Generalization Bounds using Data-Dependent Fractal Dimensions

Benjamin Dupuis [1 2 3] **George Deligiannidis** [4 5] **Umut Şimşekli** [1 2 3 6]

Abstract

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 work, 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.

[*]Equal contribution [1]Inria [2]Ecole Normale Supérieure, Paris, France [3]PSL Research University, Paris, France [4]Department of Statistics, University of Oxford, Oxford, UK [5]The Alan Turing Institute, London, UK [6]CNRS. Correspondence to: Benjamin Dupuis <benjamin.dupuis@inria.fr>, Umut Şimşekli <umut.simsekli@inria.fr>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

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 population risk minimization problem, given as follows:

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

where $x \in \mathcal{X}$ denotes the features, $y \in \mathcal{Y}$ denotes the labels, $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ denotes the data space endowed with an unknown probability measure μ_z, 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, and ℓ is the composition of the loss and the predictor, i.e. $\ell(w,z) = \ell(w,(x,y)) = \mathcal{L}(h_w(x),y)$, which will also be referred to as 'loss', with a slight abuse of notation. As μ_z is unknown, in practice one resorts to the minimization of the empirical risk, given as follows:

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

where $S := (z_i)_{1\le i\le n} \sim \mu_z^{\otimes n}$ 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 $\mathcal{W} \subset \mathbb{R}^d$:

$$\mathcal{G}(S) := \sup_{w\in\mathcal{W}}\big(\mathcal{R}(w) - \hat{\mathcal{R}}_S(w)\big). \tag{2}$$

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 suggesting that typically $\mathcal{G} \approx \sqrt{d/n}$ (Anthony & 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 (1), the optimization trajectories can exhibit a fractal structure (Şimşekli et al., 2021; Camuto et al., 2021; Birdal et al., 2021; Hodgkinson et al., 2022). Under the assumption that ℓ is uniformly bounded by some B and uniformly L-Lipschitz with respect to w, their results informally implies the following: with probability $1-\zeta$, we have that

$$\mathcal{G} \lesssim LB\sqrt{\frac{\bar{d}(\mathcal{W}) + I_\infty(\mathcal{W}, S) + \log(1/\zeta)}{n}}, \qquad (3)$$

where $\mathcal{W}$ 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 the data S and the hypothesis set $\mathcal{W}$. These notions will be formally defined in Section 2[1]. 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 (3) provides an explanation on why overparameterized networks might not overfit in practice.

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 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 the how the loss ℓ behaves over $\mathcal{W}$. 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.

[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.

1.1. Contributions

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 Appendix A.2), we achieve this goal by making use of a *data-dependent* pseudo-metric on the hypothesis set $\mathcal{W}$. Our contributions are as follows:

- We prove bounds (Theorems 3.4 and 3.5) on the worst-case generalization error of the following form:

$$\mathcal{G} \lesssim B\sqrt{\frac{\bar{d}_S(\mathcal{W}) + I + \log(1/\zeta)}{n}}, \qquad (4)$$

  where $\bar{d}_S$ denotes a notion of *data-dependent* fractal dimension and I is a (total) mutual information term (see Section 2.2). As opposed to prior work, this bound does not require any Lipschitz assumption and therefore applies to more general settings. However, this improvement comes with the expense of having a more complicated mutual information term compared to the one in (3).
- To provide more understanding about the newly introduced mutual information term I and highlight its links to prior work, we introduce a notion of 'geometric stability' and without requiring Lipschitz continuity, we prove an almost identical bound to the one in Equation (3) (with a potentially slightly worse rate in n).
- In order to be able to compute the data-dependent fractal dimension, we build on (Birdal et al., 2021) and prove that our dimension can also be computed by using numerically efficient topological data analysis tools (Carlsson, 2014; Pérez-Fernández et al., 2021).

Finally, we illustrate our bounds on experiments using various neural networks. In addition to not requiring Lipschitz continuity, we show that our data-dependent dimension provides improved correlations with the actual generalization error. All the proofs are provided in the Appendix. Python code for numerical experiments is available at `https://github.com/benjiDupuis/data_dependent_dimensions`.

2. Technical Background

2.1. Learning framework

We formalize the learning algorithm as follows. The data (probability) space is denoted by $(\mathcal{Z}, \mathcal{F}, \mu_z)$[2]. A learning algorithm $\mathcal{A}$ is 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 random-

[2] For technical measure-theoretic reasons (see Section A.6), it is best to assume $\mathcal{Z} \subseteq \mathbb{R}^N$ for some N.

ness U takes values in some probability space $(\Omega_U, \mathcal{F}_U, \mu_u)$, which means that U is $\mathcal{F}_U$-measurable 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 class of closed sets of $\mathbb{R}^d$ endowed with the Effrös σ-algebra, as in (Molchanov, 2017), the algorithm will be thought as a measurable map:

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

This formulation encompasses several settings, such as the following two examples.

Example 2.1. 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 & 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 2.2. 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.

2.2. Information theoretic quantities

Recently, one popular approach to prove generalization bounds has been based on information theory. In this context, Xu & Raginsky (2017); Russo & 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 & 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.

The classic mutual information between two random elements X and Y is defined in terms of the Kullback-Leibler (KL) divergence $I(X,Y) := \mathrm{KL}(\mathbb{P}_{X,Y} || \mathbb{P}_X \otimes \mathbb{P}_Y)$. It is well known that mutual information can be used as a decoupling tool (Xu & Raginsky, 2017); yet, in our setup, we will need to consider the *total mutual information*, which is defined as follows, $\mathbb{P}_X$ denoting the law of the variable X:

$$I_\infty(X,Y) := \log\left(\sup_B \frac{\mathbb{P}_{X,Y}(B)}{\mathbb{P}_X \otimes \mathbb{P}_Y(B)}\right). \tag{6}$$

Hodgkinson et al. (2022) used total mutual information to decouple the data and the optimization trajectory, they defined it as a limit of α-mutual information, which is equivalent, see (van Erven & Harremoës, 2014, Theorem 6).

2.3. The upper box-counting dimension

Fractal geometry (Falconer, 2014) and dimension theory have been successful tools in the study of dynamical systems and stochastic processes (Pesin, 1997; Xiao, 2004). In our setting, we will be interested in the *upper box-counting dimension*, defined as follows. Given a (pseudo-)metric space (X, ρ) and $\delta > 0$, we first define the closed δ-ball centered in $x \in X$ by $B_\delta^\rho(x) = \{y \in X,\ \rho(x,y) \le \delta\}$ and a *minimal covering* $N_\delta^\rho(X)$ as a minimal set of points of X such that $X \subset \bigcup_{y \in N_\delta^\rho(X)} B_\delta^\rho(y)$. We can then define the upper box-counting dimension as follows:

$$\overline{\dim}_B^\rho(X) := \limsup_{\delta \to 0} \frac{\log |N_\delta^\rho(X)|}{\log(1/\delta)}, \tag{7}$$

where $|A|$ denotes the cardinality of a set A.

Under the Lipschitz loss assumption, Şimşekli et al. (2021); Birdal et al. (2021); Camuto et al. (2021); Hodgkinson et al. (2022) related different kinds of fractal dimensions, computed with *the Euclidean distance* $\rho(w,w') = \mathrm{Eucl}(w,w') := \|w - w'\|_2$, to the generalization error. Our approach in this study will be based on using a *data-dependent* pseudo-metric ρ, which will enable us to remove the Lipschitz assumption.

3. Main Results

In this section we present our main theoretical results; our aim is to relate the worst-case generalization error of (5) with the upper box-counting dimension computed based on the following random pseudo-metric:

$$\rho_S(w,w') := \frac{1}{n}\sum_{i=1}^{n} |\ell(w,z_i) - \ell(w',z_i)|. \tag{8}$$

We insist on the fact that it is only a pseudo-metric because in practice we can have $\rho_S(w,w') = 0$ while $w \neq w'$, for example due to the internal symmetries of a neural network.

3.1. Main assumptions

A key component of our work is that we do not use any Lipschitz assumption on ℓ as for example in (Şimşekli et al., 2021; Hodgkinson et al., 2022). The only regularity assumption we impose is the following:

Assumption 3.1. The loss $\ell : \mathbb{R}^d \times \mathcal{Z} \longrightarrow \mathbb{R}$ is continuous in both variables and uniformly bounded by some $B > 0$.

We note that the box-counting dimension with respect to the pseudo-metric (8) involves minimal coverings, which we denote $N_\delta^{\rho_S}(A)$ for some set A. The boundedness assumption is essential to ensure that minimal coverings are finite and $\overline{\dim}_B^{\rho_S}$ is also finite. Therefore our boundedness assumption cannot be replaced with a subgaussian assumption, as opposed to (Şimşekli et al., 2021).

We also assume that we can construct minimal coverings which are random closed (finite) sets in the sense of (Molchanov, 2017, Definition 1.1.1); this is made precise with the following assumption:

Assumption 3.2. Let $C \subset \mathbb{R}^d$ be any closed set, $\delta > 0$, $S \in \mathcal{Z}^n$ and $S' \in \mathcal{Z}^m$. We can construct minimal δ-coverings $N_\delta^{\rho_{S'}}(C \cap \mathcal{W}_{S,U})$ which are random finite sets with respect to the product σ-algebra $\mathcal{F}^{\otimes n} \otimes \mathcal{F}^{\otimes n} \otimes \mathcal{F}_U$ (measurability with respect to S, S', U). We denote by $\mathcal{N}_\delta(C \cap \mathcal{W}_{S,U})$ the family of all those random minimal coverings.

Remark 3.3. Assumption 3.2 essentially enables us to avoid technical measurability complications. The main message is that we assume that we are able to construct "measurable coverings". This assumption can be cast as a *selection* property; indeed for each realization of (S, S', U) there may be a wide range of possible minimal coverings: what we assume is that we can select one of them for each (S, S', U) so that the obtained random set is measurable.

Assumption 3.2 is actually much stronger than what is needed to make our results valid. Indeed, we are able to show that, under the assumption that the mapping $\mathcal{A}$ of (5) is measurable with respect to the Effrös σ-algebra, we can construct coverings $(S, U) \longmapsto N_\delta^{\rho_S}(\mathcal{W}_{S,U})$ which are measurable and, if not minimal, yield the same upper box-counting dimension as minimal coverings, when computing the limit (7). To avoid too much technical considerations, this discussion is presented in Appendix A.6, as an additional technical contribution.

As the upper box-counting dimension (7) may be written as a countable limit, the measurability assumption 3.2 also implies that $\overline{\dim}_B^{\rho_S}(\mathcal{W}_{S,U})$ is a random variable. Continuity of the loss in Assumption 3.1 is there for technical purposes, e.g., to make quantities of the form $\sup_{w \in \mathcal{W}_{S,U}} \left(\mathcal{R}(w) - \hat{\mathcal{R}}_S(w)\right)$ well-defined random variables (see (Molchanov, 2017, Theorem 1.3.28) and Section A.6 in the Appendix).

3.2. Warm-up: fixed hypothesis spaces

In this subsection we fix a *deterministic* closed set $\mathcal{W} \subset \mathbb{R}^d$ and consider its upper box-counting dimension with respect to the data-dependent pseudo-metric (8), which we denote by $d(S) := \overline{\dim}_B^{\rho_S}(\mathcal{W})$. Our goal is to bound the worst-case generalization error as defined in (2). The next theorem is an extension of the classical covering bounds of Rademacher complexity (Barlett & Mendelson, 2002; Rebeschini, 2020).

Theorem 3.4. *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 $\mu_z^{\otimes n}$, for all $\delta < \delta_{n,\gamma,\epsilon}$ we have:*

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

Theorem 3.4 is therefore similar to (Şimşekli et al., 2021, Theorem 1), which used a fractal dimension based on the Euclidean distance on $\mathbb{R}^d$, $\|w - w'\|_2$ and a fixed hypothesis space. The improvement here is in the absence of Lipschitz assumption. Moreover, as detailed in Appendix B.6, in case of a Lipschitz ℓ, we recover, from our proofs, bounds in term of the upper box-counting dimension based on the Euclidean distance on the hypothesis set, which is the one used in prior works (Şimşekli et al., 2021). Therefore, our methods based on a data-dependent fractal dimension are more general than previous studies.

However, Theorem 3.4 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.

3.3. Random hypothesis spaces

Theorem 3.4 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.

For notational purposes let us denote the upper box-counting dimension of $\mathcal{W}_{S,U}$ induced by pseudo-metric (8) by $d(S, U) := \overline{\dim}_B^{\rho_S}(\mathcal{W}_{S,U})$, and denote the worst-case generalization error by

$$\mathcal{G}(S, U) := \sup_{w \in \mathcal{W}_{S,U}} (\mathcal{R}(w) - \hat{\mathcal{R}}_S(w)). \tag{9}$$

Here again, note that $d(S, U)$ can be written as a countable limit of random variables and therefore defines a random variable thanks to Assumption 3.2.

The main difficulty here is that classical arguments based on the Rademacher complexity cannot be applied in this case as $\mathcal{W}_{S,U}$ depends on the data sample S. 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 (cf. Section 2.3) and rely on the following decomposition: for any $\delta > 0$ and $w' \in N_\delta^{\rho_S}(\mathcal{W}_{S,U})$ we have that

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

In the above inequality, the first term can be controlled by standard techniques 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[3] for some $K \in \mathbb{N}_+$:

$$R_S^j := \mathcal{W}_{S,U} \cap \mathcal{R}^{-1}\left(\left[\frac{jB}{K}, \frac{(j+1)B}{K}\right]\right), \quad (10)$$

where $j = 0, \ldots, K-1$ and $\mathcal{R}^{-1}$ denotes the inverse image of $\mathcal{R}$. Let $N_{\delta,j}$ collect the centers of a minimal δ-cover of R_S^j relatively to ρ_S[4]. The next theorem provides a generalization bound for random hypothesis sets.

Theorem 3.5. *Let us set $K = \lfloor \sqrt{n} \rfloor$ and define $I_{n,\delta} := \max_{0 \le j \le \lfloor \sqrt{n} \rfloor} I_\infty(S, N_{\delta,j})$. Then, for all $\epsilon, \gamma, \eta > 0$, there exists $\delta_{n,\gamma,\epsilon} > 0$ such that with probability at least $1-\eta-\gamma$ under $\mu_z^{\otimes n} \otimes \mu_u$, for all $\delta < \delta_{n,\gamma,\epsilon}$ we have:*

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

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 3.5 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 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).

3.4. 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 3.5 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 the 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:

$$\forall w \in \mathcal{W}_{S,U}, \exists w' \in \mathcal{W}_{S',U}, \forall z \in \mathcal{Z}, \quad |\ell(w,z) - \ell(w',z)| \le \bar{\beta}. \quad (11)$$

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

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 (8) 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 Assumption 3.2 which states that we can make such a selection of $N_\delta(A,S)$, making it a well-defined random set, we add the fact that this selection can be made regular enough in the following sense.

Definition 3.6. 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 data-dependent Hausdorff distance, by which we mean:

$$\forall w \in N_\delta(S,A), \exists w' \in N_\delta(S',A), \quad \sup_{z \in \mathcal{Z}} |\ell(w,z) - \ell(w',z)| \le \frac{\beta}{n^\alpha}. \quad (12)$$

Based on this definition, we assume the following condition.

Assumption 3.7. 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 $\mathcal{W}_{S,U} \cap \mathcal{R}^{-1}\left(\left[\frac{jB}{K}, \frac{(j+1)B}{K}\right]\right)$ are geometrically stable with parameters (α, β).

Assumption 3.7 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 3.8. *Let $d(S,U)$ and $\mathcal{G}(S,U)$ be as in Theorem 3.5 and further define $I := I_\infty(S, \mathcal{W}_{S,U})$. Suppose that Assumption 3.7 holds. Then there exists a constant $n_\alpha, \delta_{\gamma,\epsilon,n} > 0$ such that for all $n \ge n_\alpha$, with probability*

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

[4] Assumption 3.2 extends to the randomness of those sets $N_{\delta,j}$.

$1-\gamma-\eta$, *and for all* $\delta \leq \delta_{\gamma,\epsilon,n}$, *the following inequality holds:*

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

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

While Assumption 3.7 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 3.8 is much more interpretable compared to the corresponding terms in Theorem 3.5, 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 3.8, we observe that getting closer to an optimal rate ($\alpha \approx \frac{3}{2}$) implies a larger n_α, rendering our bound asymptotic.

3.5. One step toward lower bounds

As an additional theoretical result, we present an attempt to prove a lower bound involving the introduced data-dependent fractal dimension. For this purpose, let us consider again the case of a fixed (non-random) closed hypothesis set $\mathcal{W} \subset \mathbb{R}^d$. As in Section 3.2, we make Assumption 3.1. We also introduce the *data-dependent lower box-counting dimension* as:

$$\underline{d}(S) = \underline{\dim}_B^{\rho_S}(\mathcal{W}) := \liminf_{\delta \to 0} \frac{\log |N_\delta^{\rho_S}(\mathcal{W})|}{\log(1/\delta)}. \tag{13}$$

As proving lower bounds is not the main goal of this paper, we restrict ourselves to this specific setting. The next theorem is based on very classical arguments involving Gaussian complexity and Sudakov's theorem (Vershynin, 2020, Theorem 7.4.1). This lower bound requires a slightly different definition of the worst-case generalization error:

$$\overline{\mathcal{G}}(S) := \sup_{w \in \mathcal{W}} \left| \mathcal{R}(w) - \hat{\mathcal{R}}_S(w) \right|. \tag{14}$$

Theorem 3.9. *We further assume that* $\underline{d}(S) > 0$ *almost surely. Then, for all* $\epsilon, \gamma, \eta > 0$, *there is an absolute constant* $c > 0$ *and some* $\delta_{n,\gamma,\zeta} > 0$ *such that, with probability at least* $1-\zeta-\gamma$, *for all* $\delta \leq \delta_{n,\gamma,\zeta}$ *we have:*

$$\overline{\mathcal{G}}(S) \geq \frac{c}{4}\sqrt{\frac{\delta^2 \log(1/\delta)\underline{d}(S)}{2n\log(n)}} - B\sqrt{\frac{\log(2) + 9\log(1/\zeta)}{n}}$$

Theorem 3.9 gives a lower bound that is probably less tight compared to the upper bounds we presented in this work. One could even note that the right hand side of the bound may be negative in some contexts. However, we believe that the techniques used to derive this bound are classical and may be useful for future research.

4. Computational Aspects

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).

4.1. Persistent homology

Persistent homology (PH) is a well known notion in TDA typically used for point cloud analysis (Edelsbrunner & 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 $\mathcal{W}_{S,U}$. Here we extend their approach to the case of data-dependent pseudo-metrics, which lays the ground for our experimental analysis.

The formal definition of PH is rather technical and is not essential to our problematic; hence, we only provide a high-level description here, and provide a more detailed description in Section A.4 (for a formal introduction, see (Boissonat et al., 2018; Memoli & Singhal, 2019). In essence, given a point cloud $W \subset \mathbb{R}^d$, 'PH of degree 0', denoted by PH^0 keeps track of the *connected components* in W, as we examine W at a gradually decreasing resolution.

Given a bounded (pseudo-)metric space (X, ρ), by using PH^0, one can introduce another notion of fractal dimension, called the *persistent homology dimension*, which we denote by $\dim_{\mathrm{PH}^0}^{\rho}(X)$ (see Section A.4 and (Schweinhart, 2019, Definition 4)).

Our particular interest in $\dim_{\mathrm{PH}^0}^{\rho}(X)$ in the case where ρ is a proper metric comes from an important result (Kozma et al., 2005; Schweinhart, 2020) stating that for any bounded metric space (X, ρ) we have the following identity.

$$\overline{\dim}_B^{\rho}(X) = \dim_{\mathrm{PH}^0}^{\rho}(X). \tag{15}$$

Several studies used this property to numerically evaluate the upper box-counting dimension (Adams et al., 2020; Birdal et al., 2021). In particular Birdal et al. (2021) combined it with the results from (Şimşekli et al., 2021) and showed that $\dim_{\mathrm{PH}^0}^{\mathrm{Eucl}}(X)$ associated with the Euclidean metric on the parameter space, can be linked to the generaliza-

tion error under the Lipschitz loss condition.

4.2. PH dimension in pseudo-metric spaces

In order to extend the aforementioned analysis to our data-dependent dimension, we must first prove that the equality (15) extends to pseudo-metric spaces, which is established in the following theorem:

Theorem 4.1. *Let (X, ρ) be a bounded pseudo-metric space, we have:* $\overline{\dim}_B^{\rho}(X) = \dim_{PH^0}^{\rho}(X)$.

This theorem shows that, similar to $\dim_{\mathrm{PH}^0}^{\mathrm{Eucl}}(X)$, our proposed dimension $\dim_{\mathrm{PH}^0}^{\rho_S}(X)$ can also be computed by using numerically efficient TDA tools. Moreover, Theorem 3.4 now (informally) implies that with probability $1-\zeta$:

$$\mathcal{G}(S) \lesssim \sqrt{\frac{\dim_{\mathrm{PH}^0}^{\rho_S}(W)\log(1/\delta) + \log(1/\zeta)}{n}} + \delta. \quad (16)$$

Theorems 3.5 and 3.8 can be adapted similarly.

5. Experiments

Experimental setup. In our experiments, we closely follow the setting used in (Birdal et al., 2021). 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[5]. 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). The main difference between our approach and (Birdal et al., 2021) is that we replace the Euclidean metric with the pseudo-metric ρ_S to compute the PH dimension.

Here is a brief description of the method: given a neural network, its loss $\ell(w, z)$, and a dataset $S = (z_1, \ldots, z_n)$, 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 $\{w_{K^\star+1}, \ldots, w_{K^\star+5000}\}$. 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.

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 & 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

[5]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.

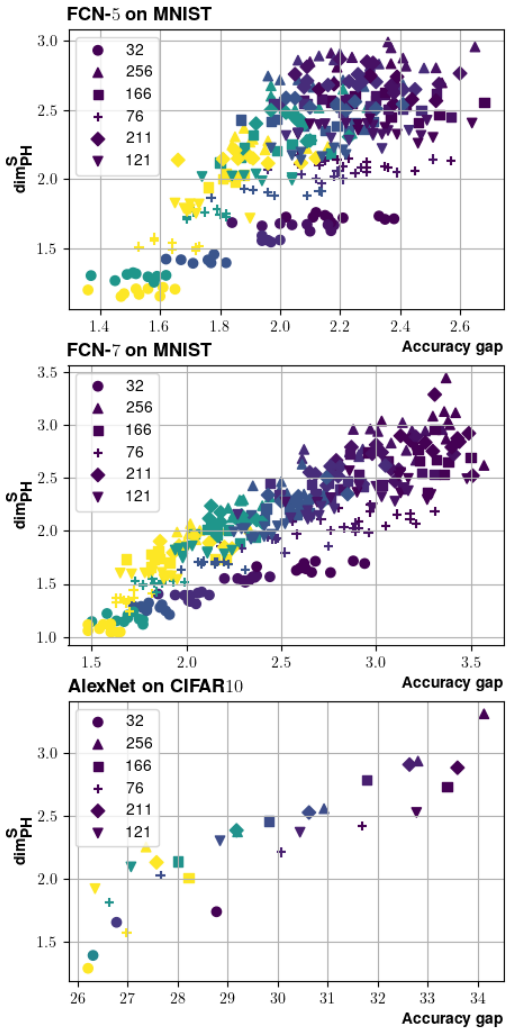

Figure 1. $\dim_{\mathrm{PH}^0}^{\rho_S}$ (denoted $\dim_{\mathrm{PH}^0}^{S}$ in the figure) versus accuracy gap for FCN-5 (*top*), FCN-7 (*middle*) on MNIST and AlexNet (*bottom*) on CIFAR-10 Different colors indicate different learning rates and different markers indicate different batch sizes.

(Krizhevsky et al., 2014). More experiments are shown in the appendix Section D. 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. All hyperparameter configurations are available in Section C.

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 (8). 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 $\dim_{\mathrm{PH}^0}^{\rho_S}$ 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

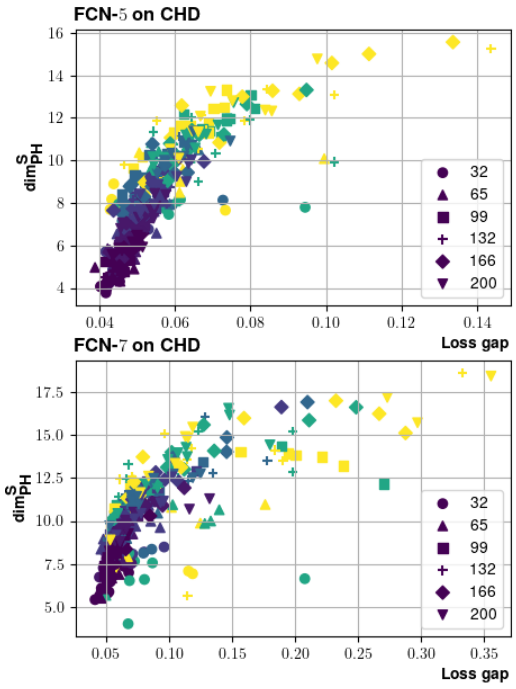

Figure 2. $\dim_{\mathrm{PH}^0}^{\rho S}$ (denoted $\dim_{\mathrm{PH}^0}^{S}$ in the figure) versus generalization gap for FCN-5 (*top*) and FCN-7 (*bottom*) trained on CHD. Different colors indicate different learning rates and different markers indicate different batch sizes.

sake of completeness, we provide how $\dim_{\mathrm{PH}^0}^{\rho S}$ behaves with respect to the the actual *loss gap* in Section D.

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 & 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 C.1. Therefore $(\rho, \mathbf{\Psi}, \tau)$ are our main indicators of performance. The values of each granulated Kendall's coefficient are reported in Section D[6].

Figures 1 and 2 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. We also observe that the highest learning rates and lowest 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 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.

[6]All those coefficients are between -1 and 1, where the value of 1 indicating a perfect positive correlation.

Table 1. Correlation coefficients on CHD

MODEL	DIM.	ρ	$\mathbf{\Psi}$	τ
FCN-5	$\dim_{\mathrm{PH}^0}^{\mathrm{EUCL}}$	$0.77_{\pm 0.08}$	$0.54_{\pm 0.11}$	$0.59_{\pm 0.07}$
FCN-5	$\dim_{\mathrm{PH}^0}^{\rho S}$	$\mathbf{0.87}_{\pm 0.05}$	$\mathbf{0.68}_{\pm 0.10}$	$\mathbf{0.71}_{\pm 0.09}$
FCN-7	$\dim_{\mathrm{PH}^0}^{\mathrm{EUCL}}$	$0.40_{\pm 0.09}$	$0.16_{\pm 0.08}$	$0.28_{\pm 0.07}$
FCN-7	$\dim_{\mathrm{PH}^0}^{\rho S}$	$\mathbf{0.77}_{\pm 0.08}$	$\mathbf{0.62}_{\pm 0.06}$	$\mathbf{0.77}_{\pm 0.08}$

Table 2. Correlation coefficients on MNIST

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

Table 3. Correlation coefficients with AlexNet on CIFAR-10

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

Next, we report the correlation coefficients for the same experiments in Tables 1, 2 and 3. The results show that on average our proposed dimension always yields improved metrics compared to the dimension introduced in (Birdal et al., 2021). The improvement is particularly better in the regression experiment we performed (as the classification task yields larger variations in the metrics, see Table 2). This may indicate that the proposed dimension may be particularly pertinent in specific settings. Moreover, increasing the size of the model, in all 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. We provide more experimental results in Section D.

Robustness analysis. The computation of $\rho_S(w, w')$ requires the exact evaluation of the loss function on every data point $\{z_1, \ldots, z_n\}$ 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 $\dim_{\mathrm{PH}^0}^{\rho S}$. 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

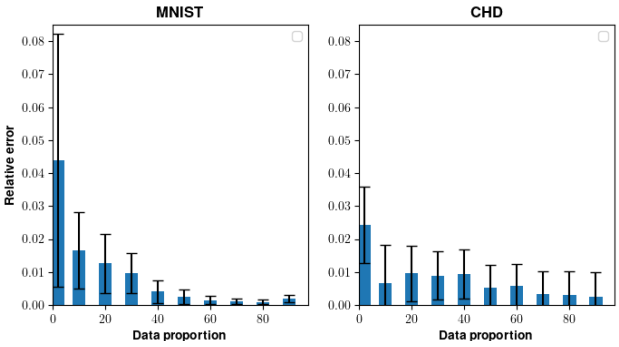

Figure 3. Robustness experiment using a FCNN trained on MNIST (*Left*) and CHD (*Right*). x-axis represents the proportion of the data T used to compute the metric, y-axis is the relative error with respect to the full dataset based dimension.

$$\rho_S(w, w') \approx \rho_T(w, w') := \frac{1}{|T|} \sum_{z \in T} |\ell(w, z) - \ell(w', z)|.$$

We now conduct experiments to analyze the robustness of the computation of $\dim_{\mathrm{PH}^0}^{\rho_S}$ 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 3 presents results on the MNIST and CHD datasets in term of the relative error, i.e., $|\dim_{\mathrm{PH}^0}^{\rho_T} - \dim_{\mathrm{PH}^0}^{\rho_S}| / \dim_{\mathrm{PH}^0}^{\rho_S}$. 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.

6. Conclusion

In this paper, 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). Refining our proof techniques, for example using the chaining method (Ledoux & Talagrand, 1991; Clerico et al., 2022), could help us improve our theoretical results or weaken the assumptions.

Acknowledgments

U.Ş. is partially supported by the French government under management of Agence Nationale de la Recherche as part of the “Investissements d’avenir” program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute). B.D. and U.Ş. are partially supported by the European Research Council Starting Grant DYNASTY – 101039676.

References

Adams, H., Aminian, M., Farnell, E., Kirby, M., Peterson, C., Mirth, J., Neville, R., Shipman, P., and Shonkwiler, C. A fractal dimension for measures via persistent homology. *Topological Data Analysis, Abel Symposia, vol. 15*, pp. 1–31, 2020. doi: 10.1007/978-3-030-43408-3_1.

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

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

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

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

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

Bauer, U. Ripser: Efficient computation of Vietoris-Rips persistence barcodes. *Journal of Applied and Computational Topology*, 5(3):391–423, September 2021. ISSN 2367-1726, 2367-1734. doi: 10.1007/s41468-021-00071-5.

Belkin, M., Hsu, D., Ma, S., and Mandal, S. Reconciling modern machine learning practice and the bias-variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, August 2019. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1903070116.

Birdal, T., Lou, A., Guibas, L., and Şimşekli, U. Intrinsic Dimension, Persistent Homology and Generalization in Neural Networks. *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, November 2021.

Bogachev, V. I. *Measure Theory*, volume Volume 1. Springer, 2007.

Boissonat, J.-D., Chazal, F., and Yvinec, M. *Geometrical and Topological Inference*. Cambridge Texts in Applied Mathematics. Cambridge University Press, 2018.

Bousquet, O. Stability and generalization. *Journal of Machine Learning Research*, 2002.

Bousquet, O., Klochkov, Y., and Zhivotovskiy, N. Sharper bounds for uniformly stable algorithms. *Proceedings of Thirty Third Conference on Learning Theory*, May 2020.

Camuto, A., Deligiannidis, G., Erdogdu, M. A., Gürbüzbalaban, M., Şimşekli, U., and Zhu, L. Fractal Structure and Generalization Properties of Stochastic Optimization Algorithms. *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, June 2021.

Carlsson, G. Topological pattern recognition for point cloud data*. *Acta Numerica*, 23:289–368, May 2014. ISSN 0962-4929, 1474-0508. doi: 10.1017/S0962492914000051.

Chandramoorthy, N., Loukas, A., Gatmiry, K., and Jegelka, S. On the generalization of learning algorithms that do not converge. *Thirty-Sixth Conference on Neural Information Processing Systems (Neurips 2022)*, August 2022.

Chaudhari, P. and Soatto, S. Stochastic gradient descent performs variational inference, converges to limit cycles for deep networks. *2018 Information Theory and Applications Workshop (ITA)*, January 2018.

Clerico, E., Shidani, A., Deligiannidis, G., and Doucet, A. Chained Generalisation Bounds. *Proceedings of Thirty Fifth Conference on Learning Theory*, June 2022.

Corneanu, C., Madadi, M., Escalera, S., and Martinez, A. Computing the Testing Error without a Testing Set. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, May 2020.

Edelsbrunner, H. and Harer, J. Computational Topology - an Introduction — Semantic Scholar. *American Mathematical Society*, 2010.

Falconer, K. *Fractal Geometry - Mathematical Foundations and Applications - Third Edition*. Wiley, 2014.

Foster, D. J., Greenberg, S., Kale, S., Luo, H., Mohri, M., and Sridharan, K. Hypothesis Set Stability and Generalization. *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, October 2020.

Harutyunyan, H., Raginsky, M., Steeg, G. V., and Galstyan, A. Information-theoretic generalization bounds for black-box learning algorithms. *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, October 2021.

Herrera, C., Krach, F., and Teichmann, J. Estimating Full Lipschitz Constants of Deep Neural Networks. *Estimating Full Lipschitz Constants of Deep Neural Networks*, June 2020.

Hodgkinson, L., Şimşekli, U., Khanna, R., and Mahoney, M. W. Generalization Bounds using Lower Tail Exponents in Stochastic Optimizers. *Proceedings of the 39th International Conference on Machine Learning*, July 2022.

Hofer, C., Kwitt, R., Niethammer, M., and Uhl, A. Deep Learning with Topological Signatures. *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, February 2018.

Hofer, C., Kwitt, R., Dixit, M., and Niethammer, M. Connectivity-Optimized Representation Learning via Persistent Homology. *Proceedings of the 36th International Conference on Machine Learning*, June 2019.

Hu, W., Li, C. J., Li, L., and Liu, J.-G. On the diffusion approximation of nonconvex stochastic gradient descent. *Annals of Mathematical Sciences and Applications*, March 2018.

Jastrzebski, S., Kenton, Z., Arpit, D., Ballas, N., Fischer, A., Bengio, Y., and Storkey, A. Three Factors Influencing Minima in SGD, September 2018.

Jiang, Y., Neyshabur, B., Mobahi, H., Krishnan, D., and Bengio, S. Fantastic Generalization Measures and Where to Find Them. *ICLR 2020*, December 2019.

Kechris, A. S. *Classical Descriptive Set Theory*. Graduate Texts in Mathematics. Springer, 1995.

Kelley Pace, R. and Barry, R. Sparse spatial autoregressions. *Statistics & Probability Letters*, 33(3):291–297, May 1997. ISSN 0167-7152. doi: 10.1016/S0167-7152(96)00140-X.

Kendall, M. G. A new reasure of rank correlation. *Biometrika*, 1938.

Kendall, M. G. and Stuart, A. *The Advanced Theory of Statistics*. Griffin, 1973. ISBN 978-0-85264-069-2.

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., and Tang, P. T. P. On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima. *ICLR 2017*, February 2017.

Kozma, G., Lotker, Z., and Stupp, G. The minimal spanning tree and the upper box dimension. *Proceedings of the American Mathematical Society*, 134(4):1183–1187, September 2005. ISSN 0002-9939, 1088-6826. doi: 10.1090/S0002-9939-05-08061-5.

Krizhevsky, A., Nair, V., and Hinton, G. E. The cifar-10 dataset, 2014.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, May 2017. ISSN 0001-0782, 1557-7317. doi: 10.1145/3065386.

Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, November 1998. ISSN 1558-2256. doi: 10.1109/5.726791.

Ledoux, M. and Talagrand, M. *Probability in Banach Spaces - Isoperimetry and Processes*. Classics in Mathematics. Springer, 1991.

Mandt, S., Hoffman, M. D., and Blei, D. M. A Variational Analysis of Stochastic Gradient Algorithms. *Proceedings of The 33rd International Conference on Machine Learning*, February 2016.

Memoli, F. and Singhal, K. A Primer on Persistent Homology of Finite Metric Spaces. *Bulletin of Mathematical Biology*, 81(7):2074–2116, July 2019. ISSN 0092-8240, 1522-9602. doi: 10.1007/s11538-019-00614-z.

Molchanov, I. *Theory of Random Sets*. Number 87 in Probability Theory and Stochastic Modeling. Springer, second edition edition, 2017.

Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B., and Sutskever, I. Deep Double Descent: Where Bigger Models and More Data Hurt. *ICLR 2020*, December 2019.

Negrea, J., Haghifam, M., Dziugaite, G. K., Khisti, A., and Roy, D. M. Information-Theoretic Generalization Bounds for SGLD via Data-Dependent Estimates. *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, November 2019.

Pensia, A., Jog, V., and Loh, P.-L. Generalization Error Bounds for Noisy, Iterative Algorithms. *2018 IEEE International Symposium on Information Theory (ISIT)*, January 2018.

Pérez, J. B., Hauke, S., Lupo, U., Caorsi, M., and Dassatti, A. Giotto-ph: A Python Library for High-Performance Computation of Persistent Homology of Vietoris-Rips Filtrations, August 2021.

Pérez-Fernández, D., Gutiérrez-Fandiño, A., Armengol-Estapé, J., and Villegas, M. Characterizing and Measuring the Similarity of Neural Networks with Persistent Homology. *CoRR*, May 2021.

Pesin, Y. B. *Dimension Theory in Dynamical Systems - Contemporary Views and Applications*. Chicago Lectures in Mathematics. The University of Chicago Press, 1997.

Rebeschini, P. Algorithmic fundations of learning, 2020.

Rieck, B., Togninalli, M., Bock, C., Moor, M., Horn, M., Gumbsch, T., and Borgwardt, K. Neural Persistence: A Complexity Measure for Deep Neural Networks Using Algebraic Topology. *ICLR*, pp. 25 p., February 2019. doi: 10.3929/ethz-b-000327207.

Russo, D. and Zou, J. How much does your data exploration overfit? Controlling bias via information usage. *IEEE Transactions on Information Theory*, October 2019.

Schweinhart, B. Persistent Homology and the Upper Box Dimension. *Discrete & Computational Geometry volume 65, pages 331–364*, July 2019.

Schweinhart, B. Fractal Dimension and the Persistent Homology of Random Geometric Complexes. *Advances in Mathematics*, June 2020.

Şimşekli, U., Sener, O., Deligiannidis, G., and Erdogdu, M. A. Hausdorff Dimension, Heavy Tails, and Generalization in Neural Networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124014, December 2021. ISSN 1742-5468. doi: 10.1088/1742-5468/ac3ae7.

Steinke, T. and Zakynthinou, L. Reasoning About Generalization via Conditional Mutual Information. *Proceedings of Thirty Third Conference on Learning Theory*, June 2020.

Suzuki, T., Abe, H., and Nishimura, T. Compression based bound for non-compressed network: Unified generalization error analysis of large compressible deep neural network. *ICLR 2020*, June 2020.

van Erven, T. and Harremoës, P. Renyi Divergence and Kullback-Leibler Divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, July 2014. ISSN 0018-9448, 1557-9654. doi: 10.1109/TIT.2014.2320500.

Vershynin, R. *High-Dimensional Probability - An Introduction with Application in Data Science*. University of California - Irvine, 2020.

Xiao, Y. Random fractals and Markov processes. *Fractal Geometry and Applications: A jubilee of Benoît Mandelbrot - American Mathematical Society*, 72.2:261–338, 2004. doi: 10.1090/pspum/072.2/2112126.

Xu, A. and Raginsky, M. Information-theoretic analysis of generalization capability of learning algorithms. *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, November 2017.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. *ICLR 2017*, February 2017.

The outline of the appendix is as follows:

- Section A: Additional technical background related to information theory, Rademacher complexity, Egoroff's Theorem and persistent homology.
- Section B: Postponed proofs of the theoretical results.
- Section C: Additional experimental details.
- Section D: Additional experimental results, including full statistic of experiments presented in the main part of the paper, as well as additional experiments on different datasets.

A. Additional technical background

A.1. Information theoretic quantities

We recall there some basics concepts of information theory that we use throughout the paper. The absolute continuity of a probability measure with respect to another one will be denoted with symbol $\ll$.

Definition A.1. 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:

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

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

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

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 & 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:

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

The following property will be useful to perform decoupling of two random variables, a proof can be found in (van Erven & Harremoës, 2014).

Theorem A.2. *With the same notations as above, we have:*

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

We can then define the following notions of mutual information:

Definition A.3. 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:

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

I_∞ will be called the *total mutual information*. Note that thanks to Theorem A.2, we recover the definition of total mutual information that we wrote in Equation (6).

Those quantities satisfy the data processing inequality, given in the following proposition.

Proposition A.4 (Data-processing inequality)**.** *If $X \longrightarrow Y \longrightarrow Z$ is a Markov chain and $\alpha \in [1, +\infty]$, then:*

$$I_\alpha(X, Z) \leq I_\alpha(X, Y).$$

We are interested in those quantities because of their decoupling properties, summarized up in the following lemmas.:

Lemma A.5 (Lemma 1 in (Xu & 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:*

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

We end this subsection by stating the decoupling in probability result that we will use several times in the proofs: Combining the definition of total mutual information with Theorem A.2, we immediately obtain:

Lemma A.6 (Lemma 1 in (Hodgkinson et al., 2022))**.** *For every measurable set B we have:*

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

A.2. Rademacher complexity

We call Rademacher random variables a tuple $(\sigma_1 \ldots, \sigma_n)$ of mutually independent Bernoulli distributions with values in the set $\{-1, 1\}$.

Definition A.7. Let us consider a fixed set $A \subset \mathbb{R}^n$ and $\boldsymbol{\sigma} := (\sigma_1 \ldots, \sigma_n)$ some Rademacher random variables, the Rademacher complexity of A is defined as:

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

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

$$\ell(\mathcal{W}, S) = \{(\ell(w, z_i)_{1 \leq i \leq n} \in \mathbb{R}^n, \ w \in \mathcal{W}\}. \tag{17}$$

Remark A.8. One could legitimately inquire about the measurability of $\mathbf{Rad}(\ell(\mathcal{W}, S))$ with respect to $\mathcal{F}^{\otimes n}$ (recall that the data space is denoted $(\mathcal{Z}, \mathcal{F}, \mu_z)$). Thanks to the closedness 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 $\mathbf{Rad}(\ell(\mathcal{W}, S))$ is integrable (and measurable) thanks to the first part of Fubini's theorem.

Rademacher complexity is linked to the worst case generalization error via the following proposition (see for example (Rebeschini, 2020)):

Proposition A.9. *Assume that the loss is uniformly bounded by B. Then, for all $\eta > 0$, we have with probability $1 - 2\eta$ that:*

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

We state the proof of this result for the sake of completeness. It is based on two classical arguments: symmetrization and Mc-Diarmid inequality.

Proof. Let us write:

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

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:

$$\begin{aligned}
\mathbb{E}[\mathcal{G}(S)] &= \mathbb{E}\left[\sup_{w\in\mathcal{W}}\left(\frac{1}{n}\sum_{i=1}^{n}\mathcal{R}(w) - \ell(w, z_i)\right)\right] \\
&= \mathbb{E}\left[\sup_{w\in\mathcal{W}}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\ell(w,\tilde{z}_i) - \ell(w, z_i)|\tilde{S}]\right] \\
&\leq \mathbb{E}\left[\mathbb{E}\left[\sup_{w\in\mathcal{W}}\frac{1}{n}\sum_{i=1}^{n}(\ell(w,\tilde{z}_i) - \ell(w, z_i))\middle|\tilde{S}\right]\right] \\
&= \mathbb{E}\left[\sup_{w\in\mathcal{W}}\frac{1}{n}\sum_{i=1}^{n}(\ell(w,\tilde{z}_i) - \ell(w, z_i))\right] \\
&= \mathbb{E}\left[\sup_{w\in\mathcal{W}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i(\ell(w, z_i) - \ell(w,\tilde{z}_i))\right] \\
&\leq 2\mathbb{E}\left[\sup_{w\in\mathcal{W}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\ell(w, z_i)\right] \\
&= 2\mathbb{E}[\mathbf{Rad}(\ell(\mathcal{W}, S))].
\end{aligned} \tag{18}$$

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 $\epsilon > 0$:

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

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:

$$\mathcal{G}(S) \leq \mathbb{E}[\mathcal{G}(S)] + \sqrt{\frac{2B^2}{n}\log(1/\eta)}. \tag{19}$$

Moreover we can also write:

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

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

$$\mathbb{E}[\mathbf{Rad}(\ell(\mathcal{W}, S))] \leq \mathbf{Rad}(\ell(\mathcal{W}, S)) + \sqrt{\frac{2B^2}{n}\log(1/\eta)}. \tag{20}$$

Therefore combining equations 18, 19 and 20 gives us that with probability at least $1 - 2\eta$:

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

□

Another important result for us is the well-known Massart's lemma, presented here in a slightly simplified version which is enough for our work:

Lemma A.10 (Massart's lemma)**.** *Let $T \subseteq \mathbb{R}^n$ be a finite set, then:*

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

Where $|T|$ denotes the cardinal of T as usual.

Example A.11. Consider the setting where we have a fixed finite hypothesis set $\mathcal{W}$. 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 A.10 gives us

$$\mathbf{Rad}(\ell(\mathcal{W}, S)) \leq B\sqrt{\frac{2\log(|\mathcal{W}|)}{n}}. \tag{21}$$

A.3. Egoroff's Theorem

Egoroff's Theorem is an essential result in our theory which states that pointwise convergence in a probability space can be made uniform on measurable sets of arbitrary high probability. It was already used in (Şimşekli et al., 2021; Camuto et al., 2021) to make the convergence of the limit defining some fractal dimension uniform up certain probability.

Theorem A.12 (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.

A.4. Persistent Homology

Persistent homology (PH) is a well known notion in TDA typically used for point cloud analysis (Edelsbrunner & 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 A.13 (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.

The 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 A.14. 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:

$$\forall k\{w_1, \ldots, w_k\} \in \mathbf{Rips}(W, \delta) \iff \forall i, j,\ d(p_i, p_j) \leq \delta. \tag{22}$$

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.

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:

$$C_k(K) := \text{span}\big(\sum_i \epsilon_i s_i,\ \forall i,\ \dim(s_i) = k\big). \tag{23}$$

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:

$$\partial(s) = \sum_{i=0}^{k} s_{\setminus i}. \tag{24}$$

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 A.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: $\beta_k(K) = \dim(H_k(K))$.

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 A.16. 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 A.14, we collect the 'birth' and 'death' of each hole, the multiset[7] of those pairs (*birth*, *death*) will be the definition of persistent homology.

Remark A.17. In all the following, the parameter δ used in the definition of the Vietoris-Rips filtration will be seen as a time parameter.

While the concept of persistent homology can be extended to arbitrary orders (see (Boissonat et al., 2018)), here, for the sake of simplicity, we only define Persistent homology of degree 0, which is much simpler and is the only one we need in our work.

Persistent homology of degree 0:

The persistent homology of degree 0, denoted $\mathbf{PH}^0$ is the multiset of the distances δ used to build the Vietoris-Rips filtration of W for which a connected component is lost.

More formally, 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 $\delta_0 = 0$. With those notations, $\mathbf{PH}^0$ is the multiset of all the δ_i corresponding to a complex $K^{\delta_i,j}$ which has one less connected component than the preceding complex in the above filtration.

To stick with the usual notations, we actually define $\mathbf{PH}^0$ as the multiset of the $(0, \delta_i)$, where the 0 correspond to the 'birth' of a connected component, while the δ_i, as described above, corresponds to the 'death' of this connected component.

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

Definition A.18 (Persistent homology dimension)**.** For any $\alpha \geq 0$ we define:

$$E_\alpha(W) := \sum_{(b,d)\in \mathrm{PH}^0(\mathbf{Rips}(W))} (d-b)^\alpha. \tag{25}$$

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

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

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

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

Proposition A.19. *For any bounded metric space X, we have $\overline{\dim}_B(X) = \dim_{PH^0}(X)$.*

Proposition A.19 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 A.18 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 $\mathcal{W}$.

A.5. Numerical estimation of the PH dimension

In this section we 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.

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.

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 $X_n = \{x_1, \ldots, x_n\}$ of X we have:

$$\log E_\alpha(X_n) \leq \log D_{\alpha,\epsilon} + \left(1 - \frac{\alpha}{\Delta + \epsilon}\right)\log(n). \tag{26}$$

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 (26), which gives us $\Delta \simeq \frac{\alpha}{1-a}$.

Remark A.20. The aforementioned algorithm *works in pseudo metric spaces*. Indeed as we tried to explain formally in the proof of proposition B.9, 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 $\dim^{\rho_S}_{\mathrm{PH}^0}(X/\sim)$ which is proven in lemma B.9 to be equal to $\dim^{\rho_S}_{\mathrm{PH}^0}(X)$. See those notations in the next subsection.

A.6. Measurable coverings and additional technical lemmas

In this section we briefly discuss a few technical measure theoretic points that are worth mentioning. Essentially, we argue that our measurability assumptions ensure that the manipulations we make in our proofs on complicated random variables are valid and meaningful. We then show that it is possible to construct the measurable coverings that we need in our proofs.

A.6.1. Some nice consequences of our measurability assumptions

The worst-case generalization error takes the general form:

$$\mathcal{G}(S,U) := \sup_{\mathcal{W}_{S,U}} \left(\mathcal{R}(w) - \hat{\mathcal{R}}_S(w)\right). \tag{27}$$

Here we require $\mathcal{W}_{S,U} \subset \mathbb{R}^d$ to be a random closed set. 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 $E = \mathbb{R}^d$.

Remark A.21. As highlighted by (Molchanov, 2017), we can develop the following theory in the more general case where E is a locally compact Hausdorff second countable space, but we avoid those technical considerations.

The definition of a random closed set is the following:

Definition A.22 (Random closed set)**.** Consider a map $W : \Omega \longrightarrow \mathbf{CL}(E)$, 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 A.23 (Effrös σ-algebra and Fell topology)**.** The Effrös σ-algebra is the one generated by the sets $\{W \in \mathbf{CL}(E), W \cap K \neq \emptyset\}$ for K going over all compact sets in $\mathbb{R}^d$.

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

One can show that the Effrös σ-algebra on $\mathbf{CL}(E)$ 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 A.22 is equivalent to asking the measurability of W with respect to $\mathfrak{E}(E)$.

The assumption that we made on our learning algorithm is the following:

Assumption A.24. 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:

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

is measurable with respect to the Effrös σ-algebra.

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

Proposition A.25 (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 A.26. 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 3.1) we have that the worst case generalization error defined by Equation (27) is a well-defined random variable.

While Example A.26 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 the decoupling results and write:

$$\mathbb{P}_{W,S}\left(\sup_{w \in W} \mathcal{R}(w) - \hat{\mathcal{R}}_S(w) \geq \epsilon \right) \leq e^{I_\infty(W,S)} \mathbb{P}_W \otimes \mathbb{P}_S \left(\sup_{w \in W} \mathcal{R}(w) - \hat{\mathcal{R}}_S(w) \geq \epsilon \right). \tag{28}$$

In order for the decoupling lemmas 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[8].

Lemma A.27. *As before, let $(\mathbf{CL}(\mathbb{R}^d), \mathfrak{E}(\mathbb{R}^d))$ 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*

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

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

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 $\mathcal{Z}$ (that it is a sub-metric space of some $\mathbb{R}^N$).

Lemma A.28. *Assume that Ω is a Polish space with a dense countable subset D and that ζ is continuous in both variables. Then the function:*

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

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.

□

A.6.2. Construction of measurable coverings

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?

Recall that we defined a δ-covering of some set X as a minimal set of points N_δ of X such that:

$$X \subseteq \bigcup_{w \in N_\delta} \bar{B}_\delta^\rho(w)$$

[8]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

The fact that we ask the coverings to be in X is for technical reasons and does not change the values of the dimensions.

We first need a technical lemma to ensure that it is equivalent to cover dense countable subsets:

Lemma A.29 (Closure property of coverings)**.** *Let W be a closed set and $\mathcal{C}$ be a countable dense subset of W. Under Assumption 3.1 we have that any covering of $\mathcal{C}$ is a covering of W for pseudo-metric ρ_S. Moreover we have, for all $\delta > 0$:*

$$|N_{2\delta}^{\rho_S}(\mathcal{C})| \leq |N_{\delta}^{\rho_S}(W)| \leq |N_{\delta}^{\rho_S}(\mathcal{C})|. \tag{29}$$

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 the triangle inequality we have:

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

□

Let us first prove that we can construct measurable coverings in the case of fixed hypothesis sets. 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 required for the high probability bounds in our results to make sense. This kind of measurability condition is 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 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 $\mathcal{F}$ is the corresponding Borel σ-algebra. With that in mind we prove the following theorem:

Theorem A.30 (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[9] of $\mathcal{W}$ and $\delta > 0$. Under Assumption 3.1, 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.

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

$$\{S \in \mathcal{Z}^n,\ |N_{\delta}^{\rho_S}(\mathcal{C})| \leq M\} = \bigcup_{F \in \mathfrak{F}_{\leq M}(\mathcal{C})} \bigcap_{k \in \mathbb{N}} \bigcup_{c \in F} \{S,\ \rho_S(c, w_k) \leq \delta\}. \tag{30}$$

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.

□

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

Remark A.31. Given any positive sequence δ_k, decreasing and converging to 0, thanks to Lemma A.29 the upper box-counting dimension can be written as

$$\overline{\dim}_B^{\rho_S}(\mathcal{W}) = \limsup_{k\to+\infty} \frac{\log |N_{\delta_k}^{\rho_S}(\mathcal{C})|}{\log(1/\delta_k)}, \tag{31}$$

which, by Theorem A.30, implies that $\overline{\dim}_B^{\rho_S}(\mathcal{W})$ is a random variable as countable upper limit of random variables.

We now come to the case of random hypothesis sets, we begin by introducing 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 A.32 (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:*

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

Equipped with this result, we can easily extend Theorem A.30 to the measurability of the covering numbers associated to a Castaing's representation of the hypothesis set:

Theorem A.33 (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 3.1, we have that the mapping between probability spaces

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

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 A.30 except that now have a Castaing's representation $(\xi_n)_{n\geq 1}$ of W.

By the same proof than Equation (30), 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 A.29.

□

Thanks to Theorem A.33, we are actually able to prove the much stronger result that we *can* build measurable coverings.

Theorem A.34 (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 3.1, 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 A.30, 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:

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

Note that i_0 is finite because $\{(\ell(w,z_i)_{1\leq i\leq n}),\ w_i nW\}$ 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 A.23.

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 $\mathfrak{F}_N(\mathbb{N}_+)$:

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

By Theorem A.33, 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 A.29, $N_\delta(S,\omega)$ also defines a covering of W and we have:

$$\overline{\dim}_B^{\rho_S}(W(\omega))=\limsup_{\delta\to 0}\frac{|N_\delta(S,\omega)|}{\log(1/\delta)}. \tag{32}$$

□

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

Remark A.35. Theorem A.34 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 always 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.This is formalized by Assumption 3.2 in the main part of the paper.

B. Postponed proofs

B.1. Proof of Theorem 3.4

This proof essentially uses classical arguments related to Rademacher complexity.

Proof. **Step** 0**:** First of all, as $\mathcal{W}$ is closed, we can consider a dense countable subset $\mathcal{C}$[10]. 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 A.29, 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 A.9 we have:

$$G(S) \leq 2\mathbf{Rad}(\ell(\mathcal{W}, S)) + 3\sqrt{\frac{2B^2}{n}\log(1/\eta)}. \tag{33}$$

Step 2**:**

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

$$\overline{\dim}_B^{\rho_S}(\mathcal{W}) := \limsup_{r \to 0} \frac{\log(|N_r|)}{\log(1/r)}. \tag{34}$$

Thanks to Theorem A.30 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:

$$\overline{\dim}_B^{\rho_S}(\mathcal{W}) := \limsup_{k \to \infty} \frac{\log(|N_{r_k}|)}{\log(1/r_k)}. \tag{35}$$

Let $\gamma > 0$, by Egoroff's Theorem A.12 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:

$$\forall \delta \leq \delta_{n,\gamma,\epsilon},\ \log(|N_\delta|) \leq (\epsilon + \overline{\dim}_B^{\rho_S}(\mathcal{W}))\log(1/\delta). \tag{36}$$

[10] 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 A.30, it does not change anything to the proof.

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 A.10):

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

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

$$\mathbf{Rad}(\ell(\mathcal{W}, S)) \leq \delta + \mathbb{E}_\sigma\left[\max_{w\in N_r}\frac{1}{n}\sigma^T\ell(w, S)\right] \leq \delta + B\sqrt{\frac{2\log(1/\delta)}{n}(\epsilon + \overline{\dim}_B^{\rho_S}(\mathcal{W}))}. \tag{37}$$

Putting together equations 33 and 37 we get that with probability at least $1-2\eta-\gamma$, for $\delta \leq \delta_{n,\gamma,\epsilon}$:

$$G(S) \leq 2\delta + 2B\sqrt{\frac{4(\epsilon + d(S))\log(1/\delta) + 9\log(1/\eta)}{n}}. \tag{38}$$

□

Remark B.1. An important remark can be made at this point. One can see that δ is still appearing in Equation (38), this is due to the possible lack of uniformity in the limit defined in Equation (34). That way the quantity $\log(1/\delta_{n,\gamma,\epsilon})$ may be seen as a sort of speed of convergence of the upper box-counting dimension. Theorem 3.4, as well as our other main results (Theorems 3.5 and 3.8) may be made uniform in n by further assumption of uniformity in n on the convergence of the limit defining the upper box-counting dimension, i.e. Equation (34), meaning that in that case $\delta_{n,\gamma,\epsilon}$ will not depend on n. This would allow us to proceed as in the proof of Lemma $S1$ in (Şimşekli et al., 2021) and set $\delta = \delta_n := 1/\sqrt{n}$, at the cost of making the bound asymptotic in n.

B.2. Proof of Theorem 3.5

Before going to the proof, let us make a few remarks on the introduced approximated level sets of the empirical risk.

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 A.1 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[11] for some $K \in \mathbb{N}_+$:

$$R_S^j := \mathcal{W}_{S,U} \cap \mathcal{R}^{-1}\left(\left[\frac{jB}{K}, \frac{(j+1)B}{K}\right]\right), \tag{39}$$

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 $N_{\delta,j}$.

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

Remark B.2. 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 our result 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 A.6 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 A.34 or we can restrict ourselves to Castaing's representations of $\mathcal{W}_{S,U}$.

As can already be noted in Remark B.2, 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.

Proof. 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 [12] $N_{\delta,j}$ corresponding to the centers of a minimal covering[13] 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:

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

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

$$\begin{aligned} \big(\mathcal{R}(w) - \hat{\mathcal{R}}_S(w)\big) &\le \big(\mathcal{R}(w') - \hat{\mathcal{R}}_S(w')\big) + \rho_S(w, w') + |\mathcal{R}(w) - \mathcal{R}(w')| \\ &\le \big(\mathcal{R}(w') - \hat{\mathcal{R}}_S(w')\big) + \delta + \frac{B}{K}. \end{aligned} \tag{40}$$

So that we get:

$$\sup_{w \in \mathcal{W}_S} \big(\mathcal{R}(w) - \hat{\mathcal{R}}_S(w)\big) \le \delta + \frac{B}{K} + \max_{0 \le j \le K-1} \max_{w \in N_{\delta,j}} \big(\mathcal{R}(w) - \hat{\mathcal{R}}_S(w)\big). \tag{41}$$

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

[13] Without loss of generality we can always assume that those sets 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 change the mutual information term in the final results (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 $-\infty$.

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

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

We have by the decoupling lemma A.6 along with Fubini's Theorem, Hoeffding inequality and a union bound:

$$\begin{aligned}
&\mathbb{P}\Big(\max_{0\le j\le K-1}\max_{w\in N_{\delta,j}}\big(\mathcal{R}(w)-\hat{\mathcal{R}}_S(w)\big) \ge \epsilon(\max_j |N_{\delta,j}|,\ \max_j I_\infty(S, N_{\delta,j}))\Big)\\
&\qquad\le \sum_{j=0}^{K-1}\mathbb{P}\Big(\max_{w\in N_{\delta,j}}\big(\mathcal{R}(w)-\hat{\mathcal{R}}_S(w)\big) \ge \epsilon(|N_{\delta,j}|, I_\infty(S, N_{\delta,j}))\Big)\\
&\qquad\le \sum_{j=0}^{K-1} e^{I_\infty(S,N_{\delta,j})}\mathbb{P}_{N_{\delta,j}}\otimes\mathbb{P}_S\Big(\max_{w\in N_{\delta,j}}\big(\mathcal{R}(w)-\hat{\mathcal{R}}_S(w)\big) \ge \epsilon(|N_{\delta,j}|, I_\infty(S, N_\delta^j))\Big)\\
&\qquad\le \sum_{j=0}^{K-1} e^{I_\infty(S,N_{\delta,j})}\mathbb{E}_{N_{\delta,j}}\Big[\mathbb{P}_S\Big(\max_{w\in N_{\delta,j}}\big(\mathcal{R}(w)-\hat{\mathcal{R}}_S(w)\big) \ge \epsilon(|N_{\delta,j}|, I_\infty(S, N_\delta^j))\Big)\Big]\\
&\qquad\le \sum_{j=0}^{K-1} e^{I_\infty(S,N_{\delta,j})}\mathbb{E}_{N_{\delta,j}}\Big[\sum_{w\in N_{\delta,j}}\mathbb{P}_S\Big(\big(\mathcal{R}(w)-\hat{\mathcal{R}}_S(w)\big) \ge \epsilon(|N_{\delta,j}|, I_\infty(S, N_\delta^j))\Big)\Big]\\
&\qquad\le \sum_{j=0}^{K-1} e^{I_\infty(S,N_{\delta,j})}\mathbb{E}_{N_{\delta,j}}\Big[|N_{\delta,j}|\exp\Big\{-\frac{n\epsilon(|N_{\delta,j}|, I_\infty(S, N_\delta^j))^2}{2B^2}\Big\}\Big]\\
&\qquad\le \sum_{j=0}^{K-1} e^{I_\infty(S,N_{\delta,j})}\mathbb{E}_{N_{\delta,j}}\Big[\frac{\eta}{K}e^{-I_\infty(S,N_{\delta,j})}\Big]\\
&\qquad= \eta.
\end{aligned} \tag{42}$$

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 $\forall j,\ |N_{\delta,j}| \le |N_{\delta/2}|$, and thus:

$$\max_{0\le j\le K-1}|N_{\delta,j}| \le |N_{\delta/2}|. \tag{43}$$

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 $\delta \le \delta_{n,\gamma,\epsilon}$ we have:

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

Therefore combining equations 40, 42, 43, we get that with probability at least $1-\gamma-\eta$, for all $\delta \leq \delta_{n,\gamma,\epsilon}$:

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

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

$$\sup_{w\in\mathcal{W}_S}\left(\mathcal{R}(w)-\hat{\mathcal{R}}_S(w)\right) \leq \delta + \frac{B}{\sqrt{n}-1} + \sqrt{2}B\sqrt{\frac{\log(\sqrt{n}/\eta)+\log(2/\delta)(\epsilon+\overline{\dim}_B^{\rho_S}(\mathcal{W}_S))+I_{n,\delta}}{n}}. \tag{44}$$

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.

□

B.3. Proof of Theorem 3.8

Here we present the proof of Theorem 3.8. 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 $J_1, \ldots, J_H$ 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 when the latter is possible and then derive two corollaries to deal with this technical issue, mostly based on the boundedness assumption. Theorem 3.8 will be the second corollary.

Remark B.3. 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 3.8 to a convergence rate in $n^{-1/2}$ which is optimal anyway.

Let us start with the main result of this section:

Proposition B.4. *We make assumptions 3.1, 3.2 and 3.7 with the same notations than in Theorem 3.8. 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:

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

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 3.5, we introduce some approximate level sets R_S^j for $j \in \{0, \ldots, K-1\}$. 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 remark we made in the proof of theorem 3.5 about the assumptions that R_S^j are non-empty without loss of generality still holds in the setting described hereafter.

The proof starts by introducing the "level-sets" of the population risk as in proof of Theorem 3.5. We define R_S^j exactly in the same way.

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:

$$\sup_{w\in\mathcal{W}_{s,U}} |\mathcal{R}(w) - \hat{\mathcal{R}}_S(w)| \leq \delta + \frac{B}{K} + \max_{0\leq j\leq K-1} \underbrace{\max_{w\in N_{\delta,j}(S,S,U)} |\mathcal{R}(w) - \hat{\mathcal{R}}_S(w)|}_{:=E_j}. \tag{45}$$

Thanks to our stability assumption 3.7, 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 term 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 (45) we get:

$$\begin{aligned}
\sup_{w\in\mathcal{W}_{s,U}} |\mathcal{R}(w) - \hat{\mathcal{R}}_S(w)| &\leq \delta + \frac{B}{K} + \frac{2J\beta}{n^\alpha} + \max_{0\leq j\leq K-1} \sum_{k=1}^{H} \max_{w\in N_{\delta,j}(S,S^{\setminus J_k},U)} \frac{1}{n} \left| \sum_{i\in J_k} \big(\ell(w,z_i) - \mathcal{R}(w)\big) \right| \\
&\leq \delta + \frac{B}{K} + \frac{2J\beta}{n^\alpha} + H \max_{0\leq j\leq K-1} \max_{1\leq k\leq H} \underbrace{\max_{w\in N_{\delta,j}(S,S^{\setminus J_k},U)} \frac{1}{n} \left| \sum_{i\in J_k} \big(\ell(w,z_i) - \mathcal{R}(w)\big) \right|}_{:=M_{j,k}(S,U)}.
\end{aligned} \tag{46}$$

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:

$$\begin{aligned}
\mathbb{P}\big(M_{j,k}(S,U) \geq \epsilon\big) &\leq e^{I_\infty(N_{\delta,j}(S,S^{\setminus J_k},U),S_{J_k})}\mathbb{P}_{N_{\delta,j}(S,S^{\setminus J_k},U)} \otimes \mathbb{P}_{S_{J_k}}\big(M_{j,k}(S,U) \geq \epsilon\big) \\
&\leq e^{I_\infty(N_{\delta,j}(S,S^{\setminus J_k},U),S_{J_k})}\mathbb{E}_{N_{\delta,j}(S,S^{\setminus J_k},U)}\Big[\mathbb{P}_{S_{J_k}}\big(M_{j,k}(S,U) \geq \epsilon\big)\Big] \\
&\leq e^{I_\infty(N_{\delta,j}(S,S^{\setminus J_k},U),S_{J_k})} \\
&\quad \times \mathbb{E}_{N_{\delta,j}(S,S^{\setminus J_k},U)}\Bigg[\mathbb{P}_{S_{J_k}}\Bigg(\bigcup_{w\in N_{\delta,j}(S,S^{\setminus J_k},U)}\Bigg\{\frac{1}{n}\Bigg|\sum_{i\in J_k}\big(\ell(w,z_i)-\mathcal{R}(w)\big)\Bigg| \geq \epsilon\Bigg\}\Bigg)\Bigg] \\
&\leq e^{I_\infty(N_{\delta,j}(S,S^{\setminus J_k},U),S_{J_k})}\mathbb{E}\Big[|N_{\delta,j}(S,S^{\setminus J_k},U)|e^{-\frac{2\epsilon^2 n^2}{JB^2}}\Big].
\end{aligned} \tag{47}$$

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:

$$S_{J_k} \longrightarrow \mathcal{W}_{S,U} \longrightarrow N_{\delta,j}(S,S^{\setminus J_k},U). \tag{48}$$

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:

$$I_\infty(N_{\delta,j}(S,S^{\setminus J_k},U),S_{J_k}) \leq I_\infty(S,\mathcal{W}_{S,U}). \tag{49}$$

Note that the mutual information term appearing in equation (49) is the same than the one appearing in (Hodgkinson et al., 2022).

Thus:

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

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 an 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 than in the proof of Theorem 3.5, 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:

$$|N_{\delta,j}(S, S^{\setminus J_k}, U)| \leq |N_{\frac{\delta}{4}}(S, S, U)|. \tag{50}$$

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:

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

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

$$\begin{aligned}\sup_{w \in \mathcal{W}_{s,U}} |\mathcal{R}(w) - \hat{\mathcal{R}}_S(w)| \leq \delta + \frac{B}{K} + \frac{2J\beta}{n^\alpha} \\ + H\sqrt{\frac{JB^2}{2n^2}\left(\left(\epsilon + \overline{\dim}_B^{d_S}(\mathcal{W}_{S,U})\right)\log(4/\delta) + \log(HK/\eta) + I_\infty(S, \mathcal{W}_{S,U})\right)}.\end{aligned} \tag{51}$$

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

□

Corollary B.5. *With the exact same setting than in proposition B.4, 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:*

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

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 (51) 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:

$$\lambda = \frac{2\alpha}{3}. \tag{52}$$

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:

$$\sup_{w\in\mathcal{W}_{s,U}} |\mathcal{R}(w) - \hat{\mathcal{R}}_S(w)| \leq \delta + \frac{B+2\beta}{n^{\alpha/3}} + B\left\{\frac{\log(1/\eta) + \left(1-\frac{\alpha}{3}\right)\log(n) + I + \left(\epsilon + d(S,U)\right)\log(4/\delta)}{2n^{\frac{2\alpha}{3}}}\right\}^{\frac{1}{2}}. \quad (53)$$

□

Remark B.6. The asymptoticity in δ defined by $\delta_{\gamma,\epsilon,n}$ above accounts for the asymptoticity coming both from the stability assumption (definition 3.6) and the convergence of the limit defining the upper box-counting dimension.

Now we prove Theorem 3.8 which is based on the same idea than the previous corollary, but when $n^{\alpha/3} \notin \mathbb{N}$.

Theorem B.7. *under the same assumptions and notations than proposition B.4. We have that for* $n \geq C(\alpha) := \max\{2^{\frac{3}{2\alpha}}, 2^{1+\frac{3}{3-2\alpha}}\}$*, with probability* $1-\gamma-\eta$*, for all* δ *smaller than some* $\delta_{\gamma,\epsilon,n} > 0$ *we have:*

$$\sup_{w\in\mathcal{W}_{s,U}} |\mathcal{R}(w) - \hat{\mathcal{R}}_S(w)| \leq \delta + \frac{3B+2\beta}{n^{\alpha/3}} + B\sqrt{\frac{\log(1/\eta) + \left(1-\frac{\alpha}{3}\right)\log(n) + I + \left(\epsilon + d(S,U)\right)\log(4/\delta)}{2n^{\frac{2\alpha}{3}}}}. \quad (54)$$

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:

$$|\hat{\mathcal{R}}_S(w) - \mathcal{R}(w)| \leq \frac{n-\tilde{n}}{n}B + \frac{\tilde{n}}{n}\left|\frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}} \ell(w,z_i) - \mathcal{R}(w)\right|. \quad (55)$$

For the first term we write:

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

The idea is to apply the proof of Theorem B.4 to the last term of equation (55), replacing d_{S_n} with $d_{S_{\tilde{n}}}$. For clarity we still denote $S = (z_1, \ldots, z_n)$ and $S_{\tilde{n}} = (z_1, \ldots, z_{\tilde{n}})$

There are several terms we need to consider:
The mutual information term: The two data processing inequality we apply to prove equation (49) 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 (50), we have:

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

We have:

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

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 (51), 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 (55). 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:

$$n \geq C(\alpha) := \max\{2^{\frac{3}{2\alpha}}, 2^{1+\frac{3}{3-2\alpha}}\}. \tag{56}$$

Putting all of this together, we get that for $n \geq C(\alpha)$ (defined in equation (56)), with probability $1-\gamma-\eta$, for all δ smaller than some $\delta_{\gamma,\epsilon,n} > 0$ we have:

$$\sup_{w\in\mathcal{W}_{s,U}} |\mathcal{R}(w) - \hat{\mathcal{R}}_S(w)| \leq \delta + \frac{3B+2\beta}{n^{\alpha/3}} + B\sqrt{\frac{\log(1/\eta) + \left(1-\frac{\alpha}{3}\right)\log(n) + I + \left(\epsilon + d(S,U)\right)\log(4/\delta)}{2n^{\frac{2\alpha}{3}}}}. \tag{57}$$

□

B.4. Proof of Theorem 4.1

Let (X,ρ) be a pseudo-metric space, we introduce the equivalence relation:

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

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

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

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

We prove that upper box-counting dimension and persistent homology dimension are 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 B.8 (Upper-box dimension with pseudo metric)**.**

$$\overline{\dim}_B(X) = \overline{\dim}_B(X/\sim). \tag{58}$$

Let $N_{\delta}^{d}(F)$ denote the minimum number of **closed δ-balls** coverings of F for the (pseudo)-metric d.

Proof. Let $F \subset X$, bounded. Let $\{x_1,\ldots,x_n\}$ 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. □

The result is also quite obvious for the persistent homology dimension, even though it is a bit more complicated to write it. For more details on persistent homology please refer to (Boissonat et al., 2018; Memoli & Singhal, 2019; Schweinhart, 2019).

Lemma B.9 (Persistent homology dimension in pseudo metric spaces)**.**

$$\dim_{PH^0}(X) = \dim_{PH^0}(X/\sim). \tag{59}$$

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:

$$\pi([a_0, \dots, a_s]) := [\pi(a_0), \dots, \pi(a_s)]. \tag{60}$$

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 (60). 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:

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

where $0 \leq \delta_1 < \cdots < \delta_C$ 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 $i \in 0, \ldots, C$ and $j \in 1, \ldots, \alpha_i$ 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 $\pi \circ s = \text{Id}$, 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:

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

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 (25). 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.

□

Combination of Equation (15), lemma B.8 and Lemma B.9 immediately gives the proof of Theorem 4.1.

B.5. Proof of Theorem 3.9

In this subsection, we show how we can leverage very classical tools from high dimensional probability to give one first step toward proving lower bounds, even though the obtained lower bound may look a bit disappointing. We combine two tools, namely Gaussian complexity and Sudakov's theorem.

Definition B.10 (Gaussian complexity)**.** Given a set $A \subset \mathbb{R}^n$, and $g_1, \ldots, g_n \sim \mathcal{N}(0, 1)$ independent, the Gaussian complexity of A is defined by:

$$\Gamma(A) := \frac{1}{n}\mathbb{E}_g\left[\sup_{a \in A} \sum_{i=1}^{n} g_i a_i\right]$$

As before we will denote, for $S \in \mathcal{Z}^n$:

$$\Gamma(\ell(\mathcal{W}, S)) := \frac{1}{n}\mathbb{E}_g\left[\sup_{w \in \mathcal{W}} \sum_{i=1}^{n} g_i \ell(w, z_i)\right]$$

We have the following lower bound of Rademacher complexity

Lemma B.11. *We have:*

$$\mathbf{Rad}(A) \geq \frac{1}{2\sqrt{\log(n)}}\Gamma(A)$$

Proof. For Rademacher random variables $\sigma_1, \dots, \sigma_n$, let us define the following function, for $\alpha = (\alpha_1, \dots, \alpha_n) \in [0,1]^n$:

$$f(\alpha) := \mathbb{E}_\sigma \left[\sup_{a \in A} \sum_{i=1}^n \sigma_i \alpha_i a_i \right].$$

it is easy to see that f is convex and continuous on the compact set $[0,1]^n$. Therefore we know that f attains its maximum for some $\alpha^0 \in [0, 1^n.]$. We denote the constant one vector by $\mathbb{1}_n \in \mathbb{R}^n$. For some $0 \leq \lambda \leq 1$, let us write $\alpha^0 = \lambda\alpha + (1-\lambda)\mathbb{1}^n$, for some α. We have, by convexity:

$$f(\alpha^0) \leq \lambda f(\alpha) + (1-\lambda) f(\mathbb{1}^n) \leq \lambda f(\alpha^0) + (1-\lambda)\mathbf{Rad}(A),$$

which implies that:

$$\mathbb{E}_\sigma \left[\sup_{a \in A} \sum_{i=1}^n \sigma_i \alpha_i a_i \right] \leq \mathbf{Rad}(A) \tag{61}$$

Now let $g_1, \dots, g_n \sim \mathcal{N}(0,1)$ be independent normal random variables. Let $g_\infty := \max_i(|g_i|)$. It is possible to write the following decomposition: $\forall i,\ g_i = |g_i|\sigma_i$ where the σ_i are Rademacher random variables **independent of** $|g_i|$.

As $g_\infty > 0$ almost surely, we have:

$$\begin{aligned}
\Gamma(A) &:= \frac{1}{n}\mathbb{E}_g \left[\sup_{a \in A} \sum_{i=1}^n g_i a_i \right] \\
&\leq \frac{1}{n}\mathbb{E}_{g_\infty} \left[g_\infty \mathbb{E}_\sigma \left[\sup_{a \in A} \sum_{i=1}^n \sigma_i \frac{|g_i|}{g_\infty} a_i \right] \right], \quad \text{(because } \mathbf{Rad} \text{ is non negative)} \\
&\leq \frac{1}{n}\mathbb{E}_{g_\infty} \left[g_\infty \mathbb{E}_\sigma \left[\sup_{a \in A} \sum_{i=1}^n \sigma_i a_i \right] \right], \quad \text{(by Equation (61))} \\
&= \frac{1}{n}\mathbb{E}_{g_\infty}[g_\infty]\mathbf{Rad}(A), \quad \text{(Fubini's theorem).}
\end{aligned}$$

We conclude by using that $\mathbb{E}[g_\infty] \leq 2\sqrt{\log(n)}$.

□

Remark B.12. It is also possible to prove that $\mathbf{Rad}(A) \leq \sqrt{\frac{\pi}{2}}\Gamma(A)$

The key ingredient for the lower bound is Sudakov's theorem, see (Vershynin, 2020, Section 7):

Theorem B.13 (Sudakov's theorem)**.** *Let $(X_t)_{t \in T}$ be a mean zero gaussian process, then for any $\epsilon > 0$ we have:*

$$\mathbb{E}\left[\sup_{t \in T} X_t \right] \geq C\delta\sqrt{N_\delta(T,d)},$$

where C is an absolute constant and $N_\epsilon(T,d)$ the covering number of T for the following pseudo-metric:

$$d(t,s)^2 := \mathbb{E}[(X_t - X_s)^2]$$

To prove our lower bound, we first need a lower bound of the expected worst case generalization error in terms of Rademacher complexity:

Proposition B.14. *Desymmetrization inequality Assume that the loss ℓ is in the interval $[0, B]$ (this does not make us lose generality). Then we have:*

$$\mathbb{E}\left[\sup_{w \in \mathcal{W}} \left| \mathcal{R}(w) - \hat{\mathcal{R}}_S(w) \right| \right] \geq \frac{1}{2}\mathbb{E}\big[\mathbf{Rad}(\ell(\mathcal{W}, S))\big] - B\sqrt{\frac{\log(2)}{2n}}.$$

While this result is classical, we present a proof for the sake of completeness, and to exhibit the absolute constants that we get in our case.

Proof. Similarly to the symmetrization inequality, we write, with $(z_i')_i$ an independent copy of $(z_i)_i$ and $(\sigma_i)_i$ independent Rademacher random variables:

$$\begin{aligned}
\mathbb{E}\big[\mathbf{Rad}(\ell(\mathcal{W},S))\big] &\leq \mathbb{E}\bigg[\frac{1}{n}\sup_{w\in\mathcal{W}}\sum_{i=1}^n \sigma_i\big(\ell(w,z_i)-\mathcal{R}(w)\big)\bigg] + \mathbb{E}\bigg[\frac{1}{n}\sup_{w\in\mathcal{W}}\sum_{i=1}^n \sigma_i\mathcal{R}(w)\bigg] \\
&\leq \mathbb{E}\bigg[\frac{1}{n}\sup_{w\in\mathcal{W}}\sum_{i=1}^n \sigma_i\big(\ell(w,z_i)-\ell(w,z_i')\big)\bigg] + B\mathbb{E}\bigg[\frac{1}{n}\bigg|\sum_{i=1}^n\sigma_i\bigg|\bigg], \quad \text{(Jensen's inequality)} \\
&\leq \mathbb{E}\bigg[\frac{1}{n}\sup_{w\in\mathcal{W}}\sum_{i=1}^n \big(\ell(w,z_i)-\ell(w,z_i')\big)\bigg] + B\mathbb{E}\bigg[\frac{1}{n}\bigg|\sum_{i=1}^n\sigma_i\bigg|\bigg], \quad \text{(Symmetrization argument)} \\
&\leq 2\mathbb{E}\bigg[\sup_{w\in\mathcal{W}}\big|\mathcal{R}(w)-\hat{\mathcal{R}}_S(w)\big|\bigg] + B\mathbb{E}\bigg[\frac{1}{n}\bigg|\sum_{i=1}^n\sigma_i\bigg|\bigg], \quad \text{(Triangle inequality)} \\
&\leq 2\mathbb{E}\bigg[\sup_{w\in\mathcal{W}}\big|\mathcal{R}(w)-\hat{\mathcal{R}}_S(w)\big|\bigg] + B\sqrt{\frac{2\log(2)}{n}}, \quad \text{(Simple case of Massart's lemma)},
\end{aligned}$$

hence the result.

□

Then we can prove the following result:

Proposition B.15. *Lower bound in term of covering numbers Assume that the loss ℓ is bounded by $B>0$, then there is an absolute constant $c>0$ (the one coming from Sudakov theorem) such that with probability at least $1-\zeta$, for all $\delta>0$ we have*

$$\sup_{w\in\mathcal{W}}\big|\mathcal{R}(w)-\hat{\mathcal{R}}_S(w)\big| \geq \frac{c}{4}\sqrt{\frac{\delta^2\log|N_\delta^{\rho_S}(\mathcal{W})|}{n\log(n)}} - B\sqrt{\frac{\log(2)+9\log(1/\zeta)}{n}},$$

where ρ_S is the data-dependent metric already used before in this project (based on an L^1 empirical mean).

Proof. Using the same reasoning, based on Mc-Diarmid's inequality, than for proving the upper bound, we write successively that with probability at least $1-\zeta$:

$$\begin{aligned}
\sup_{w\in\mathcal{W}}\big|\mathcal{R}(w)-\hat{\mathcal{R}}_S(w)\big| &\geq \mathbb{E}\bigg[\sup_{w\in\mathcal{W}}\big|\mathcal{R}(w)-\hat{\mathcal{R}}_S(w)\big|\bigg] - B\sqrt{\frac{2\log(1/\zeta)}{n}}, \quad \text{(Mc-Diarmid's inequality)} \\
&\geq \frac{1}{2}\mathbb{E}\big[\mathbf{Rad}(\ell(\mathcal{W},S))\big] - B\sqrt{\frac{\log(2)}{2n}} - \sqrt{\frac{2\log(1/\zeta)}{n}} \\
&\geq \frac{1}{2}\mathbf{Rad}(\ell(\mathcal{W},S)) - B\sqrt{\frac{\log(2)}{2n}} - \frac{3}{2}\sqrt{\frac{2\log(1/\zeta)}{n}}, \quad \text{(Mc-Diarmid's inequality)} \\
&\geq \frac{1}{4\sqrt{\log(n)}}\Gamma(\ell(\mathcal{W},S)) - B\sqrt{\frac{\log(2)}{2n}} - 3\sqrt{\frac{\log(1/\zeta)}{2n}}
\end{aligned}$$

Now we note that:

$$\Gamma(\ell(\mathcal{W},S)) := \frac{1}{n}\mathbb{E}_g\bigg[\sup_{w\in\mathcal{W}}\sum_{i=1}^n g_i\ell(w,z_i)\bigg],$$

and introduce the following gaussian process:

$$\forall w\in\mathcal{W},\ X_w := \frac{1}{\sqrt{n}}\sum_{i=1}^n g_i\ell(w,z_i).$$

The L^2 distance induced by this gaussian process on $\mathcal{W}$ can be computed by:

$$\begin{aligned} d(w,w')^2 &= \frac{1}{n}\mathbb{E}\left[\left(\sum_{i=1}^n g_i(\ell(w,z_i)-\ell(w',z_i))\right)^2\right] \\ &= \frac{1}{n}\sum_{i=1}^n(\ell(w,z_i)-\ell(w',z_i))^2 \\ &\geq \rho_S(w,w')^2, \quad \text{(Cauchy-Schwarz's inequality)} \end{aligned}$$

where

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

is the data-dependent pseudo-metric we used previously in this work. The result then follows by applying Sudakov's theorem.

□

Using this proposition, we can prove the following result:

Theorem B.16 (Lower bound with data-dependent fractal dimension)**.** *Assume that the loss ℓ is bounded by $B>0$ and that almost surely we have $\underline{\dim}_B^{\rho_S}(\mathcal{W})>0$. Then, for all $\gamma,\zeta>0$ there is an absolute constant $c>0$ and some $\delta_{n,\gamma,\zeta}>0$ such that, with probability at least $1-\zeta-\gamma$, for all $\delta\leq\delta_{n,\gamma,\zeta}$ we have:*

$$\sup_{w\in\mathcal{W}}\left|\mathcal{R}(w)-\hat{\mathcal{R}}_S(w)\right| \geq \frac{c}{4}\sqrt{\frac{\delta^2\log(1/\delta)d(S)}{2n\log(n)}} - B\sqrt{\frac{\log(2)+9\log(1/\zeta)}{n}}.$$

Remark B.17. As many of our results, the more interesting part of this result is the underlying covering numbers bound, the annoying asymptoticity in δ being introduced when we go from the covering numbers to the data-dependent fractal dimensions.

Proof. Let us fix $\gamma,\zeta\in(0,1)$. Using the definition of the lower box-counting dimension and the fact that $\underline{\dim}_B^{\rho_S}(\mathcal{W})>0$ almost surely, we can write:

$$\liminf_{\delta\to 0}\frac{\log|N_\delta^{\rho_S}(\mathcal{W})|}{\underline{\dim}_B^{\rho_S}(\mathcal{W})\log(1/\delta)}=1,$$

we can invoke Egoroff's theorem, as in previous proofs, to argue that there exists $\Omega_\gamma\in\mathcal{F}^{\otimes n}$, such that $\mu_z^{\otimes n}(\Omega_\gamma)\geq 1-\gamma$, on which the above convergence is uniform. As $\underline{\dim}_B^{\rho_S}(\mathcal{W})>0$ almost surely, we can assume without loss of generality that this is also the case on Ω_γ.

This implies that, on Ω_γ, for δ smaller than some $\delta_{n,\gamma,\zeta}$, we have:

$$\log|N_\delta^{\rho_S}(\mathcal{W})|\geq\frac{1}{2}\log(1/\delta)\underline{\dim}_B^{\rho_S}(\mathcal{W}).$$

Then, the result immediately follows from the previous proposition. □

B.6. Lipschitz case

As mentioned in the introduction, several authors (Şimşekli et al., 2021; Camuto et al., 2021; Hodgkinson et al., 2022) have proven worst-case generalization bounds involving the Hausdorff dimension of the hypothesis set, computed based on the Euclidean distance. Their is in particular based on a 'Lipschitz loss ℓ' assumption. It is therefore natural to ask whether we can find a similar result, i.e. involving the Euclidean based dimension, from our results.

Therefore, in this section, we assume that the function $(w,z)\longmapsto\ell(w,z)$ is L-Lipschitz in w, uniformly with respect to z, with $L>0$ a constant.

To simplify, we demonstrate the case of a fixed hypothesis set $\mathcal{W}\subset\mathbb{R}^d$, more general cases being derived in a similar fashion.

Then we can bound the pseudo-metric ρ_S by ($\|\cdot\|$ is the Euclidean norm):

$$\begin{aligned}\forall w, w' \in \mathbb{R}^d,\ \rho_S(w, w') &= \frac{1}{n}\sum_{i=1}^{n} |\ell(w, z_i) - \ell(w', z_i)| \\ &\leq \frac{1}{n}\sum_{i=1}^{n} L\|w - w'\| \\ &= L\|w - w'\|.\end{aligned}$$

From this observation, denoting N^e the coverings associated to the Euclidean metric, we deduce that:

$$N_\delta^{\rho_S}(\mathcal{W}) \leq N_{\delta/L}^e(\mathcal{W}). \tag{62}$$

The proof of Theorem 3.4 therefore leads us to write, instead of Equation (38), that, with probability at least $1 - 2\eta$:

$$\sup_{w\in\mathcal{W}} \left(\mathcal{R}(w) - \hat{\mathcal{R}}_S(w)\right) \leq 2\delta + 2B\sqrt{\frac{2\log|N_{\delta/L}^e(\mathcal{W})|}{n}} + 3B\sqrt{\frac{2\log(1/\eta)}{n}}.$$

Then we can use proof techniques similar to that in (Şimşekli et al., 2021) in the case of a fixed hypothesis set. More precisely, let us fix some $\epsilon > 0$, using the definition of upper box-counting dimension along with Egoroff's theorem, we have that there exists $\Omega_\gamma \in \mathcal{F}^{\otimes n}$, such that $\mu_z^{\otimes n}(\Omega_\gamma) \geq 1 - \gamma$, on which, for δ smaller than some $\delta_{\gamma,\epsilon}$ (which is independent of n, because the metric does not depend on the data anymore) , we have:

$$\log|N_{\delta/L}^e(\mathcal{W})| \leq \left(\epsilon + \overline{\dim}_B^e(\mathcal{W})\right)\log(L/\delta).$$

Because $\delta_{\gamma,\epsilon}$ does not depend on n, it is possible to set $\epsilon = \overline{\dim}_B^e(\mathcal{W})$ and set:

$$\delta = \delta_n := \frac{2}{\sqrt{n}}.$$

Therefore, we have that, with probability $1 - 2\eta - \gamma$ for n big enough:

$$\sup_{w\in\mathcal{W}} \left(\mathcal{R}(w) - \hat{\mathcal{R}}_S(w)\right) \leq \frac{4}{\sqrt{n}} + 2B\sqrt{\frac{2\overline{\dim}_B^e(\mathcal{W})\log(L\sqrt{n})}{n}} + 3B\sqrt{\frac{2\log(1/\eta)}{n}}.$$

Thus, we recover a result analogous to (Şimşekli et al., 2021), up to potentially absolute constants (coming from the fact that the proof technique is different). Their bound can therefore be seen as a particular case of our result.

C. Additional experimental details

C.1. Granulated Kendall's coefficients

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 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:

$$\tau((g_i)_i, (d_i)_i) = \frac{1}{\binom{n}{2}}\sum_{i<j} \text{sign}(g_i - g_j)\text{sign}(d_i - d_j) \tag{63}$$

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 make 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:

$$\psi_\eta := \frac{1}{|\Theta_B|} \sum_{b \in \Theta_B} \tau\big((g(\eta, b), d(\eta, b))_{\eta \in \Theta_L}\big), \quad \psi_B := \frac{1}{|\Theta_L|} \sum_{b \in \Theta_L} \tau\big((g(\eta, b), d(\eta, b))_{b \in \Theta_B}\big), \tag{64}$$

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:

$$\mathbf{\Psi} := \frac{\psi_1 + \psi_2}{2} \tag{65}$$

Remark C.1. 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 for all the results presented in this work.

Our analysis also report Spearman's rank correlation coefficient (Kendall & Stuart, 1973), denoted ρ, which is another widely used correlation statistics.

C.2. Hyperparameters and experimental setting

Here we present some additional experimental details concerning the experiments of the main part of the paper. Note that all experiments were realized using the same random seed while we were making vary the hyperparameters (e.g. learning rate and batch size). For each experiment both hyperparameters vary in a set of 6 values, making a total of 36 points if all experiment converge.

All Fully Connected Networks (FCN) have standard ReLU activation.

Classification experiments: We trained FCN-5 and FCN-7 networks of width 200 (for each inner layer) on the full training set of MNIST images until we reach 100% accuracy. Learning rate vary in the set $[5.10^{-3}, 10^{-1}]$ and batch size vary in $[32, 256]$.

The stopping criterion in those experiments is reaching 100% accuracy, given that the model is evaluated on all data points every 10000 iterations. To compute the PH dimension the last 5000 iterations were considered (this number essentially comes from computational and time constraints). The model was evaluated on each data point for all those 5000 iterations, producing a point cloud in $\mathbb{R}^{5000 \times n}$ Persistent homology was computed on 20 subset of the generated point cloud with sizes varying in $[1000, 5000]$ in order to apply the method from (Birdal et al., 2021).

Additional classification experiments, presented in Figures 1, 6 and 7 and Tables 6 and 7 involve AlexNet and LeNet networks trained on both MNIST and CIFAR-10 dataset within the same ranges of hyperparameters as described above.

Regression experiments on California Housing Dataset: We trained FCN-5 and FCN-7 of width 200 (for each inner layer) on a training set corresponding to a random subset of 80% of the 20640 points of the California Housing Dataset, using the remaining 20% for validation. Learning rate vary in the set $[1.10^{-3}, 10^{-2}]$ and batch size vary in $[32, 200]$.

The stopping criterion for regression experiments is the following: We periodically (every 2000 iterations in practice) evaluate the empirical risk on the whole training set and stop the training when the relative difference difference between two evaluations becomes smaller than some proportion, set to 0.5% in those experiments. Note that this choice may affect the results. Indeed if we wait to long before stopping the training in a regression experiment, it is possible that the geometry of the point cloud becomes trivial, so we need to ensure convergence while stopping training when the losses ℓ are still "moving enough" to get interesting fractal geometry, both for our dimension and the one of (Birdal et al., 2021).

To compute the PH dimension the last 5000 iterations were considered. The model was evaluated on each data point for all those 5000 iterations, producing a point cloud in $\mathbb{R}^{5000 \times n}$ Persistent homology was computed on 20 subset of the generated

point cloud with sizes varying in $[1000, 5000]$ in order to apply the method from (Birdal et al., 2021).

Robustness experiment: For the robustness experiment presented in figure 3, we used the exact same hyperparameters and random seed than in experiment on MNIST and California Housing Dataset as above. For proportion η varying in $[2\%, 10\%, 20\%, \ldots, 90\%, 99\%]$ we randomly select a subset T of the dataset S such that $|T|/|S| = \eta$ and compute the PH dimensions corresponding to pseudo-metric ρ_T, presented in equation (8). Note that the PH dimension computation involves sampling different subsets of the last iterates (see above), of course this sampling has been done with the same random seed for all values of η so that the observe difference in the dimensional value can only come from the selection of subset $T \subset S$.

D. Additional experimental results

D.1. More details on the experiments presented in Section 5

As mentioned above, for the experiments on MNIST and California Housing Dataset we performed 360 trainings with various seeds, learning rates and batch sizes. This allowed us to compute various statistics, namely granulated Kendall's coefficients ψ_{lr} and ψ_{bs} for learning rate and batch size respectively, Average Kendall's coefficient $\mathbf{\Psi}$, Kendall's tau τ and Spearman's rho ρ, which are all indicators of correlation. Tables 4 and 5 contain all those statistics (same data than the tables in the main part of the paper but with additional coefficients displayed, for space issues). The variation of the seed allows for displaying standard deviation of all those coefficients.

Table 4. Correlation coefficients on CHD

MODEL	DIM.	ρ	ψ_{LR}	ψ_{BS}	$\mathbf{\Psi}$	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	$\dim_{\text{PH}^0}^{\rho_S}$	$\mathbf{0.87 \pm 0.05}$	$\mathbf{0.75 \pm 0.10}$	$\mathbf{0.61 \pm 0.13}$	$\mathbf{0.68 \pm 0.10}$	$\mathbf{0.71 \pm 0.69}$
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	$\dim_{\text{PH}^0}^{\rho_S}$	$\mathbf{0.77 \pm 0.08}$	$\mathbf{0.63 \pm 0.05}$	$\mathbf{0.58 \pm 0.10}$	$\mathbf{0.62 \pm 0.06}$	$\mathbf{0.77 \pm 0.08}$

Table 5. Correlation coefficients on MNIST

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

In Table 6 we also report the full metrics on one experiment on AlexNet trained on CIFAR-10.

Table 6. Correlation coefficients with AlexNet on CIFAR-10

MODEL	DIM.	ρ	ψ_{LR}	ψ_{BS}	$\mathbf{\Psi}$	τ
ALEXNET	$\dim_{\text{PH}^0}^{\text{EUCL}}$	0.86	0.78	**0.84**	0.81	0.68
ALEXNET	$\dim_{\text{PH}^0}^{\rho_S}$	**0.93**	**0.87**	0.81	**0.84**	**0.78**

Table 7. Correlation coefficients with convolutional models on MNIST

MODEL	DIM.	ρ	ψ_{LR}	ψ_{BS}	$\mathbf{\Psi}$	τ
ALEXNET	$\dim_{\text{PH}^0}^{\text{EUCL}}$	0.85	**0.78**	**0.77**	**0.77**	0.67
ALEXNET	$\dim_{\text{PH}^0}^{\rho_S}$	**0.88**	**0.78**	**0.77**	**0.77**	**0.70**
LENET	$\dim_{\text{PH}^0}^{\text{EUCL}}$	0.74	0.78	**0.77**	0.78	0.57
LENET	$\dim_{\text{PH}^0}^{\rho_S}$	**0.80**	**0.80**	**0.77**	**0.79**	**0.62**

In Figures 5 and 4 we plot the values of $\dim_{\mathrm{PH}^0}^{\rho_S}$ against the actual loss gap (computed based on the cross entropy loss). 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, we observe on figure 5 and 4 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 most all experiments, again, the proposed dimension in close or better than the one proposed in (Birdal et al., 2021).

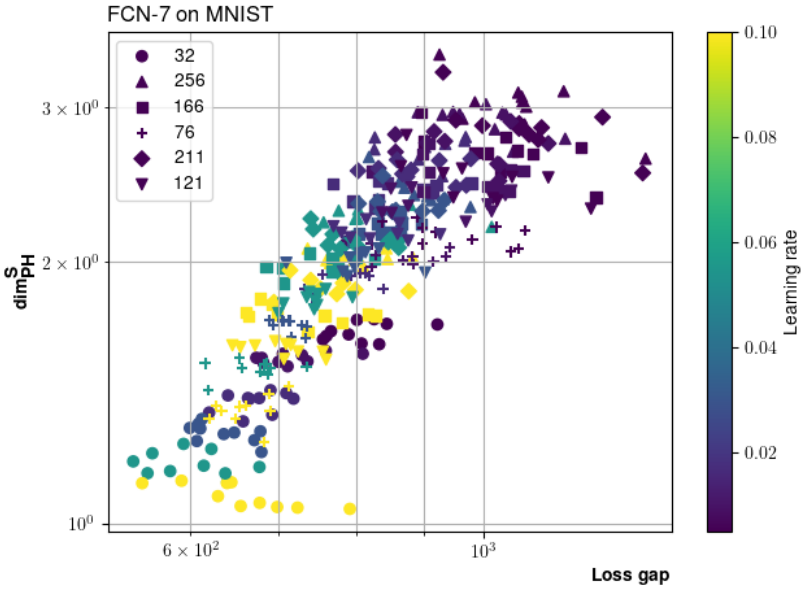

Figure 4. Plots of $\dim_{\mathrm{PH}^0}^{\rho_S}$ against the loss gap (as opposed to the accuracy gap) for a FCN-7 trained on MNIST dataset.

On Table 8 we report the correlations coefficients $(\rho, \psi_{\mathrm{lr}}, \psi_{\mathrm{bs}}, \mathbf{\Psi}, \tau)$ between our data-dependent intrinsic dimension and the actual loss gap in the same classification experiments than in Figures 5 and 4.

Table 8. Correlation coefficients on MNIST, with respect to loss gap

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

Figures 6 and 7, as well as Table 7 show experimental results obtained by training Convolutional Neural Networks (CNN) on the MNIST dataset, namely AlexNet (Krizhevsky et al., 2017) and LeNet (Lecun et al., 1998) networks. It further highlights the pertinence of our intrinsic dimension, which is well correlated with the accuracy gap in those experiments.

D.2. Evaluation of the computable part of the bounds

Recent studies showed that, even when the overall scale of a generalization bound does not match the scale of the actual generalization error, the bound can still be useful (e.g., for hyperparameter selection) if it correlates well with the generalization error.

Our study of the correlation between our data-dependent intrinsic dimension the generalization error is inspired by Jiang et al.

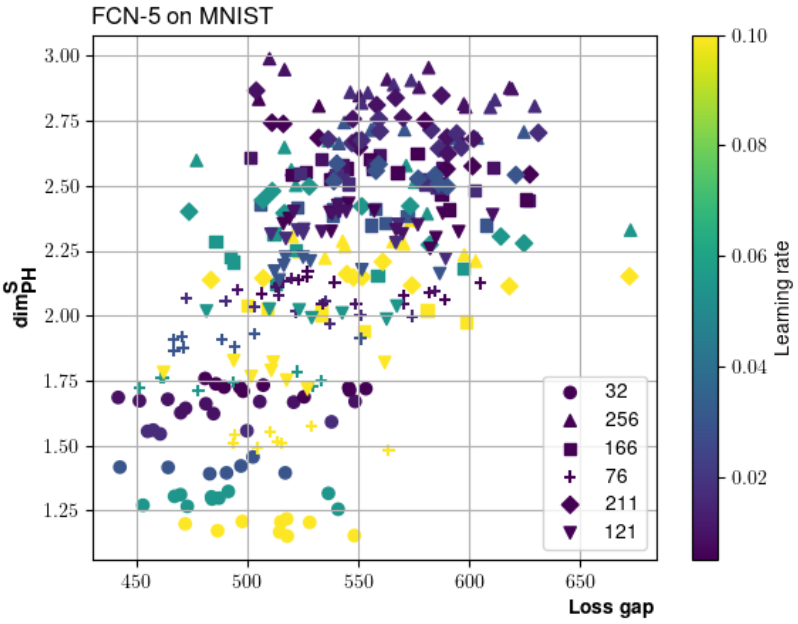

Figure 5. Plots of $\dim_{\mathrm{PH}^0}^{\rho_S}$ against the loss gap (as opposed to the accuracy gap) for a FCN-5 trained on MNIST dataset.

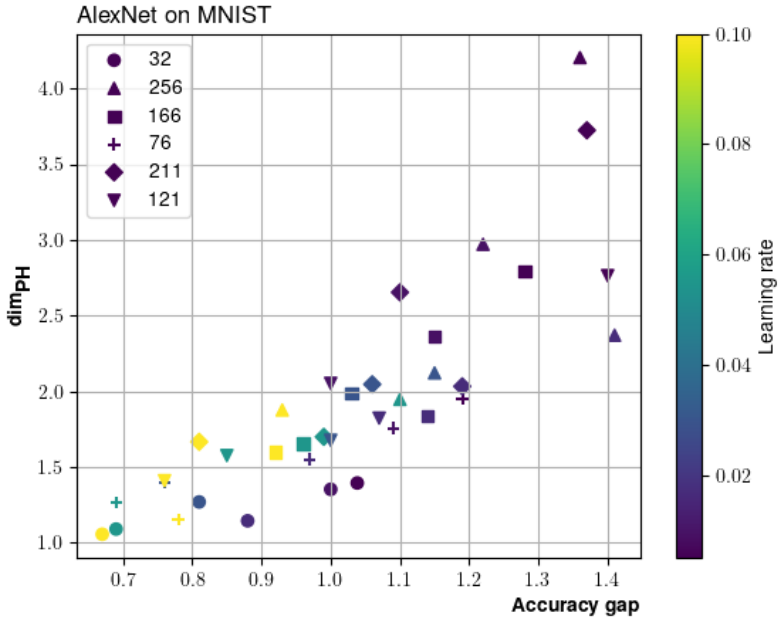

Figure 6. Plots of $\dim_{\mathrm{PH}^0}^{\rho_S}$ against the accuracy gap for an AlexNet trained on MNIST dataset.

(2019), who pointed out that correlation can be an efficient measure of the performance of different complexity measures. Moreover, in our study, we use the Granulated Kendall's coefficients, that Jiang et al. (2019) introduced, to better capture the causal relationship in the correlation between the fractal dimension and the generalization.

That being said, one could argue that it would be better to plot the full bound and compare it to the generalization error.

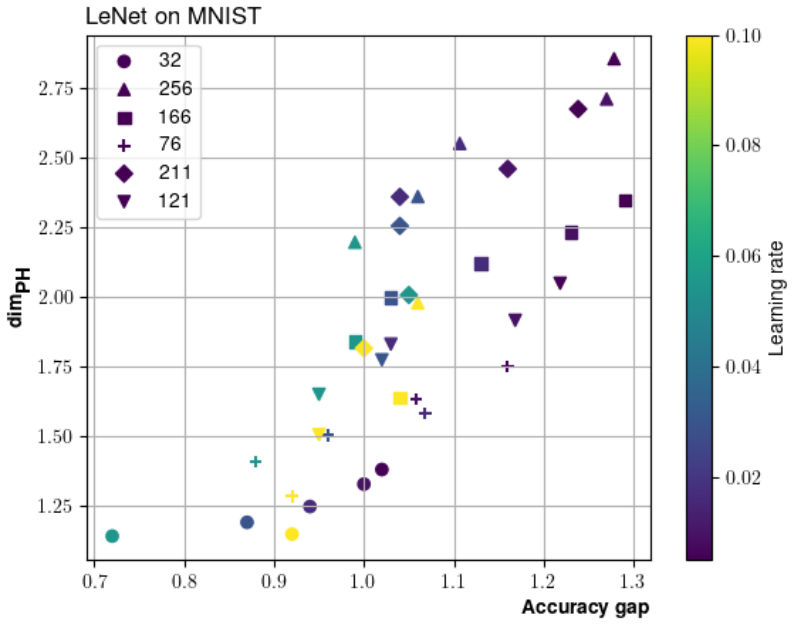

Figure 7. Plots of $\dim_{\mathrm{PH}^0}^{\rho_S}$ against the accuracy gap for a LeNet trained on MNIST dataset.

However, as we are aiming to compute the bounds in Theorems 3.5 and 3.8, this is a notoriously difficult, and often impossible task due to the presence of the mutual information (MI) terms. Hence the full computation of the bound is unfortunately not possible. Yet, we would like to underline that this has been the case for most information theoretic bounds, and fractal geometric bounds (Şimşekli et al., 2021; Birdal et al., 2021; Hodgkinson et al., 2022).

As an intermediate solution towards this direction, we propose the following experimental setting, which will be added to the next version of the paper. Since our MI terms, especially $I_\infty(S, \mathcal{W}_{S,U})$ appearing in Theorem 3.8, can be seen as similar as terms appearing in previous fractal geometric works (Şimşekli et al., 2021; Birdal et al., 2021; Hodgkinson et al., 2022), we can aim at plotting the remaining terms of the bound and still provide a meaningful experiment.

Hence, in an attempt to provide further experimental results, we can compare the generalization error to the 'computable' part of the bound, whose main term is of the form:

$$\delta + \frac{B}{\sqrt{n}-1} + \sqrt{2}B\sqrt{\frac{\dim_{PH}(\mathcal{W}_{S,U})\log(1/\delta) + \log(\sqrt{n}/\eta)}{n}}.$$

This expression can be approximated thanks to the persistent homology tools described in the paper. We will include these experiments in the paper, where we will compute the full bounds of the prior art in the same way (which will require estimating the Lipschitz constant).

One particular point of attention is that the loss functions we consider for the experiments are in practice not bounded. Despite this fact, to allow for the experimentation to take place, we set the value of the constant B to the maximum loss reached on one data-point in the whole trajectory, over all experiments. For a fully connected networks trained on MNIST, we get Figures 8 and 9.

The figures report the value of the above formula compared to the actual generalization error. The correlation with the generalization is quite well. However, an offset can be observed, corresponding to the scaling to what appears to be a rather small absolute constant. We believe that this is due to the fact that the way we estimate B is a bad estimation of the potential sub-Gaussian character of the loss (indeed, the statement of Theorem 3.5 can be easily extended for sub-Gaussian losses). As the bound scales linearly with this quantity, we see that we can easily get pretty close to the true generalization gap.

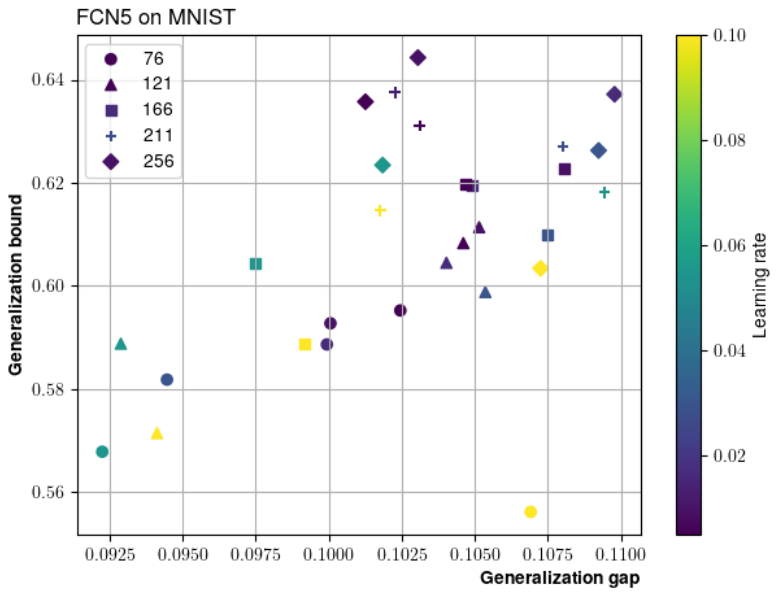

Figure 8. Plot of the generalization error against the computable part of our bounds for a 5-layer fully-connected network trained on MNIST

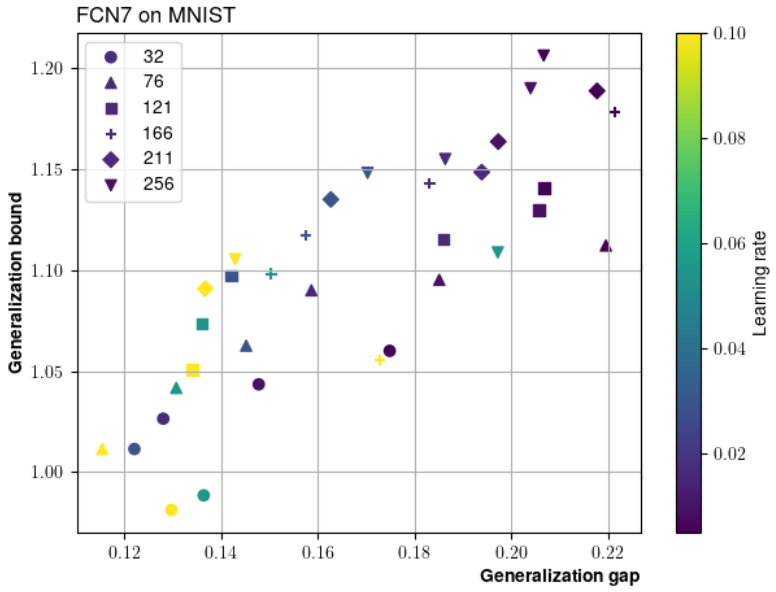

Figure 9. Plot of the generalization error against the computable part of our bounds for a 7-layer fully-connected network trained on MNIST

D.3. Experiments with bigger models and datasets

Most experiment presented in Sections 5 and D.1 are made on relatively small datasets and/or neural network models. For the sake of completeness, we present here similar experiments computed with a Resnet-18 model on CIFAR10 and CIFAR100 datasets.

Note that the main difficulty to perform such experiments is that the computation of $\dim_{\mathrm{PH}^0}^{\rho_S}(\mathcal{W}_{S,U})$ requires the evaluation of the model on *every* training data point, and the corresponding distance matrix. However, to be able to make this experiment in a reasonable amount of time, and according to our computational resources, we leveraged the ideas from Section 5, regarding the robustness analysis, and used only a subset of the dataset for the computation of $\dim_{\mathrm{PH}^0}^{\rho_S}(\mathcal{W}_{S,U})$ (while the whole dataset is obviously used for training). Moreover, note that one advantage of the proposed data-dependent intrinsic dimension is that it requires much less memory to be computed than the one proposed in (Birdal et al., 2021). Indeed, to compute this last one, we would need to store all the weights of the network, for a few thousand iterations.

Hyperparameters details Both experiments (on CIFAR10 and CIFAR100) where realized on a 4×4 grid of hyperparameters with the batch size varying from 32 to 256 and the learning rate varying from 0.1 to 0.001. In both experiments, the persistent homology dimension has been computed using the last 2000 iterates and 5% of the dataset for the computation of the associated distance matrix. For the experiment on CIFAR10, the network was trained until 100% accuracy before computing the persistent homology dimensions. For the CIFAR100 experiment, because we were not able to train the model until 100% accuracy, we stopped the training after 100000 iterations.

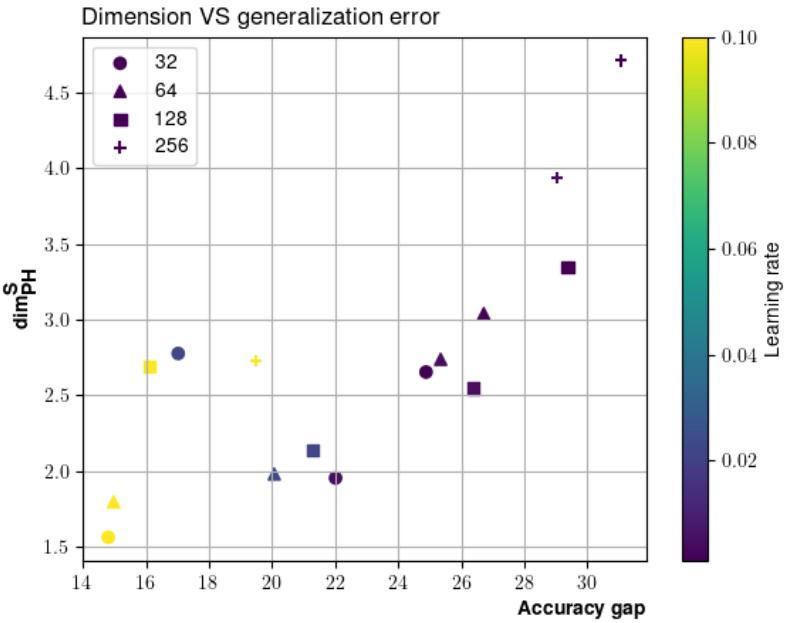

Figure 10. Plots of $\dim_{\mathrm{PH}^0}^{\rho_S}$ against the accuracy gap for Resnet-18 network, trained on CIFAR10.

Figure 10 displays our results for the CIFAR10 experiments. As in previous experiments, we observe a very satisfying correlation for lower learning rates and high batch sizes.

Results on the CIFAR100 dataset are shown on Figures 11 and 12. Something interesting is observed here; as the network didn't reach, in our experiment, an accuracy close to 100% for all hyperparameter settings, we observe two regimes regarding the correlation between the accuracy gap and the data-dependent persistent homology dimension:

- For experiments achieving very good training accuracy, the correlation is excellent, as shown in Figure 12.
- Experiments with less training accuracy look like 'out of distribution ' experiments, this particular behavior is illustrated on Figure 11.

This shows that the fractal behavior seems to be particularly pertinent in the 'permanent regime' of training, i.e. when the distribution of the parameters reaches a stable distribution.

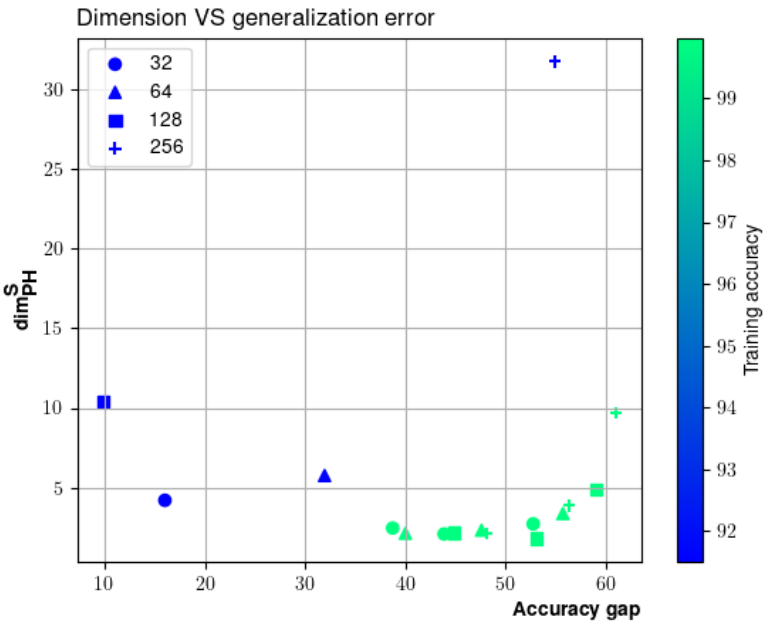

Figure 11. Plots of $\dim_{\mathrm{PH}^0}^{\rho_S}$ against the accuracy gap for Resnet-18 network, trained on CIFAR100, the training accuracy is shown to highlight the importance of convergence for the correlation.

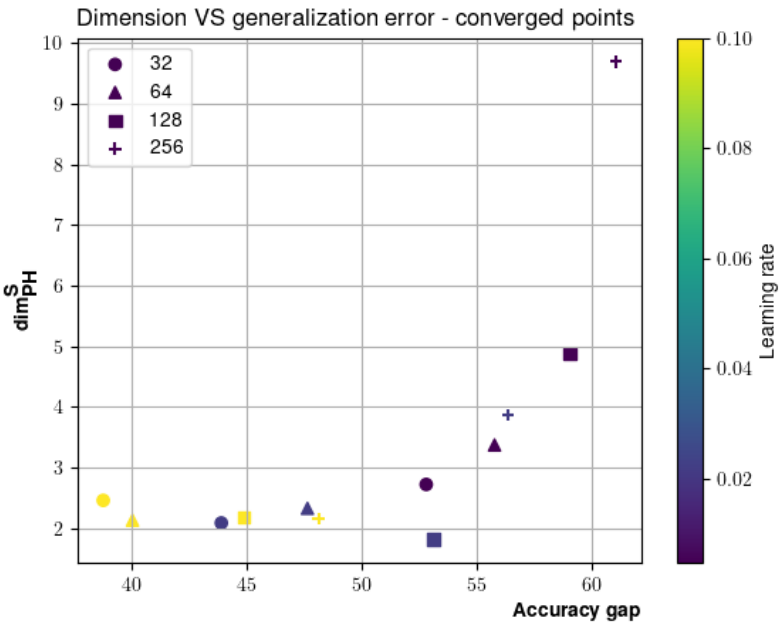

Figure 12. Plots of $\dim_{\mathrm{PH}^0}^{\rho_S}$ against the accuracy gap for Resnet-18 network, trained on CIFAR100, displaying only points for which 98% accuracy has been reached during training.