Generalization Guarantees via Algorithm-dependent Rademacher Complexity

Sarah Sachs S.C.SACHS@UVA.NL
Tim van Erven TIM@TIMVANERVEN.NL
Korteweg-de Vries Institute for Mathematics, University of Amsterdam

Liam Hodgkinson LHODGKINSON@UNIMELB.EDU.AU
School of Mathematics and Statistics, University of Melbourne

Rajiv Khanna RAJIVAK@PURDUE.EDU
Department of Computer Science, Purdue University

Umut Şimşekli UMUT.SIMSEKLI@INRIA.FR
Inria, CNRS, Ecole Normale Supérieure, PSL Research University

Editors: Gergely Neu and Lorenzo Rosasco

Abstract

Algorithm- and data-dependent generalization bounds are required to explain the generalization behavior of modern machine learning algorithms. In this context, there exists information theoretic generalization bounds that involve (various forms of) mutual information, as well as bounds based on hypothesis set stability. We propose a conceptually related, but technically distinct complexity measure to control generalization error, which is the empirical Rademacher complexity of an algorithm- and data-dependent hypothesis class. Combining standard properties of Rademacher complexity with the convenient structure of this class, we are able to (i) obtain novel bounds based on the *finite fractal dimension*, which (a) extend previous fractal dimension-type bounds from continuous to finite hypothesis classes, and (b) avoid a mutual information term that was required in prior work; (ii) we greatly simplify the proof of a recent dimension-independent generalization bound for stochastic gradient descent; and (iii) we easily recover results for VC classes and compression schemes, similar to approaches based on conditional mutual information.

Keywords: Generalization error, Rademacher complexity, Fractal geometry.

1. Introduction

The generalization error of a learning algorithm is the gap between its average loss (empirical risk) on a training sample and its expected loss (risk) on a fresh data point from the same probability distribution. If the algorithm selects its parameter estimates $\hat{\theta}$ from a set Θ, then the classical approach to control generalization error is to derive deviation bounds that hold uniformly over all $\theta \in \Theta$ (Shalev-Shwartz and Ben-David, 2014). However, Zhang et al. (2021) empirically illustrate that in modern machine learning settings such as neural networks, such an approach yields overly pessimistic, and sometimes vacuous error bounds of limited practical value. Hence it has been made clear that generalization bounds that would reflect the practical observations should take into account the effects of the data sample S^n and also the choice of the learning algorithm (e.g., stochastic gradient descent).

Russo and Zou (2016); Xu and Raginsky (2017) initiated a fertile line of research by developing algorithm-dependent generalization bounds with information-theoretic tools. Let $R(\hat{\theta})$ and $\hat{R}(\hat{\theta})$

denote the risk and empirical risk of an algorithm. They show that the generalization error is at most

$$\mathbb{E}\left[R(\hat{\theta}) - \hat{R}(\hat{\theta}, S^n)\right] \lesssim \sqrt{\frac{I(\hat{\theta}; S^n)}{n}}, \tag{1}$$

where $I(\theta; S^n)$ denotes the *mutual information* (MI) between the data sample S^n and the output of the algorithm $\hat{\theta}$. This shows that a weak statistical dependence between the data sample and the algorithm output implies better generalization. Recently, by using tools from rate-distortion theory, the bound (1) was linked to compression, implying that if the algorithm output is compressible in some sense, it implies good generalization (Sefidgaran et al., 2022), in line with the results of (Arora et al., 2018; Suzuki et al., 2020a,b; Barsbey et al., 2021).

Looking at the problem from a different angle, Simsekli et al. (2020) take into account the topological structure of the outputs of the learning algorithm by using tools from fractal geometry (Falconer, 2004). More precisely, let $\Theta_{S^n} \subset \Theta$ denote the full trajectory of a continuous time version of stochastic gradient descent on the sample S^n. Then they prove generalization bounds based on a fractal dimension of Θ_{S^n}, which were later extended and improved by Birdal et al. (2021); Camuto et al. (2021); Hodgkinson et al. (2022); Dupuis et al. (2023). Their bounds are of the following general form[1]: with probability at least $1 - \delta$,

$$\sup_{\theta \in \Theta_{S^n}} \left| R(\theta) - \hat{R}(\theta, S^n) \right| \lesssim \sqrt{\frac{\dim \Theta_{S^n} + I_\infty(\Theta_{S^n}; S^n) + \log(1/\delta)}{n}}, \tag{2}$$

where I_∞ denotes the total MI, which is larger than the regular MI, and $\dim$ denotes some notion of fractal dimension (e.g., the Minkowski, Hausdorff, or persistent homology dimension). In addition to the statistical dependence between the data sample and the random hypothesis set Θ_{S^n} as measured by I_∞, these bounds imply that the worst-case error can be controlled by the fractal dimension of Θ_{S^n}. This fractal dimension is linked to the statistical or topological properties of the learning algorithm; in particular, if the algorithm is a stochastic optimizer such as stochastic gradient descent.

While these bounds help to shed light on modern learning problems from different viewpoints, the MI terms that they contain can be troublesome for several reasons. First, MI can be infinite, which renders the bounds vacuous (Bassily et al., 2017, Section 5). Secondly, while the fractal dimension in (2) can be linked to concrete and computable properties of the learning algorithm, the MI term typically cannot be given a topological interpretation, which means the bounds as a whole also do not have a fully topological interpretation.

In order to address the first shortcoming, Steinke and Zakynthinou (2020) introduced the *conditional mutual information* (CMI) which in contrast to MI is always finite. They show that the CMI implies generalization under much weaker assumptions than MI: for instance, it can be controlled if the learning algorithm is a compression scheme (Littlestone and Warmuth, 2003) or under distributional stability assumptions such as differential privacy (Bassily et al., 2016; Dwork et al., 2015). As an alternative, Sefidgaran et al. (2022) introduced the notion of *lossy compressible learners*, which also circumvents the cases where MI can be infinite. Despite these improvements, it remains unclear

[1] Hodgkinson et al. (2022) also proved an in-expectation version of (2) which involved the weaker $I(\Theta_{S^n}; S^n)$ instead of $I_\infty(\Theta_{S^n}; S^n)$.

how to relate MI-based bounds to topological concepts.[2] Furthermore, Haghifam et al. (2022) recently showed that it is impossible to obtain minimax rates for gradient descent by using the current information-theoretic frameworks, and argued that new frameworks need to be developed.

Contributions In this study, we propose an alternative mathematical framework for analyzing algorithm- and data-dependent hypotheses. We make the following contributions:

- We prove a generalization bound with respect to an *algorithm-dependent Rademacher complexity* (ARC), Lemma 2 in Section 3. Interestingly, our construction is conceptually related to the conditioning in CMI. It is also technically similar to a special case of the Rademacher complexity for data-dependent hypothesis sets introduced by Foster et al. (2019). This special case arises when their hypothesis sets are instantiated as singletons that contain the output of a learning algorithm on the sample. For this case, our result is a refinement of their Theorem 1.[3] Both of these relations are discussed in more detail below Definition 1 in Section 3.
- Our main contribution is to demonstrate the flexibility of the ARC. In Section 4, we derive several new generalization results and re-obtain known results. More precisely,
  - In Section 4.1 we link ARC to fractal dimensions using the tools developed by Alonso (2015). This allows us to extend previous fractal dimension-type bounds from continuous to finite hypothesis classes without introducing any mutual information term as in (2).
  - For stochastic gradient descent on strongly convex and smooth losses, we use ARC to obtain a greatly simplified proof of a dimension-independent generalization bound by Park et al. (2022) (Section 4.2).
  - For learning algorithms that are compression schemes or produce output in a VC class, we show that we can obtain the same generalization properties as those obtained for CMI by Steinke and Zakynthinou (2020) (Sections 4.3 and 4.4).[4]

We believe that the proposed framework provides a promising alternative to information-theoretic approaches and opens up future directions, which we discuss in Section 5. Some of the proofs are delegated to the appendix.

2. Preliminaries

Setting Given a sample $S^n = (Z_1, \dots, Z_n) \in \mathcal{Z}^n$ of independent, identically distributed (i.i.d.) observations with common distribution $\mathcal{D}$, and a loss function $\ell : \Theta \times \mathcal{Z} \to \mathbb{R}$, let

$$R(\theta) = \mathop{\mathbb{E}}_{Z\sim\mathcal{D}} [\ell(\theta, Z)] \qquad \text{and} \qquad \hat{R}(\theta, S^n) = \frac{1}{n}\sum_{i=1}^{n} \ell(\theta, Z_i)$$

denote the risk and the empirical risk, respectively. We assume throughout that Θ is a subset of a complete separable metric space $(\mathcal{X}, \mathtt{d})$, and that ℓ is measurable. A common instantiation in

[2] We note that Sefidgaran et al. (2022, Corollary 7) link MI to fractal geometry through rate-distortion theory (Kawabata and Dembo, 1994). But their result involves the *marginal* distribution of $\hat{\theta}$, which has limited practical interest.

[3] We note that our setting is not the main focus of Foster et al. (2019), who mostly consider stability properties for data-dependent hypothesis classes.

[4] Steinke and Zakynthinou (2020) can also obtain guarantees for differentially private algorithms. While we have not investigated whether such results can also be obtained via ARC, we suspect that information-theoretic tools might be more natural to analyze differentially private algorithms.

supervised learning is that $Z = (X, Y) \in \mathcal{X} \times \mathcal{Y}$ and θ indexes a class of hypotheses $\mathcal{H} = \{h_\theta : \mathcal{X} \to \mathcal{Y} \mid \theta \in \Theta\}$. Then the loss is typically defined as a composite loss function via $\ell(\theta, z) := \tilde{\ell}(h_\theta(x), y)$ for some choice of $\tilde{\ell} : \mathcal{Y} \times \mathcal{Y}$. A learning algorithm $\mathcal{A}$ is a measurable function that maps a sample S^n to an estimate $\hat{\theta} = \mathcal{A}(S^n) \in \Theta$. We assume that $\mathcal{A}$ is a deterministic algorithm for now; at the start of Section 3.1 we discuss how our results can be extended to randomized algorithms.

Rademacher complexity A standard approach to control the generalization error relies on the (empirical) Rademacher complexity of the whole class Θ of possible outputs of the algorithm:

$$\operatorname{Rad}(\Theta, S^n) = \frac{1}{n} \mathop{\mathbb{E}}_{\sigma} \left[\max_{\theta \in \Theta} \sum_{i=1}^{n} \sigma_i \ell(\theta, Z_i) \right], \tag{3}$$

where the expectation is over Rademacher random variables $\sigma = (\sigma_1, \ldots, \sigma_n) \in \{-1, +1\}^n$, which are i.i.d. with $\Pr(\sigma_i = -1) = \Pr(\sigma_i = +1) = 1/2$. It is well known that the Rademacher complexity can be upper bounded in terms of a covering number.

Covering numbers For any $C \subset \mathcal{X}$, we will denote its ϵ-covering number by $\operatorname{Cov}(C, \mathtt{d}, \epsilon)$ and the corresponding ϵ-cover by $\operatorname{SCover}(C, \mathtt{d}, \epsilon)$. The box-covering number refers to the special case that $\mathcal{X} = \mathbb{R}^k$ and $\mathtt{d}$ is the distance induced by the ℓ_∞-norm.

3. Generalization via Algorithm-dependent Rademacher Complexity

In this section, we refine the standard Rademacher bounds on the generalization error by measuring the Rademacher complexity not on Θ but on a smaller set $\hat{\Theta}^n$ that depends on the algorithm and the data, which is defined as follows: consider two independent samples $S_-^n = (Z_1^{-1}, \ldots, Z_n^{-1})$ and $S_+^n = (Z_1^{+1}, \ldots, Z_n^{+1})$, and, for any $\sigma \in \{-1, +1\}^n$, let $S_\sigma^n = (Z_1^{\sigma_1}, \ldots, Z_n^{\sigma_n})$ denote a combined sample where σ_i determines whether to take $Z_i^{\sigma_i}$ from sample S_-^n or from sample S_+^n. Then

$$\hat{\Theta}^n := \left\{ \mathcal{A}(S_\sigma^n) : \sigma \in \{-1, +1\}^n \right\} \subset \Theta,$$

which depends on S_-^n and S_+^n and contains all possible outputs of the algorithm $\mathcal{A}$ that can be obtained by combining them with different choices of σ.

Definition 1 *We define the* Algorithm-dependent Rademacher Complexity *(ARC) as the Rademacher complexity* $\operatorname{Rad}(\hat{\Theta}^n, S_+^n)$ *of the algorithm- and data-dependent set* $\hat{\Theta}^n$*.*

In our analysis, S_-^n acts as a *ghost sample*, which is independent of S_+^n. This allows us to shrink the effective size of the function class from all possible functions indexed by the parameters Θ to a finite class of functions indexed by $\hat{\Theta}^n$ that can be realized by the algorithm by exchanging data points between S_-^n and S_+^n according to Rademacher variables σ. The ARC can therefore be seen as measuring the complexity of the algorithm when only σ is unknown, conditional on the supersample (S_-^n, S_+^n). This is conceptually similar to the conditional mutual information of Steinke and Zakynthinou (2020), except that we use Rademacher complexity where they use mutual information and therefore we get into a technically fully distinct analysis. There is also a strong connection to the Rademacher complexity for data-dependent hypothesis sets introduced by Foster et al. (2019). When their Theorem 1 is specialized to an algorithm-dependent result for our setting, it gives a bound in terms of the larger class $\bar{\Theta}^n = \{\mathcal{A}(S) : S \subset (S_-^n, S_+^n) \text{ such that } |S| = n\}$.[5] Since $\hat{\Theta}^n \subset \bar{\Theta}^n$, our result is strictly better, but it is not evident that the improvement is very large. For instance, our analysis of SGD in Section 4.2 will still go through even with the larger set $\bar{\Theta}^n$ at the cost of an additional $\log n$ factor in the bound. See Remark 9.

The following key technical result proved in Appendix A, shows that the ARC can control the generalization error in the same way as the classical Rademacher complexity does for fixed hypothesis classes. To state it, we let $\operatorname{ess\,sup} X = \inf\{a : \Pr(X > a) = 0\}$ denote the *essential supremum* of any random variable X.

Lemma 2 (Key technical lemma) *The expected generalization error of any (deterministic) algorithm $\mathcal{A} : \mathcal{Z}^n \to \Theta$ with output $\hat{\theta} = \mathcal{A}(S^n)$ is bounded by*

$$\mathop{\mathbb{E}}_{S^n}\left[R(\hat{\theta}) - \hat{R}(\hat{\theta}, S^n)\right] \leq 2 \mathop{\mathbb{E}}_{S_-^n, S_+^n}\left[\operatorname{Rad}(\hat{\Theta}^n, S_+^n)\right]. \tag{4}$$

Moreover, if there exist a b and a function $h : \Theta \to \mathbb{R}$ such that the loss $\ell(\theta, z)$ takes values in the bounded interval $[h(\theta), h(\theta) + b]$ for all $\theta \in \Theta$ and $z \in \mathcal{Z}$, then, for any $\delta \in (0, 1]$,

$$R(\hat{\theta}) - \hat{R}(\hat{\theta}, S^n) \leq 4 \operatorname*{ess\,sup}_{S_-, S_+}\left\{\operatorname{Rad}(\hat{\Theta}^n, S_+^n)\right\} + b\sqrt{\frac{8\log(2/\delta)}{n}} \tag{5}$$

with probability at least $1 - \delta$.

The proofs of both results mimic standard Rademacher complexity bounds on the generalization error, except that we do not start with the standard upper bound $R(\hat{\theta}) - \hat{R}(\hat{\theta}, S^n) \leq \sup_{\theta \in \Theta}\{R(\theta) - \hat{R}(\theta, S^n)\}$, but we instead replace $\hat{\theta}$ by the maximum over $\theta \in \hat{\Theta}^n$ *after* symmetrization by the ghost sample S_-^n. This allows $\hat{\Theta}^n$ to depend on the algorithm $\mathcal{A}$ as well as S^n and S_-^n. For notational symmetry, we then denote the original sample S^n by S_+^n in the right-hand side of both results. Although the main idea behind both proofs is the same, it is not the case that (5) follows directly from (4), because $\hat{\theta}$ may be highly unstable, so we cannot apply McDiarmid's inequality to relate $R(\hat{\theta}) - \hat{R}(\hat{\theta}, S^n)$ to its expectation. We therefore prove both results separately: the in-expectation proof is a variation on the standard in-expectation argument, which can be found in e.g. Lemma A.5 of Bartlett et al. (2005). The in-probability derivation is patterned after the proof of Theorem 4.3 in the textbook by Anthony and Bartlett (2002).

3.1. Consequences

Randomized algorithms Although Lemma 2 is stated only for deterministic algorithms, it can also be applied to algorithms that randomize. This is possible by viewing an algorithm as a function of two arguments $\mathcal{A}(S^n, \xi)$, where S^n is the sample and ξ is an independent random variable (e.g. a number of random bits) that provides the randomness. By applying Lemma 2 conditional on ξ, we obtain a generalization bound that holds for any value of ξ and hence also almost surely when ξ is drawn at random. This is the approach we take to analyze stochastic gradient descent in Section 4.2.

[5] In the notation of Foster et al. (2019), our setting corresponds to the case where $\mathcal{H}_S = \{\mathcal{A}(S)\}$, $m = n$ and $U = (S, T)$. Then their $\bar{\mathcal{H}}_{U,m} = \mathcal{H}_{S,T}$, which, in our notation equals the class of hypotheses indexed by $\bar{\Theta}^n$.

Exploiting standard properties of Rademacher complexity We may use all the well-known properties of standard Rademacher complexity to upper bound the ARC. In particular, throughout the paper we will repeatedly rely on the following upper bound in terms of the covering number of $\hat{\Theta}^n$, which holds for bounded, Lipschitz continuous losses:

Proposition 3 *Suppose that $\ell(\theta, z)$ is L-Lipschitz continuous in θ for any z and takes values in $[a, a+b]$. Then*

$$\mathrm{Rad}(\hat{\Theta}^n, S^n) \leq L\epsilon + b\sqrt{\frac{\log \mathrm{Cov}(\hat{\Theta}^n, \mathtt{d}, \epsilon)}{n}} \qquad \textit{for all } \epsilon > 0. \tag{6}$$

Proof For any $\epsilon > 0$, Lipschitzness of the loss implies that

$$\mathrm{Rad}(\hat{\Theta}^n, S^n) \leq \epsilon L + \mathrm{Rad}(\mathrm{SCover}(\hat{\Theta}^n, \mathtt{d}, \epsilon), S^n) \leq \epsilon L + b\sqrt{\frac{2\ln \mathrm{Cov}(\hat{\Theta}^n, \mathtt{d}, \epsilon)}{n}},$$

where the second inequality follows from Massart's lemma (Shalev-Shwartz and Ben-David, 2014). ■

4. Applications: controlling algorithm-dependent Rademacher complexity

In this section, we derive new results and recover known results by controlling the ARC defined in the previous section. First, in Section 4.1, we provide a new bound with respect to fractal dimensions. This result allows for control of the generalization error based on the topological properties of $\hat{\Theta}^n$ or its limiting set $\hat{\Theta}$ as n increases. Second, in Section 4.2, we consider the projected stochastic gradient descent algorithm and recover the results of Park et al. (2022) in a simple way by bounding the ARC. Third, in Sections 4.3 and 4.4, similar to Steinke and Zakynthinou (2020), we show that generalization guarantees under a compressibility condition and for VC-classes can be easily obtained via ARC as well.

4.1. Finite fractal dimensions

In this section, we provide a bound on the generalization error with respect to a *finite Minkowski dimension*, in the vein of recent results connecting error to fractal geometry (cf. (Simsekli et al., 2020; Birdal et al., 2021; Hodgkinson et al., 2022; Dupuis et al., 2023)). The finite Minkowski dimension was introduced in (Alonso, 2015) as an extension of the classical *Minkowski dimension* to finite sets. A comprehensive summary of the formal definitions for the finite Minkowski dimension and the relevant existing results is provided in Appendix B. For brevity, here we consider a simplified definition of the finite Minkowski dimension. Under some small additional assumptions which exclude notorious edge cases, these simplified definitions coincide with the original definitions in (Alonso, 2015).

Definition 4 (Diameters) *Let C be a finite set in a metric space $(\mathcal{X}, \mathtt{d})$ with $|C| \geq 2$, and let $\nu_C : C \to \mathbb{R}$ map points in C to the distance to their nearest neighbor, that is, $\nu_C(a) = \min\{\mathtt{d}(a,b) : b \in C \setminus \{a\}\}$ for $a \in C$. The covering diameter $\nabla(C)$ and the diameter $\Delta(C)$ of C are then*

$$\nabla(C) := \max_{a \in C} \nu_C(a)\,, \qquad \textit{and} \qquad \Delta(C) := \max\{\mathtt{d}(a,b) : a, b \in C\}\,.$$

A set C is called non-focal *if $\nabla(C) < \Delta(C)$, and* focal *otherwise.*

For example, the set of vertices of a simplex comprises a focal set, while a set with no equally distant points is non-focal. Our results will be for non-focal sets only, which rules out pathological edge cases. Hence, *throughout this section*, we assume that $\hat{\Theta}^n$ is non-focal. The definition of the finite Minkowski dimension follows a similar box-counting construction as the standard Minkowski dimension. The main twist which yields nontrivial values for finite sets is to consider covers that contain at least two points.

Definition 5 (finite Minkowski dimension) *A family of sets $\mathcal{U} = \{U_i\}$ is a 2-cover of a finite set C if each $U_i \subseteq C$, $|U_i| \geq 2$, and $C = \bigcup_{U \in \mathcal{U}} U$. For any non-focal set C and parameter $a \geq \nabla(C)$, the covering cardinality is*

$$T_a(C) = \min\{|\mathcal{U}| : \mathcal{U} \textit{ is a 2-cover and } U : \Delta(U) \leq a \textit{ for all } U \in \mathcal{U}\}.$$

For $T_{\nabla(C)}(C)$ we write $T(C)$. If C is finite and non-focal with $|C| > 2$, then the finite Minkowski dimension of C is

$$\dim_{\mathrm{fM}}(C) = \frac{\log T(C)}{\log \frac{\Delta(C)}{\nabla(C)}}.$$

Although it may not be obvious from the definition, Alonso (2015) shows that the finite Minkowski dimension is a natural finite analog of the classical Minkowski dimension (cf. Equation (8) below and (Falconer, 2004)), as the two definitions are consistent under appropriate limits. Most importantly, like the classical Minkowski dimension, if f is Hölder continuous such that $c_1\|x - y\|^\beta \leq |f(x) - f(y)| \leq c_2\|x - y\|^\beta$, then the finite Minkowski dimension of $f(C)$ satisfies $\dim_{\mathrm{fM}}(f(C)) = \beta \dim_{\mathrm{fM}}(C)$. In this sense, the finite Minkowski dimension is well-suited to measure local clustering as in (Hodgkinson et al., 2022). Our primary fractal dimension generalization bound is shown in Theorem 6, which arises by taking $\epsilon = \nabla(\hat{\Theta}^n)$ in (6).

Theorem 6 *If ℓ is L-Lipschitz in θ and bounded by b, $n > 2$, and $\hat{\Theta}^n \subset \mathbb{R}^k$ is non-focal, then for any set $F \supseteq \hat{\Theta}^n$,*

$$\mathrm{Rad}(\hat{\Theta}^n, S_+^n) \leq \mathcal{D}_n(F) := L\nabla(F) + b\sqrt{\frac{\dim_{\mathrm{fM}}(F)}{n} \log \frac{\Delta(F)}{\nabla(F)}}.$$

There are a few variants and consequences of Theorem 6 worth mentioning.

- **Trivial upper bound:** for $\delta > 0$ sufficiently small, consider the set $\hat{\Theta}_\delta^n = \hat{\Theta}^n \cup (\hat{\Theta}^n + \delta \mathbf{1})$, which satisfies $\nabla(\hat{\Theta}_\delta^n) = \delta$. By (Alonso, 2015, Proposition 5.3), for any finite set F,

$$\dim_{\mathrm{fM}}(F) \leq (\log(|F|) - 1) / \log \frac{\Delta(F)}{\nabla(F)}$$

and so $\mathcal{D}_n(\hat{\Theta}_\delta^n) \leq L\delta + b\sqrt{\log(2|\hat{\Theta}^n|)/n}$. Taking $\delta \to 0^+$, this implies that

$$\mathrm{Rad}(\hat{\Theta}^n, S_+^n) \leq b\sqrt{\log(2|\hat{\Theta}^n|)/n}.$$

- **No outliers:** if ϵ^* minimizing (6) satisfies $\nabla(\hat{\Theta}^n) \leq \epsilon^*$, then

$$\mathrm{Rad}(\hat{\Theta}^n, S^n_+) \leq 2b\sqrt{\frac{\dim_{\mathrm{fM}}\left(\hat{\Theta}^n\right)}{n} \log \frac{\Delta(\hat{\Theta}^n)}{\nabla(\hat{\Theta}^n)}}. \tag{7}$$

- **Steiner points:** in Appendix C, we show that it is possible to construct a set of points P with $|P| \leq |\hat{\Theta}^n|$ (which we refer to as *Steiner points*, following a similar concept in graph theory) such that $\nabla(\hat{\Theta}^n \cup P) \leq \epsilon^*$ and so the "no outliers" case (7) follows for $\hat{\Theta}^n \cup P$. This simplifies the bound at the cost of extending the set $\hat{\Theta}^n$ by a finite number of additional points.

A very interesting simplification compared to Theorem 6 arises in the limit as $n \to \infty$. Let $\hat{\Theta} = \bigcup_{n=1}^{\infty} \hat{\Theta}^n$ be the collection of all output sets of the algorithm obtained for different n. Since $\hat{\Theta}$ is infinite, we may now consider its upper Minkowski dimension, defined by

$$\overline{\dim}_{\mathcal{M}}(\hat{\Theta}) = \limsup_{\delta \to 0^+} \frac{\log \mathrm{Cov}(\hat{\Theta}, \mathtt{d}, \delta)}{\log(1/\delta)}. \tag{8}$$

Theorem 7 *Suppose ℓ is L-Lipschitz continuous in θ and takes values in an interval of length b. Then*

$$\limsup_{n \to \infty} \frac{\mathrm{Rad}(\hat{\Theta}^n, S^n_+)}{\sqrt{\log(n)/n}} \leq b\sqrt{\frac{\overline{\dim}_{\mathcal{M}}(\hat{\Theta})}{2}}.$$

The proof is a straightforward consequence of (6) and the definition of the upper Minkowski dimension; see Appendix C. It would be possible to adapt the proof to go through with $\hat{\Theta}$ replaced by $\hat{\Theta}' = \bigcup_{n=n'}^{\infty} \hat{\Theta}^n$ for some finite integer n', which at first sight looks like it gives a stronger conclusion. On closer inspection, however, the two results turn out to be equivalent, since $\hat{\Theta} \setminus \hat{\Theta}' = \bigcup_{n=1}^{n'-1} \hat{\Theta}^n$ is a finite set, which implies that $\overline{\dim}_{\mathcal{M}}(\hat{\Theta}') = \overline{\dim}_{\mathcal{M}}(\hat{\Theta})$.

Using Fatou's lemma, Lemma 2 and Theorem 7 together imply that the expected generalization error of any (deterministic) algorithm $\mathcal{A} : \mathcal{Z}^n \to \Theta$ is $\mathcal{O}(\sqrt{\log(n)/n})$ whenever $\mathbb{E}[\overline{\dim}_{\mathcal{M}}(\hat{\Theta})]$ is finite, and satisfies

$$\limsup_{n \to \infty} \frac{\mathbb{E}_{S^n}\left[R(\hat{\theta}) - \hat{R}(\hat{\theta}, S^n)\right]}{\sqrt{\log(n)/n}} \leq b\sqrt{2\,\mathbb{E}[\overline{\dim}_{\mathcal{M}}(\hat{\Theta})]}. \tag{9}$$

Hodgkinson et al. (2022) and Dupuis et al. (2023) both also obtain $\mathcal{O}(n^{-1/2})$ bounds on the generalization error involving Minkowski dimensions of algorithm-dependent sets, but they assume Ahlfors regularity and/or incorporate a potentially vacuous mutual information term, while ours requires neither.

4.2. Dimension-independent generalization for SGD

Stochastic gradient descent (SGD) is a randomized algorithm with iterative updates of the form $\hat{\theta}_{t+1} = \Phi_{i_t}(\hat{\theta}_t)$, where

$$\Phi_i(\theta) = \Pi_\Theta(\theta - \eta\nabla\ell(\theta, Z_i)),$$

$\Pi_\Theta(\theta)$ denotes the projection of θ onto Θ, and the algorithm depends on the choice of an initialization point $\theta_1 \in \Theta$, a step size $\eta > 0$, and indices i_t that are chosen uniformly at random from $\{1, \ldots, n\}$. In the sequel, assume that $\Theta \subset \mathbb{R}^k$ is compact and convex, and let the final output of the algorithm after T updates be $\hat{\theta} = \hat{\theta}_{T+1}$. For the loss $\ell(\theta, z)$, we assume that it is differentiable in θ, and satisfies the following conditions for all $z \in \mathcal{Z}$:

- *α-Strong convexity*: for any $\theta, \theta' \in \Theta$, $(\nabla_\theta \ell(\theta, z) - \nabla_{\theta'} \ell(\theta', z)) \cdot (\theta - \theta') \geq \alpha \|\theta - \theta'\|^2$,
- *β-Smoothness*: for any $\theta, \theta' \in \Theta$, $\|\nabla_\theta \ell(\theta, z) - \nabla_{\theta'} \ell(\theta', z)\| \leq \beta \|\theta - \theta'\|$.

We further assume:

- *L-Weak Lipschitz continuity*: For $L > 0$, there exists $h : \Theta \to \mathbb{R}$ such that, for any $\theta, \theta' \in \Theta$ and any $z \in \mathcal{Z}$, $|\ell(\theta, z) - h(\theta) - (\ell(\theta', z) - h(\theta'))| \leq L \|\theta - \theta'\|$.

Under these assumptions, Park et al. (2022) obtain generalization bounds for SGD that do not depend explicitly on the ambient dimension k. The key step in their analysis is that Φ_i is a γ-contractive operator for $\gamma = \sqrt{1 - 2\alpha\eta + \alpha\beta\eta^2}$; i.e., $\|\Phi_i(\theta) - \Phi_i(\theta')\| \leq \gamma \|\theta - \theta'\|$ for any parameters $\theta, \theta' \in \Theta$. This causes SGD to forget about previous iterates at a rate that depends on γ. Using the same idea, we can recover their first main result, presented as (Park et al., 2022, Theorem 2.1), with a much simpler proof based on ARC.

Theorem 8 *If the loss ℓ takes values in an interval $[a, a+b]$, is α-strongly convex, β-smooth and L-weakly Lipschitz continuous, and if Θ is compact and convex with diameter at most $\Delta(\Theta) \leq R$, then for any initialization $\theta_1 \in \Theta$, any $\eta \in (0, 2/\beta)$, any indices $i_1, \ldots, i_T$, and any $\delta \in (0, 1]$, the generalization error for stochastic gradient descent is at most*

$$R(\hat{\theta}) - \hat{R}(\hat{\theta}, S^n) \leq 4b \sqrt{\frac{\max\left\{ \left\lceil \frac{\ln(2Rn)}{\ln(1/\gamma)} \right\rceil \ln 2, 0 \right\}}{n}} + b\sqrt{\frac{8\ln(2/\delta)}{n}} + \frac{2L}{n} \tag{10}$$

with probability at least $1 - \delta$, where $\gamma = \sqrt{1 - 2\alpha\eta + \alpha\beta\eta^2}$.

This recovers Theorem 2.1 of Park et al. (2022) while shaving off a $\log n$ factor. The argument can be extended to losses that are only piecewise strongly convex and smooth to obtain an analog of their second main result, Park et al. (2022, Theorem 2.2), but we omit the details. We further remark that Theorem 8 does not require that T should be large enough for $\hat{\theta}$ to be a good approximation of the empirical risk minimizer (ERM) $\bar{\theta} = \operatorname{argmin}_{\theta \in \Theta} \hat{R}(\theta, S^n)$. However, if this is the case, then strong convexity and Lipschitzness of the loss would imply a better bound on the generalization error, because the ERM is $\frac{2L^2}{\alpha n}$-uniformly stable and consequently satisfies $R(\bar{\theta}) - \hat{R}(\bar{\theta}, S^n) \leq \frac{4L^2}{\delta \alpha n}$ (Shalev-Shwartz et al., 2010, Theorem 5).

Proof Since the losses are α-strongly convex and β-smooth, and $\eta \in (0, \beta/2)$, the operator Φ_i is γ-contractive (Park et al., 2022, Lemma A.4). This implies that for any parameters $\theta, \theta' \in \Theta$, applying the last $m \leq T$ iterations of SGD results in points that are close together. To formalize this, let $\Phi^m = \Phi_{i_T} \circ \cdots \circ \Phi_{i_{T-m+1}}$. Then

$$\|\Phi^m(\theta) - \Phi^m(\theta')\| \leq \gamma^m \|\theta - \theta'\| \leq \gamma^m R.$$

Given $\epsilon > 0$ yet to be determined, take $m = \max\left\{0, \left\lceil \frac{\ln \frac{R}{\epsilon}}{\ln(1/\gamma)} \right\rceil \right\}$ so that this bound is less than ϵ.

There are then two cases for m:

- $m \leq T$: For any $\sigma, \sigma' \in \{-1,+1\}^n$, let $(\hat{\theta}_t)_{t=1,\dots,T+1}$ and $(\hat{\theta}'_t)_{t=1,\dots,T+1}$ denote the iterates of SGD on the corresponding samples S^n_σ and $S^n_{\sigma'}$. If $\sigma_{i_t} = \sigma'_{i_t}$ for $t = T-m+1,\dots,T$, i.e. the last m iterations of SGD are the same, then we have $\|\hat{\theta}_{T+1} - \hat{\theta}'_{T+1}\| \leq \epsilon$ by the argument above. It follows that the covering number for $\hat{\Theta}^n$ at radius ϵ is at most

$$\mathrm{Cov}(\hat{\Theta}^n, \mathtt{d}, \epsilon) \leq |\{(\sigma_{i_t})_{t=T-m+1,\dots,T} \,:\, \sigma \in \{-1,+1\}^n\}| = 2^m.$$

- $m > T$: Then the argument above does not apply, but, since the output of SGD only depends on the T data points that it actually visits,

$$\mathrm{Cov}(\hat{\Theta}^n, \mathtt{d}, \epsilon) \leq |\hat{\Theta}^n| \leq |\{(\sigma_{i_t})_{t=1,\dots,T} \,:\, \sigma \in \{-1,+1\}^n\}| = 2^T < 2^m.$$

Both cases, therefore, lead to the same upper bound on the covering number. We now note that the generalization error $R(\theta) - \hat{R}(\theta, S^n)$ does not change if we replace the loss $\ell(\theta, z)$ by $\bar{\ell}(\theta, z) := \ell(\theta, z) - h(\theta)$, and therefore we may assume without loss of generality that the loss is Lipschitz continuous instead of only weakly Lipschitz continuous (replacing ℓ by $\bar{\ell}$ if necessary). It then follows from (6) that

$$\mathrm{Rad}(\hat{\Theta}^n, S^n_+) \leq L\epsilon + b\sqrt{\frac{m \ln 2}{n}} = L\epsilon + b\sqrt{\frac{\max\left\{\left\lceil \frac{\ln \frac{R}{\epsilon}}{\ln(1/\gamma)} \right\rceil \ln 2, 0\right\}}{n}}.$$

The proof is completed by plugging in $\epsilon = 1/(2n)$, and combining this with Lemma 2. ■

Remark 9 *The previous proof can be adapted to work with the larger class* $\bar{\Theta}^n$ *instead of* $\hat{\Theta}^n$*, as discussed in Section 3. This leads to a slightly worse bound on* $\mathrm{Cov}(\hat{\Theta}^n, \mathtt{d}, \epsilon)$ *of* $\binom{2n}{m} \leq (\frac{e2n}{m})^m$ *instead of our current* 2^m*, and as a result we would get* $m \ln(\frac{e2n}{m})$ *in place of the current* $m \ln 2$*, which is only an* $O(\log n)$ *factor worse.*

4.3. Generalization for compression schemes

Using conditional mutual information, Steinke and Zakynthinou (2020) show that the generalization error for the output of a k-compression scheme can be upper bounded by a quantity of order $\mathcal{O}(\sqrt{k \log n / n})$. We show this result is easily recovered using the ARC. An algorithm $\mathcal{A}$ is a *k-compression scheme* if $\mathcal{A}(S^n) = \mathcal{A}_2(\mathcal{A}_1(S^n))$, where $\mathcal{A}_1 : \mathcal{Z}^n \to \mathcal{Z}^k$ maps any sample S^n of size n to a subsample $S^k \subset S^n$ of size $k \leq n$, and $\mathcal{A}_2 : \mathcal{Z}^k \to \Theta$ deterministically determines the final output based only on this subsample.

Theorem 10 *Suppose* $\mathcal{A}$ *is a k-compression scheme and losses take values in* $[0,1]$*. Then, for any* S^n_- *and* S^n_+*,*

$$\mathrm{Rad}(\hat{\Theta}^n, S^n_+) \leq \sqrt{\frac{k \log \frac{2en}{k}}{2n}} = \mathcal{O}\Big(\sqrt{\frac{k \log n}{n}}\Big).$$

Proof Note that the total number of subsamples of length k from $\bigcup_{\sigma\in\{-1,+1\}^n} S_\sigma^n$ is $\binom{2n}{k}$, the number of subsamples of length k from $S_-^n \cup S_+^n$. Consequently, there are at most $\binom{2n}{k}$ possible parameters in $\hat{\Theta}^n$:

$$|\hat{\Theta}^n| = |\{\mathcal{A}(S_\sigma^n) : \sigma \in \{-1,+1\}^n\}\}| \leq |\{\mathcal{A}_1(S_\sigma^n) : \sigma \in \{-1,+1\}^n\}| \leq \binom{2n}{k}.$$

By Massart's lemma (Shalev-Shwartz and Ben-David, 2014), $\mathrm{Rad}(\hat{\Theta}^n, S_+) \leq \sqrt{\frac{\log|\hat{\Theta}^n|}{2n}} \leq \sqrt{\frac{\log\binom{2n}{k}}{2n}} \leq \sqrt{\frac{k}{2n}\log\frac{2en}{k}}$, as required. ■

4.4. Generalization for VC classes

For binary classification, Steinke and Zakynthinou (2020) further show that, if Θ indexes a class of finite VC dimension V, then there exists a version of the empirical risk minimizer (ERM) over that class for which the conditional mutual information is bounded by $\mathcal{O}(V\log n)$, leading to a bound on the generalization error of $\mathcal{O}(\sqrt{V\log n/n})$. An analogous result, which works for any version of the ERM, can trivially be obtained from the ARC, because

$$\mathrm{Rad}(\hat{\Theta}^n, S_+^n) \leq \mathrm{Rad}(\Theta, S_+^n) = \mathcal{O}\left(\sqrt{\frac{V\log n}{n}}\right),$$

where the first inequality follows from $\hat{\Theta}^n \subset \Theta$ and the second is a standard result for Rademacher complexity, obtained by bounding the Rademacher complexity using the growth function, which is then controlled using Sauer's lemma (Shalev-Shwartz and Ben-David, 2014).

5. Conclusion and Future Work

In this work, we considered algorithm-dependent Rademacher complexity as an approach to obtain algorithm-dependent generalization bounds. Circumventing the information-theoretic route, the proposed complexity notion on the one hand allowed us to derive and unify several known results with little effort, such as the generalization bound for SGD; on the other hand it enabled us to link the generalization error to topological properties of the learning algorithm using the finite Minkowski dimension (Alonso, 2015).

We believe that our work opens up several interesting future research directions. Given the conceptual similarities in conditioning on a supersample between ARC and CMI, it is natural to ask how the two complexity measures compare in general. In which cases is one preferable to the other? Another important connection would be the relation with algorithmic stability, which is known to play a fundamental role in generalization with a uniform rate of convergence over all distributions (Shalev-Shwartz et al., 2010).

The concept of finite fractal dimensions, introduced by Alonso (2015), turned out to be a fruitful tool to provide bounds with respect to interpretable topological properties without a mutual information term. We believe this is a promising direction, specifically when measuring the generalization bound with respect to a finite hypothesis class. It would be interesting to understand whether the definitions of a finite Minkowski dimension can be further adapted for the specific needs in generalization bounds. In particular, it would be interesting to understand if it is possible to relax the definition of a 2-cover further while preserving the limiting behavior.

Acknowledgments

Sachs and Van Erven were supported by the Netherlands Organization for Scientific Research (NWO) under grant number VI.Vidi.192.095. Şimşekli 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-P3IA0001 (PRAIRIE 3IA Institute) and by the European Research Council Starting Grant DYNASTY – 101039676. Part of the research was performed while Sachs was visiting INRIA Paris within the scope of the DYNASTY grant.

References

Juan M Alonso. A Hausdorff dimension for finite sets. *arXiv preprint arXiv:1508.02946*, 2015.

Martin Anthony and Peter L. Bartlett. *Neural Network Learning - Theoretical Foundations.* Cambridge University Press, 2002. ISBN 978-0-521-57353-5.

Sanjeev Arora, Rong Ge, Behnam Neyshabur, and Yi Zhang. Stronger generalization bounds for deep nets via a compression approach. In *International Conference on Machine Learning*, pages 254–263. PMLR, 2018.

Melih Barsbey, Milad Sefidgaran, Murat A Erdogdu, Gael Richard, and Umut Simsekli. Heavy tails in SGD and compressibility of overparametrized neural networks. *Advances in Neural Information Processing Systems*, 34:29364–29378, 2021.

Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local Rademacher complexities. *The Annals of Statistics*, 33(4):1497 – 1537, 2005. doi: 10.1214/009053605000000282.

Raef Bassily, Kobbi Nissim, Adam Smith, Thomas Steinke, Uri Stemmer, and Jonathan Ullman. Algorithmic stability for adaptive data analysis. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*. ACM, jun 2016. doi: 10.1145/2897518.2897566.

Raef Bassily, Shay Moran, Ido Nachum, Jonathan Shafer, and Amir Yehudayoff. Learners that use little information. In *International Conference on Algorithmic Learning Theory*, 2017.

Tolga Birdal, Aaron Lou, Leonidas J Guibas, and Umut Simsekli. Intrinsic dimension, persistent homology and generalization in neural networks. *Advances in Neural Information Processing Systems*, 34:6776–6789, 2021.

Alexander Camuto, George Deligiannidis, Murat A Erdogdu, Mert Gurbuzbalaban, Umut Simsekli, and Lingjiong Zhu. Fractal structure and generalization properties of stochastic optimization algorithms. *Advances in Neural Information Processing Systems*, 34:18774–18788, 2021.

Benjamin Dupuis, George Deligiannidis, and Umut Şimşekli. Generalization bounds with data-dependent fractal dimensions, 2023.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Leon Roth. Preserving statistical validity in adaptive data analysis. In *Proceedings of the forty-seventh annual ACM symposium on Theory of Computing*. ACM, jun 2015. doi: 10.1145/2746539.2746580.

Kenneth Falconer. *Fractal geometry: mathematical foundations and applications*. John Wiley & Sons, 2004.

Dylan J Foster, Spencer Greenberg, Satyen Kale, Haipeng Luo, Mehryar Mohri, and Karthik Sridharan. Hypothesis set stability and generalization. In *Advances in Neural Information Processing Systems 32*, volume 32, pages 6729–6739, 2019. URL https://proceedings.neurips.cc/paper/2019/file/300d1539c3b6aa1793b5678b857732cf-Paper.pdf.

Mahdi Haghifam, Borja Rodríguez-Gálvez, Ragnar Thobaben, Mikael Skoglund, Daniel M Roy, and Gintare Karolina Dziugaite. Limitations of information-theoretic generalization bounds for gradient descent methods in stochastic convex optimization. *arXiv preprint arXiv:2212.13556*, 2022.

Liam Hodgkinson, Umut Simsekli, Rajiv Khanna, and Michael Mahoney. Generalization bounds using lower tail exponents in stochastic optimizers. In *International Conference on Machine Learning*, pages 8774–8795. PMLR, 2022.

Tsutomu Kawabata and Amir Dembo. The rate-distortion dimension of sets and measures. *IEEE Transactions on Information Theory*, 40(5):1564–1572, 1994.

Nick Littlestone and Manfred K. Warmuth. Relating data compression and learnability. Unpublished manuscript, 2003.

Sejun Park, Umut Şimşekli, and Murat A. Erdogdu. Generalization bounds for stochastic gradient descent via localized ε-covers. In *Advances in Neural Information Processing Systems*, 2022.

Daniel Russo and James Zou. Controlling bias in adaptive data analysis using information theory. In *Artificial Intelligence and Statistics*, pages 1232–1240. PMLR, 2016.

Milad Sefidgaran, Amin Gohari, Gaël Richard, and Umut Simsekli. Rate-distortion theoretic generalization bounds for stochastic learning algorithms. In Po-Ling Loh and Maxim Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 4416–4463. PMLR, 02–05 Jul 2022.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Learnability, stability and uniform convergence. *Journal of Machine Learning Research*, 11(90):2635–2670, 2010. URL http://jmlr.org/papers/v11/shalev-shwartz10a.html.

Umut Simsekli, Ozan Sener, George Deligiannidis, and Murat A Erdogdu. Hausdorff dimension, heavy tails, and generalization in neural networks. *Advances in Neural Information Processing Systems*, 33:5138–5151, 2020.

Thomas Steinke and Lydia Zakynthinou. Reasoning About Generalization via Conditional Mutual Information. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 3437–3452. PMLR, 09–12 Jul 2020.

Taiji Suzuki, Hiroshi Abe, Tomoya Murata, Shingo Horiuchi, Kotaro Ito, Tokuma Wachi, So Hirai, Masatoshi Yukishima, and Tomoaki Nishimura. Spectral pruning: Compressing deep neural networks via spectral analysis and its generalization error. *AISTATS*, 2020a.

Taiji Suzuki, Hiroshi Abe, and Tomoaki Nishimura. Compression based bound for non-compressed network: unified generalization error analysis of large compressible deep neural network. *ICLR*, 2020b.

Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. *Advances in Neural Information Processing Systems*, 30, 2017.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

Appendix A. Proofs from Section 3

Proof [Lemma 2] In order to lighten notation, within this proof we do not write superscripts n, so that $S_-^n = S_-$, $\hat{\Theta}^n = \hat{\Theta}$, etc. We further identify S with S_+ and we abbreviate $\hat{R}_\xi(\theta) \equiv \hat{R}(\theta, S_\xi)$ for any ξ.

A common part to both results in the theorem is the following:

$$\begin{aligned}
\mathop{\mathbb{E}}_{\sigma}\left[\max_{\theta\in\hat{\Theta}}\left\{\hat{R}_{-\sigma}(\theta)-\hat{R}_{\sigma}(\theta)\right\}\right] &= \frac{1}{n}\mathop{\mathbb{E}}_{\sigma}\left[\max_{\theta\in\hat{\Theta}}\sum_{i=1}^{n}\sigma_i\big(\ell(Z_i^{-1},\theta)-\ell(Z_i^{+1},\theta)\big)\right]\\
&\le \frac{1}{n}\mathop{\mathbb{E}}_{\sigma}\left[\max_{\theta\in\hat{\Theta}}\sum_{i=1}^{n}\sigma_i\ell(Z_i^{-1},\theta)+\max_{\theta\in\hat{\Theta}}\sum_{i=1}^{n}-\sigma_i\ell(Z_i^{+1},\theta)\right]\\
&= \mathrm{Rad}(\hat{\Theta},S_-)+\mathrm{Rad}(\hat{\Theta},S_+). \qquad (11)
\end{aligned}$$

We now start by proving the in-expectation result:

$$\begin{aligned}
\mathop{\mathbb{E}}_{S_+}\left[R(\hat{\theta}(S_+))-\hat{R}_+(\hat{\theta}(S_+))\right] &= \mathop{\mathbb{E}}_{S_-,S_+}\left[\hat{R}_-(\hat{\theta}(S_+))-\hat{R}_+(\hat{\theta}(S_+))\right]\\
&= \mathop{\mathbb{E}}_{S_-,S_+,\sigma}\left[\hat{R}_{-\sigma}(\hat{\theta}(S_\sigma))-\hat{R}_\sigma(\hat{\theta}(S_\sigma))\right]\\
&\le \mathop{\mathbb{E}}_{S_-,S_+,\sigma}\left[\max_{\theta\in\hat{\Theta}}\left\{\hat{R}_{-\sigma}(\theta)-\hat{R}_\sigma(\theta)\right\}\right]\\
&\le \mathop{\mathbb{E}}_{S_-,S_+}\left[\mathrm{Rad}(\hat{\Theta},S_-)+\mathrm{Rad}(\hat{\Theta},S_+)\right]\\
&= 2\mathop{\mathbb{E}}_{S_-,S_+}\left[\mathrm{Rad}(\hat{\Theta},S_+)\right],
\end{aligned}$$

where the second inequality follows from (11). This completes the proof of (4).

We proceed to prove the in-probability result. To this end, let $\epsilon > 0$ be chosen later, and define the following two events:

$$\begin{aligned}
\mathcal{E}_1 &= \left\{S_+ : R(\hat{\theta}(S_+))-\hat{R}_+(\hat{\theta}(S_+))\ge\epsilon\right\},\\
\mathcal{E}_2 &= \left\{(S_-,S_+) : \hat{R}_-(\hat{\theta}(S_+))-\hat{R}_+(\hat{\theta}(S_+))\ge\frac{\epsilon}{2}\right\}.
\end{aligned}$$

Our goal will be to bound $\Pr(\mathcal{E}_1)$ and we will start by showing that

$$\Pr(\mathcal{E}_1) \leq 2\Pr(\mathcal{E}_2). \tag{12}$$

This can be established as follows: Since $R(\hat{\theta}(S_+)) - \hat{R}_+(\hat{\theta}(S_+)) \geq \epsilon$ and $R(\hat{\theta}(S_+)) - \hat{R}_-(\hat{\theta}(S_+)) \leq \frac{\epsilon}{2}$ together imply $\mathcal{E}_2$, we have

$$\begin{aligned}
\Pr(\mathcal{E}_2) &\geq \Pr\left(R(\hat{\theta}(S_+)) - \hat{R}_+(\hat{\theta}(S_+)) \geq \epsilon \quad \text{and} \quad R(\hat{\theta}(S_+)) - \hat{R}_-(\hat{\theta}(S_+)) \leq \frac{\epsilon}{2}\right) \\
&= \mathop{\mathbb{E}}_{S_-,S_+}\left[\mathbb{1}[R(\hat{\theta}(S_+)) - \hat{R}_+(\hat{\theta}(S_+)) \geq \epsilon]\mathbb{1}[R(\hat{\theta}(S_+)) - \hat{R}_-(\hat{\theta}(S_+)) \leq \frac{\epsilon}{2}]\right] \\
&= \mathop{\mathbb{E}}_{S_+}\left[\mathbb{1}[R(\hat{\theta}(S_+)) - \hat{R}_+(\hat{\theta}(S_+)) \geq \epsilon] \Pr_{S_-}\left(R(\hat{\theta}(S_+)) - \hat{R}_-(\hat{\theta}(S_+)) \leq \frac{\epsilon}{2}\right)\right] \\
&\geq \mathop{\mathbb{E}}_{S_+}\left[\mathbb{1}[R(\hat{\theta}(S_+)) - \hat{R}_+(\hat{\theta}(S_+)) \geq \epsilon] \times \frac{1}{2}\right] \\
&= \frac{1}{2}\Pr(\mathcal{E}_1),
\end{aligned}$$

where, for the last inequality to hold, we restrict attention to $\epsilon \geq b\sqrt{\frac{2\log(2)}{n}}$.[6] The last inequality then holds because, for any fixed θ,

$$\Pr_{S_-}\left(R(\theta) - \hat{R}_-(\theta) \leq \frac{\epsilon}{2}\right) \geq 1/2.$$

To see this, note that by Hoeffding's inequality the probability of the event's complement is at most

$$\Pr_{S_-}\left(R(\theta) - \hat{R}_-(\theta) > \frac{\epsilon}{2}\right) \leq \exp\left(-\frac{n\epsilon^2}{2b^2}\right) \leq 1/2.$$

This completes the proof of (12).

We proceed to work on the right-hand side of (12) by rewriting the probability of $\mathcal{E}_2$ as follows:

$$\begin{aligned}
\Pr(\mathcal{E}_2) &= \mathop{\mathbb{E}}_{S_-,S_+}\left[\mathbb{1}[\hat{R}_-(\hat{\theta}(S_+)) - \hat{R}_+(\hat{\theta}(S_+)) \geq \frac{\epsilon}{2}]\right] \\
&= \mathop{\mathbb{E}}_{S_-,S_+,\sigma}\left[\mathbb{1}[\hat{R}_{-\sigma}(\hat{\theta}(S_\sigma)) - \hat{R}_\sigma(\hat{\theta}(S_\sigma)) \geq \frac{\epsilon}{2}]\right] \\
&= \mathop{\mathbb{E}}_{S_-,S_+}\left[\Pr_\sigma\left(\hat{R}_{-\sigma}(\hat{\theta}(S_\sigma)) - \hat{R}_\sigma(\hat{\theta}(S_\sigma)) \geq \frac{\epsilon}{2}\right)\right].
\end{aligned}$$

We now restrict attention to ϵ that are at least

$$\epsilon \geq 2\mathop{\mathbb{E}}_\sigma\left[\max_{\theta\in\hat{\Theta}}\left(\hat{R}_{-\sigma}(\theta) - \hat{R}_\sigma(\theta)\right)\right] + b\sqrt{\frac{8\log(2/\delta)}{n}} \qquad \text{almost surely,} \tag{13}$$

so that

$$\Pr_\sigma\left(\hat{R}_{-\sigma}(\hat{\theta}(S_\sigma)) - \hat{R}_\sigma(\hat{\theta}(S_\sigma)) \geq \frac{\epsilon}{2}\right) \leq \Pr_\sigma\left(\max_{\theta\in\hat{\Theta}}\left(\hat{R}_{-\sigma}(\theta) - \hat{R}_\sigma(\theta)\right) \geq \frac{\epsilon}{2}\right) \leq \frac{\delta}{2},$$

[6] Anthony and Bartlett (2002) relax this to $\epsilon \geq b\sqrt{2/n}$ using a more involved argument based on Chebyshev's instead of Hoeffding's inequality, but this provides no benefit here, because we will use a large enough value of ϵ anyway.

where the last bound holds by McDiarmid's inequality, which applies because $\max_{\theta \in \hat{\Theta}} \big(\hat{R}_{-\sigma}(\theta) - \hat{R}_{\sigma}(\theta)\big)$ has $\frac{2b}{n}$-bounded differences. Putting everything together, we have shown that $\Pr(\mathcal{E}_1) \leq \delta$ for any (non-random) ϵ that satisfies $\epsilon \geq b\sqrt{2\log(2)/n}$ and (13). We will show that this is the case for

$$\epsilon = 4 \operatorname*{ess\,sup}_{S_-, S_+} \{\mathrm{Rad}(\hat{\Theta}, S_+)\} + b\sqrt{\frac{8\log(2/\delta)}{n}},$$

from which the intended result (5) then follows.

The first constraint on ϵ is easiest, because Rademacher complexity is always non-negative:

$$\epsilon \geq b\sqrt{\frac{8\log(2/\delta)}{n}} \geq b\sqrt{\frac{2\log(2)}{n}}.$$

It remains to check (13), which follows from (11) by

$$\mathop{\mathbb{E}}_{\sigma}\left[\max_{\theta \in \hat{\Theta}} \big(\hat{R}_{-\sigma}(\theta) - \hat{R}_{\sigma}(\theta)\big)\right] \leq \mathrm{Rad}(\hat{\Theta}, S_-) + \mathrm{Rad}(\hat{\Theta}, S_+) \leq 2 \operatorname*{ess\,sup}_{S_-, S_+} \{\mathrm{Rad}(\hat{\Theta}, S_+)\} \quad \text{almost surely.}$$

This completes the proof. ■

Appendix B. Definition and basic properties of finite Minkowski dimension

B.1. Complete definition of finite Minkowski dimension

In this section, we restate the complete definition of the finite Minkowski dimension for the convenience of the reader. In the main part of the paper, we excluded e.g. focal or empty sets via several additional assumptions. These assumptions are not needed when considering the more technical original definition by Alonso (2015).

Recall the definition of a 2-cover from the main part: a family of sets $\mathcal{U}$ is a 2-cover of a finite set C if $\mathcal{U} = \{U_i : i \in \mathbb{N}\}$ where each $U_i \subseteq C$, $|U_i| \geq 2$, and $C \subseteq \bigcup_{U \in \mathcal{U}} U$. Here we will denote the set of all such 2-covers for C as $K(C)$. Further, let $\Delta(\mathcal{U}) = \max\{\Delta(U_i) : U_i \in \mathcal{U}\}$ and $K_\delta(C) = \{\mathcal{U} \in K(C) : \Delta(\mathcal{U}) \leq \delta\}$. Now define

$$K^1(C) = \{\mathcal{U} \in K(C) : \Delta(\mathcal{U}) < \Delta(C)\} \qquad \text{and} \qquad K^1_\delta(C) = K^1(C) \cap K_\delta(C).$$

Note that in the main part, we used the following result to simplify the notation and definition of the finite Minkowski dimension.

Theorem 11 *(Alonso, 2015, Theorem 2.14) Let C be finite. Then the following are equivalent:*

1. *C has no focal point.*
2. *$K^1(C) \neq \emptyset$.*
3. *$\nabla(C) < \Delta(C)$.*

Next, define a covering, similar to the box covering for the definition of the Minkowski dimension.

Definition 12 *(Alonso, 2015, Definition 4.1) For $\mathcal{U} \in K(C)$, set*

$$B^s_{\mathcal{U}}(C) = |\mathcal{U}|\, \Delta(\mathcal{U})^s.$$

For $\delta \geq \nabla(C)$, set

$$B^s_\delta(C) = \begin{cases} \min\left\{B^s_{\mathcal{U}} : \mathcal{U} \in K^1_\delta(C)\right\}, & \text{if } K^1(C) \neq \emptyset \\ \min\left\{B^s_{\mathcal{U}} : \mathcal{U} \in K(C)\right\}, & \text{if } K^1(C) = \emptyset \end{cases},$$

and

$$B^s(C) = \max\left\{B^s_\delta(C) : \delta \geq \nabla(C)\right\}.$$

As in Section 4.1 in Alonso (2015) the finite Minkowski dimension is defined as follows:

Definition 13 (finite Minkowski dimension) *Let $s \in (0, \infty)$ such that*

$$B^s(C) = \Delta(C)^s. \tag{14}$$

The finite Minkowski dimension of a non-empty set C is

$$\dim_{\text{fM}}(C) = \begin{cases} 0 & \text{if } |C| = 1 \\ +\infty & \text{if } C \text{ is focal} \\ s \text{ satisfying (14)} & \text{otherwise.} \end{cases}$$

In Section 4.1 we used the following theorem to simplify the definition of the finite Minkowski dimension.

Theorem 14 *(Alonso, 2015, Theorem 4.11) Let C be non-empty and finite. If C is non-focal then*

$$\dim_{\text{fM}}(C) = \frac{\log T(C)}{\log \frac{\Delta(C)}{\nabla(C)}}.$$

Theorem 15 *(Alonso, 2015, Theorem 4.12) Let $\eta : \mathcal{X} \to \mathcal{X}'$ be (r, β)-Hölder continuous and $C \subseteq \mathcal{X}$ finite. Then $\beta \dim_{\text{fM}}(\eta(C)) = \dim_{\text{fM}}(C)$.*

We used the following result to derive the trivial upper bound in Section 4.1, Theorem 6.

Theorem 16 *(Alonso, 2015, Proposition 5.3) Let C be finite. Then*

$$\dim_{\text{fM}}(C) \leq \frac{\log(|C| - 1)}{\log \frac{\Delta(C)}{\nabla(C)}}.$$

Theorem 17 *(Alonso, 2015, Theorem 7.17) Let $C \subseteq \mathbb{R}^n$ be compact, with $\nabla(C) = 0 < \Delta(C)$. Then there exists a sequence of sets $\{F_k\}_{k \in \mathbb{N}}$ with $F_k \to C$ and $\lim_{k\to\infty} \dim_{\text{fM}}(F_k) = \overline{\dim}_{\mathcal{M}}(X)$.*

Appendix C. Proofs from Section 4.1

For this section, we need some additional notation. Since these definitions are only needed for the proofs within this section, we define them locally for better readability of the main section.

Notation: For any set $S \in \mathcal{X}$, we let $\mathrm{Conv}\,(S)$ denote its convex hull, formally, $\mathrm{Conv}\,(S)$ is the unique minimal convex set such that $S \subseteq \mathrm{Conv}\,(S)$.

The proof of (7) follows immediately from Lemma 18.

Lemma 18 (Existence of Optimal Steiner Points) *Suppose $\hat{\Theta}^n \subset \mathbb{R}^k$ is any finite non-focal set. Then for any $0 < \epsilon < \Delta(\hat{\Theta}^n)$, there exists a set P with the following properties.*

1. $|P| < |\hat{\Theta}^n|$.
2. $\nabla(\hat{\Theta}^n \cup P) = \epsilon$ *and* $\Delta(\hat{\Theta}^n \cup P) = \Delta(\hat{\Theta}^n)$.
3. $\mathrm{Cov}(\hat{\Theta}^n, \mathtt{d}, \epsilon) = \mathrm{Cov}(\hat{\Theta}^n \cup P, \mathtt{d}, \epsilon)$.

Proof Let S be the set corresponding to the covering number $\mathrm{Cov}(\hat{\Theta}^n, \mathtt{d}, \epsilon)$, i.e., S contains the centers of the ϵ-covers for $\hat{\Theta}^n$ and $|S| = \mathrm{Cov}(\hat{\Theta}^n, \mathtt{d}, \epsilon)$. Recall $\mathbb{B}\,(c, r, \mathtt{d}) = \{x : \mathtt{d}(c, x) \leq r\}$ denotes a ball with center c and radius r. Now consider $P \subset \cup_{a \in S} \mathbb{B}\,(a, \epsilon, \mathtt{d})$, then 3 is satisfied. It remains to show that we can choose $P \subset \cup_{a \in S} \mathbb{B}\,(a, \epsilon, \mathtt{d})$ such that 1 and 2 are satisfied. By restricting the choice of P further to $P \subset \cup_{a \in S} \mathbb{B}\,(a, \epsilon, \mathtt{d}) \cap \mathrm{Conv}\left(\hat{\Theta}^n\right)$ we can ensure that the diameter $\Delta(\hat{\Theta}^n \cup P) = \Delta(\hat{\Theta}^n)$. Due to the assumptions on ϵ, the conditions $|P| < |\hat{\Theta}^n|$ and $\nabla(\hat{\Theta}^n \cup P) \leq \epsilon$ can also be satisfied by choosing the points in P from the boundaries of the epsilon balls, i.e., $P \subset \cup_{a \in S} \mathrm{bd}\,(\mathbb{B}\,(a, \epsilon, \mathtt{d})) \cap \mathrm{Conv}\left(\hat{\Theta}^n\right)$. ■

We are now ready to prove Theorem 7.

Proof [Theorem 7] Let $\epsilon_n = \alpha\sqrt{\frac{\log n}{n}}$ for any constant $\alpha > 0$. Then, by (6),

$$\mathrm{Rad}(\hat{\Theta}^n, S_+^n) \leq L\epsilon_n + b\sqrt{\frac{\log \mathrm{Cov}(\hat{\Theta}^n, \mathtt{d}, \epsilon_n)}{n}} \leq L\epsilon_n + b\sqrt{\frac{\log \mathrm{Cov}(\hat{\Theta}, \mathtt{d}, \epsilon_n)}{n}}.$$

Hence

$$\begin{aligned}\limsup_{n\to\infty} \frac{\mathrm{Rad}(\hat{\Theta}^n, S_+^n)}{\sqrt{\log(n)/n}} &\leq \alpha L + \limsup_{n\to\infty} b\sqrt{\frac{\log \mathrm{Cov}(\hat{\Theta}, \mathtt{d}, \epsilon_n)}{\log n}} \\ &= \alpha L + \limsup_{n\to\infty} b\sqrt{\frac{\log \mathrm{Cov}(\hat{\Theta}, \mathtt{d}, \epsilon_n)}{-\log \epsilon_n}}\sqrt{\frac{-\log(\epsilon_n)}{\log n}} \\ &= \alpha L + b\sqrt{\frac{\overline{\dim}_{\mathcal{M}}(\hat{\Theta})}{2}},\end{aligned}$$

where the last equality follows by the definition of the upper Minkowski dimension (8) and because

$$\frac{-\log(\epsilon_n)}{\log n} = \frac{-\log\alpha - \frac{1}{2}\log\log n + \frac{1}{2}\log n}{\log n} \to \frac{1}{2} \qquad \text{as } n \to \infty.$$

The result follows by letting α tend to 0. ■