurable) thanks to the first part of Fubini's theorem.

The quantity $\mathbf{Rad}(\ell(\mathcal{W}, S))$ is linked to generalization via the following proposition (see for example (Rebeschini, 2020)):

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

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

The proof of proposition 3, which is written below, is based on two essential arguments: first a well-known *symmetrization argument* is used to bound the expected worst-case generalization with the expected Rademacher complexity, then those two expectations are moved by applying twice Mc-Diarmid inequality (Theorem 2.2).

Proof. Let us write:

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

We introduce $\tilde{S} = \{\tilde{z}_1, \dots, \tilde{z}_n\} \sim \mu_z^{\otimes n}$ an independent copy of S and some Rademacher random variables $(\sigma_1, \dots, \sigma_n)$, using properties of conditional expectation and Fubini's theorem we have:

$$\begin{aligned}
\mathbb{E}[\mathcal{G}(S)] &= \mathbb{E}\bigg[\sup_{w \in \mathcal{W}} \bigg(\frac{1}{n} \sum_{i=1}^{n} \mathcal{R}(w) - \ell(w, z_i)\bigg)\bigg] \\
&= \mathbb{E}\bigg[\sup_{w \in \mathcal{W}} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[\ell(w, \tilde{z}_i) - \ell(w, z_i) | \tilde{S}]\bigg] \\
&\leq \mathbb{E}\bigg[\mathbb{E}\bigg[\sup_{w \in \mathcal{W}} \frac{1}{n} \sum_{i=1}^{n} (\ell(w, \tilde{z}_i) - \ell(w, z_i)) \bigg| \tilde{S}\bigg]\bigg] \\
&= \mathbb{E}\bigg[\sup_{w \in \mathcal{W}} \frac{1}{n} \sum_{i=1}^{n} (\ell(w, \tilde{z}_i) - \ell(w, z_i))\bigg] \\
&= \mathbb{E}\bigg[\sup_{w \in \mathcal{W}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i(\ell(w, z_i) - \ell(w, \tilde{z}_i))\bigg] \\
&\leq 2\mathbb{E}\bigg[\sup_{w \in \mathcal{W}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \ell(w, z_i)\bigg] \\
&= 2\mathbb{E}[\mathbf{Rad}(\ell(\mathcal{W}, S))].
\end{aligned} \tag{11}$$

On the other hand if we denote $S^i = (z_1, \dots, z_{i-1}, \tilde{z}_i, z_{i+1}, \dots z_n)$ we have that:

$$|\mathcal{G}(S) - \mathcal{G}(S^i)| \leq \frac{2B}{n},$$

And therefore by Mc-Diarmid inequality for any $\epsilon > 0$:

$$\mathbb{P}\bigg(\mathcal{G}(S) - \mathbb{E}[\mathcal{G}(S)] \geq \epsilon\bigg) \leq \exp\bigg\{-\frac{n\epsilon^2}{2B^2}\bigg\}.$$

By taking any $\eta \in (0,1)$ we can make a clever choice for ϵ and deduce that with probability at least $1-\eta$ we have:

$$\mathcal{G}(S) \leq \mathbb{E}[\mathcal{G}(S)] + \sqrt{\frac{2B^2}{n}\log(1/\eta)}. \tag{12}$$

Moreover we can also write:

$$|\mathbf{Rad}(\ell(\mathcal{W},S)) - \mathbf{Rad}(\ell(\mathcal{W},S^i))| \leq \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{w\in\mathcal{W}} \frac{1}{n}\left|\sigma_i(\ell(w,z_i)-\ell(w,\tilde{z}_i))\right|\right] \leq \frac{2B}{n},$$

so that by Mc-Diarmid and the exact same reasoning than above we have that with probability at least $1-\eta$:

$$\mathbb{E}[\mathbf{Rad}(\ell(\mathcal{W},S))] \leq \mathbf{Rad}(\ell(\mathcal{W},S)) + \sqrt{\frac{2B^2}{n}\log(1/\eta)}. \tag{13}$$

Therefore combining equations 11, 12 and 13 gives us that with probability at least $1-2\eta$:

$$\mathcal{G}(S) \leq 2\mathbf{Rad}(\ell(\mathcal{W},S)) + 3\sqrt{\frac{2B^2}{n}\log(1/\eta)}.$$

□

Now that we can efficiently bound the generalization error by a term involving the Rademacher complexity, the natural question arises, of whether we are able to bound the Rademacher complexity itself. It appears that, thanks to our bounded loss assumption, a result may be obtained in the case of finite hypothesis sets. This is a consequence of the following concentration lemma:

Proposition 3 (Massart's lemma, (Massart, 2000))**.** *Let $T \subseteq \mathbb{R}^n$ be a finite set, then:*

$$\boldsymbol{Rad}(T) \leq \max_{t\in T}(\|t\|_2)\frac{\sqrt{2\log(|T|)}}{n}.$$

Proof. Let us take $\lambda > 0$, $\boldsymbol{\sigma} = (\sigma_1,\ldots,\sigma_n)$ some Rademacher random variables and denote $M := \max_{t\in T}(\|t\|_2)$, we have

$$\begin{aligned}
\exp(n\lambda\mathbf{Rad}(T)) &\leq \mathbb{E}_{\boldsymbol{\sigma}}\left[\exp\left(\max_{t\in T}\lambda\sum_{i=1}^n \sigma_i t_i\right)\right] \quad \text{(Jensen's inequality)}\\
&\leq \sum_{t\in T}\mathbb{E}_{\boldsymbol{\sigma}}\left[\exp\left(\lambda\sum_{i=1}^n \sigma_i t_i\right)\right]\\
&\leq \sum_{t\in T}\prod_{i=1}^n \mathbb{E}_{\sigma_i}[e^{\lambda\sigma_i t_i}], \quad \text{(Independence and Fubini's theorem)}\\
&\leq \sum_{t\in T}\prod_{i=1}^n \exp\left\{\frac{\lambda^2 t_i^2}{2}\right\} \quad \text{Hoeffding's lemma}\\
&\leq \sum_{t\in T}\exp\left\{\frac{\lambda^2}{2}\sum_{i=1}^n t_i^2\right\}\\
&\leq |T|\exp\left\{\frac{\lambda^2 M^2}{2}\right\}.
\end{aligned}$$

And therefore:

$$\mathbf{Rad}(T) \leq \frac{\log(|T|)}{\lambda n} + \lambda\frac{M^2}{2n}.$$

The minimum of the function $\lambda \mapsto a/\lambda + b\lambda$ being reached for $\lambda = \sqrt{a/b}$, we immediately get the result:

$$\mathbf{Rad}(T) \leq M\frac{\sqrt{2\log(|T|)}}{n}.$$

□

Example 5. Let's get back to a setting with a finite hypothesis set $\mathcal{W}$, as in Example 4. In that case we have that $\max_{w\in\mathcal{W}}(\|(\ell(w,z_i))_i\|_2) \leq B\sqrt{n}$, thanks to the boundedness assumption. Thus Massart's lemma 3 gives us

$$\mathbf{Rad}(\ell(\mathcal{W},S)) \leq B\sqrt{\frac{2\log(|\mathcal{W}|)}{n}}. \tag{14}$$

This result is classically involved in a more general *covering* argument applying to a more general hypothesis set $\mathcal{W}$. We will present and extend this argument to get our first result in Section 4.

2.3 Information theoretic quantities

Recently, one popular approach to prove generalization bounds has been based on information theory. In this context, Xu and Raginsky (2017); Russo and Zou (2019) proved particularly interesting generalization bounds in terms of the *mutual information* between input and output of the model. Other authors refined this argument in various settings (Pensia *et al.*, 2018; Negrea *et al.*, 2019; Steinke and Zakynthinou, 2020; Harutyunyan *et al.*, 2021) while Asadi *et al.* (2019) combined mutual information and chaining to tighten the bounds.

In our work we will use the total mutual information to specify the dependence between the data and the fractal properties of the hypothesis set.

2.3.1 Mutual information and decoupling

Let us first recall some basics concepts of information theory, the reader may consult (van Erven and Harremoës, 2014) for more details on those notions. The following definition introduces Kullback-Leibler (KL) and Renyi divergences, which are the basic building blocks of the aforementioned approaches. The absolute continuity of one probability measure with respect to another will be denoted by $\ll$.

Definition 4 (Kullback-Leibler and Renyi divergences)**.** *Let us consider a probability space* $(\Omega,\mathcal{F})$ *and two probability distributions* π *and* ρ*, with* $\pi \ll \rho$*. We define the* Kullback-Leibler divergence *of those distributions as:*

$$\boldsymbol{KL}(\pi||\rho) = \int \log\left(\frac{d\pi}{d\rho}\right) d\pi.$$

For $\alpha > 1$*, we define their* α-Renyi divergence *as:*

$$D_\alpha(\pi||\rho) = \frac{1}{\alpha-1}\log\int\left(\frac{d\pi}{d\rho}\right)^\alpha d\rho.$$

We set those two quantities to $+\infty$ *if the absolute continuity condition is not verified.*

Note that by convention we often consider that $D_1 = \mathbf{KL}$ and that Renyi divergences may also be defined for orders $\alpha < 1$ (van Erven and Harremoës, 2014), but we won't need it here.

It is easy to prove that D_α is increasing in α and it is therefore natural to define:

$$D_\infty(\pi||\rho) = \lim_{\alpha\to 0} D_\alpha(\pi||\rho).$$

Intuitively, those divergences ma be interpreted as distances over probability distributions. As it is well known, for two random variables X and Y, that X and Y are independent if and only if their joint law satisfies $\mathbb{P}_{X,Y} = \mathbb{P}_X \otimes \mathbb{P}_Y$, it is natural to further measure the statistical dependence between X and Y by computing the divergences between those two distributions.

Definition 5 (Mutual information)**.** *Let* X, Y *be two random variables on* Ω*, we define for* $\alpha \in [1,\infty]$*:*

$$I_\alpha(X,Y) := D_\alpha(\mathbb{P}_{X,Y}||\mathbb{P}_X \otimes \mathbb{P}_Y),$$

with in particular:

$$I(X,Y) := I_1(X,Y) = \boldsymbol{KL}(\mathbb{P}_{X,Y}||\mathbb{P}_X \otimes \mathbb{P}_Y).$$

I_∞ *will be called the total mutual information.*

One of the main properties of those notions of mutual information, which we shall use intensively in Section 5.3, is to satisfy the so-called *data processing inequality*, which we state in the following theorem.

Theorem 2.3. Data processing inequalities

Given a probability space $(\Omega, \mathcal{F})$, $\mathcal{G}$ a sub-σ-algebra of $\mathcal{F}$ and π, ρ two probability distributions, one has:

$$D_\alpha(\pi|_\mathcal{G}||\rho|_\mathcal{G}) \leq D_\alpha(\pi||\rho).$$

If $X \longrightarrow Y \longrightarrow Z$ is a Markov chain, then one has the *data processing inequality for mutual information*, for any $\alpha \in [1, +\infty]$: $I_\alpha(X, Z) \leq I_\alpha(X, Y)$.

Theorem 2.3.1 implies that if the conditional distribution of a random variable Z with respect to (X, Y) is independent of X, then Z and X 'share less information' than X and Y. See (van Erven and Harremoës, 2014) for a proof of this result.

Our particular interest in those quantities resides in their decoupling properties. In particular Hodgkinson *et al.* (2022) used the following proposition in their generalization bounds and introduce a mutual information term between the data and the hypothesis set, an approach from which we take inspiration in our work, which would correspond to $I_\infty(S, \mathcal{W}_{S,U})$ in our notations.

Proposition 4 (Decoupling in probability, lemma 1 in (Hodgkinson *et al.*, 2022))**.** *Let* $(\Omega, \mathcal{F}, \mathbb{P})$ *be a probability space and* X, Y *be two random variables with values in two measurable spaces* $(\Omega_X, \mathcal{F}_X)$, $(\Omega_Y, \mathcal{F}_Y)$*, then for each* $B \in \mathcal{F}_X \otimes \mathcal{F}_Y$ *we have:*

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

Our proof follows the same idea than (Hodgkinson *et al.*, 2022), perhaps in a more precise way.

Proof. Let $\mathcal{B}(p)$ denote the Bernoulli distribution with parameter p. Let us consider $\mathcal{G} = \sigma(\{B\})$ the sub σ-algebra of $\mathcal{F}_X \otimes \mathcal{F}_Y$ generated by B.
It is clear that $\mathbb{P}_{X,Y}|_\mathcal{G}$ induces the same distribution than $\mathcal{B}(\mathbb{P}_{X,Y}(B))$, and similarly for $\mathbb{P}_X \otimes \mathbb{P}_Y|_\mathcal{G}$ and $\mathcal{B}(\mathbb{P}_X \otimes \mathbb{P}_Y(B))$.
Therefore, for any $\alpha > 1$ we have by the data-processing inequality:

$$D_\alpha(\mathcal{B}(\mathbb{P}_{X,Y}(B))||\mathcal{B}(\mathbb{P}_X \otimes \mathbb{P}_Y(B))) \leq I_\alpha(X, Y) \leq I_\infty(X, Y).$$

Given $a, b \in (0, 1)$ we can compute analytically the divergence as:

$$\begin{aligned} D_\alpha(\mathcal{B}(a), \mathcal{B}(b)) &= \frac{1}{\alpha - 1} \log \int \left(\frac{\mathrm{d}\mathcal{B}(a)}{\mathrm{d}\mathcal{B}(b)} \right)^\alpha \mathrm{d}\mathcal{B}(b) \\ &= \frac{1}{\alpha - 1} \log \left(b \big(\frac{a}{b}\big)^\alpha + (1 - b) \big(\frac{1 - a}{1 - b}\big)^\alpha \right) \\ &\underset{\alpha \to \infty}{\longrightarrow} \log \max \Big\{ \frac{a}{b}, \frac{1 - a}{1 - b} \Big\} \\ &\geq \log \big(\frac{a}{b} \big). \end{aligned}$$

Hence the result if $\mathbb{P}_{X,Y}(B), \mathbb{P}_X \otimes \mathbb{P}_Y(B) \in (0, 1)$.
If $\mathbb{P}_{X,Y}(B) = 0$ then the result is obvious.
If $\mathbb{P}_{X,Y}(B) = 1$ and $\mathbb{P}_X \otimes \mathbb{P}_Y(B) > 0$, then we do the same computation but we notice that the Radon-Nykodym derivative in the integral is $\frac{\mathrm{d}\mathcal{B}(1)}{\mathrm{d}\mathcal{B}(b)}(x) = \frac{1}{b} \mathbb{1}_1(x)$ on $\{0, 1\}$.
The last case is when $\mathbb{P}_{X,Y}(B) = 1$ and $\mathbb{P}_X \otimes \mathbb{P}_Y(B) = 0$. In that case we note that the total mutual information is infinite.

□

We actually have an even stronger result (van Erven and Harremoës, 2014, Theorem 6):

Theorem 2.4. Total divergence as an essential supremum

With the same notations we have:

$$D_\infty(\pi||\rho) = \log\left(\sup_{B\in\mathcal{F}} \frac{\pi(B)}{\rho(B)}\right).$$

We see that Proposition 4 is an immediate corollary of Theorem 2.4.

2.3.2 Digression: a mutual information bound

In this subsection, we quickly present a decoupling method, based on mutual information between the input of a learning algorithm and its output, leading to a famous generalization bound from Xu and Raginsky (2017). Our goal is to be able to discuss the comparison between the mutual information terms appearing in our bounds (see Section 5) and the one more classically appearing in the literature.

The aforementioned decoupling technique is based on the celebrated Donsker-Varadhan's formula, see (Boucheron *et al.*, 2013), which is also known for its use in PAC-Bayesian generalization bounds.

Proposition 5 (Donsker-Varadhan's formula)**.** *Let π, ρ be two probability measures on some probability space, with $\pi \ll \rho$, we have the following variational formulation of the KL divergence:*

$$\boldsymbol{KL}(\pi||\rho) = \sup_f \left(\int f\,d\pi - \log\int e^f\,d\rho\right),$$

where the supremum is taken over measurable functions such that $\mathbb{E}_\rho[e^f] < +\infty$.

Proof. Let us consider a probability measure ν equivalent to ρ (i.e. $\nu \ll \rho$ and $\rho \ll \nu$), then by the chain rules of Radon-Nykodym derivatives we have

$$\begin{aligned}\mathbf{KL}(\pi||\rho) &= \mathbf{KL}(\pi||\nu) + \int \log\left(\frac{\mathrm{d}\nu}{\mathrm{d}\rho}\right)\mathrm{d}\pi \\ &\geq \int \log\left(\frac{\mathrm{d}\nu}{\mathrm{d}\rho}\right)\mathrm{d}\pi.\end{aligned}$$

As ν is equivalent to ρ, it has almost everywhere the form $\frac{\mathrm{d}\nu}{\mathrm{d}\rho} = Ce^f$, where f is some measurable function. In order for ν to be a probability measure we must have $C^{-1} = \mathbb{E}_\rho[e^f]$. Therefore we obtain:

$$\mathbf{KL}(\pi||\rho) \geq \int f\mathrm{d}\pi - \log\int e^f\mathrm{d}\rho.$$

Now just note that equality in Proposition 5 is obtained when $\mathbf{KL}(\pi||\nu) = 0$, which leads to $\frac{\mathrm{d}\pi}{\mathrm{d}\rho} = \frac{e^f}{\mathbb{E}_\rho[e^f]}$. □

This allows us to prove a first decoupling result, which is a very powerful tool widely used in the generalization literature to relate the generalization error of a learning algorithm to the mutual information between its inputs and outputs (see (Xu and Raginsky, 2017), (Negrea *et al.*, 2019), (Pensia *et al.*, 2018), (Russo and Zou, 2019)).

Proposition 6 (Lemma 1 in (Xu and Raginsky, 2017))**.** *Let X, Y be two random variables and $f(.,.)$ a measurable function. We consider $\bar{X}$ and $\bar{Y}$ two copies of X and Y which are independent. Then if $f(\bar{X},\bar{Y}) - \mathbb{E}[f(\bar{X},\bar{Y})]$ is σ^2-subGaussian, we have:*

$$|\mathbb{E}[f(X,Y)] - \mathbb{E}[f(\bar{X},\bar{Y})]| \leq \sqrt{2\sigma^2 I(X,Y)}.$$

Proof. By independence we have $\mathbb{P}_{X,Y} = \mathbb{P}_{\bar{X}} \otimes \mathbb{P}_{\bar{Y}}$. Therefore, applying proposition 5 we can write, for every $\lambda \in \mathbb{R}$:

$$\begin{aligned}I(X,Y) &= \mathbf{KL}(\mathbb{P}_{X,Y}||\mathbb{P}_X \otimes \mathbb{P}_Y) \\ &\geq \lambda\mathbb{E}_{X,Y}[f(X,Y)] - \log\mathbb{E}_{\bar{X},\bar{Y}}[e^{\lambda f(\bar{X},\bar{Y})}].\end{aligned}$$

Now using the subGaussian property we get:

$$I(X,Y) \geq \lambda \mathbb{E}_{X,Y}[f(X,Y)] - \lambda \mathbb{E}_{\bar{X},\bar{Y}}[f(\bar{X},\bar{Y})] - \log \mathbb{E}_{\bar{X},\bar{Y}}\left[e^{\lambda f(\bar{X},\bar{Y}) - \lambda \mathbb{E}[f(\bar{X},\bar{Y})]}\right]$$
$$\geq \lambda \mathbb{E}_{X,Y}[f(X,Y)] - \lambda \mathbb{E}_{\bar{X},\bar{Y}}[f(\bar{X},\bar{Y})] - \lambda^2 \frac{\sigma^2}{2}.$$

Which gives us a second order polynomial equation whose discriminant must be $\Delta \leq 0$. Thus:

$$\left(\mathbb{E}_{X,Y}[f(X,Y)] - \mathbb{E}_{\bar{X},\bar{Y}}[f(\bar{X},\bar{Y})]\right)^2 - 2\sigma^2 I(X,Y)^2 \leq 0,$$

Which gives us the result. □

Example 6. Let us consider the learning algorithm as taking as input a random dataset S and outputs a random variable W, corresponding to the trained parameters (so that W *depends* on S in general). If the loss is assumed to be σ^2 subGaussian, then the above result immediately gives us a bound on the expected generalization error:

$$|\mathbb{E}[\mathbb{R}(W) - \hat{\mathcal{R}}_S(W)]| \leq \sqrt{\frac{2\sigma^2}{n} I(S,W)}. \tag{15}$$

This is a well-known result (Xu and Raginsky, 2017, Theorem 1).

We end this section by making a remark to assess the differences between this classical mutual information bound and the mutual information terms appearing in our work, as well as previous works. The existence of those kind of bounds may shed more light on the more technical mutual information terms that we will introduce in Sections 4 and 5.

Remark 2. *In previous works proving fractal based generalization bounds (Şimşekli* et al.*, 2021; Hodgkinson* et al.*, 2022) and informally summarized by Equation 4, one can see more complex total mutual information terms appearing, namely terms of the form*[2] $I_\infty(S, \mathcal{W}_{S,U})$*, where* $\mathcal{W}_{S,U}$ *is a (random) closed set as defined by learning algorithm* (5)*. What can actually make the latter smaller than a mutual information between the data S and the output of the learning algorithm, as in Equation* (15)*, is that, in* $\mathcal{W}_{S,U}$*,* no temporal information *is contained, we just consider it as a geometric set. Here is a thought experiment to better grasp this behavior: consider a learning algorithm whose trajectory is supported on the unit sphere* $\mathbb{S}^1$ *of* $\mathbb{R}^2$*, such that it is continuous and always perform at least one full revolution. In that case* $\mathcal{W}_{S,U}$ *always correspond to* $\mathbb{S}^1$ *and therefore convey no information about the data, i.e.* $I_\infty(S, \mathcal{W}_{S,U}) = 0$*, while the last point of the trajectory (the output of the learning algorithm) clearly contains information about the data.*

[2] Actually in (Şimşekli *et al.*, 2021), the total mutual information is computed between the data and the 'coverings', we simplified the presentation in this remark. The notion of random closed set giving meaning to all of this will be studied in Section 5.1.

3 Fractal geometry

The notion of fractal, as originally conceived by Mandelbrot (1982), refers to utterly complicated geometric objects which exhibit similar behavior at any scale. One of the most striking property of those objects is that some natural extensions of the notion of dimension may induce non-integer dimensions. The rigorous study of this dimension theories is the main purpose of *fractal geometry*, from which a detailed introduction can be found in (Falconer, 2014).

Dimension theory has been a successful tool to analyze dynamical systems (Pesin, 1997) and stochastic processes (Xiao, 2004). More recently, several works have used these notions in the context of generalization (Şimşekli *et al.*, 2021; Adams *et al.*, 2020; Birdal *et al.*, 2021; Hodgkinson *et al.*, 2022). In this section we will introduce some basic concepts of fractal geometry and explain how those tools can be related to generalization theory, an idea which we aim to extend in our work.

3.1 Dimension theory

3.1.1 Box-counting dimensions

Let us fix some complete Hausdorff metric space (X, d) (we consider in this subsection that $X = \mathbb{R}^N$ for some $N \in \mathbb{N}_+$) and a compactly contained set $F \subseteq X$. For any $\delta > 0$, we will define a δ-cover as being a set $N_\delta^d(F) \subseteq F$ of the centers of *closed* δ-balls covering F, namely:

$$F \subset \bigcup_{x \in N_\delta^d(F)} \bar{B}_\delta^d(x), \tag{16}$$

where the closed balls are denoted by $\bar{B}_\delta^d(x) = \{y \in X, \ d(x,y) \leq \delta\}$. A δ-cover will be called *minimal* if $|N_\delta^d|$ is minimal. We define the *covering numbers*, denoted $|N_\delta^d|$ to be the cardinal of a minimal cover. In all this subsection we assume that the set $F \subseteq X$ admits finite covers for all δ. When we apply those notions to statistical learning theory we will explain why this is indeed the case.

Remark 3. *As this work is interested in the fractal behavior of an hypothesis set $\mathcal{W}_{S,U}$, defined by Equation* (5)*, under different metrics, we explicit the dependence of the covers on the metric. It will be omitted when the metric is obvious (e.g. Euclidean).*

Remark 4. *Our definition of covering requires the centers of a covering of a set F to be in F. This will simplify our proofs and does not change the values of the box-counting dimensions defined from those covering numbers. It also introduces a few minor technical complications which are discussed in Sections 4 and 5. Those complications are solved by the following fact, immediate consequence of the triangle inequality if $G \subseteq F$, then we have $|N_{2\delta}(G)| \leq |N_\delta(F)|$.*

One of the most basic notion of dimension, called *box-counting dimension* or sometimes *Minkowski's dimension* is based on the observation that for simple geometric objects, the covering numbers scale with δ roughly as $C.\delta^{-\dim(F)}$, which leads to the following definitions.

Definition 6 (Box-counting dimensions)**.** *Consider a bounded set $F \subset X$, we define respectively the upper and lower box-counting dimensions as:*

$$\overline{\dim}_B^d(F) := \limsup_{\delta \to 0} \frac{\log |N_\delta^d(F)|}{\log(1/\delta)}, \qquad \underline{\dim}_B^d(F) := \liminf_{\delta \to 0} \frac{\log |N_\delta^d(F)|}{\log(1/\delta)}.$$

When the above limits are equal, their value will be called the box-counting dimension (or Minkowski dimension):

$$\dim_B^d(F) = \lim_{\delta \to 0} \frac{\log |N_\delta^d(F)|}{\log(1/\delta)}.$$

In particular, our main results will be stated in terms of the *upper-box counting dimension $\mathcal{W}_{S,U}$*, therefore denoted $\overline{\dim}_B(\mathcal{W}_{S,U})$ which is our main object of interest. However, other authors pointed out that under some regularity conditions the latter can be equal to some other fractal dimensions, namely the well-known *Hausdorff dimension* (Şimşekli *et al.*, 2021; Hodgkinson *et al.*, 2022), which we will describe briefly.

Remark 5. *As highlighted in (Falconer, 2014, Definition* 2.1*), we could equivalently define box-counting dimensions by using instead of closed balls the sets of diameters at most δ, moreover other equivalent definitions exist. Here we restrict ourselves to closed balls to ease notations and future proofs.*

One of the most important feature of the above dimensions is their behavior with respect to Hölder mappings:

Proposition 7 (Dimension and Hölder functional)**.** *Let (Y,ρ) be another metric space and $f : X \longrightarrow Y$ and α-Hölder continuous map with $\alpha > 0$. Then, for $F \subset X$ we have:*

$$\overline{\dim}^{\rho}_{B}(f(F)) \leq \frac{1}{\alpha}\overline{\dim}^{d}_{B}(F), \quad \textit{and:} \quad \underline{\dim}^{\rho}_{B}(f(F)) \leq \frac{1}{\alpha}\underline{\dim}^{d}_{B}(F).$$

Proof. Let us consider a minimal δ-cover of F, then we have $f(F) \subset \bigcup_{x\in N^{d}_{\delta}(F)} f(\bar{B}_{\delta}(x))$. By Hölder continuity we can write that for any $x \in N^{d}_{\delta}(F)$ and $x' \in \bar{B}_{\delta}(x)$ we have $\rho(f(x), f(x')) \leq d(x,x')^{\alpha}$. Therefore

$$f(F) \subset \bigcup_{x\in N^{d}_{\delta}(F)} f(\bar{B}_{\delta^{\alpha}}(f(x))),$$

which leads to:

$$\frac{\log |N^{\rho}_{\delta^{\alpha}}(f(F))|}{\log(1/\delta^{\alpha})} \leq \frac{1}{\alpha}\frac{\log |N^{d}_{\delta}(F)|}{\log(1/\delta)},$$

hence the result. □

Example 7. Let us consider a standard Brownian Motion $(B_t)_{t\in[0,1]}$ in $\mathbb{R}^m$ with $m \geq 2$. As a consequence of Kolmogorov continuity theorem, it is known that $t \mapsto B_t$ is almost surely α-Hölder continuous for all $\alpha \in (0, 1/2)$. As a consequence of proposition 7 and taking the limit $\alpha \to 1/2$ we get that

$$\overline{\dim}_{B}(\{B_t,\ t \in [0,1]\}) \leq 2. \tag{17}$$

Equation (17) shows that some stochastic processes may exhibit low fractal dimensions, which do not scale with the dimension of the ambient space. This phenomenon has been widely studied Xiao (2004) and is an argument in favor of the fractal based generalization theory developed in previous works (Şimşekli *et al.*, 2021) and described by Equation (4).

Let us finish this subsection by stating a few properties of box-counting dimensions which will be useful in further sections.

Proposition 8. *Let F be a set admitting finite δ-covers for d for all $\delta > 0$, as assumed earlier. We have the following properties:*

i ***Finite stability for the upper box-counting dimension:*** *If we write F as any finite union $F = \bigcup_{i=1}^{m} F_i$, we have*

$$\overline{\dim}_{B}(F) = \max_{1\leq i\leq m} \overline{\dim}_{B}(F_i).$$

ii ***Closure stability:*** *Lower and upper box-counting dimensions are stable by closure, namely:*

$$\overline{\dim}_{B}(\bar{F}) = \overline{\dim}_{B}(F), \quad \textit{and: } \underline{\dim}_{B}(\bar{F}) = \underline{\dim}_{B}(F).$$

iii ***Euclidean bound:*** *If the metric space is the Euclidean space $\mathbb{R}^D$, endowed with the usual Euclidean distance, and F is bounded, then we have*

$$0 \leq \underline{\dim}_{B}(F) \leq \overline{\dim}_{B}(F) \leq D.$$

Proof. **Finite stability:** Consider $F = \bigcup_{i=1}^{m} F_i$, then by the previous points it is obvious that $\forall i \in \{1,\ldots,m\}$, $\overline{\dim}^{d}_{B}(F_i) \leq \overline{\dim}^{d}_{B}(F)$ and therefore $\max_{1\leq i\leq m} \overline{\dim}^{d}_{B}(F_i) \leq \overline{\dim}^{d}_{B}(F)$.
Now if $\delta > 0$ and N^{i}_{δ} is a δ-cover of F_i. Let $\bar{d} := \max_{1\leq i\leq m} \overline{\dim}^{d}_{B}(F_i) \leq \overline{\dim}^{d}_{B}(F)$ and $\epsilon > 0$. By definition of the upper-limit and as we have a finite number of sets, there exists $\delta_0 > 0$ such that for

all $0 < \delta < \delta_0$ we have:

$$\forall i,\ |N_\delta^i| \leq \big(\frac{1}{\delta}\big)^{\bar{d}+\epsilon}$$

Therefore, as $\bigcup_i N_\delta^i$ is a δ-cover of F for $0 < \delta < \delta_0$:

$$\frac{\log(|N_\delta|)}{\log(1/\delta)} \leq \frac{\log(|N_\delta^1| + \cdots + |N_\delta^m|)}{\log(1/\delta)} \leq \frac{\log(\max_{1\leq i\leq m}|N_\delta^i|)}{\log(1/\delta)} + \frac{\log(m)}{\log(1/\delta)} \leq \bar{d} + \epsilon + \frac{\log(m)}{\log(1/\delta)}$$

As m is finite and ϵ is arbitrary, we get the inequality.

Closure stability: By the fact that we consider coverings by closed balls, a δ-cover of F is also a covering of its closure. Conversely, by Remark 4 we have $|N_{2\delta}(F)| \leq |N_\delta(\bar{F})|$. The result follows by taking upper and lower limits.

Euclidean bound: We use in that case an equivalent definition of the box-counting dimensions given in (Falconer, 2014, Definition 2.1) which is that we can use cubes of edge size δ. We get the result by noting that, as F is bounded, it is included in a cube of edge size B for some $B > 0$, which can be covered by $\lceil (B/\delta)^D \rceil$. Taking the logarithm gives the inequality.

□

Remark 6. *The closure stability property of box-counting dimension, displayed in Proposition 8, may be seen as a justification of our choice that the learning algorithm* (5) *generates closed hypothesis sets. Indeed, if that were not the case we could modify it by composing it with the closure operation, as we are interested in the quantity* $\overline{\dim}_B(\mathcal{W}_{S,U})$, *this would not make us loose any generality.*

3.1.2 Hausdorff dimension

The main drawback of the box-counting dimension is that there is no measure associated to it, so it tells us about the dimension of objects and not their size. The concept of Hausdorff dimension aims at defining a notion of dimension based on a generalization of the Lebesgue measure. Even though Hausdorff dimension is not our main object of interest, we mention it to relate our approach to previous works and describe how it could help improve the presented results.

Given $F \subset X$, $s \geq 0$ and $\delta > 0$ we define:

$$\mathcal{H}_\delta^s(F) := \inf \left\{ \sum_{i=1}^{+\infty} \mathrm{diam}(U_i)^s,\ F \subset \bigcup_i U_i,\ \mathrm{diam}(U_i) \leq \delta \right\}, \tag{18}$$

where $\mathrm{diam}(U_i)$ denotes the diameter of U_i. It is obvious from the definition that $\mathcal{H}_\cdot^s$ is decreasing in δ and therefore it makes sense to define:

Definition 7 (Hausdorff Measure)**.** *With the same notations as above, we define the Hausdorff measure of F as:*

$$\mathcal{H}^s(F) := \lim_{\delta\to 0} \mathcal{H}_\delta^s(F).$$

It may be proven that $\mathcal{H}^s(\cdot)$ indeed defines a measure on the Borel σ-algebra of X, under mild conditions which hold for $\mathbb{R}^n$, see (Pesin, 1997, Section 1.1*) or (Falconer, 2014, Section* 3.1*).*

It can be shown that, up to a proportionality factor, $\mathcal{H}^s$ corresponds to the s-dimensional Lebesgue measure of $\mathbb{R}^m$.

Using an argument from (Falconer, 2014), we see that if for some s we have $\mathcal{H}^s(F) < \infty$ then for any $s' > s$, taking a δ-cover (U_i) we get:

$$\sum_i \mathrm{diam}(U_i)^{s'} \leq \delta^{s'-s} \sum_i \mathrm{diam}(U_i)^s.$$

Then, by taking infimum over the covers and letting $\delta \to 0$ we have $\mathcal{H}^{s'}(F) = 0$. Therefore there exists a critical value of s for which the graph of $s \longmapsto \mathcal{H}^s(F)$ transitions from ∞ to 0. This motivates the following definition:

Definition 8 (Hausdorff dimension)**.** *The critical value* $\inf s \geq 0,\ \mathcal{H}^s(F) = 0$ *is called the Hausdorff dimension of F and will be denoted* $\dim_H^d(F)$. *If in addition this value is in* $(0, +\infty)$, *F will be called a s-set.*

Example 8. Let $(X_t)_{t\in[0,1]}$ be a Markov process in $\mathbb{R}^d$ with transition function $P(t,x,A)$. One can prove (Xiao, 2004, Theorem 4.2) that under some regularity conditions we have $\mathbb{P}^x$-almost surely that

$$\overline{\dim}_B(\{X_t,\ t\in[0,1]\}) = \sup\left\{\alpha \geq 0,\ \limsup_{r\to 0} \frac{1}{r^\alpha}\int_0^1 P(0,t,B(0,r))dt < +\infty\right\}. \tag{19}$$

Equation (19) gives a generic formula for the Hausdorff dimension of stochastic processes whose 'range' $\{X_t,\ t\in[0,1]\}$ can correspond to many learning algorithms described in Section 1.1. A wide literature study improvements of Equation (19), especially in the case of stochastic dynamics as described in Example 1, which is one of the key tools of earlier works relating generalization of learning algorithms to their fractal behavior (Şimşekli *et al.*, 2021).

Hausdorff dimension has the same behavior with respect to Hölder mappings as described in Proposition 7, however it is not stable by closure, which in particular implies that it cannot always coincide with the box-counting dimension[3]. Another notable difference between both notion is the resides in the following property of local stability:

Proposition 9 (Local stability of Hausdorff dimension)**.** *Let $(F_i)_{i\in I}$ be a countable collection of sets, we have:*

$$\dim_H\left(\bigcup_{i\in I} F_i\right) = \sup_{i\in I}\{\dim_H(F_i)\}.$$

While in general box-counting and Hausdorff dimensions are not equal, but we have the following inequalities (Falconer, 2014, Proposition 3.4):

$$\dim_{\mathrm{H}}(\cdot) \leq \underline{\dim}_{\mathrm{B}}(\cdot) \leq \overline{\dim}_B(\cdot). \tag{20}$$

Finally, we end our discussion of dimension theory by stating under which regularity condition the Hausdorff dimension and the box-counting dimension coincide.

Proposition 10 (Mattila (1999, Theorem 5.7))**.** *Let $F \subset \mathbb{R}^d$ be a bounded set and assume there exists a* finite *measure μ on the Borel σ-algebra of $\mathbb{R}^d$ such that $\mu(F) > 0$ as well as $a, b, s, r_0 > 0$ such that:*

$$\forall x \in F,\ \forall 0 < r \leq r_0,\ 0 < ar_s \leq \mu\big(B(x,r)\big) \leq br_s < +\infty.$$

Then both dimension are equal and the box-counting dimension is defined: $\dim_B(F) = \dim_H(F)$.

The intuition behind this result is that if there exists a measure on a set such that the volume of balls scale as r^s with the radius r of the ball, the 'dimension' of the set must be s.

The kind of regularity conditions described in Proposition 10 is used in previous works (Şimşekli *et al.*, 2021; Camuto *et al.*, 2021; Hodgkinson *et al.*, 2022) to assert that their bounds relating generalization error to the upper-box counting dimension of the hypothesis set $\mathcal{W}_{S,U}$ can be modified with its Hausdorff dimension, which has better properties. In this work we chose to not make any of those regularity assumptions, however this may open a future research direction.

3.2 Previous fractal-based generalization bounds

Now that we have quickly described the main tools of fractal geometry that we need, this subsection is devoted to describe how it was used in previous works to relate it to generalization error.

Motivation: This idea initially comes from the work of Şimşekli *et al.* (2021), who was in particular interested in modeling the behavior of stochastic optimization algorithm (e.g. stochastic gradient descent) with *heavy tailed dynamics* of the form:

$$\mathrm{d}W_t = -\nabla f(W_t)\mathrm{d}t + \Sigma_1(W_t)\mathrm{d}B_t + \Sigma_2(W_t)\mathrm{d}L_t^\alpha, \tag{21}$$

where W_t represents the parameters of the model, B_t is a Brownian motion and L_t^α is a 'stable Levy process' (Schilling, 2016). The precise study of such dynamics is beyond the scope of this project, however, it connects statistical learning theory to the widely studied fractal behavior of such models. In particular, it is shown in (Schilling, 1998, Theorem 4) that the Hausdorff dimension in that case may be expressed in terms of the tail properties of the process, which may also be numerically computed.

[3] Consider for example the set $\mathbb{Q}\cap[0,1]$ which has upper box-counting dimension 1 but Hausdorff dimension 0.

Remark 7. *Note that by considering $\mathcal{W}_{S,U} = \overline{\{W_t,\ t \in [0,T]\}}$, the model described by Equation* (21) *falls into the scope of the learning algorithm formalized in Section 1.1.*

We will now explain how worst-case generalization error may be related to the fractal dimensions *induced by the Euclidean distance* on the hypothesis set. Let us first consider the particular case on a fixed hypothesis set, denoted $\mathcal{W}$ and being a closed set in $\mathbb{R}^d$. The remaining of the setting and notations are exactly the same than in Section 1.

The intuition is the following: given two points $w, w' \in \mathcal{W}$ and $S \in \mathcal{Z}^n$, one can always write the following decomposition:

$$|\mathcal{R}(w) - \hat{\mathcal{R}}_S(w)| \leq |\mathcal{R}(w') - \hat{\mathcal{R}}_S(w')| + |\mathcal{R}(w) - \mathcal{R}(w')| + |\hat{\mathcal{R}}_S(w) - \hat{\mathcal{R}}_S(w')|. \tag{22}$$

Now, if we take w, w' to be in the same small ball in $\mathbb{R}^d$, a regularity on ℓ (e.g. Lipschitz continuity) would allow us to bound the last two terms of Equation (22) while the first term could be handled with concentration inequalities. This is the idea of the proof of the following theorem, first proven by Şimşekli *et al.* (2021) and which is the first result linking generalization and fractal geometry, it is here stated and proven within our notations.

Theorem 3.1. Euclidean fractal generalization bound for fixed hypothesis sets

We assume that the loss $(w, z) \mapsto \ell(w, z)$ is uniformly bounded by B ($|\ell| \leq B$) and is L-Lipschitz continuous in w uniformly with respect to z, with $B, L > 0$. Moreover we make the assumption that W is bounded.
Then, for all $\zeta \in (0, 1)$, with probability at least $1 - \zeta$ over $\mu_z^{\otimes n}$, there exists $N \in \mathbb{N}_+$ such that for all $n \geq N$ we have

$$\sup_{w \in \mathcal{W}} |\mathcal{R}(w) - \hat{\mathcal{R}}_S(w)| \leq \frac{2}{\sqrt{n}} + B\sqrt{\frac{2\overline{\dim}_B(\mathcal{W}) \log(nL^2) + 2\log(1/\zeta)}{n}}. \tag{23}$$

Proof. Let us fix some $\delta > 0$ and consider N_δ a minimal cover of $\mathcal{W} \subset \mathbb{R}^d$ for the Euclidean distance. let's recall that in Section 3.1 we have defined such a cover as a minimal set of centers of closed δ-balls covering $\mathcal{W}$.
Now let us take any $w \in \mathcal{W}$, then by definition of the coverings, there exists $w' \in N_\delta$ such that $\|w - w'\| \leq \delta$.
For any $S \in \mathcal{Z}^n$, we note that, by linearity, the functions $\mathcal{R}(\cdot)$ and $\hat{\mathcal{R}}_S(\cdot)$ are L-Lipschitz continuous. Therefore we can rewrite the decomposition of Equation (22) as:

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

Therefore we have

$$\sup_{w \in \mathcal{W}} |\mathcal{R}(w) - \hat{\mathcal{R}}_S(w)| \leq 2L\delta + \max_{w \in N_\delta} |\mathcal{R}(w) - \hat{\mathcal{R}}_S(w)|. \tag{24}$$

We can bound the first term of the right hand side of (24) by leveraging a union bound and Hoeffding's inequality; for any $\epsilon > 0$, we have:

$$\begin{aligned}
\mu_z^{\otimes n}\left(\max_{w \in N_\delta} |\mathcal{R}(w) - \hat{\mathcal{R}}_S(w)| \geq \epsilon \right) &= \mu_z^{\otimes n}\left(\bigcup_{w \in N_\delta} \left\{ \left| \mathbb{E}_z[\ell(w,z)] - \frac{1}{n}\sum_{i=1}^n \ell(w, z_i) \right| \geq \epsilon \right\} \right) \\
&\leq \sum_{w \in N_\delta} \mu_z^{\otimes n}\left(\left\{ \left| \mathbb{E}_z[\ell(w,z)] - \frac{1}{n}\sum_{i=1}^n \ell(w, z_i) \right| \geq \epsilon \right\} \right) \\
&\leq 2 \sum_{w \in N_\delta} \exp\left\{ -\frac{2\epsilon^2}{n(B/n)^2} \right\} \\
&= 2|N_\delta| \exp\left\{ -\frac{n\epsilon^2}{2B^2} \right\}.
\end{aligned} \tag{25}$$

Now let $\zeta \in (0,1)$, as we want a result true with probability at least $1-\zeta$, we make the following choice for ϵ:

$$\epsilon := B\sqrt{\frac{2\log(2|N_\delta|) + 2\log(1/\zeta)}{n}}, \tag{26}$$

so that

$$\mu_z^{\otimes n}\left(\max_{w \in N_\delta} |\mathcal{R}(w) - \hat{\mathcal{R}}_S(w)| \geq B\sqrt{\frac{2\log(2|N_\delta|) + 2\log(1/\zeta)}{n}}\right) \leq \zeta.$$

Now, using Definition 6 we have

$$\overline{\dim}_B(\mathcal{W}) = \limsup_{\delta \to 0} \frac{\log(|N_\delta|)}{\log(1/\delta)} = \limsup_{\delta \to 0} \frac{\log(2|N_\delta|)}{\log(1/\delta)},$$

where the Euclidean metric is omitted in the notation of $\overline{\dim}_B$.
Let us consider a sequence $\delta_n = 1/\sqrt{nL^2}$, as $\overline{\dim}_B(\mathcal{W})$ is a constant, we have by definition of the upper limit that there exists $N \in \mathbb{N}_+$ such that:

$$\forall n \geq N,\ \log(2|N_{\delta_n}|) \leq 2\overline{\dim}_B(\mathcal{W})\log(1/\delta_n) = \overline{\dim}_B(\mathcal{W})\log(nL^2).$$

Therefore, for all $n \geq N$ we have

$$\mu_z^{\otimes n}\left(\sup_{w \in \mathcal{W}} |\mathcal{R}(w) - \hat{\mathcal{R}}_S(w)| \geq \frac{2}{\sqrt{n}} + B\sqrt{\frac{2\overline{\dim}_B(\mathcal{W})\log(nL^2) + 2\log(1/\zeta)}{n}}\right) \leq \zeta,$$

which directly implies the result.

□

Remark 8. *Instead of the bounded loss assumption in Theorem 3.1, one could have assumed that $(w,z) \mapsto \ell(w,z)$ is sub-Gaussian in z, uniformly with respect to w, in which case we would have used the general Hoeffding's inequality 2.1 and get the same result than equation (23) but with in place of B the parameter Σ appearing in the third point of Proposition 1.*

The result presented in Theorem 23 is not satisfying enough because it does not take into account the general case of a random hypothesis $\mathcal{W}_{S,U}$ set generated by (5). As U is independent of S and plays no role in the concentration inequalities used in the proof above, only the dependence of $\mathcal{W}_{S,U}$ on S has to be handled. There are two possible solutions for this:

1. Define a notion of uniform dimension over the possible realization of the hypothesis sets, such as

$$\sup_{S \in \mathcal{Z}^n} \overline{\dim}_B(\mathcal{W}_{S,U}). \tag{27}$$

2. Directly get a bound in terms of the (random) dimension $\overline{\dim}_B(\mathcal{W}_{S,U})$.

The second solution is obviously better as it conveys more information about the data S and is a tighter bound. In our results (see Sections 4 and 5) we will try to achieve this kind of results for another type of fractal dimension. However, Şimşekli *et al.* (2021) also propose the following corollary for the first solution. It is worth mentioning because, as the upper-box counting dimension is only finitely stable, in order to for the quantity (27) to appear we need the upper box-counting dimension to be equal to the Hausdorff dimension (or any other locally stable dimension), hence requiring regularity assumptions on $\mathcal{W}$, as in the following result.

Corollary 2. *We assume that ℓ is bounded by B and L-Lipschitz continuous as in Theorem 3.1 and further assume that $\mathcal{Z}$ is countable, $\mathcal{W}_{S,U}$ is uniformly bounded and that $\bigcup_{S \in \mathcal{Z}^n} \mathcal{W}_{S,U}$ satisfies the assumptions of Proposition 10.*

Then there exists $N \in \mathbb{N}_+$ such that with probability at least $1-\zeta$ over $\mu_z^{\otimes n} \otimes \mu_u$, for all $n \geq N$ we have:

$$\sup_{w \in \mathcal{W}_{S,U}} \leq \frac{2}{\sqrt{n}} + B\sqrt{\frac{2\sup_{S \in \mathcal{Z}^n} \dim_H(\mathcal{W}_{S,U})\log(nL^2) + 2\log(1/\zeta)}{n}}.$$

Proof. We simply write:

$$\sup_{w\in\mathcal{W}_{S,U}} |\mathcal{R}(w)-\hat{\mathcal{R}}_S(w)| \leq \sup_{w\in\bigcup_{S\in\mathcal{Z}^n}\mathcal{W}_{S,U}} |\mathcal{R}(w)-\hat{\mathcal{R}}_S(w)|.$$

As U is independent of S, we can apply Theorem 3.1 along with Proposition 10 for almost any U and get that with probability 1 over U and $1-\zeta$ over S, for n big enough:

$$\begin{aligned}\sup_{w\in\mathcal{W}_{S,U}} &\leq \frac{2}{\sqrt{n}} + B\sqrt{\frac{2\overline{\dim}_B\left(\bigcup_{S\in\mathcal{Z}^n}\mathcal{W}_{S,U}\right)\log(nL^2)+2\log(1/\zeta)}{n}} \\ &\leq \frac{2}{\sqrt{n}} + B\sqrt{\frac{2\dim_{\mathrm{H}}\left(\bigcup_{S\in\mathcal{Z}^n}\mathcal{W}_{S,U}\right)\log(nL^2)+2\log(1/\zeta)}{n}} \\ &\leq \frac{2}{\sqrt{n}} + B\sqrt{\frac{2\sup_{S\in\mathcal{Z}^n}\dim_{\mathrm{H}}(\mathcal{W}_{S,U})\log(nL^2)+2\log(1/\zeta)}{n}}.\end{aligned}$$

□

Remark 9. *The assumption that $\mathcal{W}_{S,U}$ is uniformly bounded is actually not necessary and can be relaxed via an an argument based on the dominated convergence theorem, see (Şimşekli* et al.*, 2021, Section* 6.2*).*

Several works attempted to get results similar to the second proposed solution (get 'pointwise' bounds directly in terms of $\overline{\dim}_B(\mathcal{W}_{S,U})$) in different settings (Şimşekli *et al.*, 2021; Camuto *et al.*, 2021; Hodgkinson *et al.*, 2022). The main message here is that to achieve such a bound, one need to deal with the statistical dependence between the data S and the random hypothesis sets $\mathcal{W}_{S,U}$. Using decoupling technique similar to those presented in Section 2.3, those authors where able to generalize Theorem 3.1 to the case of random hypothesis sets at the cost of introducing mixing or mutual information terms, leading (informally) to results stated by Equation (4).

4 From Rademacher complexity to a data-dependent intrinsic dimension

In Section 3.2 we explain how previous works combined concentration inequalities and fractal geometry to get generalization bounds involving the fractal dimensions of the hypothesis set $\mathcal{W}_{S,U}$, but those works are not satisfying enough because of the Lipschitz assumption and the lack of data-dependence of the introduced dimensions.

In this section, we will start putting together the tools described in Sections 2.1 and 3 to prove generalization bounds involving another type of dimension. First we explain how a classical covering argument on Rademacher complexity leads to the definition of a data-dependent pseudo-metric on the parameter space. We then quickly study the behavior of the box-counting dimensions induced by this pseudo-metrics before proving a generalization bound for fix hypothesis sets involving this *data-dependent* dimension.

We do not make any Lipschitz continuity assumption. However, from now on we make Assumption 1, namely $\ell(\cdot,\cdot)$ is continuous and uniformly bounded by $B > 0$.

4.1 What type of pseudo-metric can we use?

As we already mentioned, our goal is to introduce the upper box-counting dimension induced by a data-dependent metric on the parameter space. A natural question arises of what type of metric could be pertinent.

Looking at the kind of decomposition presented in Equation (22), one could think that using as a distance $(w, w') \mapsto |\mathcal{R}(w) - \hat{\mathcal{R}}_S(w) - \mathcal{R}(w') + \hat{\mathcal{R}}_S(w')|$ could work, as it allows to bound the terms that were controlled by the Lipschitz assumption in the proof of Theorem 3.1. However, this would mean that we aim to bound $\sup_w(\mathcal{R}(w) - \hat{\mathcal{R}}_S(w))$ with the box-counting dimension induced by this pseudo-metric, hence bounding the diameter of a set in terms of its dimension, which does not seem very meaningful.

Moreover we require the two following conditions on the pseudo-metric we choose:

1. The corresponding fractal dimensions must take non-trivial values ranging in a sufficiently large interval, typically if we choose $(w, w') \mapsto |\hat{\mathcal{R}}_S(w) - \hat{\mathcal{R}}_S(w')|$ the dimension will always be smaller than 1, because it would be the dimension of $\hat{\mathcal{R}}_S(\mathcal{W}_{s,U})$.
2. We want it to be directly numerically estimable from the data S, i.e. we do not want any expectation involved in the definition of the dimension.

Our choice of pseudo metric is inspired by a classical covering argument on Rademacher complexity (Rebeschini, 2020); similar techniques are also used in a more general context related to concentration inequalities (Boucheron *et al.*, 2013, Section 13).

This covering argument is as follows, let us introduce the following random pseudo-metric:

$$\forall S \in \mathcal{Z}^n,\ \forall w, w' \in \mathbb{R}^d,\ \rho_S(w, w') := \frac{1}{n}\sum_{i=1}^n |\ell(w, z_i) - \ell(w', z_i)|. \tag{28}$$

Now let us take $S \in \mathcal{Z}^n$, consider a fixed hypothesis set $\mathcal{W}$ and introduce a minimal δ-covering[4], denoted $N_\delta^{\rho_S}(\mathcal{W})$ in accordance with Section 3.1. We further introduce some Rademacher random variables $(\sigma_1, \dots, \sigma_n)$ and take $w' \in N_\delta^{\rho_S}(\mathcal{W})$ $w \in \bar{B}_\delta^{\rho_S}(w')$, we have, by the triangle inequality:

$$\frac{1}{n}\sum_{i=1}^n \sigma_i \ell(w, z_i) \leq \frac{1}{n}\sum_{i=1}^n \sigma_i \ell(w, z_i) + \rho_S(w, w') \leq \frac{1}{n}\sum_{i=1}^n \sigma_i \ell(w, z_i) + \delta.$$

Therefore:

$$\begin{aligned}\mathbf{Rad}(\ell(\mathcal{W}, S)) &:= \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{w \in \mathcal{W}} \frac{1}{n}\sum_{i=1}^n \sigma_i \ell(w, z_i)\right] \\ &\leq \delta + \mathbb{E}_{\boldsymbol{\sigma}}\left[\max_{w \in N_\delta^{\rho_S}(\mathcal{W})} \frac{1}{n}\sum_{i=1}^n \sigma_i \ell(w, z_i)\right].\end{aligned} \tag{29}$$

[4] Keep in mind that in Section 3 we defined the coverings $N_\delta^d(F)$ as the set of *centers* of the balls covering F.

From Equation (29) we make the following observations:

1. The last term of equation (29) is a Rademacher complexity over a finite set and can therefore be handled using concentration techniques such as Massart's Lemma 3.

2. This will make a random term of the form $\log(|N_\delta^{\rho_S}(\mathcal{W})|)$ appear in the bound, thus we need to make so that it is well defined and control the dependence on S of the limit defining the upper box-counting dimension.

3. Most importantly, **no Lipschitz continuity assumption** is needed to make this argument work, this is why, in the following, we will aim at extending this technique.

We will extend this covering argument in Section 4.3. Before coming to that, one should note that we are now considering covering numbers in random pseudo-metric spaces, therefore we need to quickly study the properties of the corresponding fractal dimensions.

4.2 Data-dependent fractal dimensions

In this subsection we study the properties of the fractal dimensions induced on subsets of $\mathbb{R}^d$ by the pseudo-metrics ρ_S defined by Equation (28). We immediately see that with those distances it is indeed possible to get $\rho_S(w, w') = 0$ while $w \neq w'$ (for example due to over-parameterization or the internal symmetries of a neural network), so we need to extend the theory to this framework and prove some properties which we will need afterwards.

This analysis also serves as a theoretical justification of our experimental work, presented in Section 6. Indeed, those experiments are based on results relating the upper box-counting dimension to topological data analysis (Kozma *et al.*, 2005; Schweinhart, 2019) *in metric spaces*, in order to extend prior works (Adams *et al.*, 2020; Birdal *et al.*, 2021) to numerically estimate the fractal dimension. Thus, by extending those results to pseudo-metric spaces we make possible the use of similar tools in our case.

Let us start by very simple remarks in generic pseudo-metric spaces.

Definition 9 (Pseudo-metrics and associated quantities)**.** *A pseudo-metric on set X, is an application $\rho : X \times X \longrightarrow \mathbb{R}_+$ which is null on the diagonal ($\rho(x,x) = 0$), symmetric and verify the triangle inequality. The previous definition of box-counting dimensions are naturally extended to this setting, provided that finite covering numbers are defined.*

Let us quickly assert that the main properties of the box-counting dimensions remain true with pseudo-metrics. Indeed, they do not use the fact that the metric actually separates distinct points.

Proposition 11 (Pseudo-metrics based box-counting dimension)**.** *Let F be a set admitting finite δ-covers for d for all $\delta > 0$. All statements of proposition 8 remain valid if d is a pseudo metric.*

Actually, we will show that the properties of box-counting dimension on a pseudo-metric space can actually be understood in terms of its *metric identification.*

Definition 10 (metric identification)**.** *Let (X, ρ) be a pseudo-metric space, we introduce the equivalence relation:*

$$\forall x, y \in X,\ x \sim y \iff \rho(x,y) = 0.$$

We call metric identification of X the quotient of X by this equivalence relation. The canonical projection on the quotient will be denoted as:

$$\pi : X \longrightarrow X/\sim .$$

Clearly, ρ induces a metric on $X/\sim$ that we will denote $\rho^\star := \pi_\star \rho$.

We prove that upper box-counting dimension invariant by this identification operation. Let us recall that we always consider the covers are made from closed δ-balls, even though equivalent definitions exist.

Lemma 2 (Upper-box dimension with pseudo metric)**.** *With the same notations as above, we have*

$$\overline{\dim}_B(X) = \overline{\dim}_B(X/\sim). \tag{30}$$

Proof. Let $F \subset X$, bounded. Let $\{x_1, \dots, x_n\}$ be the centers of a closed δ-balls covering of F for metric ρ. We have:

$$\forall x, x' \in B(x_i, \delta),\ \rho^\star(\pi(x), \pi(y')) = \rho(x, x') \leq \delta.$$

Therefore $\pi(B(x_i, \delta)) \subset B(\pi(x_i), \delta)$, therefore $N_\delta^\rho(F) \geq N_\delta^{\rho^\star}(\pi(F))$.
On the other hand, if $\{y_1, ..., y_n\}$ are the centers of a covering of $\bar{F} \subset X/\sim$, a similar reasoning shows that the $\pi^{-1}(B(y_i, \delta))$ give a covering of $\pi^{-1}(F)$ with (set included in) δ-balls. □

Let us now note that pseudo-metric (28) may be written as $\rho_S(w, w') = \|F_S(w) - F_s(w')\|_1$, where F_S is a random embedding defined for every $S = (z_1, \dots, z_n) \in \mathcal{Z}^n$:

$$F_S : \mathbb{R}^d \ni w \longmapsto \frac{1}{n}(\ell(w, z_i))_{1 \leq i \leq n} \in \mathbb{R}^n. \tag{31}$$

Therefore, it is worth studying the box-counting dimensions induced by pseudo-metrics having this form, which we do in the following proposition, which states that those embedding preserve the dimensions.

Proposition 12 (Covering numbers for embedding-based pseudo-metrics)**.** *Let $n \geq 1$, X be some non empty set and $L : X \longrightarrow \mathbb{R}^n$ be a function. Let us consider a norm $\| \cdot \|$ on $\mathbb{R}^n$ and denote by e the corresponding metric.*

We define on X the pseudo metric $\rho(x, y) := \|L(x) - L(y)\|$. Then for every $F \subset X$ such that $L(F)$ is bounded, the covering numbers $N_\delta^\rho(F)$ and $N_\delta^{\|\cdot\|}(L(F))$ are the same. In particular:

$$\overline{\dim}_B^\rho(F) = \overline{\dim}_B^e(L(F)).$$

Proof. Let F be as in the proposition. All the coverings mentioned hereafter exist because $L(F)$ has a compact closure in $\mathbb{R}^d$.

- Let N_δ^ρ be a closed δ-balls cover of F, then obviously:

$$L(F) \subset \bigcup_{x \in N_\delta^\rho} L(\bar{B}_\delta^\rho(x)),$$

and for all $x \in N_\delta^\rho$ and $y, y' \in L(\bar{B}_\delta\rho(x))$, there exist $x, x' \in X$ such that $y = L(x)$ and $y' = L(x')$, therefore $\|y - y'\| = \rho(x, x') \leq \delta$. Hence $N_\delta^{\|\cdot\|}(L(F)) \leq N_\delta^\rho(F)$.

- Let $N_\delta^{\|\cdot\|}$ be a closed δ-balls cover of $L(F)$, then clearly:

$$F \subset \bigcup_{y \in N_\delta^{\|\cdot\|}} L^{-1}(\bar{B}_\delta^{\|\cdot\|}(y)).$$

For each $y \in N_\delta^{\|\cdot\|}$, as $y \in L(F)$ we can chose $x \in F$ such that $x = L(y)$. By the same computation as the previous point we have that $L^{-1}(\bar{B}_\delta^{\|\cdot\|}(y)) \subseteq \bar{B}_\delta^\rho(x)$ and therefore $N_\delta^{\|\cdot\|}(L(F)) \geq N_\delta^\rho(F)$.

□

The two previous results can be summarized by the commutation of the following diagram.

$$\begin{tikzcd}
X \arrow[r, "\pi"] \arrow[d, "L"'] \arrow[dr, "\overline{\dim}_B^\rho"] & X/\sim \arrow[d, "\overline{\dim}_B^{\rho^\star}"] \\
\mathbb{R}^n \arrow[r, "\overline{\dim}_B^e"'] & \mathbb{R}_+
\end{tikzcd}$$

Remark 10. *As we have by Assumption 1 that the loss ℓ is bounded, so are the embeddings F_S given by* (31)*. Therefore, their image $F_S(W)$ for* any *$W \in \mathbb{R}^d$ are compactly contained in $\mathbb{R}^n$, which makes the covering numbers $N_\delta^{\rho_S}(\cdot)$, and therefore the upper box-counting dimension, always well-defined. This property shows that the boundedness assumption is unfortunately necessary to make our bounds defined* *and non-vacuous. This is one drawback of our work compared to some others like (Şimşekli* et al.*, 2021; Birdal* et al.*, 2021) who only require a sub-Gaussian assumption on the loss. However, in the aforementioned works, the hypothesis set $\mathcal{W}_{S,U}$ needs to be bounded to make its upper box-counting dimension finite, which is not required anymore in our work.*

Accordingly, we now always make Assumptions 1, so that the considered δ-covers are finite.

Now let us turn our attention to the particular case of pseudo-metric ρ_S defined by Equation (28). Due to the fact that we imposed, in Section 3.1, that the coverings f a set are included in this set, we have to specify the closure properties of coverings in our case, to give meaning to the measurability results we want to prove. This is done in the following lemma:

Lemma 3 (Closure property of coverings)**.** *Let W be a closed set and $\mathcal{C}$ be a countable dense subset of W. Under Assumption 1 we have that any covering of $\mathcal{C}$ is a covering of W for pseudo-metric ρ_S. Moreover we have, for all $\delta > 0$:*

$$|N_{2\delta}^{\rho_S}(\mathcal{C})| \leq |N_\delta^{\rho_S}(W)| \leq |N_\delta^{\rho_S}(\mathcal{C})|. \tag{32}$$

Proof. Let us consider some $\delta > 0$, a minimal δ-cover $\{c_1, \dots, c_K\}$ of $\mathcal{C}$ and $w \in W$. By density, there exists a sequence $(\xi_n)_n$ in $\mathcal{C}$ such that $\xi_n \to w$. As $\{c_1, \dots, c_K\}$ is a finite cover of $\mathcal{C}$, we can assume without loss of generality that, for all n, $\xi_n \in \bar{B}_\delta^{\rho_S}(c_i)$ for some i. Therefore, by continuity we have $\rho_S(w, c_i) = \lim_{n\to\infty} \rho_S(w, \xi_n) \leq \delta$. Thus:

$$|N_\delta^{\rho_S}(W)| \leq |N_\delta^{\rho_S}(\mathcal{C})|.$$

Now, by Remark 4 (which is valid because it only requires the triangle inequality) we have:

$$|N_{2\delta}^{\rho_S}(\mathcal{C})| \leq |N_\delta^{\rho_S}(W)|$$

□

Remark 11. *Lemma 3 states that we can always construct minimal coverings as being minimal coverings of dense countable subsets, this will always yields the same box-counting dimensions. Therefore, for now, we will always consider coverings which are minimal coverings of a dense countable subset. This will allow us to construct coverings of he considered sets in a measurable way, while not affecting the dimensions. This consideration may be implicit in several proofs.*

Now we finish this subsection with an important result asserting that the covering numbers we get from pseudo-metric ρ_S are *measurable* with respect to $S \in \mathcal{Z}^n$. Indeed, this is essential to ensure the fact that the upper box-counting dimension $\overline{\dim}_B^{\rho_S}$ induced by ρ_S is a well-defined random variable, which is require for a high probability bound of the form given by Equation (6) to make sense. This kind of measurability conditions are often assumed by authors dealing with potentially random covering numbers (Şimşekli *et al.*, 2021; Camuto *et al.*, 2021). In our case, we can prove this measurability under some condition.

Recall that in Section 1.1 we required that $\mathcal{Z}$ has a metric space structure, typically inherited by an inclusion in an Euclidean space $\mathbb{R}_N$ and that its σ-algebra is the corresponding Borel σ-algebra. With that in mind we prove the following theorem:

Theorem 4.1. Measurability of covering numbers in the case of fixed hypothesis set

Let $\mathcal{W}$ be a closed set, $\mathcal{C}$ be a dense countable subset[a] of $\mathcal{W}$ and $\delta > 0$. Under Assumption 1, we have that the mapping between probability spaces

$$(\mathcal{Z}^n,\ \mathcal{F}^{\otimes n}) \ni S \longmapsto |N_\delta^{\rho_S}(\mathcal{C})| \in (\mathbb{N}_+,\ \mathcal{P}(\mathbb{N}_+)),$$

is a random variable, where $\mathcal{P}(A)$ denotes the subsets of a set A.

[a] It always exists for any closed set in $\mathbb{R}^d$.

Proof. For any set X let us denote by $\mathfrak{F}_{\leq k}(X)$ the set of finite subsets of X with at most k elements. We start by noting that thanks to the continuous loss assumption, we have that $S \mapsto \rho_S(w, w')$ is

continuous for any $w, w' \in \mathbb{R}^d$. Moreover, let us denote $\mathcal{C} := \{w_k,\ k \in \mathbb{N}\}$.
Thus, to show the measurability condition, it suffices to show that for any $M \in \mathbb{N}_+$ we have: $\{S \in \mathcal{Z}^n,\ |N_\delta^{\rho_S}(\mathcal{C})| \leq M\} \in \mathcal{F}^{\otimes n}$. we can write

$$|N_\delta^{\rho_S}(\mathcal{C})| \leq M \iff \exists F \in \mathfrak{F}_{\leq M}(\mathcal{C}),\ \forall k \in \mathbb{N},\ \mathcal{C} \subset \bigcup_{c \in F} \bar{B}_\delta^{\rho_S}(c).$$

Therefore

$$\{S \in \mathcal{Z}^n,\ |N_\delta^{\rho_S}(\mathcal{C})| \leq M\} = \bigcup_{F \in \mathfrak{F}_{\leq M}(\mathcal{C})} \bigcap_{k \in \mathbb{N}} \bigcup_{c \in F} \{S,\ \rho_S(c, w_k) \leq \delta\}. \tag{33}$$

By continuity, it is clear that $\{S,\ \rho_S(c, w_k) \leq \delta\} \in \mathcal{F}^{\otimes n}$, hence we have the result by countable unions and intersections.

□

Remark 12. *Given any positive sequence δ_k, decreasing and converging to 0, thanks to Lemma 3 the upper box-counting dimension can be written as*

$$\overline{\dim}_B^{\rho_S}(\mathcal{W}) = \limsup_{k \to +\infty} \frac{\log |N_{\delta_k}^{\rho_S}(\mathcal{C})|}{\log(1/\delta_k)}, \tag{34}$$

which, by Theorem 4.1, implies that $\overline{\dim}_B^{\rho_S}(\mathcal{W})$ is a random variable as countable upper limit of random variables.

4.3 First result for a fixed hypothesis space

Thanks to the results of the last two subsections, we now have almost everything we need to prove our first main result.

We need one last ingredient, which is necessary to control the dependence in $S \in \mathcal{Z}^n$ of the limit (34). This is achieved by the following theorem from Egoroff. The use of this theorem in our work is inspired by previous studies Şimşekli *et al.* (2021); Camuto *et al.* (2021), in which it is a key element of the proofs, for similar reasons than the one we face.

Essentially, Egoroff's theorem states that the convergence of a sequence of measurable functions in a finite measure space can be made uniform with probability arbitrarily large.

Theorem 4.2. Egoroff's theorem Bogachev (2007)

Let $(\Omega, \mathcal{F}, \mu)$ be a measurable space with μ a positive finite measure. Let $f_n, f : \Omega \longrightarrow (X, d)$ be functions with values in a separable metric space X and such that μ-almost everywhere $f_n(x) \to f(x)$.
Then for all $\gamma > 0$ there exists $\Omega_\gamma \in \mathcal{F}$ such that $\mu(\Omega \backslash \Omega_\gamma) \leq \gamma$ and on Ω_γ the convergence of (f_n) to f is uniform.

Proof. From the simple convergence we have that

$$\forall x \in \Omega,\ \forall \epsilon > 0,\ \exists m \geq 1,\ \forall n \geq m,\ d(f_n(x), f(x)) \leq \epsilon.$$

Therefore let us introduce:

$$A_{m,\epsilon} := \{x \in \Omega,\ \forall n \geq m,\ d(f_n(x), f(x)) \leq \epsilon\}.$$

Clearly from the assumptions we have that $A_{m,\epsilon} \in \mathcal{F}$ and by the dominated converge theorem we have:

$$\mu(A_{m,\epsilon}) \underset{m \to \infty}{\longrightarrow} \mu(\Omega) < +\infty.$$

Let $(\epsilon_k)_k$ be a decreasing sequence of positive numbers converging to 0. We deduce that for every $k \geq 1$ there exists $m_k \geq 1$ such that:

$$\mu(A_{m_k, \epsilon_k}) \geq \mu(\Omega) - \frac{\gamma}{2^k}.$$

Therefore:

$$\mu\Big(\bigcap_{k=1}^{\infty} A_{m_k,\epsilon_k}\Big) \geq \mu(\Omega) - \sum_{k=1}^{\infty} \frac{\gamma}{2^k} = \mu(\Omega) - \gamma.$$

We conclude by noting that, by construction, the convergence is uniform on the set $\bigcap_{k=1}^{\infty} A_{m_k,\epsilon_k}$. □

The next theorem is our extension of the classical covering bounds of Rademacher complexity (Barlett and Mendelson, 2002; Rebeschini, 2020) presented in Section 4.1.

Theorem 4.3. Data-dependent fractal generalization bound for fixed hypothesis set

Let us consider a fixed closed hypothesis set $\mathcal{W}$, under Assumption 1 we have the following: For all $\epsilon, \gamma, \eta > 0$ and $n \in \mathbb{N}_+$ there exists $\delta_{n,\gamma,\epsilon} > 0$ such that with probability at least $1 - 2\eta - \gamma$ under $S \sim \mu_z^{\otimes n}$, for all $\delta < \delta_{n,\gamma,\epsilon}$ we have:

$$\sup_{w \in \mathcal{W}} \big(\mathcal{R}(w) - \hat{\mathcal{R}}_S(w)\big) \leq 2B\sqrt{\frac{4(\overline{\dim}_B^{\rho_S}(\mathcal{W}) + \epsilon)\log(1/\delta) + 9\log(1/\eta)}{n}} + 2\delta.$$

Proof. **Step 0:** First of all, as $\mathcal{W}$ is closed, we can consider a dense countable subset $\mathcal{C}^a$. Thanks to the boundedness assumption, we can find finite coverings N_r for each value of $r > 0$. The notation N_r refers in this proof to the set of the centers of a covering of $\mathcal{C}$ by closed r-balls under the pseudo-metric ρ_S. Invoking results from Lemma 3, those set N_r are also δ-coverings of $\mathcal{W}$ and induce the upper box-counting dimension of $\mathcal{W}$ under ρ_S, so that considering them does not change the dimension.
Step 1: Let us set:

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

Invoking proposition 2 we have:

$$G(S) \leq 2\mathbf{Rad}(\ell(\mathcal{W}, S)) + 3\sqrt{\frac{2B^2}{n}\log(1/\eta)}. \tag{35}$$

Step 2:
Therefore we have everywhere for $S \in \mathcal{Z}^n$:

$$\overline{\dim}_B^{\rho_S}(\mathcal{W}) := \limsup_{r \to 0} \frac{\log(|N_r|)}{\log(1/r)}. \tag{36}$$

Thanks to Theorem 4.1 we have that $\log(|N_r|)$ is a random variable. Let us consider an arbitrary positive sequence r_k decreasing and converging to 0. We have:

$$\overline{\dim}_B^{\rho_S}(\mathcal{W}) := \limsup_{k \to \infty} \frac{\log(|N_{r_k}|)}{\log(1/r_k)}. \tag{37}$$

Let $\gamma > 0$, by Egoroff's Theorem 4.2 there exist a set Ω_γ such that $\mu_z^{\otimes n}(\Omega_\gamma) \geq 1 - \gamma$, on which the above convergence is uniform. Therefore, if we fix $\epsilon > 0$, we have that there exists $K \in \mathbb{N}$ such that

$$\forall S \in \Omega_\gamma,\ \forall k \geq K,\ \sup_{0<\delta<r_k} \frac{\log(|N_\delta|)}{\log(1/\delta)} \leq \epsilon + \overline{\dim}_B^{\rho_S}(\mathcal{W}).$$

Now, setting $\delta_{n,\gamma,\epsilon} := r_K$, we have that on Ω_γ:

$$\forall \delta \leq \delta_{n,\gamma,\epsilon},\ \log(|N_\delta|) \leq (\epsilon + \overline{\dim}_B^{\rho_S}(\mathcal{W}))\log(1/\delta). \tag{38}$$

Now let us fix $S \in \Omega_\gamma$ and the associated cover N_r, for (σ_i) Rademacher random variables independent of S and N_r, taking two points w, w' such that $\rho_S(w, w') \leq r$ we can use the triangle inequality and write:

$$\frac{1}{n}\sum_{i=1}^{n} \sigma_i \ell(w, z_i) \leq r + \frac{1}{n}\sum_{i=1}^{n} \sigma_i \ell(w', z_i).$$

Therefore we have:

$$\mathbf{Rad}(\ell(\mathcal{W},S)) \leq r + \mathbb{E}_\sigma\left[\max_{w\in N_r}\frac{1}{n}\sigma^T\ell(w,S)\right].$$

As the Rademacher random variables are independent of the other random variables we have by Massart's lemma (lemma 3):

$$\mathbf{Rad}(\ell(\mathcal{W},S)) \leq r + B\sqrt{\frac{2\log(|N_r|)}{n}}.$$

Therefore if we take $\delta \leq \delta_\gamma$ we get that with probability at least $1-\gamma$:

$$\mathbf{Rad}(\ell(\mathcal{W},S)) \leq \delta + \mathbb{E}_\sigma\left[\max_{w\in N_r}\frac{1}{n}\sigma^T\ell(w,S)\right] \leq \delta + B\sqrt{\frac{2\log(1/\delta)}{n}(\epsilon + \overline{\dim}_B^{\rho_S}(\mathcal{W}))}. \tag{39}$$

Putting together equations 35 and 39 we get that with probability at least $1-2\eta-\gamma$, for $\delta \leq \delta_{n,\gamma,\epsilon}$:

$$G(S) \leq 2\delta + 2B\sqrt{\frac{4(\epsilon + d(S))\log(1/\delta) + 9\log(1/\eta)}{n}}. \tag{40}$$

□

[a] the fact that we cover a dense countable subset and not directly $\mathcal{W}$ here is just made to invoke the measurability result of Theorem 4.1, it does not change anything to the proof.

Theorem 4.3 is therefore similar to (Şimşekli *et al.*, 2021, Theorem 1) (Theorem 3.1 in this work), which used a fractal dimension based on the Euclidean distance on $\mathbb{R}^d$, $\|w-w'\|_2$ and a fixed hypothesis space. One of the main improvement here is in the absence of Lipschitz assumption.

However, Theorem 4.3 might not be sufficiently satisfying. The proof involves techniques that do not hold in the case of random hypothesis spaces, an issue which we address in the next subsection.

Let us make a *very important remark* about Theorem 4.3.

Remark 13. *As we can see in the above statement, the bound is not asymptotic in n (as it was in (Şimşekli* et al.*, 2021)) but is asymptotic in δ, moreover this asymptoticity in δ* ***depends on*** *n. Another way of writing it would be that there exists a sequence of positive numbers (δ_n) which is decreasing and convergent to 0 (depending on the fixed parameter γ and ϵ) such that for all n we have:*

$$\sup_{w\in\mathcal{W}}\left(\mathcal{R}(w) - \hat{\mathcal{R}}_S(w)\right) \leq 2B\sqrt{\frac{4(\overline{\dim}_B^{\rho_S}(\mathcal{W}) + \epsilon)\log(1/\delta_n) + 9\log(1/\eta)}{n}} + 2\delta_n.$$

And therefore we have two terms appearing in this kind of bound:

- *The fractal dimension, here represented by $\overline{\dim}_B^{\rho_S}$.*
- *The convergence rate of the limit defining this dimension, represented by δ_n.*

Note, however, that if we could assume that the convergence of the limit defining the random upper box-counting dimension, which is Equation (37) in the above proof, is ***uniform*** *in n. Then we could actually get that, by taking $\delta_n = 1/\sqrt{n}$, for n big enough we have with probability $1-2\eta-\gamma$ that:*

$$\sup_{w\in\mathcal{W}}\left(\mathcal{R}(w) - \hat{\mathcal{R}}_S(w)\right) \leq 2B\sqrt{\frac{2(\overline{\dim}_B^{\rho_S}(\mathcal{W}) + \epsilon)\log(n) + 9\log(1/\eta)}{n}} + \frac{2}{\sqrt{n}}.$$

Therefore we would get a bound even more resembling Theorem 3.1, at the cost of asymptoticity in n. Getting this uniformity is a research direction for future works.

5 General case

Theorem 4.3 is interesting because it gives a bound similar to (Şimşekli *et al.*, 2021) in the case of a fixed hypothesis set but with a new notion of data dependent intrinsic dimension. Now we come to the case where the hypothesis set $\mathcal{W}_{S,U}$ generated by the learning algorithm (5) is a random set, more precisely we want to prove a worst-case generalization bound over $\mathcal{W}_{S,U}$ in terms of the random upper box-counting dimension $\overline{\dim}_B^{\rho_S}(\mathcal{W}_{S,U})$.

However, this new general case presents several new difficulties, compared to the bound with fixed hypothesis set presented in Section 4.3. The main difficulties are the following:

- The hypothesis set $\mathcal{W}_{S,U}$ is now a 'random closed set'; to give meaning to our derivations and prove our results, we will formalize it rigorously using the theory developed in (Molchanov, 2017). This will allow us to refine our definition of learning algorithm (5), discuss a new measurability assumption on the coverings and study some nice consequences of this assumption.

- The symmetrization technique used to link Rademacher complexity to the worst-case generalization error in the proof of proposition 2 does not work anymore with random hypothesis sets. To overcome this issue we introduce a technique of 'approximate level sets of the risk population'. While this allows us to have our proof work, it also introduces very intricate technical terms, which will be discussed.

- Now that the hypothesis set is random, we have to deal with the statistical dependence between $\mathcal{W}_{S,U}$ and S, which is a similar issue than one faced in previous works (Şimşekli *et al.*, 2021; Camuto *et al.*, 2021; Hodgkinson *et al.*, 2022). This will involve information theoretic quantities which have been defined in Section 2.3.

As a reminder, we still assume in all the following that Assumption 1 holds.

5.1 Random sets formalization

In this subsection, we prove rather technical results stating that we can construct coverings, covering numbers and supremum over random sets in a measurable way, in an attempt to make the project more rigorous. The reader not interested in those technical developments may go directly to Subsection 5.2 and assume that everything is measurable.

The worst-case generalization error, in the general setting of a learning algorithm presented in Section 1.1, now takes the following form:

$$\mathcal{G}(S,U) := \sup_{\mathcal{W}_{S,U}} \big(\mathcal{R}(w) - \hat{\mathcal{R}}_S(w)\big). \tag{41}$$

Where, for now, we only required $\mathcal{W}_{S,U} \subset \mathbb{R}^d$ to be a closed set. However, this hypothesis set is also random because of its dependence in S and U. In this subsection, we will make this precise by describing basic notions of random set theory and prove a few technical results which will lay the ground of a rigorous theoretical basis for our main results. The interested reader can consult (Kechris, 1995; Molchanov, 2017). Other works mentioned similar formulation of the problem (Hodgkinson *et al.*, 2022), though with not much technical details.

Let us fix a probability space $(\Omega, \mathcal{T}, \mathbb{P})$ and denote $X = \mathbb{R}^d$.

Remark 14. *As highlighted by (Molchanov, 2017), we can develop the following theory in the more general case where X is a locally compact Hausdorff second countable space, but we avoid those technical considerations.*

Example 9. In our setting, described in Section 1, the underlying probability space is $(\mathcal{Z}^n \times \Omega_U, \mathcal{F}^{\otimes n} \otimes \mathcal{F}_U, \mu_z^{\otimes n} \otimes \mu_u)$.

The definition of a random closed set is the following:

Definition 11 (Random closed set)**.** *Consider a map $W : \Omega \longrightarrow \boldsymbol{CL}(E)$, W is said to be a random closed set if for every compact set $K \subset E$ we have:*

$$\{\omega, W(\omega) \cap K \neq \emptyset\} \in \mathcal{T}.$$

A natural question is to know whether we can cast it as a random variable defined in the usual way, the answer is yes and is formalized by the following definition.

Definition 12 (Effrös σ-algebra and Fell topology)**.** *The Effrös σ-algebra is the one generated by the sets $\{W \in \boldsymbol{CL}(E), W \cap K \neq \emptyset\}$ for K going over all compact sets in $\mathbb{R}^d$.*

The Fell topology *on $\boldsymbol{CL}(E)$ is the one generated by open sets $\{W \in \boldsymbol{CL}(E), W \cap K \neq \emptyset\}$ for K going over all compact sets and $\{W \in \boldsymbol{CL}(E), W \cap \mathcal{O} \neq \emptyset\}$ for $\mathcal{O}$ going over all open sets of $\mathbb{R}^d$.*

One can show that the Effrös σ-algebra on $\boldsymbol{CL}(E)$ corresponds to the Borel σ-algebra induced by the Fell topology (Molchanov, 2017, Chapter 1.1). The Effrös σ-algebra will be denoted by $\mathfrak{E}(E)$.

It can be shown that Definition 11 is equivalent to asking the measurability of W with respect to $\mathfrak{E}(E)$.

Let us now refine our definition of learning algorithm with the following assumption:

Assumption 2. *We assume that $\mathcal{W}_{S,U}$ is a random closed set in the sense of the above definition. It means that the mapping defining the learning algorithm:*

$$\mathcal{A} : \bigcup_{n=0}^{+\infty} \mathcal{Z}^n \times \Omega_U \longrightarrow \boldsymbol{CL}(\mathbb{R}^d),$$

is measurable with respect to the Effrös σ-algebra.

Thanks to this definition, we can already state one particularly useful result:

Proposition 13 (Theorem 1.3.28 in Molchanov (2017))**.** *Consider $(G_w)_{w\in E}$ a $\mathbb{R}$-valued, almost surely continuous, stochastic process on $E = \mathbb{R}^d$ and W a random closed set in E. Then the mapping*

$$\Omega \ni \omega \longmapsto \sup_{w\in W(\omega)} G_w(\omega)$$

is a random variable.

Example 10. If we define $\mathcal{R}(w) - \hat{\mathcal{R}}_S(w)$ and $W = \mathcal{W}_{S,U}$, then thanks to the continuity of the loss (Assumption 1) we have that the worst case generalization error defined by Equation (41) is a well-defined random variable.

While Example 10 gives us useful information, it is actually not enough for some arguments of our proofs to hold. In particular, to deal with the statistical dependence between the data and the random hypothesis set, we want to be able to perform the following operation: Given a random closed set W and $S \in \mathcal{Z}^n$ we want to apply Proposition 4 and write:

$$\mathbb{P}_{W,S}\left(\sup_{w\in W} \mathcal{R}(w) - \hat{\mathcal{R}}_S(w) \geq \epsilon\right) \leq e^{I_\infty(W,S)} \mathbb{P}_W \otimes \mathbb{P}_S\left(\sup_{w\in W} \mathcal{R}(w) - \hat{\mathcal{R}}_S(w) \geq \epsilon\right). \tag{42}$$

In order for Equation (42) to hold, we actually need the measurability of the mapping

$$\mathbf{CL}(\mathbb{R}^d) \times \mathcal{Z}^n \ni (W,S) \longmapsto \sup_{w\in W} |\mathcal{R}(w) - \hat{\mathcal{R}}_S(w)|,$$

with respect to $\mathfrak{E}(\mathbb{R}^d) \otimes \mathcal{F}^{\otimes n}$.

We show two results in this direction, the first one assuming that the data space $\mathcal{Z}$ is countable[5].

Lemma 4. *As before, let $(\boldsymbol{CL}(\mathbb{R}^d), \mathfrak{E}(\mathbb{R}^d)$ denotes the closed sets of $\mathbb{R}^d$ endowed with the Effrös σ-algebra, $(\Omega, \mathcal{T})$ be a countable measurable space (with $\mathcal{T} = \mathcal{P}(\Omega)$) and $\zeta(x,\omega)$ be an almost surely continuous stochastic process on $\mathbb{R}^d$. Then the function*

$$f : \boldsymbol{CL}(\mathbb{R}^d) \times \Omega \ni (W,\omega) \longmapsto \sup_{x\in W} \zeta(x,\omega) \in \mathbb{R}$$

is measurable with respect to $\mathfrak{E}(\mathbb{R}^d) \otimes \mathcal{T}$.

[5] This countability assumption on the dataset is found in some other works, especially in (Şimşekli *et al.*, 2021) who used it to leverage the local stability of Hausdorff dimension, their argument is reported in the proof of Corollary 2

Proof. It is enough to show that $f^{-1}(]t,+\infty[) \in \mathfrak{E}(\mathbb{R}^d) \otimes \mathcal{T}$ for any $t \in \mathbb{Q}$ as those sets $]t,+\infty[$ generate the Borel σ-algebra in $\mathbb{R}$. Let us fix some $t \in \mathbb{Q}$. Let us denote $\zeta_\omega := \zeta(\cdot,\omega)$, we have:

$$f^{-1}(]t,+\infty[) = \bigcup_{\omega\in\Omega} \Big(\{F \in \mathbf{CL}(\mathbb{R}^d),\ F \cap \zeta_\omega^{-1}(]t,+\infty[) \neq \emptyset\} \times \{\omega\} \Big).$$

By (Molchanov, 2017, Proposition 1.1.2), we have that the sets of the form $\{F \in \mathbf{CL}(\mathbb{R}^d),\ F \cap \mathcal{O} \neq \emptyset\}$ generate $\mathfrak{E}(\mathbb{R}^d)$, with $\mathcal{O}$ running through open sets of $\mathbb{R}^d$. Therefore the continuity of ζ and the countability of $\mathcal{Z}$ give us:

$$f^{-1}(]t,+\infty[) \in \mathfrak{E}(\mathbb{R}^d) \otimes \mathcal{T}.$$

□

If we want to get rid of the countability assumption on Ω, we have to introduce some metric structure on it. This approach justifies the assumptions made on $\mathcal{Z}$ in Section 1.1.

Lemma 5. *Assume that Ω is a Polish space with a dense countable subset D and that ζ is continuous in both variables. Then the function:*

$$f : \boldsymbol{CL}(\mathbb{R}^d) \times \Omega \ni (W,\omega) \longmapsto \sup_{x\in W} \zeta(x,\omega) \in \mathbb{R},$$

is measurable with respect to $\mathfrak{E}(\mathbb{R}^d) \otimes \mathcal{B}_\Omega$, where $\mathcal{B}_\Omega$ is the Borel σ-algebra on Ω.

Proof. As before, let $t \in \mathbb{Q}$, for $X, \omega \in \mathbf{CL}(\mathbb{R}^d) \times \Omega$ we have that

$$(X,\omega) \in f^{-1}(]t,+\infty[) \iff \exists x \in X,\ \exists \epsilon \in \mathbb{Q}_{>0},\ \exists \bar{d} \in D,\ \forall d \in B(\bar{d},\epsilon) \cap D, \zeta(d,x) > t,$$

and therefore

$$f^{-1}(]t,+\infty[) = \bigcup_{\bar{r}\in D} \bigcup_{\epsilon\in\mathbb{Q}_{>0}} \Big(\big(\bigcap_{d\in B(\bar{d},\epsilon)} \{F \in \mathbf{CL}(\mathbb{R}^d),\ F \cap \zeta_\omega^{-1}(]t,+\infty[) \neq \emptyset\}\big) \times B(\bar{d},\epsilon) \Big).$$

The results follows from the same arguments as in the proof of the previous lemma.

□

Those technical lemmas make the arguments made in this section, based on total mutual information, valid.

To end this technical discussion about random closed set we try to answer the following questions: are the covering numbers with respect to pseudo-metric ρ_S measurable? Moreover, can we construct coverings that are well-defined random close sets themselves[6]?

To answer those related questions, we need to introduce Castaing's representations, which are a fundamental tool to deal with random closed sets (Molchanov, 2017, Theorem 1.3.3 and Definition 1.3.6).

Proposition 14 (Castaing's representations)**.** *Let W be a random closed set in $\mathbb{R}^d$, then there exists a countable family $(\xi_n)_{n\geq 1}$ of $\mathbb{R}^d$-valued random variables whose closure is almost surely equal to W, namely:*

$$\overline{\{\xi_n,\ n \geq 1\}} = W, \ \textit{almost surely.}$$

Equipped with this result, we can easily extend Theorem 4.1 to the measurability of the covering numbers associated to a Castaing's representation of the hypothesis set:

Theorem 5.1. Measurability of covering numbers in the case of random hypothesis set

Let $W \subset \mathbb{R}^d$ be a random closed set over a probability space $(\Omega,\mathcal{T})$ and $\delta > 0$. Let us introduce a Castaing's representation $(\xi_n)_{n\geq 1}$ of W.
Then, under Assumption 1, we have that the mapping between probability spaces

$$(\mathcal{Z}^n,\ \mathcal{F}^{\otimes n}) \otimes (\Omega,\mathcal{T}) \ni (S,\omega) \longmapsto |N_\delta^{\rho_S}(\{\xi_n(\omega),\ n \geq 1\})| \in (\mathbb{N}_+,\ \mathcal{P}(\mathbb{N}_+)),$$

[6] by covering we mean the position of the centers of a cover by closed balls, as in Section 3

is a random variable, where $\mathcal{P}(A)$ denotes the subsets of a set A. In particular, the upper-box counting dimension $\overline{\dim}_B^{\rho_S}$ is a random variable.

Proof. The proof follows exactly that of Theorem 4.1 except that now have a Castaing's representation $(\xi_n)_{n\geq 1}$ of W.
By the same proof than Equation (33), we have:

$$\{(S,\omega),\ |N_\delta^{\rho_S}(\{\xi_n(\omega),\ n\geq 1\})| \leq M\} = \bigcup_{I\in\mathfrak{F}_{\leq M}(\mathbb{N}_+)} \bigcap_{k\in\mathbb{N}} \bigcup_{i\in I} \{(S,\omega),\ \rho_S(\xi_i(\omega),\xi_k(\omega)) \leq \delta\}.$$

By continuity and composition of random variables, it is clear that

$$\{(S,\omega),\ \rho_S(\xi_i(\omega),\xi_k(\omega)) \leq \delta\} \in \mathcal{F}^{\otimes n} \otimes \mathcal{T},$$

hence we have the result by countable unions and intersections.
Therefore $\overline{\dim}_B^{\rho_S}(W(\omega))$ is a random variable is a random variable as a direct consequence of Lemma 3.

□

Thanks to Theorem 5.1, we are actually able to proof the much stronger result that we *can* build measurable coverings.

Theorem 5.2. Measurable coverings

Let $W \subset \mathbb{R}^d$ be a random closed set over a probability space $(\Omega,\mathcal{T},\mathbb{P})$ and $\delta > 0$. Let $\mathfrak{F}(\mathbb{N}_+)$ denote the set of finite subsets of $\mathbb{N}_+$. Then, under Assumption 1, we can build a map:

$$N_\delta : \mathcal{Z}^n \times \Omega \longrightarrow \mathfrak{F}(\mathbb{R}^d) \subset \mathbf{CL}(\mathbb{R}^d),$$

which is measurable (with respect to the Effrös σ-algebra on the right hand-side) and such that for almost all $(S,\omega) \in \mathcal{Z}^n \times \Omega$, $N_\delta(S,\omega)$ is a finite set which is (almost surely) a covering of $W(\omega)$ with respect to pseudo-metric ρ_S and such that we have almost surely over $\mu_z^{\otimes n} \otimes \mathbb{P}$:

$$\overline{\dim}_B^{\rho_S}(W(\omega)) = \limsup_{\delta\to 0} \frac{|N_\delta(S,\omega)|}{\log(1/\delta)}.$$

Proof. Let us introduce a Castaing's representation $(\xi_k)_{k\geq 1}$ of W and denote by $\mathfrak{F}_N(\mathbb{N}_+)$ the set of finite subsets of $\mathbb{N}_+$ with exactly N elements. Again, as in Theorem 4.1, the proof is based on the idea that, thanks to the continuity of the loss ℓ defining the pseudo-metric ρ_S, a cover of $\{\xi_k,\ k\in\mathbb{N}_+\}$ covers W. Let us denote $\mathcal{C}(\omega) = \{\xi_k(\omega),\ k\in\mathbb{N}_+\}$.
As $\mathfrak{F}_N(\mathbb{N}_+)$ is countable, for each $N\in\mathbb{N}_+$, we introduce $(F_i^N)_{i\geq 1}$ an ordering of $\mathfrak{F}_N(\mathbb{N}_+)$.
Now for each $(S,\omega)\in\mathcal{Z}^n\times\Omega$, we define:

$$\forall i\in\mathbb{N}_+,\ F_i(S,\omega) := F_i^{|N_\delta^{\rho_S}(\mathcal{C}(\omega))|}.$$

Let us now introduce the minimal index of a set of indices that can cover W:

$$i_0(S,\omega) := \operatorname{argmin}\left\{ i\in\mathbb{N}_+,\ \forall k\geq 1,\ \exists j\in F_i(S,\omega),\ \rho(\xi_j(\omega),\xi_i(\omega)) \leq \delta \right\}.$$

Note that i_0 is finite because $\{(\ell(w,z_i)_{1\leq i\leq n}),\ w_i n W\}$ is compactly contained, thanks to the boundedness assumption on ℓ, i.e. the covering numbers are finite.
We can therefore build the following 'covering indices' function:

$$\mathcal{I}_\delta : \mathcal{Z}^n\times\Omega \longrightarrow \mathfrak{F}(\mathbb{N}_+),$$

defined by $\mathcal{I}_\delta(S,\omega) = F_{i_0(S,\omega)}(S,\omega)$.

Now we want to introduce an 'evaluation functional', i.e. a mapping:

$$\Xi : \Omega \times \mathfrak{F}(\mathbb{N}_+) \longrightarrow \mathfrak{F}(\mathbb{R}^d) \subset \mathbf{CL}(\mathbb{R}^d),$$

defined by $\Xi(\omega, I) = \{\xi_i(\omega),\ i \in I\}$. It is easy to see that Ξ is measurable, indeed for any compact set $K \subset \mathbb{R}^d$ we have:

$$\{(\omega, I),\ \Xi(\omega, I) \cap K \neq \emptyset\} = \bigcup_{F \in \mathfrak{F}(\mathbb{N}_+)} \bigcup_{i \in F} \{\xi_i \in K\} \times \{I\},$$

implying the measurability by countable unions and Definition 12.
The key point of the proof is that we construct the coverings as $N_\delta(S, \omega) = \Xi(\omega, \mathcal{I}_\delta(S, \omega))$, so that the measurability of N_δ reduces to that of $\mathcal{I}_\delta$. This is achieved by noting that for any non-empty $I \in \mathfrak{F}(\mathbb{N}_+)$, such that $I = F_{i_1}^N$ for some $N, i_1 \geq 1$, we have, by leveraging the countable ordering of $\mathfrak{F}_N(\mathbb{N}_+)$:

$$\begin{aligned}\mathcal{I}_\delta^{-1}(\{I\}) =& \{|N_\delta^{\rho_S}(\mathcal{C}(\omega))| = N\} \cap \left(\bigcap_{k=1}^{+\infty} \bigcup_{m \in I} \{(S, \omega),\ \rho_S(\xi_k(\omega), \xi_m(\omega)) \leq \delta\} \right) \\ & \cap \left(\bigcap_{i < i_1} \bigcup_{k=1}^{+\infty} \bigcap_{m \in F_i^N} \{(S, \omega),\ \rho_S(\xi_k(\omega), \xi_m(\omega)) > \delta\} \right).\end{aligned}$$

By Theorem 5.1, we have the measurability of $(S, \omega) \mapsto |N_\delta^{\rho_S}(W(\omega))|$, hence the measurability result follows by continuity of ℓ (and therefore $\rho_\cdot(\cdot, \cdot)$) and countable unions and intersections.
Now, using Lemma 3, $N_\delta(S, \omega)$ also defines a covering of W and we have:

$$\overline{\dim}_B^{\rho_S}(W(\omega)) = \limsup_{\delta \to 0} \frac{|N_\delta(S, \omega)|}{\log(1/\delta)}. \tag{43}$$

□

Let us make the following important remark , which summarizes most of this subsection.

Remark 15. *Theorem 5.2 shows that we can construct measurable coverings of the random closed hypothesis set under pseudo-metric ρ_S. While those coverings may not be strictly speaking minimal, they yields the same upper-box counting dimensions, which is enough for all proofs in this work to hold. Note that this technical complication of not being minimal comes from the fact that we asked the minimal coverings of a set F to be included in F, however this also removes further technical complications. If we do not impose this condition, our proof would imply that we can construct measurable minimal coverings.*

From now on, we will implicitly assume that the coverings we consider are measurable and induce correct upper box-counting dimension, the present subsection being a theoretical basis for this assumption.

5.2 A generalization bound for random hypothesis space

We now come to the proof of a generalization bound for random hypothesis spaces, under the previous assumptions. In particular, we assumed in Assumption 1 that the loss ℓ is bounded. Here, without loss of generality we assume that uniformly we have $\ell \in [0, B]$, this will simplify the notations.

For notational purposes let us denote the upper box-counting dimension of $\mathcal{W}_{S,U}$ induced by pseudo-metric ρ_S by

$$d(S, U) := \overline{\dim}_B^{\rho_S}(\mathcal{W}_{S,U}),$$

where ρ_S is the pseudo-metric define by Equation 28.

Let us also denote the worst-case generalization error by

$$\mathcal{G}(S, U) := \sup_{w \in \mathcal{W}_{S,U}} (\mathcal{R}(w) - \hat{\mathcal{R}}_S(w)), \tag{44}$$

which is a random variable thanks to the results of Section 5.1.

Moreover, as we proved in Remark 12, note that $d(S, U)$ can be written as a countable limit of random variables and therefore defines a random variable thanks to the measurability of the coverings.

The main difficulty here is that classical arguments developed in Sections 2.2 and 4.3 based on the Rademacher complexity cannot be applied in this case as $\mathcal{W}_{S,U}$ depends on the data sample S. In particular neither The symmetrization argument used nor the use of Mc-Diarmid's inequality in the proof of proposition 2 hold.

Remark 16. *As proven by Foster* et al. *(2020), some kind of 'uniform stability assumption' (see Section 5.3) on the hypothesis set $\mathcal{W}_{S,U}$ could make a reasoning similar to Mc-Diarmid inequality valid, however it does not correct the issue regarding symmetrization. Moreover, we claim that this type of reasoning essentially leads to expected bounds and make us lose too much geometrical information on $\mathcal{W}_{S,U}$.*

Hence, to be able to develop a covering argument, we first cover the set $\mathcal{W}_{S,U}$ by using the pseudo-metric ρ_S and rely on the following decomposition: for any $\delta > 0$ and $w' \in N_\delta^{\rho_S}(\mathcal{W}_{S,U})$ we have that

$$\mathcal{R}(w) - \hat{\mathcal{R}}_S(w) \leq \mathcal{R}(w') - \hat{\mathcal{R}}_S(w') + |\hat{\mathcal{R}}_S(w) - \hat{\mathcal{R}}_S(w')| + |\mathcal{R}(w) - \mathcal{R}(w')|.$$

In the above inequality, the first term can be controlled by standard techniques, namely concentration inequalities and decoupling theorems presented in Section 2 as w' lives in a finite set $N_\delta^{\rho_S}(\mathcal{W}_{S,U})$ and the second term is trivially less than δ by the definition of coverings. However, the last term cannot be bounded in an obvious way. To overcome this issue we introduce 'approximate level-sets' of the population risk, defined as follows[7] for some $K \in \mathbb{N}_+$:

$$R_S^j := \mathcal{W}_{S,U} \cap \mathcal{R}^{-1}\left(\left[\frac{jB}{K}, \frac{(j+1)B}{K}\right]\right), \tag{45}$$

where $j = 0, \ldots, K-1$ and $\mathcal{R}^{-1}$ denotes the inverse image of $\mathcal{R}$. The interval $\left[\frac{jB}{K}, \frac{(j+1)B}{K}\right]$ will be denoted I_j. Note that thanks to the

Let $N_{\delta,j}$ collect the centers of a minimal δ-cover of R_S^j relatively to ρ_S, the measurability condition on the coverings extend to the randomness of those sets $N_{\delta,j}$.

Remark 17. *Without loss of generality we can always assume that those sets R_S^j are non-empty. Indeed we can always add one deterministic point of $\mathcal{R}^{-1}(I_j)$ in each of the coverings $N_{\delta,j}$ one deterministic (always the same) element of $\mathcal{R}^{-1}(I_j)$. It won't make the mutual information term appearing in the final result of Theorem 5.3 bigger (by the data-processing inequality) and it won't change the upper box-counting dimension because of its finite stability. Moreover if some of the sets $\mathcal{R}^{-1}(I_j)$ are empty then we just need to restrict ourselves to a deterministic subset of $[0, B]$. If we don't want to do this, another way, maybe cleaner, of handling the potential empty sets would be to use the convention $\max(\emptyset) = 0$ everywhere in the proof, then we should also adapt the definition of $\epsilon(N, I)$ below to replace $\log(KN)$ by $\max(0, \log(KN))$, where $\log(0)$ is set to $-\infty$. All those manipulations would essentially lead to the same results.*

Measurability of the coverings: We proved in Section 5.1 that we can construct measurable coverings (as random sets), which are actually coverings of a dense countable subset (or a Castaing's representation) of $\mathcal{W}_{S,U}$. Therefore, without loss of generality and thanks to the continuity of the loss ℓ, we can assume in all the remaining of this work that all the considered coverings are random sets, because either they can be constructed by Theorem 5.2 or we can restrict ourselves to Castaing's representations of $\mathcal{W}_{S,U}$.

As can already be noted in Remark 17, our approximate level set technique introduces quite a lot of technical difficulties and intricate terms. We believe that this proof technique is interesting but may not be a definitive answer to the problem at hand, improving it is a direction for future research.

The next theorem provides a generalization bound for random hypothesis sets.

Theorem 5.3. Data-dependent fractal generalization bound for random hypothesis set

Let us set $K = \lfloor \sqrt{n} \rfloor$ and define

$$I_{n,\delta} := \max_{0 \leq j \leq \lfloor \sqrt{n} \rfloor} I_\infty(S, N_{\delta,j}).$$

Then, for all $\epsilon, \gamma, \eta > 0$, there exists $\delta_{n,\gamma,\epsilon} > 0$ such that with probability at least $1 - \eta - \gamma$ under

[7] As U is independent of S, we drop the dependence on it to ease the notation.

$\mu_z^{\otimes n} \otimes \mu_u$, for all $\delta < \delta_{n,\gamma,\epsilon}$ we have:

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

Proof. As mentioned above, we assume without loss of generality that that the loss takes values in $[0,B]$.
Let us fix some integer $K \in \mathbb{N}_+$ and define $I_j = [\frac{jB}{K}, \frac{(j+1)B}{K}]$, such that:

$$[0,B] = \bigcup_{j=0}^{K-1} I_j$$

Then, given $\mathcal{W}_S$ we define the set $R_S^j := \mathcal{W}_S \cap \mathcal{R}^{-1}(I_j)$.
We then introduce the random closed (finite) sets [a] $N_{\delta,j}$ corresponding to the centers of a minimal covering of R_S^j, such that $N_{\delta,j} \subset R_S^j$, for the pseudo-metric:

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

The first step is to write that almost surely:

$$\sup_{w \in \mathcal{W}_S} \big(\mathcal{R}(w) - \hat{\mathcal{R}}_S(w)\big) = \max_{0 \leq j \leq K-1} \sup_{w \in R_S^j} \big(\mathcal{R}(w) - \hat{\mathcal{R}}_S(w)\big)$$

Then, given $w, w' \in R_S^j$ such that $\rho_S(w,w') \leq \delta$ we have by the triangle inequality:

$$\begin{aligned}\big(\mathcal{R}(w) - \hat{\mathcal{R}}_S(w)\big) &\leq \big(\mathcal{R}(w') - \hat{\mathcal{R}}_S(w')\big) + \rho_S(w,w') + |\mathcal{R}(w) - \mathcal{R}(w')| \\ &\leq \big(\mathcal{R}(w') - \hat{\mathcal{R}}_S(w')\big) + \delta + \frac{B}{K}.\end{aligned} \tag{46}$$

So that we get:

$$\sup_{w \in \mathcal{W}_S} \big(\mathcal{R}(w) - \hat{\mathcal{R}}_S(w)\big) \leq \delta + \frac{B}{K} + \max_{0 \leq j \leq K-1} \max_{w \in N_{\delta,j}} \big(\mathcal{R}(w) - \hat{\mathcal{R}}_S(w)\big). \tag{47}$$

Let us introduce the random element being the concatenation of the covers defined above:

$$N_\delta^{\cup} := (N_{\delta,0}, \dots, N_{\delta,K-1})$$

Now we fix some $\eta > 0$ and just introduce the random variable ϵ as a function of two variables N and I:

$$\epsilon(N,I) := \sqrt{\frac{2B^2}{n}\bigg(\log(1/\eta) + \log\big(KN\big) + I\bigg)}.$$

We have by the decoupling Proposition 4 along with Fubini's Theorem, Hoeffding inequality and a union bound:

$$\begin{aligned}
&\mathbb{P}\Big(\max_{0\leq j\leq K-1} \max_{w\in N_{\delta,j}} \big(\mathcal{R}(w) - \hat{\mathcal{R}}_S(w)\big) \geq \epsilon(\max_j |N_{\delta,j}|,\ \max_j I_\infty(S, N_{\delta,j}))\Big) \\
&\leq \sum_{j=0}^{K-1} \mathbb{P}\Big(\max_{w\in N_{\delta,j}} \big(\mathcal{R}(w) - \hat{\mathcal{R}}_S(w)\big) \geq \epsilon(N_{\delta,j}, I_\infty(S, N_{\delta,j}))\Big) \\
&\leq \sum_{j=0}^{K-1} e^{I_\infty(S,N_{\delta,j})} \mathbb{P}_{N_{\delta,j}} \otimes \mathbb{P}_S\Big(\max_{w\in N_{\delta,j}} \big(\mathcal{R}(w) - \hat{\mathcal{R}}_S(w)\big) \geq \epsilon(N_{\delta,j}, I_\infty(S, N_\delta^j))\Big) \\
&\leq \sum_{j=0}^{K-1} e^{I_\infty(S,N_{\delta,j})} \mathbb{E}_{N_{\delta,j}} \Big[\mathbb{P}_S\Big(\max_{w\in N_{\delta,j}} \big(\mathcal{R}(w) - \hat{\mathcal{R}}_S(w)\big) \geq \epsilon(N_{\delta,j}, I_\infty(S, N_\delta^j))\Big)\Big] \\
&\leq \sum_{j=0}^{K-1} e^{I_\infty(S,N_{\delta,j})} \mathbb{E}_{N_{\delta,j}} \Big[\sum_{w\in N_{\delta,j}} \mathbb{P}_S\Big(\big(\mathcal{R}(w) - \hat{\mathcal{R}}_S(w)\big) \geq \epsilon(N_{\delta,j}, I_\infty(S, N_\delta^j))\Big)\Big] \\
&\leq \sum_{j=0}^{K-1} e^{I_\infty(S,N_{\delta,j})} \mathbb{E}_{N_{\delta,j}} \Big[|N_{\delta,j}| \exp\Big\{ -\frac{n\epsilon(N_{\delta,j}, I_\infty(S, N_\delta^j))^2}{2B^2}\Big\}\Big] \\
&\leq \sum_{j=0}^{K-1} e^{I_\infty(S,N_{\delta,j})} \mathbb{E}_{N_{\delta,j}} \Big[\frac{\eta}{K} e^{-I_\infty(S,N_{\delta,j})}\Big] \\
&= \eta.
\end{aligned} \tag{48}$$

Now let us consider a random minimal δ-cover of the whole (random) hypothesis set $\mathcal{W}_S$. Given $j \in \{0, \dots, K-1\}$, we have in particular almost surely that:

$$\mathcal{W}_S \cap R_j \subseteq \bigcup_{w\in N_\delta} B_\delta^{\rho_S}(w)$$

Where $B_\delta^{\rho_S}(w)$ denotes the closed δ-ball for metric ρ_S centered in w. Therefore there exists a non-empty subset $\tilde{N}_\delta \subseteq N_\delta$ such that for all $w \in \tilde{N}_\delta$ we have $B_\delta^{\rho_S}(w) \cap R_j \neq \emptyset$.
Therefore we can collect in some set $\tilde{N}_{\delta,j}$ one element in each $B_\delta^{\rho_S}(w) \cap R_j$ for $w \in \tilde{N}_\delta$ and the triangular inequality gives us:

$$R_S^j \subseteq \bigcup_{w\in \tilde{N}_{\delta,j}} B_{2\delta}^{\rho_S}(w)$$

This proves that almost surely $\forall j,\ |N_{\delta,j}| \leq |N_{\delta/2}|$, and thus:

$$\max_{0\leq j\leq K-1} |N_{\delta,j}| \leq |N_{\delta/2}| \tag{49}$$

We know that we have almost surely that:

$$\limsup_{\delta\to 0} \frac{\log(|N_{\delta/2}|)}{\log(2/\delta)} = \overline{\dim}_B^{\rho_S}(\mathcal{W}_S).$$

Therefore let us fix $\gamma, \epsilon > 0$. Using Egoroff's Theorem we can say that there exists $\delta_{n,\gamma,\epsilon} > 0$ such that, with probability at least $1 - \gamma$, for all $\delta \leq \delta_{n,\gamma,\epsilon}$ we have:

$$\log\big(|N_{\delta/2}|\big) \leq (\epsilon + \overline{\dim}_B^{\rho_S}(\mathcal{W}_S)) \log(2/\delta)$$

Therefore combining equations 46, 48, 49, we get that with probability at least $1 - \gamma - \eta$, for all

$\delta \leq \delta_{n,\gamma,\epsilon}$:

$$\begin{aligned}
&\sup_{w\in\mathcal{W}_S} \big(\mathcal{R}(w) - \hat{\mathcal{R}}_S(w)\big) \\
&\qquad \leq \delta + \frac{B}{K} + \sqrt{\frac{2B^2}{n}\bigg(\log(K/\eta) + \log\big(\max_j |N_{\delta,j}|\big) + \max_j I_\infty(S, N_{\delta,j})\bigg)} \\
&\qquad \leq \delta + \frac{B}{K} + \sqrt{\frac{2B^2}{n}\bigg(\log(K/\eta) + \log|N_{\delta/2}| + \max_j I_\infty(S, N_{\delta,j})\bigg)} \\
&\qquad \leq \delta + \frac{B}{K} + \sqrt{\frac{2B^2}{n}\bigg(\log(K/\eta) + \log(2/\delta)(\epsilon + \overline{\dim}_B^{\rho_S}(\mathcal{W}_S)) + \max_j I_\infty(S, N_{\delta,j})\bigg)}
\end{aligned}$$

The choice of K has not been done yet, considering the above equation the best choice is clearly: $K = K_n := \lfloor \sqrt{n} \rfloor$. Let us introduce the notation:

$$I_{n,\delta} := \max_j I_\infty(S, N_{\delta,j}).$$

This way we get that with probability at least $1 - \gamma - \eta$, for all $\delta \leq \delta_{n,\gamma,\epsilon}$:

$$\sup_{w\in\mathcal{W}_S} \big(\mathcal{R}(w) - \hat{\mathcal{R}}_S(w)\big) \leq \delta + \frac{B}{\sqrt{n}-1} + \sqrt{2}B\sqrt{\frac{\log(\sqrt{n}/\eta) + \log(2/\delta)(\epsilon + \overline{\dim}_B^{\rho_S}(\mathcal{W}_S)) + I_{n,\delta}}{n}} \tag{50}$$

Note that it is possible to set this value of K, which depends on n, at the end of the proof, because the previous limits do not depend on K.

□

[a] Note that, as mentioned earlier, in this paper we always assume that minimal coverings are random sets.

This theorem gives us a bound in the general case similar to (Şimşekli *et al.*, 2021, Theorem 2), yet without requiring Lipschitz continuity.

Moreover, also similar to (Şimşekli *et al.*, 2021; Hodgkinson *et al.*, 2022), Theorem 5.3 introduces a mutual information term $I_{n,\delta}$, which intuitively measures the local mutual dependence between the data and the coverings. This can be seen as how the data influences the 'local fractal behavior' of the hypothesis set. On the other hand, despite the similarity to prior work, $I_{n,\delta}$ might be more complex because the dependence of $N_{\delta,j}$ on S comes both from the pseudo-metric ρ_S and the hypothesis set $\mathcal{W}_{S,U}$. In the next subsection, we show that we can modify our theory in a way that it involves the simpler mutual information term proposed in (Hodgkinson *et al.*, 2022).

5.3 Uniform geometric stability and mutual information

The intricate dependence between $N_{\delta,j}$ and S makes it hard to express the term $I_{n,\delta}$ in Theorem 5.3 or bound it with standard methods (e.g. data-processing inequality). In this subsection, we introduce a notion of 'geometric stability' to obtain a more interpretable bound.

Algorithmic stability is a key notion in learning theory and has been shown to imply good generalization properties (Bousquet, 2002; Bousquet *et al.*, 2020; Chandramoorthy *et al.*, 2022). Recently, Foster *et al.* (2020) extended this notion to the stability of *hypothesis sets*, and proposed a notion of stability as a bound on a kind of Hausdorff distance between the hypothesis sets generated by neighboring datasets. In our setting this would mean that there exists some $\bar{\beta} > 0$ such that for all $S, S' \in \mathcal{Z}^n$ differing only by one element, for all $u \in \mathcal{U}$, we have:

$$\forall w \in \mathcal{W}_{S,U},\ \exists w' \in \mathcal{W}_{S',U},\ \forall z \in \mathcal{Z},\ |\ell(w,z) - \ell(w',z)| \leq \bar{\beta}. \tag{51}$$

Foster *et al.* (2020) argue that in many situations $\bar{\beta} = \mathcal{O}(1/n)$.

While Equation (51) has been proven to be powerful, especially to relate the worst-case generalization error to its expectation, we claim that it is not enough to capture the fine geometrical behavior of fractal hypothesis sets. Indeed, it gives a global information about a particular kind of Hausdorff distance between two sets, without giving local information, e.g. on the coverings.

Inspired by (Foster *et al.*, 2020), we introduce a stability notion, coined *geometric stability*, on the minimal coverings that will allow us to reduce the statistical dependence between the dataset $S \sim \mu_z^{\otimes n}$ and those coverings.

To state our stability notion, we need to refine our definition of coverings. Let $A \subset \mathbb{R}^d$ be some closed set, potentially random. For any $\delta > 0$ we define $N_\delta(A, S)$ to be the random minimal coverings of A by closed δ-balls under pseudo-metric ρ_S (28) with centers in A. Note that the dependence in S in $N_\delta(A, S)$ only refers to the *pseudo-metric* used. In addition to being able to construct measurable coverings $N_\delta(A, S)$, making it a well-defined random set, we add the fact that this selection can be made regular enough in the following sense.

Definition 13 (geometric stability)**.** *We say that a set A is geometrically stable if there exist some $\beta > 0$ and $\alpha > 0$ such that for δ small enough we can find a random covering $S \mapsto N_\delta(A, S)$ such that for all $S \in \mathcal{Z}^n$ and $S' \in \mathcal{Z}^{n-1}$ such that $S' = S \setminus \{z_i\}$ for some i, then $N_\delta(A, S)$ and $N_\delta(A, S')$ are within β/n^α distance for an uniform data-dependent Hausdorff distance, i.e.,*

$$\forall w \in N_\delta(S, A),\ \exists w' \in N_\delta(S', A),\ \sup_{z \in \mathcal{Z}} |\ell(w, z) - \ell(w', z)| \leq \frac{\beta}{n^\alpha}. \tag{52}$$

Based on this definition, we assume the following condition.

Assumption 3. *Let $K \in \mathbb{N}_+$. There exists $\alpha \in (0, 3/2)$ and $\beta > 0$ (potentially depending on K) such that all sets of the form $\mathcal{W}_{S,U} \cap \mathcal{R}^{-1}\big(\big[\frac{jB}{K}, \frac{(j+1)B}{K}\big]\big)$ are geometrically stable with parameters (α, β).*

Assumption 3 essentially imposes a *local* regularity condition on the fractal behavior of $\mathcal{W}_{S,U}$ with respect to the pseudo-metric ρ_S. Intuitively it means that we can select a regular enough covering among all coverings. Note that the geometric stability is a condition on how the coverings vary with respect to the pseudo-metric, which is fundamentally different than (Foster *et al.*, 2020).

The next theorem provides a generalization bound under the geometric stability condition.

Theorem 5.4. Data-dependent fractal generalization bound using geometric stability

Let $d(S, U)$ and $\mathcal{G}(S, U)$ be as in Theorem 5.3 and further define $I := I_\infty(S, \mathcal{W}_{S,U})$. Suppose that 3 holds. Then there exists constants $n_\alpha, \delta_{\gamma,\epsilon,n} > 0$ such that for all $n \geq n_\alpha$, with probability $1 - \gamma - \eta$, and for all $\delta \leq \delta_{\gamma,\epsilon,n}$, the following inequality holds:

$$\mathcal{G}(S, U) \leq \frac{3B + 2\beta}{n^{\alpha/3}} + \delta + B\sqrt{\frac{\big(\epsilon + d(S, U)\big)\log(4/\delta) + \log(1/\eta) + \log(n) + I}{2n^{\frac{2\alpha}{3}}}}.$$

Moreover, we have that $n_\alpha = \max\{2^{\frac{3}{2\alpha}}, 2^{1+\frac{3}{3-2\alpha}}\}$.

Let us present the proof of Theorem 5.4. The proof proceeds in two steps and is based on what we will call a *grouping technique*. 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 $JH = n$. In the end of the proof a particular choice is made.

A minor technical difficulty appears when it is not actually possible two write $JH = n$ for a pertinent choice of (J, H). Therefore we first present a result, Proposition 15, when the latter is possible and then derive two corollaries to deal with this technical issue, mostly based on the boundedness assumption. Theorem 5.4 will be the second corollary.

Remark 18. *For the sake of the proof we need to assume $\alpha \leq \frac{3}{2}$, which is just asking for a potentially weaker assumption, which is not a problem. Note that the value $\alpha \leq \frac{3}{2}$ will lead in Theorem 5.4 to a convergence rate in $n^{-1/2}$ which is optimal anyway.*

Let us start with the more interesting result of this section, which contains the main proof techniques.

Proposition 15. *Under Assumptions 1, 2 and 3 with the same notations than in Theorem 5.4, we also take arbitrary $J, H \in \mathbb{N}_+$ such that $JH = n$*

Then for all $n \geq 2^{\frac{3}{3-2\alpha}}$, with probability $1-\gamma-\eta$, for all δ smaller than some $\delta_{\gamma,\epsilon,n} > 0$ we have:

$$\begin{aligned}\sup_{w\in\mathcal{W}_{S,U}} |\mathcal{R}(w) - \hat{\mathcal{R}}_S(w)| \leq& \delta + \frac{B}{\sqrt{n}-1} + \frac{2J\beta}{n^\alpha} \\ &+ H\sqrt{\frac{JB^2}{2n^2}\Big(\big(\epsilon + d(S,U)\big)\log(4/\delta) + \log(H\sqrt{n}/\eta) + I\Big)}\end{aligned}$$

Proof. Let us first refine our notations for the coverings to make the proof clearer. Throughout this section, for any S, S' we will denote $N_\delta(S, S', U)$ the centers of a covering of $\mathcal{W}_{S,U}$ by closed δ-balls under pseudo-metric $d_{S'}$. As in the proof of Theorem 5.3, we introduce some approximate level sets R_S^j for $j \in \{0, \dots, K-1\}$. We then denote by $N_{\delta,j}(S, S', U)$ the centers of a covering of R_S^j by closed δ-balls under pseudo-metric $d_{S'}$. (note that the R_S^j still depends on U but the dependence has been dropped to ease the notations).
The proof starts by introducing the "level-sets" of the population risk as in the proof of Theorem 5.3. We define R_S^j exactly in the same way. The same remark as in proof of Theorem 5.3 about the fact that we can assume R_S^j is non-empty also holds here.
The proof starts with the same statement:

$$\sup_{w\in\mathcal{W}_{s,U}} |\mathcal{R}(w) - \hat{\mathcal{R}}_S(w)| \leq \max_{0\leq j\leq K-1} \sup_{w\in R_S^j} |\mathcal{R}(w) - \hat{\mathcal{R}}_S(w)|$$

For all j, we (minimally) cover R_S^j with δ-covers for pseudo-metric d_S, such that the centers are in R_S^j. We collect those centers in $N_{\delta,j}(S, S, U)$.
This leads us to:

$$\sup_{w\in\mathcal{W}_{s,U}} |\mathcal{R}(w) - \hat{\mathcal{R}}_S(w)| \leq \delta + \frac{B}{K} + \max_{0\leq j\leq K-1} \underbrace{\max_{w\in N_{\delta,j}(S,S,U)} |\mathcal{R}(w) - \hat{\mathcal{R}}_S(w)|}_{:=E_j} \tag{53}$$

Thanks to our stability assumption 3, we can say that for δ small enough there exists a random minimal covering such that for all $j \in \{0, \dots, K-1\}$ and all $k \in \{1, \dots, H\}$ the covering $N_{\delta,j}(S, S^{\setminus J_k}, U)$ satisfies:

$$\forall w \in N_{\delta,j}(S,S,U),\ \exists w' \in N_{\delta,j}(S, S^{\setminus J_k}, U),\ \sup_{z\in\mathcal{Z}} |\ell(w,z) - \ell(w',z)| \leq \frac{\beta J}{n^\alpha}$$

Where the J factor on the right hand side comes from the fact that our stability assumption can be seen as a Lipschitz assumption in terms of the Hausdorff distance of the coverings with respect to the Hamming distance on the datasets.
Recall that we assume that all $N_{\delta,j}$ have coverings-metrics stability with common parameters β, α.
As in the previous proposition, we split the index set $\{1, \dots, n\}$ into H groups of size J, with $HJ = n$, which allows us to write (with a similar proof):

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

Putting this back into equation (53) we get:

$$\begin{aligned}\sup_{w\in\mathcal{W}_{s,U}} |\mathcal{R}(w)-\hat{\mathcal{R}}_S(w)| &\leq \delta + \frac{B}{K} + \frac{2J\beta}{n^\alpha} \\ &\quad + \max_{0\leq j\leq K-1} \sum_{k=1}^{H} \max_{w\in N_{\delta,j}(S,S^{\setminus J_k},U)} \frac{1}{n}\left| \sum_{i\in J_k} \big(\ell(w,z_i)-\mathcal{R}(w)\big)\right| \\ &\leq \delta + \frac{B}{K} + \frac{2J\beta}{n^\alpha} \\ &\quad + H \max_{0\leq j\leq K-1} \max_{1\leq k\leq H} \underbrace{\max_{w\in N_{\delta,j}(S,S^{\setminus J_k},U)} \frac{1}{n}\left| \sum_{i\in J_k} \big(\ell(w,z_i)-\mathcal{R}(w)\big)\right|}_{:=M_{j,k}(S,U)}\end{aligned} \tag{54}$$

Let ϵ be a random variable depending on $N_{\delta,j}(S,S^{\setminus J_k},U)$ only. We use a decoupling lemma (lemma 1 in (Hodgkinson *et al.*, 2022)) along with Hoeffding's inequality to write:

$$\begin{aligned}\mathbb{P}\big(M_{j,k}(S,U)\geq\epsilon\big) &\leq e^{I_\infty(N_{\delta,j}(S,S^{\setminus J_k},U),S_{J_k})}\mathbb{P}_{N_{\delta,j}(S,S^{\setminus J_k},U)}\otimes\mathbb{P}_{S_{J_k}}\big(M_{j,k}(S,U)\geq\epsilon\big) \\ &\leq e^{I_\infty(N_{\delta,j}(S,S^{\setminus J_k},U),S_{J_k})}\mathbb{E}_{N_{\delta,j}(S,S^{\setminus J_k},U)}\Big[\mathbb{P}_{S_{J_k}}\big(M_{j,k}(S,U)\geq\epsilon\big)\Big] \\ &\leq e^{I_\infty(N_{\delta,j}(S,S^{\setminus J_k},U),S_{J_k})} \\ &\times \mathbb{E}_{N_{\delta,j}(S,S^{\setminus J_k},U)}\left[\mathbb{P}_{S_{J_k}}\left(\bigcup_{w\in N_{\delta,j}(S,S^{\setminus J_k},U)}\left\{\frac{1}{n}\left|\sum_{i\in J_k}\big(\ell(w,z_i)-\mathcal{R}(w)\big)\right|\geq\epsilon\right\}\right)\right] \\ &\leq e^{I_\infty(N_{\delta,j}(S,S^{\setminus J_k},U),S_{J_k})}\mathbb{E}\left[|N_{\delta,j}(S,S^{\setminus J_k},U)|e^{-\frac{2\epsilon^2n^2}{JB^2}}\right]\end{aligned} \tag{55}$$

The key point of the proof, and the reason for which we have introduced this strong stability assumption on the coverings is that we can now use the following Markov chain:

$$S_{J_k} \longrightarrow \mathcal{W}_{S,U} \longrightarrow N_{\delta,j}(S,S^{\setminus J_k},U). \tag{56}$$

Therefore, by the data processing inequality:

$$I_\infty(N_{\delta,j}(S,S^{\setminus J_k},U),J_{J_k}) \leq I_\infty(\mathcal{W}_{S,U},S_{J_k})$$

Now using the easier Markov chain:

$$\mathcal{W}_{S,U} \longrightarrow S \longrightarrow S_{J_k},$$

We have:

$$I_\infty(N_{\delta,j}(S,S^{\setminus J_k},U),S_{J_k}) \leq I_\infty(S,\mathcal{W}_{S,U}) \tag{57}$$

Note that the mutual information term appearing in equation (57) is the same than the one appearing in (Hodgkinson *et al.*, 2022).
Thus:

$$\mathbb{P}\big(M_{j,k}(S,U)\geq\epsilon\big) \leq e^{I_\infty(S,\mathcal{W}_{S,U})}\mathbb{E}\left[|N_{\delta,j}(S,S^{\setminus J_k},U)|e^{-\frac{2\epsilon^2n^2}{JB^2}}\right]$$

Equipped with this result we can make an informed choice for the random variable ϵ, for a fixed $\eta>0$:

$$\epsilon = \epsilon_{j,k} := \sqrt{\frac{JB^2}{2n^2}\Big(\log|N_{\delta,j}(S,S^{\setminus J_k},U)| + \log(HK/\eta) + I_\infty(S,\mathcal{W}_{S,U})\Big)},$$

Now we can apply a union bound to get:

$$\begin{aligned}\mathbb{P}\big(\max_{0\leq j\leq K-1}\max_{1\leq k\leq H} M_{j,k}(S,U) \geq \max_{0\leq j\leq K-1}\max_{1\leq k\leq H}\epsilon_{j,k}\big) &\leq \sum_{j=0}^{K-1}\sum_{k=1}^{H}\mathbb{P}\big(M_{j,k}(S,U)\geq \max_{0\leq j\leq K-1}\max_{1\leq k\leq H}\epsilon_{j,k}\big)\\ &\leq \sum_{j=0}^{K-1}\sum_{k=1}^{H}\mathbb{P}\big(M_{j,k}(S,U)\geq \epsilon_{j,k}\big)\\ &= \eta\end{aligned}$$

Now let us have a closer look at those covering numbers $|N_{\delta,j}(S,S^{\setminus J_k},U)|$. Note that we have:

$$\forall w,w'\in\mathbb{R}^d,\ d_{S^{\setminus J_k}}(w,w')\leq \frac{n}{n-J}d_S(w,w'),$$

And therefore $|N_{\delta,j}(S,S^{\setminus J_k},U)|\leq |N_{\frac{\delta(n-J)}{n},j}(S,S,U)|$.
Moreover, using the same reasoning as in the proof of Theorem 5.3, we know that we have $|N_{\delta,j}(S,S^{\setminus J_k},U)|\leq |N_{\delta/2}(S,S^{\setminus J_k},U)|$.
Thus:

$$|N_{\delta,j}(S,S^{\setminus J_k},U)|\leq |N_{\frac{\delta(n-J)}{2n}}(S,S,U)|$$

As before, we will want to solve the trade-off in the values of H and J by setting $J=n^\lambda$ for some $\lambda\in(0,1)$ (this time we do not allow the value $\lambda=1$, which will be justified later when we find the actual value of λ). A very simple calculation gives us:

$$\frac{\delta(n-J)}{2n}=\frac{\delta}{2}\left(1-\frac{1}{n^{1-\lambda}}\right)$$

Therefore we can say that if $n\geq 2^{\frac{1}{1-\lambda}}$, then $\frac{\delta(n-J)}{2n}\geq\delta/4$ and therefore:

$$|N_{\delta,j}(S,S^{\setminus J_k},U)|\leq |N_{\frac{\delta}{4}}(S,S,U)| \tag{58}$$

We know that:

$$\overline{\dim}_B^{d_S}(\mathcal{W}_{S,U})=\limsup_{\delta\to 0}\frac{|N_{\frac{\delta}{4}}(S,S,U)|}{\log(4/\delta)}.$$

If we fix $\epsilon,\gamma>0$, we can apply Egoroff's Theorem to write that with probability $1-\gamma$, we have for δ small enough:

$$|N_{\frac{\delta}{4}}(S,S,U)|\leq\big(\epsilon+\overline{\dim}_B^{d_S}(\mathcal{W}_{S,U})\big)\log(4/\delta)$$

Therefore, we can
say that with probability $1-\eta-\gamma$, we have for δ small enough:

$$\begin{aligned}\sup_{w\in\mathcal{W}_{s,U}}|\mathcal{R}(w)-\hat{\mathcal{R}}_S(w)|\leq{}&\delta+\frac{B}{K}+\frac{2J\beta}{n^\alpha}\\ &+H\sqrt{\frac{JB^2}{2n^2}\Big(\big(\epsilon+\overline{\dim}_B^{d_S}(\mathcal{W}_{S,U})\big)\log(4/\delta)+\log(HK/\eta)+I_\infty(S,\mathcal{W}_{S,U})\Big)}\end{aligned}\tag{59}$$

Setting $K=\lfloor\sqrt{n}\rfloor$ and noting that $1-\alpha/3\leq 1$ in the above equation gives us the result.

□

Corollary 3. *With the exact same setting than in proposition 15, if we assume in addition that* $n^{\alpha/3}\in$

$\mathbb{N}_+$*, then for all* $n \geq 2^{\frac{3}{3-2\alpha}}$*, with probability* $1-\gamma-\eta$*, for all* δ *smaller than some* $\delta_{\gamma,\epsilon,n} > 0$ *we have:*

$$\begin{aligned}\sup_{w\in\mathcal{W}_{s,U}} |\mathcal{R}(w) - \hat{\mathcal{R}}_S(w)| \leq & \delta + \frac{B+2\beta}{n^{\alpha/3}} \\ & + B\sqrt{\frac{\log(1/\eta) + \left(1-\frac{\alpha}{3}\right)\log(n) + I + \big(\epsilon + d(S,U)\big)\log(4/\delta)}{2n^{\frac{2\alpha}{3}}}}\end{aligned}$$

Proof. We want to write J in the form $J = n^\lambda$ with some $\lambda > 0$. We see that there is a trade-off to be solved in the values of (J,H) if we want both all terms in equation (59) to have the same order of magnitude in n, which leads to $H\sqrt{J}/n = J/n^\alpha$. Therefore we want to have $1/\sqrt{J} = J/n^\alpha$ and $\lambda/2 = \alpha - \lambda$, which implies the following important formula:

$$\lambda = \frac{2\alpha}{3} \tag{60}$$

Finally, we are left again with choosing the value of K, an obvious choice is $K = n^{\alpha/3} \in \mathbb{N}_+$ to get the same order of magnitude. Thus we get the final result: for $n \geq 2^{\frac{3}{3-2\alpha}}$, with probability $1-\gamma-\eta$, for all δ smaller than some $\delta_{\gamma,\epsilon,n} > 0$ we have:

$$\begin{aligned}\sup_{w\in\mathcal{W}_{s,U}} |\mathcal{R}(w) - \hat{\mathcal{R}}_S(w)| \leq & \delta + \frac{B+2\beta}{n^{\alpha/3}} \\ & + B\left\{\frac{\log(1/\eta) + \left(1-\frac{\alpha}{3}\right)\log(n) + I + \big(\epsilon + d(S,U)\big)\log(4/\delta)}{2n^{\frac{2\alpha}{3}}}\right\}^{\frac{1}{2}}\end{aligned} \tag{61}$$

□

Remark 19. *The asymptoticity in* δ *defined by* $\delta_{\gamma,\epsilon,n}$ *above accounts for the asymptoticity coming both from the stability assumption (definition 13) and the convergence of the limit defining the upper box-counting dimension.*

Now we prove Theorem 5.4 which is based on the same idea than the previous corollary, but when $n^{\alpha/3} \notin \mathbb{N}$. The proof is as follows:

Proof. We define $J := \lfloor n^{2\alpha/3\rfloor}$, $J := \lfloor n^{1-2\alpha/3} \rfloor$ and $\tilde{n} := JH$. We obviously have $\tilde{n} \leq n$.
Using the boundedness assumption we have:

$$|\hat{\mathcal{R}}_S(w) - \mathcal{R}(w)| \leq \frac{n-\tilde{n}}{n}B + \frac{\tilde{n}}{n}\left|\frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}} \ell(w,z_i) - \mathcal{R}(w)\right| \tag{62}$$

For the first term we write:

$$\frac{n-\tilde{n}}{n}B \leq \frac{n - \big(n^{2\alpha/3}-1\big)\big(n^{1-2\alpha/3}-1\big)}{n} = \frac{n^{2\alpha/3} + n^{\alpha/3} - 1}{n} \leq \frac{2B}{n^{\alpha/3}}$$

The idea is to apply the proof of Theorem 15 to the last term of equation (62), replacing d_{S_n} with $d_{S_{\tilde{n}}}$. For clarity we still denote $S = (z_1, \dots, z_n)$ and $S_{\tilde{n}} = (z_1, \dots, z_{\tilde{n}})$
There are several terms we need to consider:
The mutual information term: The two data processing inequality we apply to prove equation (57) still apply so we can still write $I_\infty(S, \mathcal{W}_{S,U})$ in the bound.
Dimension term: Let us denote by $d(S,S',U)$ the upper-box dimension of $\mathcal{W}_{S,U}$ for pseudo-metric $d_{S'}$. Using the same reasoning than equation (58), we have:

$$|N_\delta(S, S_{\tilde{n}}, U)| \leq |N_{\delta\frac{\tilde{n}}{n}}(S,S,U)|$$

We have:

$$\delta\frac{\tilde{n}}{n} \geq \delta\frac{\big(n^{2\alpha/3}-1\big)\big(n^{1-2\alpha/3}-1\big)}{n} \geq \delta\left(1 - \frac{1}{n^{2\alpha/3}}\right)$$

And therefore, once we have $n \geq 2^{\frac{3}{2\alpha}}$ we have:

$$|N_\delta(S, S_{\tilde{n}}, U)| \leq |N_{\frac{\delta}{2}}(S,S,U)|,$$

which implies:

$$d(S, S_{\tilde{n},U} \leq d(S,S,U).$$

Terms in n: Now we look at equation (59), where we have 4 types of term in n which are of the form:

- $1/K$
- $H\sqrt{J}/n$
- $\sqrt{\log(HK)}H\sqrt{J}/n$
- J/n^α

We do not forget that we also have to multiply those terms by the factor $\tilde{n}/n$ coming from equation (62). Setting $K := \lfloor 1+\sqrt{J} \rfloor$ we get successively:

$$\frac{\tilde{n}}{n}\frac{1}{K} \leq \frac{1}{n^{\alpha/3}}, \quad \frac{\tilde{n}}{n}H\sqrt{J}/n \leq \frac{1}{n^{\alpha/3}}, \quad \frac{\tilde{n}}{n}J/n^\alpha \leq \frac{1}{n^{\alpha/3}}.$$

For the logarithmic term we have:

$$\log(HK) \leq \log(2\sqrt{J}n^{1-2\alpha/3}) \leq \log(2n^{1-\alpha/3})$$

Moreover, if $n \geq 2^{\frac{3}{2\alpha}}$ we have:

$$\tilde{n} \geq \left(n^{2\alpha/3} - 1\right)\left(n^{1-2\alpha/3}\right) \geq n/2.$$

Therefore the condition $\tilde{n} \geq 2^{\frac{3}{3-2\alpha}}$ is implied by $n/2 \geq 2^{\frac{3}{3-2\alpha}}$. So now the condition on n becomes:

$$n \geq C(\alpha) := \max\{2^{\frac{3}{2\alpha}}, 2^{1+\frac{3}{3-2\alpha}}\} \tag{63}$$

Putting all of this together, we get that for $n \geq C(\alpha)$ (defined in equation (63)), with probability $1-\gamma-\eta$, for all δ smaller than some $\delta_{\gamma,\epsilon,n} > 0$ we have:

$$\begin{aligned} \sup_{w \in \mathcal{W}_{s,U}} |\mathcal{R}(w) - \hat{\mathcal{R}}_S(w)| \leq & \delta + \frac{3B+2\beta}{n^{\alpha/3}} \\ & + B\sqrt{\frac{\log(1/\eta) + \left(1-\frac{\alpha}{3}\right)\log(n) + I + \left(\epsilon + d(S,U)\right)\log(4/\delta)}{2n^{\frac{2\alpha}{3}}}} \end{aligned} \tag{64}$$

□

While Assumption 3 might be restrictive, our goal here is to highlight how such geometric regularity can help us deal with the statistical dependence between the data and the hypothesis set.

Note that the mutual information term appearing in Theorem 5.4 is much more interpretable compared to the corresponding terms in Theorem 5.3, and has the exact same form as the term presented in (Hodgkinson *et al.*, 2022).

We also note that, this way of controlling the dependence between the data and the hypothesis set comes at the expense of potentially losing in the convergence rate of our bound. More precisely, for a stability index of α, we get a convergence rate of $n^{-\alpha/3}$. By examining the value of constant n_α in Theorem 5.4, we observe that getting closer to an optimal rate ($\alpha \approx \frac{3}{2}$) implies a larger n_α, rendering our bound asymptotic.

6 Computational aspects and experimental results

In this section, we will illustrate how the proposed data-dependent dimension can be numerically computed, by making a rigorous connection to topological data analysis (TDA) tools Boissonat *et al.* (2018). This is achieved in Section 6.2 by extending the algorithm proposed by Birdal *et al.* (2021) to our pseudo-metric case. Their method is based on results on 'persistent homology' that we will briefly introduced in Section 6.1. For a more details introduction, the interested reader is invited to consult (Boissonat *et al.*, 2018; Memoli and Singhal, 2019).

We will then use those notions to numerically evaluate the correlation between the proposed data-dependent intrinsic dimension and the generalization error and compare it to the fractal dimensions proposed in (Şimşekli *et al.*, 2021; Birdal *et al.*, 2021), in terms of correlation statistics, including the recently introduced 'granulated Kendall's coefficients' (Jiang *et al.*, 2019).

6.1 Persistent homology

Persistent homology (PH) is a well known notion in TDA typically used for point cloud analysis Edelsbrunner and Harer (2010); Carlsson (2014). Previous works have linked neural networks and algebraic topology Rieck *et al.* (2019); Pérez-Fernández *et al.* (2021), especially in Corneanu *et al.* (2020) who established experimental evidence of a link between homology and generalization. Important progress was made in Birdal *et al.* (2021), who used PH tools to estimate the upper-box counting dimension induced by the Euclidean distance on $\mathcal{W}_{S,U}$. In this subsection, we introduce a few necessary PH tools to understand this approach.

Throughout this subsection we consider a finite set of point $W \subset \mathbb{R}^m$. We will denote by $\mathbb{K}$ the unique two elements field $\mathbb{Z}/2\mathbb{Z}$.

Definition 14 (Abstract simplicial complex and filtrations)**.** *Given a finite set V, an abstract simplicial complex (which we will often refer simply as complex) K is a subset of $\mathcal{P}(V)$ the subsets of V such that:*

- $\forall v \in V,\ \{v\} \in K$
- $\forall s \in K, \mathcal{P}(s) \subseteq K$

The elements of K are called the simplices. For any non-empty simplex s, we call the number $|s| - 1$ its dimension*, denoted* $\dim(s)$*. Given a simplicial complex K, a filtration of K is a sequence of sub-complexes increasing for the inclusion $\emptyset \subset K^0 \subset \cdots \subset K^N = K$ such that every complex is obtained by adding one simplex to to the previous one: $K^{i+1} = K^i \cup \{\sigma^{i+1}\}$. Thus a filtration of a complex induces an ordering on the simplices, which will be denoted $(s^i)_i$ by convention.*

An example of simplicial complex is provided on Figure 1. A filtration will be denoted by

$$\emptyset \longrightarrow K^0 \longrightarrow \cdots \longrightarrow K^N = K,$$

and the corresponding simplices, in the order in which they are added to the filtration, will typically be denoted by $(s_0, \ldots, s_N)$.

Example 11. The most important filtration that we shall encounter is the *Vietoris-Rips filtration* (VR filtration) $\mathbf{Rips}(W)$. For any $\delta > 0$ we first construct the Vietoris-Rips simplicial complex $\mathbf{Rips}(W, \delta)$ by the following condition:

$$\forall k \{w_1, \ldots, w_k\} \in \mathbf{Rips}(W, \delta) \iff \forall i, j,\ d(p_i, p_j) \leq \delta. \tag{65}$$

Then $\mathbf{Rips}(W)$ is formed by adding the complexes in the increasing order of δ from 0 to $+\infty$. Complexes with the same value of δ are ordered based on their dimension and ordered arbitrarily if they have the same dimension. See Figure 1.

Remark 20. *There exist other natural filtrations, notably the Cech filtration. However VR filtration and Cech filtration are equivalent to compute persistent homology of degree* 0*, which is our main interest here. Therefore all the filtrations we will consider will be VR filtrations for better clarity.*

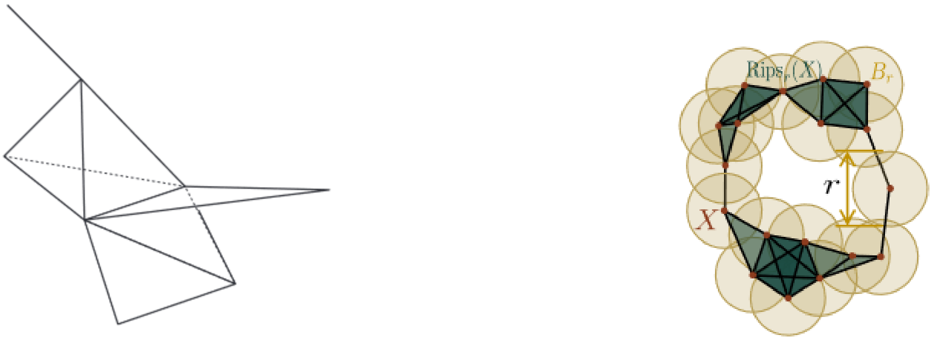


Figure 1: **Left:** Example of simplicial complex, in 3D space, figure from (Boissonat *et al.*, 2018). **Right:** Example of Vietoris-Rips complex, figure from (Birdal *et al.*, 2021).

Intuitively, Persistent homology of degree i keeps track of lifetimes of 'holes of dimension i', it is built over the concept of chains, which are a sort of linearized version of sets of simplices. More precisely, the space of k-chains $C_k(K)$ over complex K is defined as the set of formal linear combinations of the k-dimensional simplices of k:

$$C_k(K) := \mathrm{span}\Big(\sum_i \epsilon_i s_i,\ \forall i,\ \dim(s_i) = k\Big). \tag{66}$$

We will denote a simplex by its points $s = [w_0, \dots, w_k]$ and use the notation $s_{\setminus i} := [w_0, \dots, w_{i-1}, w_{i+1}, \dots, w_k]$. The *boundary operator* $\partial : C_k(K) \longrightarrow C_{k-1}(K)$ is the linear map induced by the relations on the simplices:

$$\partial(s) = \sum_{i=0}^{k} s_{\setminus i}. \tag{67}$$

Example 12. Consider a 'triangle', i.e. a simplex with 3 points $[a, b, c]$. Then the boundary operator, by Equation (67) gives us

$$\partial[a, b, c] = [a, b] + [a, c] + [b, c],$$

which corresponds to the boundary of the triangle, in the usual geometric sense. The above definition of the boundary operator is just a multidimensional extension of this simple fact.

It is easy to verify that $\partial^2 = 0$ and therefore we have an exact sequence, where $N = |W|$,

$$\{0\} \xrightarrow{\partial} C_N(K) \xrightarrow{\partial} C_{N-1}(K) \xrightarrow{\partial} \dots \xrightarrow{\partial} C_0(K) \xrightarrow{\partial} \{0\},$$

from which it is natural to define:

Definition 15 (Cycles and homology groups)**.** *With the same notations, we define:*

- *The k cycles of K:* $Z_k(K) := \ker(\partial : C_k(K) \longrightarrow C_{k-1}(K))$.
- *The k-th boundary of K:* $B_k(K) := \Im(\partial : C_{k+1}(K) \longrightarrow C_k(K))$.
- *k-th homology group (it is actually a quotient vector space):* $H_k(K) := Z_k/B_k$.

The k-th Betti *number of K is defined as the dimension of the homology group:* $\beta_k = \dim(H_k(K))$.

Those Betti numbers, β_k, correspond, in our analogy, to the number of holes of dimension k, i.e. the numbers of cycles whose 'interior' is not in the complex, and therefore corresponds to a hole.

Remark 21. *In particular, β_0 corresponds to the number of connected components in the complex.*

Now that we defined the notion of homology, we go on with the definition of *persistent homology* (PH). The intuition is the following: when we build the Vietoris-Rips filtration of the point cloud W, by increasing parameter δ in Example 11, we collect the 'birth' and 'death' of each hole, the multiset[8] of those pairs (*birth*, *death*) will be the definition of persistent homology.

[8] By multiset, we mean that it can contain several time the same element, in our case the same persistence pair.

Remark 22. *In all the following, the parameter δ used in the definition of the Vietoris-Rips filtration will be seen as a time parameter.*

Construction of persistence pairs: Let us consider the VR filtration of the set W, which we denote by the ordering of simplices with the same notations than in definition 14, so that the filtration is represented by the sequence of simplices $(s^i)_i$. We call s^i *positive* if it is part of a $\dim(s^i)$-cycle, otherwise it is *negative*. Persistence homology is based on an inductive construction of the basis of the homology groups associated to the filtration $H_k(K^i)$, two things can happen when we add a simplex s^i in the filtration (see (Boissonat *et al.*, 2018, Chapters 11, 12) for more details):

- s^i **is positive**: Then s^i is added in the basis of $H_k(K^i)$, with $k = \dim(s^i)$.
- s^i **is negative**: Then we denote $K+1 = \dim(s^i)$ and $(e_{i_1,\dots,e_{i_p}})$ a basis of $H_k(K^{i-1})$. We can show that it is possible to write:
$$\partial s^j = \sum_j \epsilon_j e_{i_j} + B,$$
With B a boundary and not all coefficients are 0. Therefore, selecting one of them arbitrarily, for example $\ell = \operatorname{argmax}\{j,\ \epsilon_j = 1\}$, we have to remove c_{i_j} from the basis of $H_k(K^{i-1})$.

Therefore positive increment the basis of the homology groups (they create holes) while negative simplicies destroy one element of the basis (they fill holes). With the same notation we say that b is the birth of element e_{i_j} while $d = i$ is its death. This way we build a sequence of *persistent pairs*. Those persistent pair are a characteristic of the complex and the multiset of lifetimes $(d - b)$ is independent of the ordering of the VR filtration or any arbitrary choice made in the construction.

Definition 16 (Persistent homology)**.** *The* persistent homology *of degree k, denoted PH^k is the multiset of all lifetime $\delta(death) - \delta(birth)$ obtained when constructing the basis of H_k, where δ is the length defining the VR complexes (equation 65).*

Persistent homology of degree 0: The persistent homology of degree 0, which is of primary interest for us, is actually very simple to describe. When constructing the VR filtration of W, all zero dimensional simplices (single points) are first added in an arbitrary order, they are obviously all positive and their equivalence class are added in the basis of H_0. Thus all of them have a birth δ equal to 0. Then each time a 1 dimensional simplex s^i is added:

- If s^i is positive: then it will impact only the basis H_1, so we don't care about it.
- Otherwise it means that $s^i =: [a, b]$ *destroys one connected component* by connecting two previously unconnected component. Therefore a and b have the same class in $H_0(K^i)$. Selecting arbitrary a or b, we define its death as the current δ.

Remark 23. *We see that in the case of $PH^0(\boldsymbol{Rips}(W))$, it would suffice to collect the multiset of death times of connected components.*

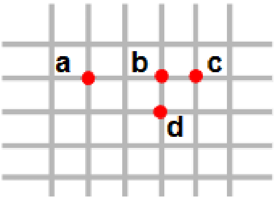


Figure 2: Simple example to illustrate $\mathrm{PH}^0(\mathbf{Rips}(W))$.

Example 13. Consider the point cloud $W = \{a, b, c, d\}$ represented on Figure 2, where the grid has unit length 1. Then we have $\mathrm{PH}^0(\mathbf{Rips}(W)) = \{\{(0,1),\ (0,1),\ (0,2)\}\}$, with multiset been denoted by $\{\{\cdot\}\}$.

Definition 17 (Persistent homology dimension)**.** *For any $\alpha \geq 0$ we define:*

$$E_\alpha(W) := \sum_{(b,d) \in PH^0(\boldsymbol{Rips}(W))} (d - b)^\alpha. \tag{68}$$

The persistent homology dimension of degree 0 *(PH dimension) of any set bounded metric space* $\mathcal{W}$ *is then defined as:*

$$\dim_{PH^0}(\mathcal{W}) := \inf\{\alpha > 0,\ \exists C > 0,\ \forall W \subset \mathcal{W} \textit{ finite},\ E_\alpha(W) < C\}.$$

Where the definition of VR filtration in finite subsets of metric spaces is naturally defined.

The importance of this dimension for our work relies in the following result (see Schweinhart (2019), Kozma *et al.* (2005)):

Proposition 16. *For any bounded metric space* X*, we have* $\overline{\dim}_B(X) = \dim_{PH^0}(X)$*.*

Proposition 16 opens the door to the numerical estimation of the upper-box dimension. Indeed, PH can be evaluated via several libraries (Bauer, 2021; Pérez *et al.*, 2021), moreover, Birdal *et al.* (2021) noted that, while Definition 17 is impossible to evaluate in practice, it can be approximated from $\mathrm{PH}^0(\mathbf{Rips}(W))$ computed on a finite number of finite subsets of the point cloud $\mathcal{W}$.

6.2 Numerical estimation of the data-dependent fractal dimension

In this subsection, we will describe how we can use the aforementioned algorithm form Birdal *et al.* (2021) to numerically approximate our data-dependent fractal dimension $\overline{\dim}_B^{\rho_S}(\mathcal{W}_{S,U})$.

As our dimension is defined through a (bounded) pseudo-metric, we must first prove that Proposition 16 extends to this setting. The key ingredient is the following Proposition, which states that the persistent homology dimension is invariant under metric identification (see Definition 10 for notations).

Proposition 17 (Persistent homology dimension in pseudo metric spaces)**.** *Consider a bounded pseudo-metric space* (X, ρ)*, then we have the equality:*

$$\dim_{PH^0}(X) = \dim_{PH^0}(X/\sim).$$

Intuitively, the proof of this result is as follows: When constructing the VR filtration in a pseudo-metric space, points within 0 pseudo-distance will only add pairs of the form $(0,0)$ in their persistence homology of degree 0, because they are created with the same value of the distance parameter δ in construction of the VR filtration.

Proof. Let K be a simplicial complex based on a finite point set $T \in X$. Let us denote by $\tilde{K} := \pi(K)$ the image of K by the canonical projection $\pi : X \longrightarrow X/\sim$, defined by its value on the simplices:

$$\pi([a_0, \dots, a_s]) := [\pi(a_0), \dots, \pi(a_s)]. \tag{69}$$

We also introduce a *section* of π, i.e. an injective application $s : X/\sim \longrightarrow X$, such that $\pi \circ s = \mathrm{Id}_{X/\sim}$. Clearly, $\tilde{K}$ is still a simplicial complex.The map π does not preserve the dimension of the simplices, as $[\pi(a_0), \dots, \pi(a_s)]$ is seen as a set, and two a_i can have the same image, but π always reduces the dimension.
Note that $\tilde{K}$ and $s(\tilde{K})$ clearly define simplicial complex, but that $s(\tilde{K})$ can only be seen as a sub-complex of K. Therefore, we define $s : \tilde{K} \longrightarrow K$ analogously to Equation (69). Actually, by injectivity of s, this allows us to identify $\tilde{K}$ with a sub-complex of K.
Thus, both π and s linear maps on the space of k-chains:

$$\pi : C_k(K) \longrightarrow C_k(\tilde{K}), \quad s : C_k(\tilde{K}) \longrightarrow C_k(K),$$

which both commute with the boundary operator, indeed, for any simplex $[a_0, \dots, a_s]$ and $\epsilon_i \in \mathbb{K}$:

$$\begin{aligned}\pi \circ \partial([a_0, \dots, a_s]) &= \pi\left(\sum_{i=0}^{s} \epsilon_i [a_0, \dots, a_{i-1}, a_{i+1}, \dots, a_s]\right) \\ &= \sum_{i=0}^{s} \epsilon_i [\pi(a_0), \dots, \pi(a_{i-1}), \pi(a_{i+1}), \dots, \pi(a_s)] \\ &= \partial \circ \pi([a_0, \dots, a_s]),\end{aligned}$$

with the exact same computation for s, so that the following diagram commutes:

$$\begin{tikzcd}
C_1(K) \arrow[r, "\partial"] \arrow[d, bend left, "\pi"] & C_0(K) \arrow[r, "\partial"] \arrow[d, bend left, "\pi"] & \{0\} \arrow[d, shift left] \\
C_1(\tilde{K}) \arrow[r, "\partial"] \arrow[u, bend left, "s"] & C_0(\tilde{K}) \arrow[r, "\partial"] \arrow[u, bend left, "s"] & \{0\} \arrow[u, shift left]
\end{tikzcd}$$

Therefore, π and s induces linear maps between the homology groups, making the following diagram commute:

$$\begin{tikzcd}
C_0(K) \arrow[r, bend left, "\pi"] \arrow[d] & C_0(\tilde{K}) \arrow[l, bend left, "s"] \arrow[d] \\
H_0(K) \arrow[r, bend left, "\bar{\pi}"] & H_0(\tilde{K}) \arrow[l, bend left, "\bar{s}"]
\end{tikzcd}$$

Now let us consider $P = \{x_1, \dots, x_n\}$ a finite set in (X, ρ) and denote accordingly $\tilde{P} := \pi(P)$. Let us introduce a Vietoris-Rips filtration of P denoted by:

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

where $0 \leq \delta_1 < \dots < \delta_C$ are the 'time-distance' indices of the filtration and for the same value of δ the simplices are ordered by their dimension and arbitrarily if they also have the same dimension. Obviously $\delta_0 = 0$.
As $\pi : P \longrightarrow \tilde{P}$ preserves distances, it is clear that, up to allowing certain complexes to appear several times in a row, the nested sequence $(\tilde{K}^{i,j})_{(0 \leq i \leq C, 1 \leq j \leq \alpha_i)}$ is a Vietoris-Rips filtration for $\tilde{P}$.
Let us fix some $i \in 0, \dots, C$ and $j \in 1, \dots, \alpha_i$ such that either $i \leq 1$ or $j = \alpha_0$. This way we have:

$$\forall a, b \in P,\ \pi(a) = \pi(b) \implies [a, b] \in K^{i,j},$$

by definition of the VR filtration (all simplices within $\delta_0 = 0$ ρ-distance have been added in the filtration). Therefore, if $\pi(a) = \pi(b)$, as $\partial[a, b] = [a] + [b]$, we have that $\overline{[a]} = \overline{[b]}$ in $H_0(K^{i,j})$. As be definition of s, for any $a \in P$ we have $\pi \circ s \circ \pi(a) = \pi(a)$, we have the following identity (the bars denote classes in homology groups):

$$\bar{s} \circ \bar{\pi}([a]) = \overline{s \circ \pi([a])} = \overline{[a]}.$$

Therefore, as also $\pi \circ s = \mathrm{Id}$, we have that $\bar{s}$ and $\bar{\pi}$ are inverse of one another, so that we have an isomorphism $H_0(K^{i,j}) \cong H_0(\tilde{K}^{i,j})$ and the following diagram:

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

As already mentioned, persistent homology of degree 0 is characterized by the multi-set of 'death times' δ_i. All death before K^{0,α_0-1} are 0 so they do not add anything the weighted life-sum of Equation (68). After K^{0,α_0-1}, the isomorphisms in the diagram show that the basis will evolve exactly in the same way so the death times will be the same, therefore the weighted sum are the same in both spaces for any P. Therefore, by definition, we have the equality between the persistent homology dimension. □

Combining Proposition 17 with some results of Section 4.2, we get the main result of this Subsection:

Theorem 6.1

For any bounded pseudo-metric space X we have: $\overline{\dim}_B(X) = \dim_{\mathrm{PH}^0}(X)$.

Proof. By Lemma 2, we have that $\overline{\dim}_B(X) = \overline{\dim}_B(X/\sim)$ and by Proposition 4.2 we have $\dim_{\mathrm{PH}^0}(X) = \dim_{\mathrm{PH}^0}(X/\sim)$, the result follows immediately. □

Numerical estimation Let us now briefly discuss how we numerically estimate the persistent homology dimension, which is essentially the algorithm presented in Birdal *et al.* (2021) where we changed the distance, which implies that we must evaluate on all data points for the last iterates. See also Adams *et al.* (2020); Schweinhart (2020) for similar ideas.

The algorithm is based on the following result, proved by proposition 2 of Birdal *et al.* (2021) and proposition 21 of Schweinhart (2020): If X is a bounded metric space with $\Delta = \dim^d_{\mathrm{PH}^0}(X)$, then for all $\epsilon > 0$ and $\alpha \in (0, \Delta + \epsilon)$ there exists $D_{\alpha,\epsilon} > 0$ such that for all finite subset $X_n = \{x_1, \dots, x_n\}$ of X we have:

$$\log E_\alpha(X_n) \leq \log D_{\alpha,\epsilon} + \left(1 - \frac{\alpha}{\Delta + \epsilon}\right) \log(n). \tag{70}$$

Then we can perform an affine regression of $\log E_\alpha(X_n)$ with respect to $\log n$ and get a slope a. Moreover it is argued in Birdal *et al.* (2021) that the slope has good chance to be approximately the one appearing in Equation (70), which gives us $\Delta \simeq \frac{\alpha}{1-a}$.

Remark 24. *The aforementioned algorithm* ***works in pseudo metric spaces****. Indeed as we tried to explain formally in the proof of Proposition 17, PH^0 in a pseudo-metric space only add some zeros to the quantities E_α computed in its metric identifications. Therefore the above algorithm is approximating $\dim^{\rho_S}_{PH^0}(X/\sim)$ which is proven in lemma 17 to be equal to $\dim^{\rho_S}_{PH^0}(X)$. See those notations in the next subsection.*

All persistent homology computation presented here have been made with the package presented in Pérez *et al.* (2021), which allows us to use more points in our persistent homology computation, e.g. Birdal *et al.* (2021) was only using between 1000 points prior to convergence for AlexNet and 200 for the other experiments. In our work we use up to 8000 points, which may allow us to better capture the fractal behavior.

6.3 Correlation statistics

Now that we are actually able to approximate the intrinsic dimensions, it use natural to ask how we can assess the quality of a generalization bounds and most importantly compare generalization bounds between each other. Indeed we would like to compare our bounds to other work, especially those considering euclidean based fractal dimensions, as in Hodgkinson *et al.* (2022), Şimşekli *et al.* (2021) and Birdal *et al.* (2021).

Kendall's coefficient, initially introduced in Kendall (1938), is a well-known statistics to assess the co-monoticity of two observations, or rank correlation. It is usually denoted with letter τ.

If we consider $((g_i, d_i)_{1 \leq i \leq n})$ a sequence of observation of two random elements, in our case the generalization error g and the intrinsic dimension d. In our setting it is very likely that both (g_i) and (d_i) will have pairwise distinct elements and that ties would therefore have little impact on the analysis. Therefore we will assume it in our presentation to make it easier. To compute Kendall's τ coefficient, denoted $\tau((g_i)_i, (d_i)_i)$, we look at all the possible pairs of couples (g_i, d_i) and count 1 if they are ordered the same way and -1 otherwise. The coefficients is then normalized by the total number of pairs which is $\binom{n}{2}$. Therefore an analytical formula is:

$$\tau((g_i)_i, (d_i)_i) = \frac{1}{\binom{n}{2}} \sum_{i<j} \mathrm{sign}(g_i - g_j)\mathrm{sign}(d_i - d_j). \tag{71}$$

However, as highlighted in Jiang *et al.* (2019), vanilla Kendall's τ may fail to capture any notion of causality in the correlation. Indeed, in our experiments we vary several hyperparameters (e.g. learning rate L and batch size B), we want to somehow measure whether the observed correlation is due to the influence of a hyperparameter on both the generalization error and the persistent homology dimension computation.

To overcome this issue, we follow the approach of Jiang *et al.* (2019), whose authors introduced a notion of *granulated Kendall's coefficient*. Let Θ_L an Θ_B denote the (finite) set in which our two hyperparameters vary. We first compute τ coefficients when fixing (all but) one hyperparameter, and then average those coefficients to get the granulated Kendall's coefficients:

$$\psi_\eta := \frac{1}{|\Theta_B|} \sum_{b \in \Theta_B} \tau\big((g(\eta, b), d(\eta, b))_{\eta \in \Theta_L}\big), \quad \psi_B := \frac{1}{|\Theta_L|} \sum_{b \in \Theta_L} \tau\big((g(\eta, b), d(\eta, b))_{b \in \Theta_B}\big), \tag{72}$$

Where $g(\eta, b)$ and $d(\eta, b)$ denote the generalization and dimension obtained with learning-rate η and batch size b. We can then average those coefficients to get one numerical measure:

$$\mathbf{\Psi} := \frac{\psi_\eta + \psi_B}{2}. \tag{73}$$

Remark 25. *Of course this analysis extends to more than* 2 *hyperparameters, but most of our experiments used only learning-rate and batch size.*

We created Python scripts to compute those granulated Kendall's coefficients in the results presented in this work.

Spearman's rank correlation coefficient: For the sake of completeness, we record a third type of correlation statistics in our experiments. Given observations $(g_1, \dots, g_n)$ and $(d_1, \dots, d_n)$ as above, we first order them (for the usual order on $\mathbb{R}$) to get ranked observations $R(g) := (r_{g,1}, \dots, r_{g,n})$ and $R(d) := (r_{d,1}, \dots, r_{d,n})$. *Spearman's coefficient* ρ is computed as the statistical correlation between the ranked observations, namely:

$$\rho(g, d) = \frac{\operatorname{cov}(R(g), R(d))}{\sigma(R(g))\sigma(R(d))}.$$

6.4 Experimental analysis

6.4.1 Experimental setup

In our experiments, we follow the setting presented in Section 6.2, extending the one of Birdal *et al.* (2021), except that we replace the Euclidean metric with the pseudo-metric ρ_S to compute the PH dimension.

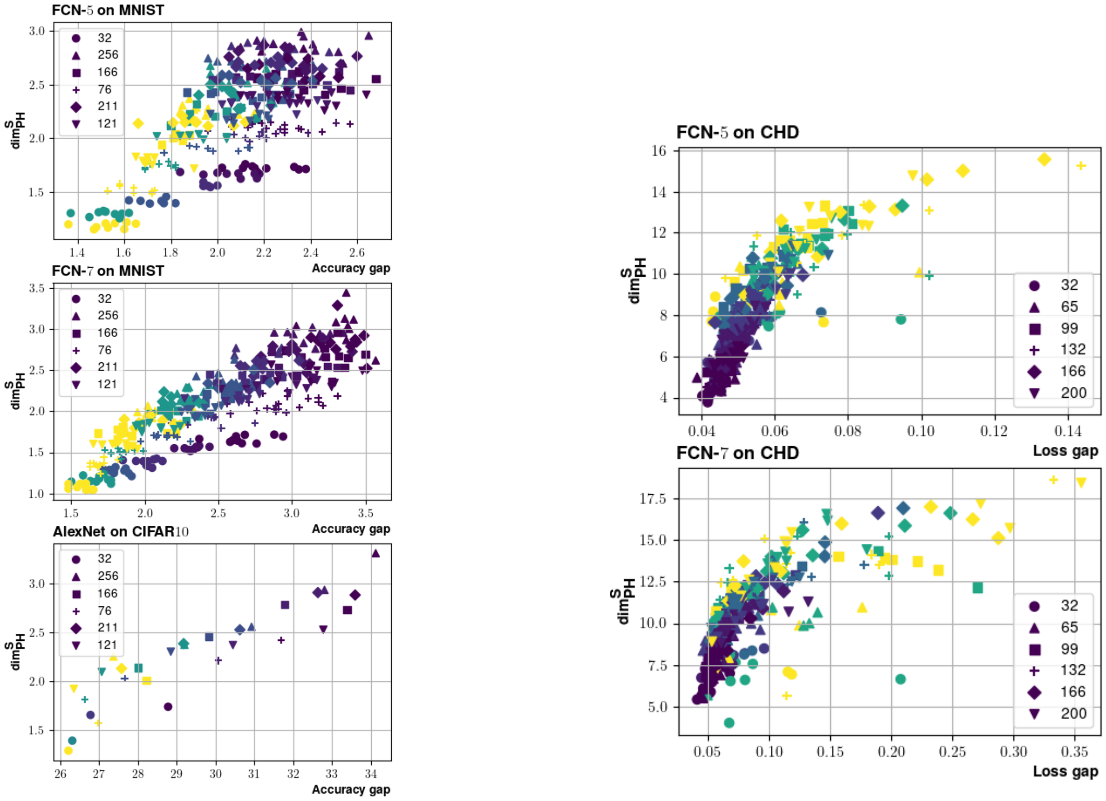
In particular, we consider learning a neural network by using SGD, and choose the hypothesis set $\mathcal{W}_{S,U}$ as the *optimization trajectory* near the local minimum found by SGD[9]. Then, we numerically estimate $\dim_{\mathrm{PH}^0}^{\rho_S}(\mathcal{W}_{S,U})$ by using the PH software provided in Pérez *et al.* (2021).

Here is a brief description of the method: given a neural network, its loss $\ell(w, z)$, and a dataset $S = (z_1, \dots, z_n)$, we compute the iterations of SGD for $K^\star$ iterations, $(w_k)_{k=0}^{K^\star}$, such that $w_{K^\star}$ reaches near a local minimum. We then run SGD for 5000 more iterations and set $\mathcal{W}_{S,U}$ to $\{w_{K^\star+1}, \dots, w_{K^\star+8000}\}$. We then approximate $\dim_{\mathrm{PH}^0}^{\rho_S}(\mathcal{W}_{S,U})$ by using the algorithm proposed in Birdal *et al.* (2021) by replacing the Euclidean distance with ρ_S, as presented in Section 6.2. Thanks to Theorem

We experimentally evaluate $\dim_{\mathrm{PH}^0}^{\rho_S}(\mathcal{W}_{S,U})$ in different settings: (i) regression experiment with Fully Connected Networks of 5 (FCN-5) and 7 (FCN-7) layers trained on the California Housing Dataset (CHD) Kelley Pace and Barry (1997), (ii) training FCN-5 and FCN-7 networks on the MNIST dataset Lecun *et al.* (1998) and (iii) training AlexNet Krizhevsky *et al.* (2017) on the CIFAR-10 dataset Krizhevsky *et al.* (2014). All the experiments use standard ReLU activation and vanilla SGD with constant step-size. We made both learning rate and batch size vary across a 6×6 grid. For experiments on CHD and MNIST we also used 10 different random seeds. Batch sizes vary between 32 and 256 while learning rates vary between 10^{-3} and 10^{-1}.

Note that in the case of a classification experiment, one could not compute $\dim_{\mathrm{PH}^0}^{\rho_S}$ using a zero-one loss in Equation (28). Indeed, it would be equivalent to computing PH on the *finite* set $\{0,1\}^n \subset \mathbb{R}^n$, which trivially gives an upper box-counting dimension of 0. To overcome this issue, we compute $\dim_{\mathrm{PH}^0}^{\rho_S}$

[9] Note that as the trajectories collected by SGD will only contain finitely many points, its dimension will be trivially 0. However, as in Birdal *et al.* (2021), we treat this finite set an approximation to the full trajectory. This is justified since even for infinite X, $\dim_{\mathrm{PH}^0}^{\rho_S}(X)$ is computed based on *finite* subsets of X.



(a) $\dim_{\mathrm{PH}^0}^{\rho_S}$ (denoted $\dim_{\mathrm{PH}^0}^{S}$ in the figure) versus accuracy gap for FCN-5 (*top*), FCN-7 (*middle*) on MNIST and AlexNet (*bottom*) on CIFAR-10.

(b) $\dim_{\mathrm{PH}^0}^{\rho_S}$ (denoted $\dim_{\mathrm{PH}^0}^{S}$ in the figure) versus generalization gap for FCN-5 (*top*) and FCN-7 (*bottom*) trained on CHD.

Figure 3: $\dim_{\mathrm{PH}^0}^{\rho_S}$ versus generalization error in various settings. Different colors indicate different learning rates and different markers indicate different batch sizes.

using the surrogate loss (cross entropy in our case) and illustrate that it is still a good predictor of the gap between the training and testing accuracies. For the sake of completeness, the behavior of $\dim_{\mathrm{PH}^0}^{\rho_S}$ with respect to the actual *loss gap* is reported in Figures 4a and 4b, even though it may be of no practical interest, as opposed to using the accuracy gap as in Figure 3a.

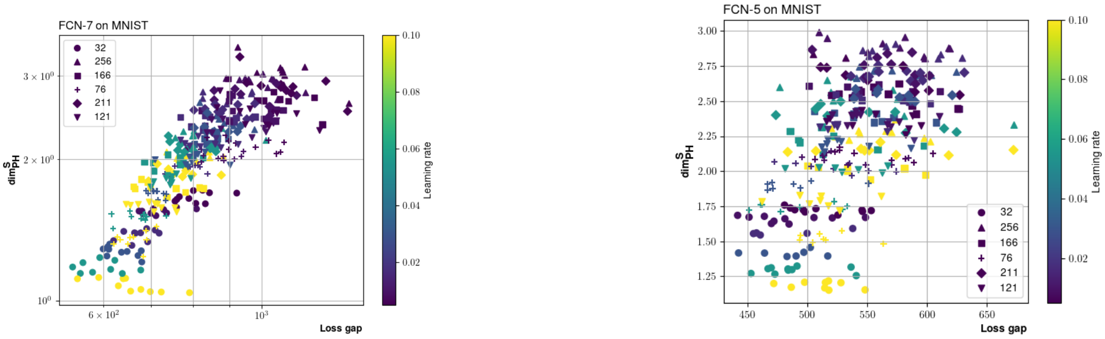
6.4.2 Numerical results

In order to compare our data-dependent intrinsic dimension with the one introduced in Birdal *et al.* (2021), which is the PH dimension induced by the Euclidean distance on the trajectory and denoted $\dim_{\mathrm{PH}^0}^{\mathrm{Eucl}}$, we compute various correlation statistics, namely the Spearman's rank correlation coefficient ρ Kendall and Stuart (1973) and Kendall's coefficient τ Kendall (1938). We also use the *mean Granulated Kendall's Coefficient* $\mathbf{\Psi}$ introduced in Jiang *et al.* (2019), which aims at isolating the influence of each hyperparameter and according to the authors could better capture the causal relationships between the generalization and the proposed complexity metric (the intrinsic dimension in our case). For more details on the exact computation of these coefficients, please refer to Section 6.3. Therefore $(\rho, \mathbf{\Psi}, \tau)$ are our main indicators of performance[10].

Figures 3a and 3b depict the data-dependent dimension versus the generalization gap, as computed in different settings. We observe that, in all cases, we have a strong correlation between $\dim_{\mathrm{PH}^0}^{\rho_S}(\mathcal{W}_{S,U})$ and the generalization gap, for a wide range of hyperparameters. Additional experiments, using bigger convolutional models and therefore computed with less experiments are presented in Figures 5a and 5b.

We also observe that the highest learning rates and smallest batch sizes seem to give less correlation, which is similar to what was observed in Birdal *et al.* (2021) as well. This might be caused by the increased

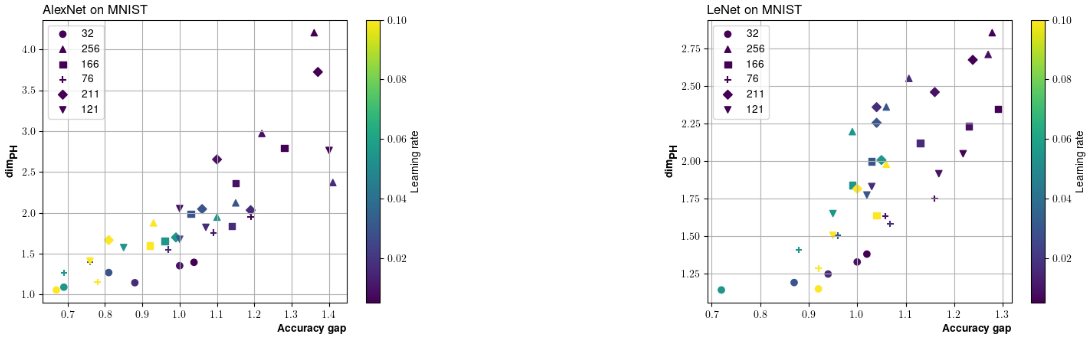
[10] All those coefficients are between −1 and 1, where the value of 1 indicates a perfect positive correlation.



(a) Plots of $\dim^{\rho_S}_{\mathrm{PH}^0}$ against the loss gap (as opposed to the accuracy gap) for a FCN-7 trained on MNIST dataset.

(b) Plots of $\dim^{\rho_S}_{\mathrm{PH}^0}$ against the loss gap (as opposed to the accuracy gap) for a FCN-5 trained on MNIST dataset.

Figure 4: $\dim^{\rho_S}_{\mathrm{PH}^0}$ versus loss gaps computed with the surrogate loss used during training (i.e. cross-entropy loss) for the same classification experiments than Figure 3a. Different colors indicate different learning rates and different markers indicate different batch sizes.



(a) Plots of $\dim^{\rho_S}_{\mathrm{PH}^0}$ against the accuracy gap for an AlexNet trained on MNIST dataset.

(b) Plots of $\dim^{\rho_S}_{\mathrm{PH}^0}$ against the accuracy gap for a LeNet trained on MNIST dataset.

Figure 5: $\dim^{\rho_S}_{\mathrm{PH}^0}$ versus accuracy gap for convolutional networks trained on the MNIST dataset.

noise as we suspect that the point clouds in those settings show more complex fractal structures and hence require more points for a precise computation of the PH dimension.

Next, correlation coefficients, for the same experiments, are reported in Tables 3, 1 and 4. The results show that on average our proposed dimension always yields improved metrics compared to the dimension introduced in Birdal *et al.* (2021).

Table 1: Correlation coefficients on MNIST

Model	Dim.	ρ	ψ_{LR}	ψ_{BS}	$\boldsymbol{\Psi}$	τ
FCN-5	$\dim^{\mathrm{EUCL}}_{\mathrm{PH}^0}$	0.62 ± 0.10	0.78 ± 0.07	$\mathbf{0.80 \pm 0.10}$	0.78 ± 0.08	0.47 ± 0.07
FCN-5	$\dim^{\rho_S}_{\mathrm{PH}^0}$	$\mathbf{0.73 \pm 0.07}$	$\mathbf{0.84 \pm 0.06}$	0.78 ± 0.10	$\mathbf{0.81 \pm 0.07}$	$\mathbf{0.56 \pm 0.06}$
FCN-7	$\dim^{\mathrm{EUCL}}_{\mathrm{PH}^0}$	0.80 ± 0.04	0.92 ± 0.07	$\mathbf{0.85 \pm 0.11}$	0.88 ± 0.04	0.62 ± 0.04
FCN-7	$\dim^{\rho_S}_{\mathrm{PH}^0}$	$\mathbf{0.89 \pm 0.02}$	$\mathbf{0.96 \pm 0.05}$	0.84 ± 0.05	$\mathbf{0.90 \pm 0.04}$	$\mathbf{0.73 \pm 0.03}$

The improvement is particularly better in the regression experiment we performed (as the classification task yields larger variations in the metrics, see Table 1). This may indicate that the proposed dimension may be particularly pertinent in specific settings. Moreover, increasing the size of the model, in all

Table 2: Correlation coefficients on MNIST, with respect to loss gap

Model	Dim.	ρ	ψ_{LR}	ψ_{BS}	$\boldsymbol{\Psi}$	τ
FCN-5	$\dim_{\text{PH}^0}^{\text{EUCL}}$	$\mathbf{0.76 \pm 0.06}$	$\mathbf{0.33 \pm 0.18}$	$\mathbf{0.75 \pm 0.09}$	$\mathbf{0.54 \pm 0.11}$	$\mathbf{0.58 \pm 0.05}$
FCN-5	$\dim_{\text{PH}^0}^{\rho_S}$	0.73 ± 0.09	0.30 ± 0.20	$\mathbf{0.75 \pm 0.09}$	0.52 ± 0.12	0.57 ± 0.07
FCN-7	$\dim_{\text{PH}^0}^{\text{EUCL}}$	0.86 ± 0.05	0.77 ± 0.12	$\mathbf{0.80 \pm 0.08}$	0.79 ± 0.06	0.69 ± 0.06
FCN-7	$\dim_{\text{PH}^0}^{\rho_S}$	$\mathbf{0.90 \pm 0.03}$	$\mathbf{0.80 \pm 0.10}$	0.79 ± 0.06	$\mathbf{0.80 \pm 0.06}$	$\mathbf{0.75 \pm 0.05}$

Table 3: Correlation coefficients on CHD

Model	Dim.	ρ	ψ_{LR}	ψ_{BS}	$\boldsymbol{\Psi}$	τ
FCN-5	$\dim_{\text{PH}^0}^{\text{EUCL}}$	0.77 ± 0.08	0.62 ± 0.11	0.46 ± 0.14	0.54 ± 0.11	0.59 ± 0.07
FCN-5	$\dim_{\text{PH}^0}^{\rho_S}$	$\mathbf{0.87 \pm 0.05}$	$\mathbf{0.75 \pm 0.10}$	$\mathbf{0.61 \pm 0.13}$	$\mathbf{0.68 \pm 0.10}$	$\mathbf{0.71 \pm 0.09}$
FCN-7	$\dim_{\text{PH}^0}^{\text{EUCL}}$	0.40 ± 0.09	0.07 ± 0.13	0.25 ± 0.11	0.16 ± 0.08	0.28 ± 0.07
FCN-7	$\dim_{\text{PH}^0}^{\rho_S}$	$\mathbf{0.77 \pm 0.08}$	$\mathbf{0.63 \pm 0.05}$	$\mathbf{0.58 \pm 0.10}$	$\mathbf{0.62 \pm 0.06}$	$\mathbf{0.77 \pm 0.08}$

experiments, seems to have a positive impact on the correlation. We suspect that this might be due to the increasing local-Lipschitz constant of the network.

The values of $\dim_{\text{PH}^0}^{\rho_S}$ against the actual loss gap (computed based on the cross entropy loss) are plotted in Figures 4b and 4a. The corresponding correlation statistics are reported in Table 2. While this has probably little practical interest compared to the plots shown in the main part of the paper, it highlights the fact that the correlation is indeed still there. As before, we note that low batch sizes and high learning rates yields better results, but that the correlation is very good for middle range values of those hyperparameters. As in the regression experiment, one can observe on Figures 4b and 4a that a bigger network gives better empirical correlation between the data-dependent dimension and the generalization error. Another interesting observation is that there seems to be more noise in the coefficients with respect to the loss gap than with respect to the accuracy gap. In almost all experiments, again, the proposed dimension is close or better than the one proposed in (Birdal *et al.*, 2021).

6.4.3 Robustness analysis

The computation of $\rho_S(w, w')$ requires the exact evaluation of the loss function on every data point $\{z_1, \dots, z_n\}$ for every $w, w' \in \mathcal{W}_{S,U}$. This introduces a computational bottleneck in case where n is excessively large. To address this issue, in this section we will explore an approximate way of computing $\dim_{\text{PH}^0}^{\rho_S}$. Similar to the computation of a stochastic gradient, instead of computing the distance on every data point, we will first draw a random subset of data points $T \subset S$, with $|T| \ll n$ and use the following approximation $\rho_S(w, w') \approx \rho_T(w, w') := \frac{1}{|T|} \sum_{z \in T} |\ell(w, z) - \ell(w', z)|$.

We now conduct experiments to analyze the robustness of the computation of $\dim_{\text{PH}^0}^{\rho_S}$ with respect to varying size of random subsets T. More precisely, we randomly select a subset $T \subset S$ whose size varies between 2% and 99% of the size dataset S and compute the PH dimension using the approximate pseudo-metric. Note that the whole dataset S is of course still used to produce the SGD iterates. Figure 6 presents results on the MNIST and CHD datasets in terms of the relative error, i.e., $|\dim_{\text{PH}^0}^{\rho_T} - \dim_{\text{PH}^0}^{\rho_S}| / \dim_{\text{PH}^0}^{\rho_S}$. The results show that the proposed dimension is significantly robust to the approximation of the pseudo-metric: even with 40% of the data, we achieve almost identical results as using the full dataset.

Table 4: Correlation coefficients with AlexNet on CIFAR-10

Model	Dim.	ρ	ψ_{LR}	ψ_{BS}	$\boldsymbol{\Psi}$	τ
AlexNet	$\dim_{\mathrm{PH}^0}^{\mathrm{EUCL}}$	0.86	0.78	**0.84**	0.81	0.68
AlexNet	$\dim_{\mathrm{PH}^0}^{\rho_S}$	**0.93**	**0.87**	0.81	**0.84**	**0.78**

Table 5: Correlation coefficients with convolutional models on MNIST

Model	Dim.	ρ	ψ_{LR}	ψ_{BS}	$\boldsymbol{\Psi}$	τ
AlexNet	$\dim_{\mathrm{PH}^0}^{\mathrm{EUCL}}$	0.85	**0.78**	**0.77**	**0.77**	0.67
AlexNet	$\dim_{\mathrm{PH}^0}^{\rho_S}$	**0.88**	**0.78**	**0.77**	**0.77**	**0.70**
LeNet	$\dim_{\mathrm{PH}^0}^{\mathrm{EUCL}}$	0.74	0.78	**0.77**	0.78	0.57
LeNet	$\dim_{\mathrm{PH}^0}^{\rho_S}$	**0.80**	**0.80**	**0.77**	**0.79**	**0.62**

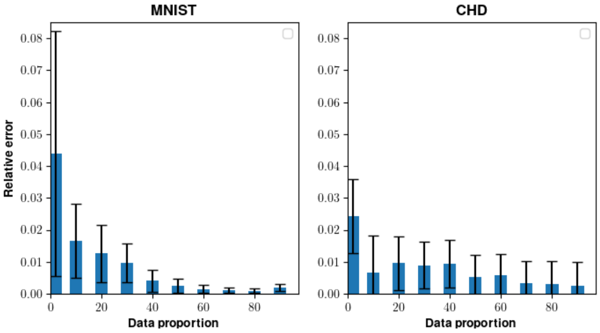


Figure 6: Robustness experiment using a FCNN trained on MNIST (*Left*) and CHD (*Right*). x-axis represents the proportion of the data T used to compute the metric, y-axis is the relative error with respect to the full dataset based dimension.

7 Conclusion

In this project, we proved generalization bounds that do not require the Lipschitz continuity of the loss, which can be crucial in modern neural network settings. We linked the generalization error to a data-dependent fractal dimension of the random hypothesis set. We first extended some classical covering arguments to state a bound in the case of a fixed hypothesis set and then proved a result in a general learning setting. While some intricate mutual information terms between the geometry and the data appeared in this bound, we presented a possible workaround by the introduction of a stability property for the coverings of the hypothesis set. Finally, we made a connection to persistent homology, which allowed us to numerically approximate the intrinsic dimension and thus support our theory with experiments.

Certain points remain to be studied concerning our results. First the existence of differentiable persistent homology libraries Hofer *et al.* (2018, 2019) open the door to the use of our intrinsic dimension as a regularization term as in Birdal *et al.* (2021). From a theoretical perspective, as in previous works, the main drawbacks of this work remain the presence of non-estimable mutual information terms. However, introducing those terms allowed us to display interesting proof techniques. We hope that future work may bring a better understanding on this issue, which could be achieved by leveraging stability notions as in Section 5.3 or by considering the convergence of the random pseudo-metric ρ_S to its expectation. Those research directions remain to be explored. Finally, our proof techniques could be refined, for example by using the celebrated chaining method Ledoux and Talagrand (1991); Clerico *et al.* (2022), to improve our theoretical results or weaken the assumptions.

Acknowledgments

I would like to thank my supervisor at INRIA, Umut Şimşekli, for helping to develop the ideas and writing this project , and George Deligiannidis for his useful advice. I also thank Clement Hongler for accepting to supervise this project at EPFL and allowing me to work on this subject as my Master Thesis.

References

Adams, H., Aminian, M., Farnell, E., Kirby, M., Peterson, C., Mirth, J., Neville, R., Shipman, P. and Shonkwiler, C. (2020) A fractal dimension for measures via persistent homology. *Topological Data Analysis, Abel Symposia, vol. 15* pp. 1–31.

Anthony, M. and Barlett, P. L. (1999) *Neural Network Learning: Theoretical Foundations.* Cambridge University Press.

Arora, S., Ge, R., Neyshabur, B. and Zhang, Y. (2018) Stronger generalization bounds for deep nets via a compression approach. *Proceedings of the 35th International Conference on Machine Learning* .

Asadi, A. R., Abbe, E. and Verdú, S. (2019) Chaining Mutual Information and Tightening Generalization Bounds. *Advances in Neural Information Processing Systems 31 (NeurIPS 2018)* .

Barlett, P. L. and Mendelson, S. (2002) Rademacher and Gaussian Complexities: Risk Bounds and Structural Result. *Journal of Machine Learning Research* .

Barsbey, M., Sefidgaran, M., Erdogdu, M. A., Richard, G. and Şimşekli, U. (2021) Heavy Tails in SGD and Compressibility of Overparametrized Neural Networks. *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)* .

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

Belkin, M., Hsu, D., Ma, S. and Mandal, S. (2019) Reconciling modern machine learning practice and the bias-variance trade-off. *Proceedings of the National Academy of Sciences* **116**(32), 15849–15854.

Birdal, T., Lou, A., Guibas, L. and Şimşekli, U. (2021) Intrinsic Dimension, Persistent Homology and Generalization in Neural Networks. *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)* .

Bogachev, V. I. (2007) *Measure Theory.* Volume Volume 1. Springer.

Boissonat, J.-D., Chazal, F. and Yvinec, M. (2018) *Geometrical and Topological Inference.* Cambridge Texts in Applied Mathematics. Cambridge University Press.

Boucheron, S., Lugosi, G. and Massart, P. (2013) *Concentration Inequalities - A Non-asymptotic Theory of Independence.* Oxford university press edition.

Bousquet, O. (2002) Stability and generalization. *Journal of Machine Learning Research* .

Bousquet, O., Klochkov, Y. and Zhivotovskiy, N. (2020) Sharper bounds for uniformly stable algorithms. *Proceedings of Thirty Third Conference on Learning Theory* .

Camuto, A., Deligiannidis, G., Erdogdu, M. A., Gürbüzbalaban, M., Şimşekli, U. and Zhu, L. (2021) Fractal Structure and Generalization Properties of Stochastic Optimization Algorithms. *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)* .

Carlsson, G. (2014) Topological pattern recognition for point cloud data*. *Acta Numerica* **23**, 289–368.

Chandramoorthy, N., Loukas, A., Gatmiry, K. and Jegelka, S. (2022) On the generalization of learning algorithms that do not converge. *Thirty-Sixth Conference on Neural Information Processing Systems (Neurips 2022)* .

Chaudhari, P. and Soatto, S. (2018) Stochastic gradient descent performs variational inference, converges to limit cycles for deep networks. *2018 Information Theory and Applications Workshop (ITA)* .

Clerico, E., Shidani, A., Deligiannidis, G. and Doucet, A. (2022) Chained Generalisation Bounds. *Proceedings of Thirty Fifth Conference on Learning Theory* .

Corneanu, C., Madadi, M., Escalera, S. and Martinez, A. (2020) Computing the Testing Error without a Testing Set. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* .

Dupuis, B., Deligiannidis, G. and Şimşekli, U. (2023) Generalization Bounds with Data-dependent Fractal Dimensions.

Edelsbrunner, H. and Harer, J. (2010) Computational Topology - an Introduction | Semantic Scholar. *American Mathematical Society* .

Falconer, K. (2014) *Fractal Geometry - Mathematical Foundations and Applications - Third Edition.* Wiley.

Foster, D. J., Greenberg, S., Kale, S., Luo, H., Mohri, M. and Sridharan, K. (2020) Hypothesis Set Stability and Generalization. *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)* .

Harutyunyan, H., Raginsky, M., Steeg, G. V. and Galstyan, A. (2021) Information-theoretic generalization bounds for black-box learning algorithms. *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)* .

Herrera, C., Krach, F. and Teichmann, J. (2020) Estimating Full Lipschitz Constants of Deep Neural Networks. *Estimating Full Lipschitz Constants of Deep Neural Networks* .

Hodgkinson, L., Şimşekli, U., Khanna, R. and Mahoney, M. W. (2022) Generalization Bounds using Lower Tail Exponents in Stochastic Optimizers. *Proceedings of the 39th International Conference on Machine Learning* .

Hoeffding, W. (1963) Probability Inequalities for Sums of Bounded Random Variables: Journal of the American Statistical Association: Vol 58, No 301. https://www.tandfonline.com/doi/abs/10.1080/01621459.1963.10500830.

Hofer, C., Kwitt, R., Dixit, M. and Niethammer, M. (2019) Connectivity-Optimized Representation Learning via Persistent Homology. *Proceedings of the 36th International Conference on Machine Learning* .

Hofer, C., Kwitt, R., Niethammer, M. and Uhl, A. (2018) Deep Learning with Topological Signatures. *Advances in Neural Information Processing Systems 30 (NIPS 2017)* .

Hu, W., Li, C. J., Li, L. and Liu, J.-G. (2018) On the diffusion approximation of nonconvex stochastic gradient descent. *Annals of Mathematical Sciences and Applications* .

Jastrzebski, S., Kenton, Z., Arpit, D., Ballas, N., Fischer, A., Bengio, Y. and Storkey, A. (2018) Three Factors Influencing Minima in SGD.

Jiang, Y., Neyshabur, B., Mobahi, H., Krishnan, D. and Bengio, S. (2019) Fantastic Generalization Measures and Where to Find Them. *ICLR 2020* .

Kechris, A. S. (1995) *Classical Descriptive Set Theory.* Graduate Texts in Mathematics. Springer.

Kelley Pace, R. and Barry, R. (1997) Sparse spatial autoregressions. *Statistics & Probability Letters* **33**(3), 291–297.

Kendall, M. G. (1938) A new reasure of rank correlation. *Biometrika* .

Kendall, M. G. and Stuart, A. (1973) *The Advanced Theory of Statistics.* Griffin. ISBN 978-0-85264-069-2.

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M. and Tang, P. T. P. (2017) On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima. *ICLR 2017* .

Kozma, G., Lotker, Z. and Stupp, G. (2005) The minimal spanning tree and the upper box dimension. *Proceedings of the American Mathematical Society* **134**(4), 1183–1187.

Krizhevsky, A., Nair, V. and Hinton, G. E. (2014) The cifar-10 dataset.

Krizhevsky, A., Sutskever, I. and Hinton, G. E. (2017) ImageNet classification with deep convolutional neural networks. *Communications of the ACM* **60**(6), 84–90.

Lecun, Y., Bottou, L., Bengio, Y. and Haffner, P. (1998) Gradient-based learning applied to document recognition. *Proceedings of the IEEE* **86**(11), 2278–2324.

Ledoux, M. and Talagrand, M. (1991) *Probability in Banach Spaces - Isoperimetry and Processes.* Classics in Mathematics. Springer.

Mandelbrot, B. (1982) *The Fractal Geometry of Nature.* W. H. Freeman and Co.

Mandt, S., Hoffman, M. D. and Blei, D. M. (2016) A Variational Analysis of Stochastic Gradient Algorithms. *Proceedings of The 33rd International Conference on Machine Learning* .

Massart, P. (2000) Some applications of concentration inequalities to statistics. *Annales de la Faculté des sciences de Toulouse: Mathématiques* .

Mattila, P. (1999) *Geometry of Sets and Measures in Euclidean Spaces.* Cambridge University Press.

Memoli, F. and Singhal, K. (2019) A Primer on Persistent Homology of Finite Metric Spaces. *Bulletin of Mathematical Biology* **81**(7), 2074–2116.

Molchanov, I. (2017) *Theory of Random Sets.* Second edition edition. Number 87 in Probability Theory and Stochastic Modeling. Springer.

Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B. and Sutskever, I. (2019) Deep Double Descent: Where Bigger Models and More Data Hurt. *ICLR 2020* .

Negrea, J., Haghifam, M., Dziugaite, G. K., Khisti, A. and Roy, D. M. (2019) Information-Theoretic Generalization Bounds for SGLD via Data-Dependent Estimates. *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)* .

Pensia, A., Jog, V. and Loh, P.-L. (2018) Generalization Error Bounds for Noisy, Iterative Algorithms. *2018 IEEE International Symposium on Information Theory (ISIT)* .

Pérez, J. B., Hauke, S., Lupo, U., Caorsi, M. and Dassatti, A. (2021) Giotto-ph: A Python Library for High-Performance Computation of Persistent Homology of Vietoris-Rips Filtrations.

Pérez-Fernández, D., Gutiérrez-Fandiño, A., Armengol-Estapé, J. and Villegas, M. (2021) Characterizing and Measuring the Similarity of Neural Networks with Persistent Homology. *CoRR* .

Pesin, Y. B. (1997) *Dimension Theory in Dynamical Systems - Contemporary Views and Applications.* Chicago Lectures in Mathematics. The University of Chicago Press.

Rebeschini, P. (2020) Algorithmic fundations of learning.

Rieck, B., Togninalli, M., Bock, C., Moor, M., Horn, M., Gumbsch, T. and Borgwardt, K. (2019) Neural Persistence: A Complexity Measure for Deep Neural Networks Using Algebraic Topology. *ICLR* p. 25 p.

Russo, D. and Zou, J. (2019) How much does your data exploration overfit? Controlling bias via information usage. *IEEE Transactions on Information Theory* .

Schilling, R. L. (1998) Feller Processes Generated by Pseudo-Differential Operators: On the Hausdorff Dimension of Their Sample Paths | SpringerLink. *Journal of Theoretical Probability* .

Schilling, R. L. (2016) An Introduction to levy and Feller Processes. Advanced Courses in Mathematics - CRM Barcelona 2014.

Schweinhart, B. (2019) Persistent Homology and the Upper Box Dimension. *Discrete & Computational Geometry volume 65, pages 331–364* .

Schweinhart, B. (2020) Fractal Dimension and the Persistent Homology of Random Geometric Complexes. *Advances in Mathematics* .

Şimşekli, U., Sener, O., Deligiannidis, G. and Erdogdu, M. A. (2021) Hausdorff Dimension, Heavy Tails, and Generalization in Neural Networks. *Journal of Statistical Mechanics: Theory and Experiment* **2021**(12), 124014.

Steinke, T. and Zakynthinou, L. (2020) Reasoning About Generalization via Conditional Mutual Information. *Proceedings of Thirty Third Conference on Learning Theory* .

Suzuki, T., Abe, H. and Nishimura, T. (2020) Compression based bound for non-compressed network: Unified generalization error analysis of large compressible deep neural network. *ICLR 2020* .

van Erven, T. and Harremoës, P. (2014) Renyi Divergence and Kullback-Leibler Divergence. *IEEE Transactions on Information Theory* **60**(7), 3797–3820.

Vapnik, V. N. and Chervonenkis, A. Ya. (2015) On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities. In *Measures of Complexity: Festschrift for Alexey Chervonenkis*, eds V. Vovk, H. Papadopoulos and A. Gammerman, pp. 11–30. Cham: Springer International Publishing. ISBN 978-3-319-21852-6.

Vershynin, R. (2020) *High-Dimensional Probability - An Introduction with Application in Data Science.* University of California - Irvine: .

Xiao, Y. (2004) Random fractals and Markov processes. *Fractal Geometry and Applications: A jubilee of Benoît Mandelbrot - American Mathematical Society* **72.2**, 261–338.

Xu, A. and Raginsky, M. (2017) Information-theoretic analysis of generalization capability of learning algorithms. *Advances in Neural Information Processing Systems 30 (NIPS 2017)* .

Zhang, C., Bengio, S., Hardt, M., Recht, B. and Vinyals, O. (2017) Understanding deep learning requires rethinking generalization. *ICLR 2017* .

Appendix A Notations

Nomenclature

$\dim_B^d$ Box counting dimension computed with (pseudo-)metric d.

$\ell : \mathbb{R}^d \times \mathcal{Z} \longrightarrow \mathbb{R}$ Model 'loss' function, understood as the composition of a loss and a parametric model.

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

$\dim_{\mathrm{H}}^d$ Hausdorff dimension computed with (pseudo-)metric d.

$\underline{\dim}_{\mathrm{B}}^d$ Lower box counting dimension computed with (pseudo-)metric d.

$\mathcal{B}_X$ Borel σ-algebra associated to a topological space X.

$\mathcal{H}^s$ s-Hausdorff measure.

$\mathcal{R}(\cdot)$ Population risk.

$\mathcal{W}_{S,U}$ Sample path generated by the learning algorithm.

$\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ Data space, made of features $\mathcal{X}$ and targets $\mathcal{Y}$.

μ_z Data probability distribution over $\mathcal{Z}$.

$\mathrm{diam}(U)$ Diameter of the set U.

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

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

$|A|$ Cardinal of A if A is a finite set.

N_δ^ρ Centers of a covering by closed δ-balls for (pseudo-)metric ρ, if ρ is omitted it corresponds to the euclidean distance.

$S = (z_1, \dots, z_n) \in \mathcal{Z}^n$ Dataset of i.i.d. random variables following the data distribution, i.e. $S \sim \mu_z^{\otimes n}$.

Appendix B Additional theoretical results

B.1 An expected bound using uniform hypothesis set stability

In this subsection, we present a slightly different bound than the ones presented in the main part of this document.

This bound uses the uniform hypothesis set stability assumption, introduced by Foster *et al.* (2020) and explicitly described by Equation (51). Moreover, we still make Assumption 1 of bounded continuous losses.

The particularity of this bound is that it involves an expected upper-box counting dimension, for which the pseudo-metric, still define by Equation (28), is computed using a *validation dataset* independent from the training dataset S and still following the data distribution μ_z. Moreover, using results from (Foster *et al.*, 2020), this bound can easily be turned into a high probability bound, thanks to the hypothesis set stability, but it would still involve the aforementioned expected dimension.

While this bound may have less practical interest than the ones presented in Section 4 and 5, we discuss it for the following reasons:

1. It does not involve **any mutual information term**.
2. The assumptions on which it relies are more realistic than the setting of Section 5.3.

3. It is a particularly interesting example of the *grouping technique* introduced in Section 5.3 and how it can help to deal with the statistical dependence between the hypothesis set and the data. We also want to highlight that this proof technique may be more general and extend to different type of results, for instance it could be used in the Lipschitz setting of Şimşekli *et al.* (2021).

Let us first restate the hypothesis set stability assumption from (Foster *et al.*, 2020), by changing to more suitable notations.

Assumption 4. *There exists $\beta > 0$ and $\alpha > 0$ such that for all $S, S' \in \mathcal{Z}^n$ differing only by one element, for all $u \in \mathcal{U}$, we have:*

$$\forall w \in \mathcal{W}_{S,U},\ \exists w' \in \mathcal{W}_{S',U},\ \forall z \in \mathcal{Z},\ |\ell(w,z) - \ell(w',z)| \leq \frac{\beta}{n^\alpha}$$

Remark 26. *Following Foster* et al. *(2020), one should think of parameter α as being $\alpha \simeq 1$. Regarding the following theorem, it means that we will probably lose in the convergence rate in n, but with a great benefit which is to dismiss any mutual information term.*

In the proof of this section, we will separate the set of indices $\{1, \dots, n\}$ into H groups of size J, as we did in Section 5.3. We will always assume that $HJ = n$, the discussion of what happens when this is not the case can be made by following exactly the same steps than in Section 5.3, leveraging the bounded loss assumption, the details are left to the reader to focus on the main proof ideas.

Theorem B.1. Expected bound without mutual information

We make assumptions 1, 2 and 4. We define $J = n^{2\alpha/3}$ and assuming $n^{\alpha/3}$ is an integer we consider $\tilde{S}_J \sim \mu_z^{\otimes J}$, independent of S.
Then for all $\delta, \gamma, \epsilon > 0$, there exists $\delta_{\gamma,\epsilon,n} > 0$ such that for all $\delta \leq \delta_{\gamma,\epsilon,n}$ we have the expected bound:

$$\mathbb{E}\Bigg[\sup_{w \in \mathcal{W}_{S,U}} \big(\mathcal{R}(w) - \hat{\mathcal{R}}_S(w)\big)\Bigg] \leq 2(\delta + B\gamma) + 2\sqrt{\frac{\beta^2 + 4B^2\big(\epsilon + \mathbb{E}\big[\overline{\dim}_B^{d_{\tilde{S}_J}}(\mathcal{W}_{S,U})\big]\big)\log(1/\delta)}{n^{\frac{2\alpha}{3}}}}$$

Proof. Let us introduce $\tilde{S} := \{\tilde{z}_1, \dots, \tilde{z}_n\} \sim \mu_z^{\otimes n}$ independent of S and separate the set of indices $\{1, \dots, n\}$ into H groups of size J such that $HJ = n$, $(H, J \in \mathbb{N}_+)$, so that we have a disjoint union:

$$\{1, \dots, n\} = \bigcup_{1 \leq k \leq H} J_k.$$

Let us write, for $w \in \mathcal{W}_{S,U}$:

$$\mathcal{R}(w) - \hat{\mathcal{R}}_S(w) = \frac{1}{n}\sum_{k=1}^{H}\sum_{i \in J_k}\big(\ell(w, z_i) - \mathcal{R}(w)\big). \tag{74}$$

Let us denote $S_{J_k} := (z_i)_{i \in J_k}$ and $S(J_k) \in \mathcal{Z}^n$ such that for all i, $S(J_k)_i = z_i$ if $i \in J_k$ and $S(J_k)_i = \tilde{z}_i$ if $i \in J_k$.
By an immediate recursion we have:

$$\forall w \in \mathcal{W}_{S,U},\ \exists w'_k \in \mathcal{W}_{S(J_k),U},\ \forall z \in \mathcal{Z},\ |\ell(w,z) - \ell(w'_k,z)| \leq \frac{J\beta}{n^\alpha}. \tag{75}$$

Therefore, taking such $w, w'_1, \dots, w'_H$ we get by the triangle inequality that

$$\begin{aligned}\mathcal{R}(w) - \hat{\mathcal{R}}_S(w) &\leq \frac{1}{n}\sum_{k=1}^{H}\sum_{i \in J_k}\Big\{|\ell(w,z_i) - \ell(w'_k,z_i)| + \mathbb{E}_{z \sim \mu_z}\big[\ell(w,z) - \ell(w'_k,z)\big] + \mathcal{R}(w'_k) - \ell(w'_k,z_i)\Big\} \\ &\leq 2\frac{J^2 H\beta}{n^{1+\alpha}} + \sum_{k=1}^{H}\sup_{w \in \mathcal{W}_{S(J_k),U}}\frac{1}{n}\sum_{i \in J_k}\big(\mathcal{R}(w) - \ell(w,z_i)\big),\end{aligned}$$

nd therefore

$$\sup_{w\in\mathcal{W}_{S,U}} \big(\mathcal{R}(w)-\hat{\mathcal{R}}_S(w)\big) \leq \frac{2\beta J}{n^\alpha} + \sum_{k=1}^H \underbrace{\sup_{w\in\mathcal{W}_{S(J_k),U}} \frac{1}{n}\sum_{i\in J_k}\big(\mathcal{R}(w)-\ell(w,z_i)\big)}_{:=\Sigma_k}. \tag{76}$$

Now let us temporary introduce a third dataset $S'\sim\mu_z^{\otimes n}$ independent of S and $\tilde{S}$. We will bound the expectation of Σ_k using a symmetrization technique and Rademacher complexity. The covering argument is left for the end of the proof.
Using properties of the conditional expectation we get:

$$\mathbb{E}[\Sigma_k] = \mathbb{E}\bigg[\sup_{w\in\mathcal{W}_{S(J_k),U}} \frac{1}{n}\sum_{i\in J_k}\mathbb{E}[\ell(w,z_i')-\ell(w,z_i)|S_{J_k}]\bigg].$$

Now let us introduce i.i.d. Rademacher random variables $\sigma_1,\dots,\sigma_n$ independent of all the other random variables. We have:

$$\begin{aligned}\mathbb{E}[\Sigma_k] &\leq \mathbb{E}\bigg[\mathbb{E}\bigg[\sup_{w\in\mathcal{W}_{S(J_k),U}} \frac{1}{n}\sum_{i\in J_k}\big(\ell(w,z_i')-\ell(w,z_i)\big)\bigg|S_{J_k}\bigg]\bigg]\\ &= \mathbb{E}\bigg[\sup_{w\in\mathcal{W}_{S(J_k),U}} \frac{1}{n}\sum_{i\in J_k}\big(\ell(w,z_i')-\ell(w,z_i)\big)\bigg]\\ &= \mathbb{E}\bigg[\sup_{w\in\mathcal{W}_{S(J_k),U}} \frac{1}{n}\sum_{i\in J_k}\sigma_i\big(\ell(w,z_i')-\ell(w,z_i)\big)\bigg]\\ &\leq 2\mathbb{E}\bigg[\sup_{w\in\mathcal{W}_{S(J_k),U}} \frac{1}{n}\sum_{i\in J_k}\sigma_i\ell(w,z_i)\bigg].\end{aligned}$$

Now, by independence of the σ_i with respect to the other random variables, we have by Fubini's theorem:

$$\mathbb{E}[\Sigma_k] \leq \frac{2}{H}\mathbb{E}\big[\mathbf{Rad}\big(\ell(\mathcal{W}_{S(J_k),U},S_{J_k})\big)\big]. \tag{77}$$

Given a set W and $(x_i)_{i\in A}\in\mathcal{Z}^{|A|}$ we will denote by $N_\delta(W,(x_i)_i)$ the centers of a δ-cover by closed delta balls of W with pseudo-metric: $(w,w')\mapsto\sum_{i\in A}|\ell(w,z_i)-\ell(w',z_i)|$.
A key argument to the proof is to note that by independence between $S(J_k)$ and S_{J_k}, the quantities $\mathbf{Rad}\big(\ell(\mathcal{W}_{S(J_k),U},S_{J_k})\big)$ and $\mathbf{Rad}\big(\ell(\mathcal{W}_{S,U},\tilde{S}_{J_k})\big)$ *have the same distribution* and therefore:

$$\mathbb{E}[\Sigma_k] \leq \frac{2}{H}\mathbb{E}\big[\mathbf{Rad}\big(\ell(\mathcal{W}_{S,U},\tilde{S}_J)\big)\big],$$

where $\tilde{S}_J$ arbitrary denotes one of the $\tilde{S}_{J_k}$ (they have the same distributions).
Now we now that we have almost surely:

$$\overline{\dim}_B^{d_{\tilde{S}_J}}(\mathcal{W}_{S,U}) = \limsup_{\delta\to0}\frac{\log|N_\delta(\mathcal{W}_{S,U},\tilde{S}_J)|}{\log(1/\delta)},$$

therefore, by applying Egoroff's theorem, we can get that for fixed $\epsilon,\gamma>0$, there exists a set set $\Omega_\gamma\in\mathcal{F}_{\mathcal{Z}}^{\otimes 2n}\otimes\mathcal{F}_U$ with $\mathbb{P}(\Omega_\gamma)\geq1-\gamma$ such that on Ω_γ the convergence above is uniform and we have, for δ small enough:

$$\log|N_\delta(\mathcal{W}_{S,U},\tilde{S}_J)| \leq \big(\epsilon+\overline{\dim}_B^{d_{\tilde{S}_J}}(\mathcal{W}_{S,U})\big)\log(1/\delta).$$

Now we can introduce a covering and apply Massart's lemma to write that for all $\delta>0$ we have almost surely (with a slight abuse of notation on $\tilde{S}_J$):

$$\begin{aligned}\mathbf{Rad}\big(\ell(\mathcal{W}_{S,U},\tilde{S}_J)\big) &\leq \delta+\mathbb{E}_\sigma\bigg[\sup_{w\in N_\delta(\mathcal{W}_{S,U},\tilde{S}_J)}\frac{1}{J}\sum_{i=1}^J\sigma_i\tilde{z}_i\bigg]\\ &\leq \delta+B\sqrt{\frac{2\log|N_\delta(\mathcal{W}_{S,U},\tilde{S}_J)|}{J}}.\end{aligned}$$

Hence we can write, using the fact that we can bound the Rademacher complexity by B, that for δ small enough:

$$\begin{aligned}\mathbb{E}[\Sigma_k] &\leq \frac{2}{H}\mathbb{E}\big[\mathbf{Rad}\big(\ell(\mathcal{W}_{S,U}, \tilde{S}_J)\big)\big] \\ &\leq \frac{2B}{H}\gamma + \frac{2}{H}\mathbb{E}\big[\mathbb{1}_{\Omega_\gamma}\mathbf{Rad}\big(\ell(\mathcal{W}_{S,U}, \tilde{S}_J)\big)\big] \\ &\leq \frac{2B}{H}\gamma + \frac{2B}{H}\mathbb{E}\Bigg[\mathbb{1}_{\Omega_\gamma}\sqrt{\frac{2\log|N_\delta(\mathcal{W}_{S,U}, \tilde{S}_J)|}{J}}\Bigg].\end{aligned}$$

Therefore, by Jensen's inequality:

$$\mathbb{E}[\Sigma_k] \leq \frac{2B}{H}\gamma + \frac{2B}{H}\sqrt{\frac{2\big(\epsilon + \mathbb{E}\big[\overline{\dim}_B^{d_{\tilde{S}_J}}(\mathcal{W}_{S,U})\big]\big)\log(1/\delta)}{J}}.$$

Putting everything together we get, still for δ small enough:

$$\mathbb{E}\Bigg[\sup_{w \in \mathcal{W}_{S,U}}\big(\mathcal{R}(w) - \hat{\mathcal{R}}_S(w)\big)\Bigg] \leq \frac{2\beta J}{n^\alpha} + 2\delta + 2B\gamma + B\sqrt{\frac{8}{J}}\sqrt{\big(\epsilon + \mathbb{E}\big[\overline{\dim}_B^{d_{\tilde{S}_J}}(\mathcal{W}_{S,U})\big]\big)\log(1/\delta)}. \quad (78)$$

We can see in equation 78 that there is a trade-off to be made in the values of K and J. We want J of the form $J = n^\lambda$ with $\lambda \in (0, 1]$. We want the two terms in which the appears to be of the same order magnitude in n, which leads to:

$$\frac{J}{n^\alpha} = \frac{1}{\sqrt{J}}.$$

And therefore we ask that $\alpha - \lambda = \lambda/2$ and gives us the fundamental result:

$$\boxed{\lambda = \frac{2\alpha}{3}} \quad (79)$$

So that, for δ small enough (depending on the values of n and γ):

$$\mathbb{E}\Bigg[\sup_{w \in \mathcal{W}_{S,U}}\big(\mathcal{R}(w) - \hat{\mathcal{R}}_S(w)\big)\Bigg] \leq 2(\delta + B\gamma) + 2\sqrt{\frac{\beta^2 + 4B^2\big(\epsilon + \mathbb{E}\big[\overline{\dim}_B^{d_{\tilde{S}_J}}(\mathcal{W}_{S,U})\big]\big)\log(1/\delta)}{n^{\frac{2\alpha}{3}}}}. \quad (80)$$

Thus the final result with a rate in n of $\alpha/3$. Note that if $\alpha = 1$ (which is reasonable, see Foster *et al.* (2020)), then we get a rate in $1/3$, while with a fixed hypothesis space we had a rate in $1/2$. Therefore avoiding the coupling (and therefore the mutual information term) made us lose in the convergence rate.

□

Remark 27. *As we can see in Theorem B.1, as typically $\alpha \simeq 1$, we removed the mutual information term at the cost of losing in the convergence rate in n, which is now typically in $n^{-1/3}$.*

Remark 28. *As already mentioned, Theorem B.1 could be turned into a high probability bound by using the results from (Foster* et al.*, 2020), we would still have an expected fractal dimension term though.*

Appendix C Some code samples

Obviously, it is not possible to include all the code of the experiments in this report. However, most of this code is very classic (models, training, data handling) and does not need a particular attention. We just report in this section a code sample illustrating how the data-dependent intrinsic dimension is computed from the iterations of a neural network. Part of this code is inspired by (Birdal *et al.*, 2021), but it has been significantly accelerated, especially by changing the way distance matrices are handled.

Listing 1 present a code sample used to approximate the data-dependent persistent homology dimension $\dim_{\mathrm{PH}^0}^{\rho_S}(\mathcal{W}_{S,U})$. More precisely, the function `fast_dimension_computation` takes as inputs a list of vectors $((\ell(w_1, z_i))_{1\leq i\leq n}, \dots, (\ell(w_K, z_i))_{1\leq i\leq n})$ for some K last iterations of training and computes the corresponding distance matrix $(\rho_S(w_i, w_j))_{1\leq i,j\leq K}$, it then uses it to approximate the dimension using Equation (70).

```python

import time
from pathlib import Path

import numpy as np
import torch
from loguru import logger
from tqdm import tqdm
from gph.python import ripser_parallel
from sklearn.metrics.pairwise import pairwise_distances


def sample_W(W, nSamples, isRandom=True):
    n = W.shape[0]
    random_indices = np.random.choice(n, size=nSamples, replace=False)
    return W[random_indices]


def ph_dim_from_distance_matrix(dm: np.ndarray,
                                min_points=200,
                                max_points=1000,
                                point_jump=50,
                                h_dim=0,
                                alpha: float = 1.,
                                seed: int = 42) -> float:
    """
    This functions:
     - turn W into a torch tensor
     - compute the distance matrix
     - use it to compute PH dim

    :param dm: distance matrix, should be of shape (N, N)
    """
    assert dm.ndim == 2, dm
    assert dm.shape[0] == dm.shape[1], dm.shape

    np.random.seed(seed)

    test_n = range(min_points, max_points, point_jump)
    lengths = []

    for points_number in test_n:

        sample_indices = np.random.choice(dm.shape[0], points_number, replace=False)
        dist_matrix = dm[sample_indices, :][:, sample_indices]

        diagrams = ripser_parallel(dist_matrix, maxdim=0, n_threads=-1, metric="
    precomputed")[’dgms’]

        d = diagrams[h_dim]
        d = d[d[:, 1] < np.inf]
        lengths.append(np.power((d[:, 1] - d[:, 0]), alpha).sum())  # The fact that \
    alpha = 1 appears here

    lengths = np.array(lengths)

    # compute our ph dim by running a linear least squares
    x = np.log(np.array(list(test_n)))
    y = np.log(lengths)
    N = len(x)
    m = (N * (x * y).sum() - x.sum() * y.sum()) / (N * (x ** 2).sum() - x.sum() ** 2)
    b = y.mean() - m * x.mean()

    error = ((y - (m * x + b)) ** 2).mean()

    logger.debug(f"Ph Dimension Calculation has an approximate error of: {error}.")

    return alpha / (1 - m)


def fast_dimension_computation(w: np.ndarray,
                               min_points=200,
                               max_points=1000,
                               point_jump=50,
                               h_dim=0,
                               alpha: float = 1.,
                               seed: int = 42,
                               save_dir: str = None,
                               metric: str = "euclidean"):

    assert w.shape[0] <= max_points, (w.shape[0], max_points)
    assert w.shape[0] >= min_points, (w.shape[0], min_points)

    starting_time = time.time()
    dm = pairwise_distances(w, metric=metric)
    logger.debug(f"Distance matrix computation time: {round(time.time() - starting_time, 2)}s")

    if save_dir is not None:
        save_path = Path(save_dir) / "distance_matrix.npy"
        save_path.parent.mkdir(parents=True, exist_ok=True)
        np.save(str(save_path), dm)

    return ph_dim_from_distance_matrix(dm,
                                       min_points,
                                       max_points,
                                       point_jump,
                                       h_dim,
                                       alpha,
                                       seed)
```

Listing 1: Main part of the code used to estimate our data-dependent intrinsic dimension.