Fractal Structure and Generalization Properties of Stochastic Optimization Algorithms

Alexander Camuto[1], **George Deligiannidis**[1], **Murat A. Erdogdu**[2],
Mert Gürbüzbalaban[3✉], **Umut Şimşekli**[4✉], **Lingjiong Zhu**[5]

1: University of Oxford & Alan Turing Institute **2:** University of Toronto & Vector Institute
3: Rutgers Business School **4:** INRIA & École Normale Supérieure - PSL Research University
5: Florida State University

The authors are in alphabetical order.
✉: Corresponding authors

Abstract

Understanding generalization in deep learning has been one of the major challenges in statistical learning theory over the last decade. While recent work has illustrated that the dataset and the training algorithm must be taken into account in order to obtain meaningful generalization bounds, it is still theoretically not clear which properties of the data and the algorithm determine the generalization performance. In this study, we approach this problem from a dynamical systems theory perspective and represent stochastic optimization algorithms as *random iterated function systems* (IFS). Well studied in the dynamical systems literature, under mild assumptions, such IFSs can be shown to be ergodic with an invariant measure that is often supported on sets with a *fractal structure*. As our main contribution, we prove that the generalization error of a stochastic optimization algorithm can be bounded based on the 'complexity' of the fractal structure that underlies its invariant measure. Then, by leveraging results from dynamical systems theory, we show that the generalization error can be explicitly linked to the choice of the algorithm (e.g., stochastic gradient descent – SGD), algorithm hyperparameters (e.g., step-size, batch-size), and the geometry of the problem (e.g., Hessian of the loss). We further specialize our results to specific problems (e.g., linear/logistic regression, one hidden-layered neural networks) and algorithms (e.g., SGD and preconditioned variants), and obtain analytical estimates for our bound. For modern neural networks, we develop an efficient algorithm to compute the developed bound and support our theory with various experiments on neural networks.

1 Introduction

In statistical learning, many problems can be naturally formulated as a risk minimization problem

$$\min_{w\in\mathbb{R}^d}\Big\{\mathcal{R}(w) := \mathbb{E}_{z\sim\pi}[\ell(w,z)]\Big\}, \tag{1}$$

where $z \in \mathcal{Z}$ denotes a data sample coming from an unknown distribution π, and $\ell : \mathbb{R}^d \times \mathcal{Z} \to \mathbb{R}_+$ is the composition of a loss and a function from the hypothesis class parameterized by $w \in \mathbb{R}^d$. Since the distribution π is unknown, one needs to rely on empirical risk minimization as a surrogate to (1),

$$\min_{w\in\mathbb{R}^d}\Big\{\hat{\mathcal{R}}(w,\mathbf{S}_n) := (1/n)\sum\nolimits_{i=1}^{n}\ell(w,z_i)\Big\}, \tag{2}$$

where $\mathbf{S}_n := \{z_1,\dots,z_n\}$ denotes a *training set* of n points that are independently and identically distributed (i.i.d.) and sampled from π, and model training often amounts to using an optimization algorithm to solve the above problem.

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

Statistical learning theory is mainly interested in understanding the behavior of the *generalization error*, i.e., $\hat{\mathcal{R}}(w, \mathbf{S}_n) - \mathcal{R}(w)$. While classical results suggest that models with large number of parameters should suffer from poor generalization [SSBD14, AB09], modern neural networks challenge this classical wisdom: they can fit the training data perfectly, yet manage to generalize well [ZBH+17, NBMS17]. Considering that the generalization error is influenced by many factors involved in the training process, the conventional algorithm- and data-agnostic uniform bounds are typically overly pessimistic in a deep learning setting. In order to obtain meaningful, non-vacuous bounds, the underlying data distribution and the choice of the optimization algorithm need to be incorporated in the generalization bounds [ZBH+17, DR17].

Our goal in this study is to develop novel generalization bounds that explicitly incorporate the data and the optimization dynamics, through the lens of dynamical systems theory. To motivate our approach, let us consider stochastic gradient descent (SGD), which has been one of the most popular optimization algorithms for training neural networks. It is defined by the following recursion:

$$w_k = w_{k-1} - \eta \nabla \tilde{\mathcal{R}}_k(w_{k-1}), \quad \text{where} \quad \nabla \tilde{\mathcal{R}}_k(w) := \nabla \tilde{\mathcal{R}}_{\Omega_k}(w) := (1/b) \sum\nolimits_{i \in \Omega_k} \nabla \ell(w, z_i).$$

Here, k represents the iteration counter, $\eta > 0$ is the step-size (also called the learning-rate), $\nabla \tilde{\mathcal{R}}_k$ is the stochastic gradient, b is the batch-size, and $\Omega_k \subset \{1, 2, \dots, n\}$ is a random subset drawn with or without replacement with cardinality $|\Omega_k| = b$ for all k.

Constant step-size SGD forms a Markov chain with a stationary distribution $w_\infty \sim \mu$, which exists and is unique under mild conditions [DDB20, YBVE20], and intuitively we can expect that the generalization performance of the trained model to be intimately related to the behavior of the risk $\mathcal{R}(w)$ over this limit distribution μ. In particular, the Markov chain defined by the SGD recursion can be written by using random functions h_{Ω_k} at each SGD iteration k, i.e.,

$$w_k = h_{\Omega_k}(w_{k-1}), \quad \text{with} \quad h_{\Omega_k}(w) = w - \eta \nabla \tilde{\mathcal{R}}_k(w). \tag{3}$$

Here, the randomness in h_{Ω_k} is due to the selection of the subset Ω_k. In fact, such a formulation is not specific to SGD; it can cover many other stochastic optimization algorithms if the random function h_{Ω_k} is selected accordingly, including second-order algorithms such as *preconditioned SGD* [Li17]. Such random recursions (3) and characteristics of their stationary distribution have been studied extensively under the names of *iterated random functions* [DF99] and *iterated function system* (IFS) [Fal04]. In this paper, from a high level, we relate the 'complexity of the stationary distribution' of a particular IFS to the generalization performance of the trained model.

We illustrate our context in two toy examples. In the first one, we consider a 1-dimensional quadratic problem with $n = 2$ and $\ell(w, z_1) = w^2/2$ and $\ell(w, z_2) = w^2/2 - w$. We run SGD with constant step-size η to minimize the resulting empirical risk. We simply choose $\Omega_k \subset \{1, 2\}$ uniformly random with batch-size $b = 1$, and we plot the histograms of stationary distributions for different step-size choices $\eta \in \{1/100, 1/3, 2/3\}$ in Figure 1. We observe that the *support* of the stationary distribution of SGD depends on the step-size: As the step-size increases the support becomes less dense and a *fractal structure* in the stationary distribution can be clearly observed.

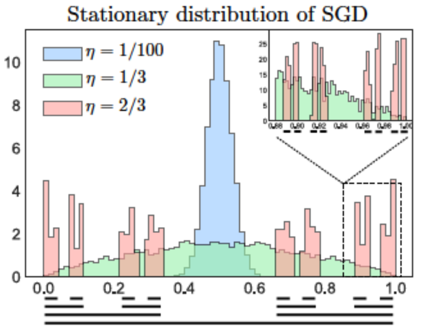

Figure 1: Middle-third Cantor set as the support of the stationary distribution of constant step-size SGD for $\ell(w, z_1) = w^2/2$ and $\ell(w, z_2) = w^2/2 - w$.

This behavior is not surprising, at least for this toy example. It is well-known that the set of points that is invariant under the resulting IFS (termed as the attractor of the IFS) for the specific choice of $\eta = 2/3$ is the famous 'middle-third Cantor set' [FW09], which coincides with the support of the stationary distribution of the SGD.

As another example, we run SGD with constant step-size in order to train an ordinary linear regression model for a dataset of $n = 5$ samples and $d = 2$ dimensions, i.e., $a_i^\top w \approx y_i$, where for $i = 1, \dots, 5$, y_i and each coordinate of a_i are drawn uniformly at random from the interval $[-1, 1]$. Figure 2 shows the heatmap of the resulting stationary distributions for different step-size choices η ranging from 0.1 to 0.9 (bright colors represent higher density). We observe that for small step-size choices, the stationary distribution is dense, whereas a fractal structure can be clearly observed as the step-size gets larger.

Fractals are complex patterns and the level of this complexity is typically measured by the *Hausdorff dimension* of the fractal, which is a notion of dimension that can take fractional values[1], and can be much smaller than the ambient dimension d. Recently, assuming that SGD trajectories can be well-approximated by a certain type of stochastic differential equations (SDE) [ŞGN+19, ŞSG19, NcGR19, ŞZTG20], it is shown that the generalization error can be controlled by the Hausdorff dimension of the trajectories of the SDE, instead of their ambient dimension d [ŞSDE20]. That is, the ambient dimension that appears in classical learning theory bounds is replaced with the Hausdorff dimension.

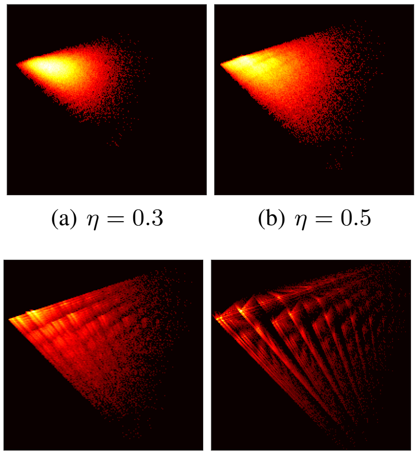
(a) $\eta = 0.3$ (b) $\eta = 0.5$ (c) $\eta = 0.7$ (d) $\eta = 0.9$

Figure 2: The stationary distribution of constant step-size SGD for linear regression, where we have $n = 5$ data points and $w \in \mathbb{R}^2$.

The fractal geometric approach presented in [ŞSDE20] can capture the 'low dimensional structure' of fractal sets and provides an alternative perspective to the compression-based approaches that aim to understand why overparametrized networks do not overfit [AGNZ18, SAN20, SAM+20, HJTW21]. However, SDE approximations for SGD often serve as mere heuristics, and guaranteeing a good approximation typically requires unrealistically small step-sizes [LTE19]. For more realistic step-sizes, theoretical concerns have been raised about the validity of conventional SDE approximations for SGD [LMA21, GŞZ21, Yai19]. Another drawback of the SDE approximation is that the bounds in [ŞSDE20] are implicit, in the sense that they cannot be related to algorithm hyperparameters, problem geometry, or data.

We address these issues and present a direct, *discrete-time* analysis by exploiting the connections between IFSs and stochastic optimization algorithms. Our contributions are summarized as follows:

- We extend [ŞSDE20] and show that the generalization error can be linked to the Hausdorff dimension of *invariant measures* (rather than the Hausdorff dimension of *sets* as in [ŞSDE20]). More precisely, under appropriate conditions, we establish a generalization bound for the stationary distribution of IFS $w_\infty \sim \mu$. That is, with probability at least $1 - 2\zeta$,

$$|\hat{\mathcal{R}}(w_\infty, \mathbf{S}_n) - \mathcal{R}(w_\infty)| \lesssim \sqrt{\frac{\overline{\dim}_{\mathrm{H}}\mu \, \log^2(n)}{n} + \frac{\log(1/\zeta)}{n}}, \tag{4}$$

  for n large enough, where $\overline{\dim}_{\mathrm{H}}\mu$ is the (upper) Hausdorff dimension of the measure μ.
- By leveraging results from IFS theory, we further link $\overline{\dim}_{\mathrm{H}}\mu$ to (i) the form of the recursion (e.g., h_{Ω_k} in (3)), (ii) algorithm hyperparameters (e.g., η, b), and (iii) problem geometry (e.g., Hessian of $\tilde{\mathcal{R}}_k$), through a single term, which encapsulates all these components and their interaction.
- We establish bounds on $\overline{\dim}_{\mathrm{H}}\mu$ for SGD and preconditioned SGD algorithms, when used to minimize various empirical risk minimization problems such as least squares, logistic regression, support vector machines. In all cases, we explicitly link the generalization performance of the model to the hyperparameters of the underlying training algorithm.
- Finally, we numerically compute key quantities that appear in our generalization bounds, and show empirically that they have a statistically significant correlation with the generalization error.

Notation and preliminaries. $B_d(x, r) \subset \mathbb{R}^d$ denotes the closed ball centered around $x \in \mathbb{R}^d$ with radius r. A function $f : \mathbb{R}^{d_1} \to \mathbb{R}^{d_2}$ is said to be (Fréchet) differentiable at $x \in \mathbb{R}^{d_1}$ if there exists a $d_1 \times d_2$ matrix $J_f(x) : \mathbb{R}^{d_1} \to \mathbb{R}^{d_2}$ such that $\lim_{\|h\|\to 0} \|f(x+h) - f(x) - J_f(x)h\|/\|h\| = 0$. The matrix $J_f(x)$ is called the differential of f, also known as the Jacobian matrix at x, and determinant of $J_f(x)$ is called the Jacobian determinant [HS74]. For real-valued functions f, g, we define $f(n) = \omega(g(n))$ if $\lim_{n\to\infty} |f(n)|/g(n) = \infty$. For a set A, $|A|$ denotes its cardinality. For a scalar-valued function $\tilde{f} : \mathbb{R} \to \mathbb{R}$, we define $\|\tilde{f}\|_\infty = \max_{r\in\mathbb{R}} |\tilde{f}(r)|$.

[1] The Hausdorff dimension of the middle-third Cantor set in Figure 1 is $\log_3(2) \approx 0.63$ whereas the ambient dimension is 1 [Fal04, Example 2.3].

2 Technical Background on Fractal Geometry

Fractal sets emerge virtually in all branches of science, and fractal-based techniques have been used in machine learning [SHTY13, MSS19, DSD20, ŞSDE20, AGZ21]. The inherent 'complexity' of a fractal set often plays an important role and it is typically measured by its *fractal dimension*, where several notions of dimension have been proposed [Fal04]. In this section, we briefly mention two important notions of fractal dimension, which will be used in our theoretical development.

Minkowski dimension of a set. The Minkowski dimension (also known as the box-counting dimension [Fal04]) is defined as follows. Let $F \subset \mathbb{R}^d$ be a set and for $\delta > 0$, let $N_\delta(F)$ denote a collection of sets that contains the smallest number of closed balls of diameter at most δ which cover F. Then the upper-Minkowski dimension of F is defined as follows:

$$\overline{\dim}_{\mathrm{M}} F := \limsup_{\delta \to 0} \Big[\log |N_\delta(F)| \, / \, \log(1/\delta) \Big]. \tag{5}$$

To visualize the upper-Minkowski dimension of a set F, consider the set F lying on an evenly spaced grid and count how many boxes are required to cover the set. The upper-Minkowski dimension measures how this number changes as the grid is made finer using a box-counting algorithm.

Hausdorff dimension of a set. An alternative to the purely geometric Minkowski dimension, the Hausdorff dimension [Hau18] is a measure theoretical notion of fractal dimension. It is based on the *Hausdorff measure*, which generalizes the traditional notions of area and volume to non-integer dimensions [Rog98]. More precisely, for $s \geq 0$, let $F \subset \mathbb{R}^d$ and $\delta > 0$, and denote $\mathcal{H}^s_\delta(F) := \inf \sum_{i=1}^\infty \mathrm{diam}(A_i)^s$, where the infimum is taken over all the δ-coverings $\{A_i\}_i$ of F, that is, $F \subset \cup_i A_i$ with $\mathrm{diam}(A_i) < \delta$ for every i. The s-dimensional Hausdorff measure of F is defined as the monotone limit $\mathcal{H}^s(F) := \lim_{\delta \to 0} \mathcal{H}^s_\delta(F)$. When $s \in \mathbb{N}$, $\mathcal{H}^s$ is the s-dimensional Lebesgue measure up to a constant factor; hence the generalization of 'volume' to fractional orders.

Based on the Hausdorff measure, the *Hausdorff dimension* of a set $F \subset \mathbb{R}^d$ is then defined as follows:

$$\dim_{\mathrm{H}} F := \sup \{s > 0 : \mathcal{H}^s(F) > 0\} = \inf \{s > 0 : \mathcal{H}^s(F) < \infty\}.$$

In other words, the Hausdorff dimension of F is the 'moment' s when $\mathcal{H}^s(F)$ drops from ∞ to 0, that is, $\mathcal{H}^r(F) = 0$ for all $r > \dim_{\mathrm{H}} F$ and $\mathcal{H}^r(F) = \infty$ for all $r < \dim_{\mathrm{H}} F$.

We always have $0 \leq \dim_{\mathrm{H}} F \leq d$, and when F is bounded, we always have $0 \leq \dim_{\mathrm{H}} F \leq \overline{\dim}_{\mathrm{M}} F \leq d$ [Fal04]. Furthermore, the Hausdorff dimension of $\mathbb{R}^d$ equals d, and the Hausdorff dimension of smooth Riemannian manifolds correspond to their intrinsic dimension, e.g. $\dim_{\mathrm{H}} \mathbb{S}^{d-1} = d-1$, where $\mathbb{S}^{d-1}$ is the unit sphere in $\mathbb{R}^d$.

Hausdorff dimension of a probability measure. IFSs generate invariant measures as the number of iterates goes to infinity, and random fractals arise from such invariant measures. There has been a growing literature that studies the structure of such random fractals [Saz00, NSB02, MS02, Ram06, FST06, JR08], where the notion of fractal dimension can be extended to measures, and our theory will rely on the Hausdorff dimension of invariant measures associated with stochastic optimization algorithms. In particular, we will mainly use the *upper Hausdorff dimension* $\overline{\dim}_{\mathrm{H}} \mu$ of a Borel probability measure μ on $\mathbb{R}^d$, which is defined as follows: $\overline{\dim}_{\mathrm{H}} \mu := \inf \{\dim_{\mathrm{H}} A : \mu(A) = 1\}$. In other words, $\overline{\dim}_{\mathrm{H}} \mu$ is the smallest Hausdorff dimension of all measurable sets with full measure.

3 Generalization Bounds for Stochastic Optimization Algorithms as IFSs

In this section, we will present our main theoretical results which relate the generalization error to the upper-Hausdorff dimension of the invariant measure associated with a stochastic optimization algorithm. We consider a standard supervised learning setting, where $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is the space of features and $\mathcal{Y}$ is the space of labels, and π is the unknown data distribution on $\mathcal{Z}$.

For mathematical convenience, in order to construct the training set with n elements, we first consider an *infinite sequence* of i.i.d. data samples from the data distribution π, then we take the first n elements from this infinite sequence. More precisely, we consider the (countable) product measure $\pi^\infty = (\pi \otimes \pi \otimes \dots)$ defined on the cylindrical sigma-algebra. Then, we consider the infinite i.i.d. data sequence as $\mathbf{S} \sim \pi^\infty$, i.e., $\mathbf{S} = (z_j)_{j \geq 1}$ with $z_j \overset{\text{i.i.d.}}{\sim} \pi$ for all $j = 1, 2, \dots$. Finally, we define the training set as $\mathbf{S}_n := (z_1, \dots, z_n)$, i.e., we take the first n elements of $\mathbf{S}$. To avoid technical complications, throughout the paper we will assume that all the encountered functions and sets are measurable. All the proofs are given in the supplement.

Given a dataset $\mathbf{S}_n$, we represent the *training algorithm* as an IFS, which is based on the following recursion: $w_k = h_{\Omega_k}(w_{k-1}; \mathbf{S}_n)$, where the mini-batch Ω_k is i.i.d. sampled according to some distribution (e.g., sampling without-replacement uniformly among all possible mini-batches). This compact representation enables us to cover a broad range of optimization algorithms with a unified notation, including SGD (see (3)), as well as preconditioned SGD, and stochastic Newton methods. For example, if we take $h_{\Omega_k}(w; \mathbf{S}_n) = w - \eta H_k(w)^{-1} \nabla \tilde{\mathcal{R}}_k(w)$ where $H_k(w)$ is an estimate of the Hessian of $\tilde{\mathcal{R}}_k$, we cover stochastic Newton methods [EM15]. Similar constructions can be made for other popular algorithms, such as SGD-momentum [Qia99], RMSProp [HSS12], or Adam [KB15].

Notice that there are only finitely many values that Ω_k can take. For example, in the case of without-replacement mini-batch sampling with batch-size b, there are in total $m_b = m_b^{\text{wo-replacement}} := \binom{n}{b}$ many subsets of $\{1, 2, \dots, n\}$ with cardinality b. Alternatively, another setup would be to divide the dataset into $m_b = m_b^{\text{batch}} := n/b$ batches with each batch having b elements, and at each iteration k, we can randomly choose one of the batches. In both examples we can enumerate as $S_1, S_2, \dots, S_{m_b}$. If the probability of sampling the mini-batch $\Omega_k = S_i$ is p_i for every i, then, with a slight abuse of notation, we can rewrite the IFS recursion as:

$$w_k = h_{U_k}(w_{k-1}; \mathbf{S}_n), \tag{6}$$

where U_k is a random variable taking values in $\{1, 2, \dots, m_b\}$ and $p_i := \mathbb{P}(U_k = i)$. If the mini-batch sampling is uniform (i.e., the default option in practice), we have $p_i = 1/m_b$; however, we are not restricted to this option, the sampling scheme is allowed to be more general. We finally call the triple $(\mathbb{R}^d, \{h_i(\cdot; \mathbf{S}_n)\}_{i=1}^{m_b}, \{p_i\}_{i=1}^{m_b})$ an iterated function system (IFS).

Given a dataset $\mathbf{S}_n$, we are interested in the limiting behavior of the training algorithm (6). We characterize this behavior by considering the invariant measure $\mu_{W|\mathbf{S}_n}$ of the IFS (also called stationary distribution), that is a Borel probability measure on $\mathbb{R}^d$, such that $w_\infty \sim \mu_{W|\mathbf{S}_n}$. To be able to work in this context, we first need to ensure that the recursion (6) admits an invariant measure, i.e., $\mu_{W|\mathbf{S}_n}$ exists. Accordingly, we require the following mild conditions on the IFS (6). Let U be a random variable with the same distribution as U_k. If the recursion (6) is *Lipschitz on average*, i.e.,

$$\mathbb{E}[L_U \mid \mathbf{S}_n] < \infty, \quad \text{with} \quad L_U := \sup_{x,y \in \mathbb{R}^d} \frac{\|h_U(x; \mathbf{S}_n) - h_U(y; \mathbf{S}_n)\|}{\|x - y\|}, \tag{7}$$

and is *contractive on average*, i.e., if

$$\mathbb{E}\left[\log(L_U) \mid \mathbf{S}_n\right] = \sum\nolimits_{i=1}^{m_b} p_i \log(L_i) < 0, \qquad \text{with} \qquad p_i > 0, \text{for any } i = 1, \dots, m_b, \tag{8}$$

then it can be shown that the process is ergodic and admits a unique invariant measure where the limit $\rho := \lim_{k\to\infty} (1/k) \log \left\| h_{U_k} h_{U_{k-1}} \cdots h_{U_1} \right\|$ exists almost surely and is a constant [Elt90], where ρ is called the *Lyapunov exponent*. Furthermore, under further technical assumptions, it can be shown that (6) is geometrically ergodic [DF99]. We note that this condition for the existence of the invariant measure is only applicable to stochastic optimization algorithms with a constant stepsize in which case the random map h_U is not time-varying. If decaying stepsize is used instead, then the limit may degenerate to be a singleton. For example, in the toy example illustrated in Figure 1 with quadratics in dimension one, if we use SGD with decaying stepsize $\eta_k = c/k$ where the positive constant c is small enough, then the limit of the iterates will be a singleton as the iterates will converge to the global minimum of the optimization objective (see e.g. [GOP21, GOP19]).

Our goal will be to relate the generalization error to $\overline{\dim}_{\mathrm{H}} \mu_{W|\mathbf{S}_n}$. To achieve this goal, at first sight, it might seem tempting to extract a full-measure set A by using the definition of $\mu_{W|\mathbf{S}_n}$, such that $\mu_{W|\mathbf{S}_n}(A) = 1$ and $\dim_{\mathrm{H}} A \approx \overline{\dim}_{\mathrm{H}} \mu_{W|\mathbf{S}_n}$, and then directly invoke the results from [ŞSDE20], which would link the generalization error to $\dim_{\mathrm{H}} A$, hence, also to $\overline{\dim}_{\mathrm{H}} \mu_{W|\mathbf{S}_n}$. However, since [ŞSDE20] does not consider an IFS framework, the conditions they require (e.g., boundedness of A, $\dim_{\mathrm{M}} A = \dim_{\mathrm{H}} A$) are not suited to IFSs, and hence prevent us from directly using their results.

As a remedy, we make a detour and show that we can find *almost full-measure* sets A, such that $\mu_{W|\mathbf{S}_n}(A) \approx 1$ and $\overline{\dim}_{\mathrm{M}} A \approx \overline{\dim}_{\mathrm{H}} \mu_{W|\mathbf{S}_n}$ (notice that in this case we directly use the Minkowski dimension of A, as opposed to its Hausdorff dimension). To achieve this goal, we require the following geometric regularity condition on the invariant measure.

H 1. *For π^∞-almost all $\mathbf{S}$ and all $n \in \mathbb{N}_+$, the recursion (6) satisfies (7) and (8) and the limit $\lim_{r\to 0}\Big[\log \mu_{W|\mathbf{S}_n}(B_d(w,r)) \;/\; \log r\Big]$ exists for $\mu_{W|\mathbf{S}_n}$-almost every w.*

This is a common condition [Pes08] and is satisfied for a large class of measures. For instance, 'sufficiently regular' measures with the property that $C_1 r^s \leq \mu(B(x,r)) \leq C_2 r^s$ for some constant $s > 0$ and positive constants C_1, C_2 will satisfy this assumption. Such measures are called Ahlfors-regular (cf. [ŞSDE20, Assumption H4] for a related condition), and it is known that IFSs that satisfy certain 'open set conditions' lead to Ahlfors regular invariant measures (see [MT10, Section 8.3]). Yet, our assumption is more general and does not immediately require Ahlfors-regularity.

Under **H**1, we now formalize our key observation, which serves as the basis for our bounds.

Proposition 1. *Assume that **H**1 holds. Then for every $\varepsilon > 0$, $n \in \mathbb{N}_+$, and π^∞-almost every $\mathbf{S}$, there exists $\delta := \delta(\varepsilon, \mathbf{S}_n) \in (0,1]$ and a bounded measurable set $A_{\mathbf{S}_n,\delta} \subset \mathbb{R}^d$, such that*

$$\mu_{W|\mathbf{S}_n}(A_{\mathbf{S}_n,\delta}) \geq 1 - \delta, \quad \textit{and} \quad \overline{\dim}_{\mathrm{M}} A_{\mathbf{S}_n,\delta} \leq \overline{\dim}_{\mathrm{H}} \mu_{W|\mathbf{S}_n} + \varepsilon, \tag{9}$$

and $\delta(\varepsilon, \mathbf{S}_n) \to 0$ as $\varepsilon \to 0$.

Thanks to this result, we can now leverage the proof technique presented in [ŞSDE20, Theorem 2], and link the generalization error to $\overline{\dim}_{\mathrm{H}}\mu_{W|\mathbf{S}_n}$ through $\overline{\dim}_{\mathrm{M}} A_{\mathbf{S}_n,\delta}$. We shall emphasize that, mainly due to the sets $A_{\mathbf{S}_n,\delta}$ not being of full-measure, our framework introduces additional non-trivial technical difficulties that we need to tackle in our proof.

We now introduce our second assumption, which roughly corresponds to a 'topological stability' condition, and is adapted from [ŞSDE20, Assumption H5]. Formally, consider the (countably infinite) collection of closed balls of radius β, whose centers are on the fixed grid $N_\beta := \Big\{\Big(\frac{(2j_1+1)\beta}{2\sqrt{d}}, \ldots, \frac{(2j_d+1)\beta}{2\sqrt{d}}\Big) : j_i \in \mathbb{Z}, i = 1, \ldots, d\Big\}$, and for a set $A \subset \mathbb{R}^d$, define $N_\beta(A) := \{x \in N_\beta : B_d(x,\beta) \cap A \neq \emptyset\}$, which is the collection of the centers of the balls that intersect A.

H 2. *Let $\mathcal{Z}^\infty := (\mathcal{Z} \times \mathcal{Z} \times \cdots)$ denote the countable product endowed with the product topology and let $\mathfrak{B}$ be the Borel σ-algebra generated by $\mathcal{Z}^\infty$. For a Borel set $A \subset \mathbb{R}^d$, let $\mathfrak{F}, \mathfrak{G}$ be the sub-σ-algebras of $\mathfrak{B}$ generated by the collections of random variables given by $\{\hat{\mathcal{R}}(w, \mathbf{S}_n) : w \in \mathbb{R}^d, n \geq 1\}$ and $\Big\{\mathbb{1}\left\{w \in N_\beta(A_{\mathbf{S}_n,\delta})\right\}, \mu_{W|\mathbf{S}_n}(A_{\mathbf{S}_n,\delta}), \overline{\dim}_{\mathrm{H}}\mu_{W|\mathbf{S}_n} : \delta, \beta \in \mathbb{Q}_{>0}, w \in N_\beta, n \geq 1\Big\}$, where $A_{\mathbf{S}_n,\delta}$ is given in Proposition 1. There exists a constant $M \geq 1$ such that for any $F \in \mathfrak{F}$, $G \in \mathfrak{G}$, we have $\mathbb{P}(F \cap G) \leq M\mathbb{P}(F)\mathbb{P}(G)$.*

H2 simply ensures that the dependence between the training error and the topological properties of the support of $\mu_{W|\mathbf{S}_n}$ can be controlled via M. Hence, it can be seen as a form of *algorithmic stability* [BE02], where M measures the level of stability of the topology of $\mu_{W|\mathbf{S}_n}$: a small M indicates that the geometrical structure of $\mu_{W|\mathbf{S}_n}$ does not heavily depend on the particular value of $\mathbf{S}_n$. The constant M is also related to the mutual information [XR17, AAV18, HŞKM21], but may be better behaved than the mutual information as it relies on very specific functions of the random variables. Similar to mutual information, a-priori there is no reason to expect M to be finite for general algorithms; intuitively, however, the more stochasticity an algorithm incorporates the more we expect the set $A_{\mathbf{S}_n,\delta}$ and the loss landscape to decouple. For example, for a purely random algorithm (which ignores $\mathbf{S}_n$) the two objects will be independent. In the other extreme, where the algorithm is deterministic given the sample may fail to be finite. Since we are controlling the generalization error on the support, which is itself a random set depending on the sample, we require **H**2 to be able to make progress.

We require one final assumption, which states that the loss ℓ is sub-exponential.

H 3. *ℓ is L-Lipschitz continuous in w, and when $z \sim \pi$, for all w, $\ell(w,z)$ is (ν,κ)-sub-exponential, that is, for all $|\lambda| < 1/\kappa$, we have $\log \mathbb{E}_{z\sim\pi}\left[\exp\left(\lambda\left(\ell(w,z) - \mathcal{R}(w)\right)\right)\right] \leq \nu^2\lambda^2/\kappa$.*

Armed with these assumptions, we can now present our main result.

Theorem 1. *Assume that **H**1 to 3 hold and $\overline{\dim}_{\mathrm{H}}\mu_{W|\mathbf{S}_n} = \omega(\log\log(n)/\log(n))$, π^∞-almost-surely. Then, the following bound holds for sufficiently large n:*

$$|\hat{\mathcal{R}}(W, \mathbf{S}_n) - \mathcal{R}(W)| \leq 8\nu\sqrt{\frac{\overline{\dim}_{\mathrm{H}}\mu_{W|\mathbf{S}_n}\log^2(nL^2)}{n} + \frac{\log(13M/\zeta)}{n}}, \tag{10}$$

with probability at least $1 - 2\zeta$ *over the joint distribution of* $\mathbf{S} \sim \pi^\infty$, $W \sim \mu_{W|\mathbf{S}_n}$.

This theorem shows that the Hausdorff dimension of the invariant measure acts as a 'capacity metric' and the generalization error is therefore directly linked to this metric, i.e., the complexity of the underlying fractal structure has close links to the generalization performance. On the other hand, the condition $\overline{\dim}_{\mathrm{H}} \mu_{W|\mathbf{S}_n} = \omega(\log\log(n)/\log(n))$ is very mild and makes sure that the dimension of the IFS does not decrease very rapidly with increasing number of data points n. We shall mention that Theorem 1 has an asymptotic nature as we do not have an explicit control on how large n should be. This is due to the fact that the notions of Minkowski and Hausdorff dimensions are essentially asymptotic, which unfortunately prevents us from obtaining any truly non-asymptotic result. However, obtaining nonasymptotic results might be possible with further assumptions on the fractal dimension of the invariant measures and their supports.

Theorem 1 enables us to access the rich theory of IFSs, where bounds on the Hausdorff dimension are readily available, and connect them to statistical learning theory. The following result is a direct corollary to Theorem 1 and [Ram06, Theorem 2.1] (see Theorem S2 in the supplementary document).

Corollary 1. *Assume that the conditions of Theorem 1 hold. Furthermore, consider the recursion (6) and assume that* h_i *are continuously differentiable with derivatives* J_{h_i} *that are* α*-Hölder continuous for some* $\alpha > 0$. *Then, there exists a constant* $M > 1$ *such that for sufficiently large* n*:*

$$|\hat{\mathcal{R}}(W, \mathbf{S}_n) - \mathcal{R}(W)| \leq 8\nu \sqrt{\frac{\mathcal{E} \log^2(nL^2)}{\left[\sum_{i=1}^{m_b} p_i \int_{\mathbb{R}^d} \log(\|J_{h_i}(w)\|) \mathrm{d}\mu_{W|\mathbf{S}_n}(w)\right] n} + \frac{\log(13M/\zeta)}{n}}, \quad (11)$$

with probability $1 - 2\zeta$ *over* $\mathbf{S} \sim \pi^\infty$, $W \sim \mu_{W|\mathbf{S}_n}$, *where* $\mathcal{E} := \sum_{i=1}^{m_b} p_i \log(p_i)$ *denotes the negative entropy of the mini-batch sampling scheme,* $\|\cdot\|$ *denotes the operator norm (with the* ℓ_2*-norm being the underlying norm), and* J_{h_i} *is the Jacobian of* h_i *defined in the notation section.*

Note that the Hölder condition is mainly used to ensure that the invariant measure exists and the constant α does not directly interact with the bound. However, it might affect the rate of convergence to the invariant measure.

By this result, we discover an interesting quantity, $\sum_{i=1}^{m_b} p_i \int_{\mathbb{R}^d} \log(\|J_{h_i}(w)\|) \mathrm{d}\mu_{W|\mathbf{S}_n}(w)$, which *simultaneously* captures the effects of the data and the algorithm[2]. To see it more clearly, let us consider the SGD recursion (3), where $\|J_{h_i}(w)\| = \|I - \eta \nabla^2 \tilde{\mathcal{R}}_{S_i}(w)\|$ and $\{S_i\}_{i=1}^{m_b}$ denotes the enumeration of the mini-batches. Then, the overall quantity becomes

$$\mathbb{E}_{U,W}\left[\log \|I - \eta \nabla^2 \tilde{\mathcal{R}}_{S_U}(W)\|\right], \quad (12)$$

where the expectation is taken over the mini-batch index $U \in \{1, \ldots, m_b\}$ with $\mathbb{P}(U = i) = p_i$, and $W \sim \mu_{W|\mathbf{S}_n}$. We clearly observe that this term depends on (i) the algorithm choice through the form of h_i, (ii) step-size η, (iii) batch-size through m_b, (iv) problem geometry through $\nabla^2 \tilde{\mathcal{R}}$, and (v) data distribution through $\mu_{W|\mathbf{S}_n}$. We believe that such a compact representation of all these constituents and their interaction is novel and will open up interesting future directions.

4 Analytical Estimates for the Hausdorff Dimension

The generalization bound presented in Theorem 1 applies to a number of stochastic optimization algorithms that can be represented with an IFS and to a large class of losses that can be non-convex or convex. It is controlled by the Hausdorff dimension of the invariant measure $\mu_{W|\mathbf{S}_n}$ which needs to be estimated. In the numerical experiments section, we will discuss how this quantity can be estimated from the dataset $\mathbf{S}_n$ and the iterates of the underlying algorithm.

Corollary 1 shows that for smooth losses, the Hausdorff dimension can be controlled with the expectation of the norm of the logarithm of the Jacobian $\log(\|J_{h_i}(w)\|)$ with respect to the invariant measure $\mu_{W|\mathbf{S}_n}$. In general, an explicit characterization of the invariant measure is not known. Nevertheless, under additional appropriate assumptions that can hold in practice, such as boundedness of the data of the loss, we next discuss that it is possible to get uniform lower and upper bounds on the quantity $\|J_{h_i}(w)\|$ which leads to analytical upper bounds on $\overline{\dim}_{\mathrm{H}} \mu_{W|\mathbf{S}_n}$.

[2] Note that thanks to [Ram06], we can allow state-dependent $p_i = p_i(w)$; yet, we do not consider this option as its application is not immediately clear.

As illustrative examples; in the following, we will consider the setting where we divide $\mathbf{S}_n$ into $m_b = m_b^{\text{batch}} = n/b$ batches with each batch having b elements, and then we discuss how analytical estimates on the (upper) Hausdorff dimension $\overline{\dim}_{\mathrm{H}}\mu_{W|\mathbf{S}_n}$ can be obtained for some particular problems including least squares, regularized logistic regression, and one hidden-layer networks. In the supplementary document, we also discuss how similar bounds can be obtained for support vector machines and other algorithms such as preconditioned SGD and stochastic Newton methods.

Least squares. We consider the least squares problem, with data points $z_i = (a_i, y_i)$ and loss

$$\ell(w, z_i) := \left(a_i^T w - y_i\right)^2/2 + \lambda\|w\|^2/2, \tag{13}$$

where $\lambda > 0$ is a regularization parameter.

Proposition 2 (Least squares)**.** *Consider the least squares problem* (13)*. Assume the step-size $\eta \in (0, \frac{1}{R^2+\lambda})$, where $R := \max_i \|a_i\|$ is finite. Then, we have the following upper bound:*

$$\overline{\dim}_{\mathrm{H}}\mu_{W|\mathbf{S}_n} \leq \frac{\log(n/b)}{\log(1/(1-\eta\lambda))}. \tag{14}$$

Note that here ℓ is only pseudo-Lipschitz $|\ell(w, z_i) - \ell(w', z_i)| \leq L_i(1 + \|w\| + \|w'\|)\|w - w'\|$ for some $L_i > 0$, rather than globally Lipschitz. However; the conditions in Proposition 2 ensure that w will stay in a bounded region, in which case ℓ becomes Lipschitz. Also note that only the logarithm of the Lipschitz constant directly enters the bound.

We observe that, for fixed n, the upper bound for $\overline{\dim}_{\mathrm{H}}\mu_{W|\mathbf{S}_n}$ is decreasing both in η and b. This behavior is not surprising: large η results in chaotic behaviors (cf. Figures 1,2), and in the extreme case where $b = n$, the algorithm becomes deterministic and hence converges to a single point, in which case the Hausdorff dimension becomes 0. However, the decrement due to b does not automatically grant good generalization performance: since the algorithm becomes deterministic, the stability constant M in **H**2 can get arbitrarily large, hence the bound in Theorem 1 could become vacuous. This outcome reveals an interesting tradeoff between the Hausdorff dimension and the constant M, through the batch-size b, and investigating this tradeoff is an interesting future direction.

We further notice that the numerator in (14) is logarithmically increasing with n, which is compensated by the factor $1/n$ in Theorem 1. Nevertheless, we can take the batch-size in proportion with n (i.e., setting m_b to a constant value), in order avoid this logarithmic growth. We also note that the input dimension d potentially affects the term R, which forms the bound for the input data, and hence the input dimension will indirectly affect the generalization bound. Finally, regarding the remaining bounds in this section, even though their forms might differ from (14), similar remarks also apply. Hence, we will omit the discussion.

Regularized logistic regression. Given the data points $z_i = (a_i, y_i)$, consider regularized logistic regression with the loss:

$$\ell(w, z_i) := \log\left(1 + \exp\left(-y_i a_i^T w\right)\right) + \lambda\|w\|^2/2, \tag{15}$$

where $\lambda > 0$ is a regularization parameter. We have the following result.

Proposition 3 (Regularized logistic regression)**.** *Consider the regularized logistic regression* (15)*. Assume the step-size $\eta < 1/\lambda$ and the input data is bounded, i.e. $R := \max_i \|a_i\| < 2\sqrt{\lambda}$. We have:*

$$\overline{\dim}_{\mathrm{H}}\mu_{W|\mathbf{S}_n} \leq \frac{\log(n/b)}{\log(1/(1-\eta\lambda+\frac{1}{4}\eta R^2))}. \tag{16}$$

Next, we consider a non-convex formulation for robust regression (see e.g. [MBM18]), with $z_i = (a_i, y_i)$ and the loss

$$\ell(w, z_i) := \rho\left(y_i - \langle w, a_i\rangle\right) + \lambda\|w\|^2/2, \tag{17}$$

where $\lambda > 0$ is a regularization parameter and ρ is a non-convex function, assumed to be twice continously differentiable, where a standard choice is *Tukey's bisquare loss* defined as $\rho_{\text{Tukey}}(t) = 1 - (1 - (t/t_0)^2)^3$ for $|t| \leq t_0$, and $\rho_{\text{Tukey}}(t) = 1$ for $|t| \geq t_0$ (see e.g. [MBM18]), and exponential squared loss: $\rho_{\exp}(t) = 1 - e^{-|t|^2/t_0}$, where $t_0 > 0$ is a tuning parameter (see e.g. [WJHZ13]).

Proposition 4 (Non-convex formulation for robust regression)**.** *Consider the non-convex formulation for robust regression* (17)*. Assume the step-size* $\eta < \frac{1}{\lambda + R^2(2/t_0)}$*, where* $R = \max_i \|a_i\| < \sqrt{\lambda t_0/2}$*. Then, we have the following upper bound for the Hausdorff dimension:*

$$\overline{\dim}_{\mathrm{H}} \mu_{W|\mathbf{S}_n} \leq \frac{\log (n/b)}{\log(1/(1 - \eta\lambda + \eta R^2 \frac{2}{t_0}))}. \tag{18}$$

One hidden-layer neural network. Given the data points $z_i = (a_i, y_i)$. Let $a_i \in \mathbb{R}^d$ be the input and $y_i \in \mathbb{R}^m$ be the corresponding output, and let $w_r \in \mathbb{R}^d$ be the weights of the r-th hidden neuron of a one hidden-layer network, and $b_r \in \mathbb{R}$ is the output weight of hidden unit r. For simplicity of the presentation, following [ZMG19, DZPS19], we only optimize the weights of the hidden layer, i.e. $w = \left[w_1^T w_2^T \dots w_m^T\right]$ is the decision variable with the regularized squared loss:

$$\ell(w, z_i) := \|y_i - \hat{y}_i\|^2 + \lambda \|w\|^2/2, \quad \hat{y}_i := \sum\nolimits_{r=1}^{m} b_r \sigma\left(w_r^T a_i\right), \tag{19}$$

where the non-linearity $\sigma : \mathbb{R} \to \mathbb{R}$ is smooth and $\lambda > 0$ is a regularization parameter.

Proposition 5 (One hidden-layer network)**.** *Consider the one hidden-layer network* (19)*. Assume the step-size* $\eta < \frac{1}{2\lambda}$*. Then, we have the following upper bound for the Hausdorff dimension:*

$$\overline{\dim}_{\mathrm{H}} \mu_{W|\mathbf{S}_n} \leq \frac{\log (n/b)}{\log(1/(1 - \eta(\lambda - C)))}, \tag{20}$$

where $C := M_y \|b\|_\infty \|\sigma''\|_\infty R^2 + (\max_j \|v_j\|_\infty)^2 < \lambda$*, where* $R := \max_i \|a_i\|$*,* $M_y := \max_i \|y_i - \hat{y}_i\|$*, and* $v_i := \left[b_1 \sigma'(w_1^T a_i) a_i \; b_2 \sigma'(w_2^T a_i) a_i \; \cdots b_m \sigma'(w_m^T a_i) a_i\right]^T$*.*

In Proposition 4, we assumed that C is uniformly bounded and $\lambda > C$ for mathematical convenience. However, in general, C by its definition might increase in m and n so that we do not expect that we can choose small λ for large data sets and wide networks.

5 Experiments

Our aim now is to empirically investigate whether the bound in Corollary 1 is informative, in that it is predictive of a neural network's generalization error. As the second term of this bound cannot be evaluated, we will assume that it is negligible and focus our efforts on the first term. Further, because the denominator of the first term is the only term that depends on the invariant measure, we want to establish that the inverse of this denominator is predictive of a neural network's generalization error. Note however that for complex models, such as modern neural networks, analytically bounding $\|J_{h_i(x)}\|$ becomes highly non-trivial.

In our experiments, we fix the algorithm to SGD and we develop a numerical method for computing the term (12). Noting that $J_{h_i}(w) = I - \eta \nabla^2 \tilde{\mathcal{R}}_i(w)$, for simplicity we denote the inverse of (12) as the 'complexity': $R = 1/\left[\sum_{i=1}^{m_b} p_i \int_{\mathbb{R}^d} \log\left(\|J_{h_i}(w)\|\right) \mathrm{d}\mu_{W|\mathbf{S}_n}(w)\right]$[3].

To approximate the expectations, we propose the following simple Monte Carlo strategy:

$$R^{-1} \approx \left[1/(N_W N_U)\right] \sum\nolimits_{i=1}^{N_W} \sum\nolimits_{j=1}^{N_U} \log\left(\|J_{h_{U_j}}(W_i)\|\right), \tag{21}$$

where U_j denotes i.i.d. random mini-batch indices that are drawn without-replacement from $\{1, \dots, n\}$ and $W_i \overset{\text{i.i.d.}}{\sim} \mu_{W|S}$. Assuming (8) is ergodic [DF99], we treat the iterates w_k as i.i.d. samples from $\mu_{W|\mathbf{S}_n}$ for large k, hence, $\log(\|J_{h_{U_j}}(W_i)\|)$ can be computed on these iterates, and (21) can be computed accordingly. Our implementation for computing $\|J_{h_{U_j}}(W_i)\|$ for neural networks with millions of parameters is detailed in the supplementary document. Though the size of J_{h_i} is very large in our experiments ($\approx$ 20M $\times$ 20M on average), our algorithm can efficiently compute the norms without constructing $J_{h_{U_j}}(W_i)$, by extending the approach presented in [YGKM20]. Our implementation is available at `https://github.com/umutsimsekli/fractal_generalization`.

[3] Note that the first term of the bound in Corollary 1 suggests computing $\sqrt{R}$ rather than R; however, both choices yield very similar results with high statistical significance.

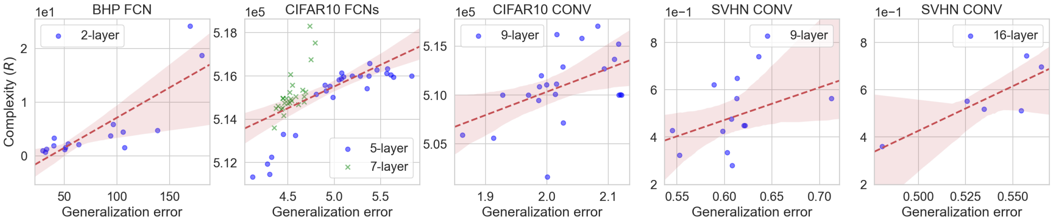

Figure 3: Estimates of R plotted against the generalization error ($|\text{training loss} - \text{test loss}|$) for VGG11 and FCNs trained on CIFAR10, SVHN and BHP with varying η, b. The linear regression of best fit is plotted in red, where shading corresponds to the 95% confidence interval. For all plots the one-sided p-value, testing whether the null hypothesis that the slope of the line is in-fact negative and not positive, is significantly less than 0.001, indicating that it is highly likely that R and the generalization error are positively correlated.

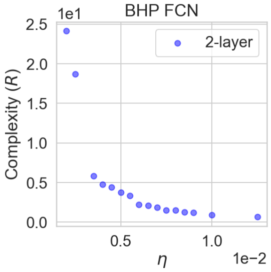

Figure 4: Estimated R plotted against η for 2 layer FCN trained on BHP.

In Figure 3 we plot the estimates of R for a variety of convolutional (CONV) and fully connected network (FCN) architectures trained on CIFAR10, SVHN and Boston House Prices (BHP). For the full details of the models, the hardware used, their run-time, and the convergence criterion used, see Section S7 in the supplement. The plot demonstrates that R and generalization error are positively correlated and that this correlation is significant (p-value $\ll 0.001$) for all model architectures. This provides evidence that the bound on the generalization error in Corollary 1 is informative.

To support our findings in Section 4 that the bound for the Hausdorff dimension $\overline{\dim}_{\mathrm{H}} \mu_{W|\mathbf{S}_n}$ is monotonically decreasing in the step-size η, we plot R against η in Figure 4 for the networks trained on BHP in Figure 3. R decreases with increasing η, clearly backing our findings. We note that these results were inconclusive for classification models trained with a cross-entropy loss, in that we could not clearly observe a negative or positive correlation. Future work will further study this lack of correlation, particular to classification models.

6 Conclusion

In this work, we investigated stochastic optimization algorithms through the lens of IFSs and studied their generalization properties. Under some assumptions, we showed that the generalization error can be controlled based on the Hausdorff dimension of the invariant measure determined by the iterations, which can lead to tighter bounds than the standard bounds based on the ambient dimension. We proposed an efficient methodology to estimate the Hausdorff dimension in deep learning settings and supported our theory with several experiments on neural networks. We also illustrated our bounds on specific problems and algorithms such as SGD and its preconditioned variants, which unveil new links between generalization, algorithm parameters and the Hessian of the loss.

Our study does not have a direct societal impact as it is largely theoretical. The limitation of our study is its asymptotic nature due to operating on invariant measures. Future work will address (i) obtaining nonasymptotic bounds in terms of the number of iterations k, (ii) including the term of (12) as a regularizer to the optimization problem, which would be an alternative to the methods that aim to "decrease the intrinsic dimension" [ZQH+18, BLGŞ21].

Acknowledgements. MAE is partially funded by CIFAR AI Chairs program, and CIFAR AI Catalyst grant, NSERC Grant [2019-06167]. MG's research is supported in part by the grants Office of Naval Research Award Number N00014-21-1-2244, National Science Foundation (NSF) CCF-1814888, NSF DMS-2053485, NSF DMS-1723085. UŞ's research is 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). LZ is grateful to the partial support from NSF DMS-2053454 and a Simons Foundation Collaboration Grant.

References

[AAV18] Amir Asadi, Emmanuel Abbe, and Sergio Verdú. Chaining mutual information and tightening generalization bounds. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 7234–7243, 2018. (Cited on page 6.)

[AB09] Martin Anthony and Peter L Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 2009. (Cited on page 2.)

[AGNZ18] Sanjeev Arora, Rong Ge, Behnam Neyshabur, and Yi Zhang. Stronger generalization bounds for deep nets via a compression approach. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 254–263. PMLR, 10–15 Jul 2018. (Cited on page 3.)

[AGZ21] Naman Agarwal, Surbhi Goel, and Cyril Zhang. Acceleration via fractal learning rate schedules. In *International Conference on Machine Learning*, 2021. (Cited on page 4.)

[Anc16] Andreas Anckar. Dimension bounds for invariant measures of bi-Lipschitz iterated function systems. *Journal of Mathematical Analysis and Applications*, 440(2):853–864, 2016. (Cited on pages 12 and 15.)

[BE02] Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2(Mar):499–526, 2002. (Cited on page 6.)

[BLGŞ21] Tolga Birdal, Aaron Lou, Leonidas Guibas, and Umut Şimşekli. Intrinsic dimension, persistent homology and generalization in neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. (Cited on page 10.)

[Bog07] Vladimir I Bogachev. *Measure Theory*, volume 1. Springer, 2007. (Cited on page 2.)

[DDB20] Aymeric Dieuleveut, Alain Durmus, and Francis Bach. Bridging the gap between constant step size stochastic gradient descent and Markov chains. *Annals of Statistics*, 48(3):1348–1382, 2020. (Cited on page 2.)

[DF99] Persi Diaconis and David Freedman. Iterated random functions. *SIAM Review*, 41(1):45–76, 1999. (Cited on pages 2, 5, and 9.)

[DR17] Gintare Karolina Dziugaite and Daniel M Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence (UAI)*, 2017. (Cited on page 2.)

[DSD20] Nadav Dym, Barak Sober, and Ingrid Daubechies. Expression of fractals through neural network functions. *IEEE Journal on Selected Areas in Information Theory*, 1(1):57–66, 2020. (Cited on page 4.)

[DZPS19] Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2019. (Cited on page 9.)

[Elt90] John H Elton. A multiplicative ergodic theorem for Lipschitz maps. *Stochastic Processes and their Applications*, 34(1):39–47, 1990. (Cited on page 5.)

[EM15] Murat A Erdogdu and Andrea Montanari. Convergence rates of sub-sampled Newton methods. *Advances in Neural Information Processing Systems*, 28:3052–3060, 2015. (Cited on page 5.)

[Fal97] Kenneth J. Falconer. *Techniques in Fractal Geometry*. Wiley, 1997. (Cited on page 2.)

[Fal04] Kenneth Falconer. *Fractal Geometry: Mathematical Foundations and Applications*. Wiley, 2004. (Cited on pages 2, 3, and 4.)

[FST06] Ai Hua Fan, Károly Simon, and Hajnal R. Tóth. Contracting on average random IFS with repelling fixed point. *Journal of Statistical Physics*, 122:169–193, 2006. (Cited on page 4.)

[FW09] De-Jun Feng and Yang Wang. On the structures of generating iterated function systems of Cantor sets. *Advances in Mathematics*, 222(6):1964–1981, 2009. (Cited on page 2.)

[GOP19] Mert Gürbüzbalaban, Asuman Ozdaglar, and Pablo A Parrilo. Convergence rate of incremental gradient and incremental Newton methods. *SIAM Journal on Optimization*, 29(4):2542–2565, 2019. (Cited on page 5.)

[GOP21] Mert Gürbüzbalaban, Asu Ozdaglar, and Pablo A Parrilo. Why random reshuffling beats stochastic gradient descent. *Mathematical Programming*, 186(1):49–84, 2021. (Cited on page 5.)

[GŞZ21] Mert Gürbüzbalaban, Umut Şimşekli, and Lingjiong Zhu. The heavy-tail phenomenon in SGD. In *International Conference on Machine Learning*, 2021. (Cited on page 3.)

[Hau18] Felix Hausdorff. Dimension und äusseres Mass. *Mathematische Annalen*, 79(1-2):157–179, 1918. (Cited on page 4.)

[HJTW21] Daniel Hsu, Ziwei Ji, Matus Telgarsky, and Lan Wang. Generalization bounds via distillation. In *International Conference on Learning Representations*, 2021. (Cited on page 3.)

[HS74] Morris W. Hirsch and Stephen Smale. *Differential Equations, Dynamical Systems, and Linear Algebra*. Academic Press, 1974. (Cited on page 3.)

[HŞKM21] Liam Hodgkinson, Umut Şimşekli, Rajiv Khanna, and Michael W Mahoney. Generalization properties of stochastic optimizers via trajectory analysis. *arXiv preprint arXiv:2108.00781*, 2021. (Cited on page 6.)

[HSS12] Geoffrey Hinton, Nitish Srivastava, and Kevin Swersky. Overview of mini-batch gradient descent. *Neural Networks for Machine Learning*, 575, 2012. (Cited on page 5.)

[JR08] Joanna Jaroszewska and Michał Rams. On the Hausdorff dimension of invariant measures of weakly contracting on average measurable IFS. *Journal of Statistical Physics*, 132:907, 2008. (Cited on page 4.)

[KB15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. (Cited on page 5.)

[Li17] Xi-Lin Li. Preconditioned stochastic gradient descent. *IEEE Transactions on Neural Networks and Learning Systems*, 29(5):1454–1466, 2017. (Cited on page 2.)

[LMA21] Zhiyuan Li, Sadhika Malladi, and Sanjeev Arora. On the validity of modeling SGD with stochastic differential equations (SDEs). *arXiv preprint arXiv:2102.12470*, 2021. (Cited on page 3.)

[LTE19] Qianxiao Li, Cheng Tai, and Weinan E. Stochastic modified equations and dynamics of stochastic gradient algorithms i: Mathematical foundations. *Journal of Machine Learning Research*, 20(40):1–47, 2019. (Cited on page 3.)

[MBM18] Song Mei, Yu Bai, and Andrea Montanari. The landscape of empirical risk for nonconvex losses. *Annals of Statistics*, 46(6A):2747–2774, 2018. (Cited on page 8.)

[MS02] Józef Myjak and Tomasz Szarek. On Hausdorff dimension of invariant measures arising from non-contractive iterated function systems. *Annali di Matematica Pura ed Applicata*, 181:223–237, 2002. (Cited on page 4.)

[MSS19] Eran Malach and Shai Shalev-Shwartz. Is deeper better only when shallow is good? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019. (Cited on page 4.)

[MT10] John M Mackay and Jeremy T Tyson. *Conformal Dimension: Theory and Application*, volume 54. American Mathematical Society, 2010. (Cited on page 6.)

[NBMS17] Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring generalization in deep learning. In *Advances in Neural Information Processing Systems (NIPS)*, pages 5947–5956, 2017. (Cited on page 2.)

[NcGR19] Thanh Huy Nguyen, Umut Şimşekli, Mert Gürbüzbalaban, and Gaël Richard. First exit time analysis of stochastic gradient descent under heavy-tailed gradient noise. In *Advances in Neural Information Processing Systems*, pages 273–283, 2019. (Cited on page 3.)

[NSB02] Matthew Nicol, Nikita Sidorov, and David Broomhead. On the fine structure of stationary measures in systems which contract-on-average. *Journal of Theoretical Probability*, 15:715–730, 2002. (Cited on page 4.)

[Pes08] Yakov B Pesin. *Dimension Theory in Dynamical Systems: Contemporary Views and Applications*. University of Chicago Press, 2008. (Cited on pages 6, 1, and 2.)

[Qia99] Ning Qian. On the momentum term in gradient descent learning algorithms. *Neural Networks*, 12(1):145–151, 1999. (Cited on page 5.)

[Ram06] Michał Rams. Dimension estimates for invariant measures of contracting-on-average iterated function systems. *arXiv preprint math/0606420*, 2006. (Cited on pages 4, 7, and 2.)

[RKM16] Farbod Roosta-Khorasani and Michael W Mahoney. Sub-sampled Newton methods II: Local convergence rates. *arXiv preprint arXiv:1601.04738*, 2016. (Cited on page 4.)

[Rog98] Claude Ambrose Rogers. *Hausdorff Measures*. Cambridge University Press, 1998. (Cited on page 4.)

[SAM$^+$20] 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. In *International Joint Conference on Artificial Intelligence*, pages 2839–2846, 2020. (Cited on page 3.)

[SAN20] Taiji Suzuki, Hiroshi Abe, and Tomoaki Nishimura. Compression based bound for non-compressed network: unified generalization error analysis of large compressible deep neural network. In *International Conference on Learning Representations*, 2020. (Cited on page 3.)

[Saz00] Tomasz Sazarek. The dimension of self-similar measures. *Bulletin of the Polish Academy of Sciences. Mathematics*, 48:293–202, 2000. (Cited on page 4.)

[ŞGN$^+$19] Umut Şimşekli, Mert Gürbüzbalaban, Thanh Huy Nguyen, Gaël Richard, and Levent Sagun. On the heavy-tailed theory of stochastic gradient descent for deep neural networks. *arXiv preprint arXiv:1912.00018*, 2019. (Cited on page 3.)

[SHTY13] Mahito Sugiyama, Eiju Hirowatari, Hideki Tsuiki, and Akihiro Yamamoto. Learning figures with the Hausdorff metric by fractals—towards computable binary classification. *Machine Learning*, 90(1):91–126, 2013. (Cited on page 4.)

[SSBD14] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014. (Cited on page 2.)

[ŞSDE20] Umut Şimşekli, Ozan Sener, George Deligiannidis, and Murat A Erdogdu. Hausdorff dimension, heavy tails, and generalization in neural networks. In *Advances in Neural Information Processing Systems*, volume 33, 2020. (Cited on pages 3, 4, 5, 6, and 8.)

[ŞSG19] Umut Şimşekli, Levent Sagun, and Mert Gürbüzbalaban. A tail-index analysis of stochastic gradient noise in deep neural networks. In *International Conference on Machine Learning*, pages 5827–5837, 2019. (Cited on page 3.)

[ŞZTG20] Umut Şimşekli, Lingjiong Zhu, Yee Whye Teh, and Mert Gürbüzbalaban. Fractional underdamped Langevin dynamics: Retargeting SGD with momentum under heavy-tailed gradient noise. *arXiv preprint arXiv:2002.05685*, 2020. (Cited on page 3.)

[Ver18] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*, volume 47. Cambridge university press, 2018. (Cited on page 5.)

[Wai19] Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, volume 48. Cambridge University Press, 2019. (Cited on page 10.)

[WJHZ13] Xueqin Wang, Yunlu Jiang, Mian Huang, and Heping Zhang. Robust variable selection with exponential squared loss. *Journal of the American Statistical Association*, 108(502):632–643, 2013. (Cited on page 8.)

[XR17] Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2524–2533, 2017. (Cited on page 6.)

[Yai19] Sho Yaida. Fluctuation-dissipation relations for stochastic gradient descent. In *International Conference on Learning Representations*, 2019. (Cited on page 3.)

[YBVE20] Lu Yu, Krishnakumar Balasubramanian, Stanislav Volgushev, and Murat A Erdogdu. An analysis of constant step size SGD in the non-convex regime: Asymptotic normality and bias. *arXiv preprint arXiv:2006.07904*, 2020. (Cited on page 2.)

[YGKM20] Zhewei Yao, Amir Gholami, Kurt Keutzer, and Michael W. Mahoney. PyHessian: Neural networks through the lens of the Hessian. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 581–590, 2020. (Cited on pages 9 and 5.)

[ZBH$^+$17] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017. (Cited on page 2.)

[ZMG19] Guodong Zhang, James Martens, and Roger Grosse. Fast convergence of natural gradient descent for overparameterized neural networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019. (Cited on pages 9 and 3.)

[ZQH$^+$18] Wei Zhu, Qiang Qiu, Jiaji Huang, Robert Calderbank, Guillermo Sapiro, and Ingrid Daubechies. LDMNet: Low dimensional manifold regularized neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2743–2751, 2018. (Cited on page 10.)

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See the Least Squares part of Section 4, the end of the Section 5 (Experiments), and Conclusion.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Conclusion.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See supplementary material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section S7 in the Supplementary Document
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Section S7 in the Supplementary Document
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] The PyHessian paper, with the only external codebase we used, is cited.
   (b) Did you mention the license of the assets? [N/A] All code used was released in an open-source manner using an MIT license.
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Fractal Structure and Generalization Properties of Stochastic Optimization Algorithms

SUPPLEMENTARY DOCUMENT

This document provides additional material for the NeurIPS 2021 submission entitled *"Fractal Structure and Generalization Properties of Stochastic Optimization Algorithms"*. The document is organized as follows:

- Technical background for the proofs.
  - In Section S1, we provide additional background on dimension theory. In particular we define the Minkowski dimension and the local dimension for a *measure*. Then, we provide three existing theoretical results that will be used in our proofs.
- Additional theoretical results.
  - In Section S2, we provide an upper-bound on the Hausdorff dimension of the invariant measure of SGD, when applied on support vector machines. This result is a continuation of the results given in Section 4.
  - In Section S3, we provide upper-bounds on the Hausdorff dimension of the invariant measure of *preconditioned* SGD on different problems.
  - In Section S4, we provide an upper-bound on the Hausdorff dimension of the invariant measure of the *stochastic Newton algorithm* applied on linear regression.
  - In Section S5, we illustrate the conditions (7) and (8) on a simple setup.
- Details of the experimental results.
  - In Section S6, we provide the details of the algorithm that we developed for computing the operator norm $\|I - \eta \nabla^2 \tilde{R}_k(w)\|$ for neural networks.
  - In Section S7, we provide the details of the SGD hyperparameters, network architectures, and information regarding hardware/run-time.
- Proofs.
  - In Section S8, we provide the proofs all the theoretical results presented in the main document and the supplementary document.

S1 Further Background on Dimension Theory

S1.1 Minkowski dimension of a measure

Based on the definition of the Minkowski dimension of sets as given in Section 2, we can define the Minkowski dimension of a finite Borel measure μ, as follows [Pes08]:

$$\overline{\dim}_{\mathrm{M}}\mu := \lim_{\delta\to 0} \inf \left\{ \overline{\dim}_{\mathrm{M}} Z : \mu(Z) \geq 1-\delta \right\}. \tag{S1}$$

Note that in general, we have

$$\overline{\dim}_{\mathrm{M}}\mu \leq \inf \left\{ \overline{\dim}_{\mathrm{M}} Z : \mu(Z) = 1 \right\},$$

where the inequality can be strict, see [Pes08, Chapter 7].

S1.2 Local dimensions of a measure

It is sometimes more convenient to consider a dimension notion that is defined in a pointwise manner. Let μ be a finite Borel regular measure on $\mathbb{R}^d$. The lower and upper local (or pointwise) dimensions of μ at $x \in \mathbb{R}^d$ are respectively defined as follows:

$$\underline{\dim}_{\mathrm{loc}}\mu(x) := \liminf_{r\to 0} \frac{\log \mu(B(x,r))}{\log r}, \tag{S2}$$

$$\overline{\dim}_{\mathrm{loc}}\mu(x) := \limsup_{r\to 0} \frac{\log \mu(B(x,r))}{\log r}, \tag{S3}$$

where $B(x, r)$ denotes the ball with radius r about x. When the values of these dimensions agree, the common value is called the local (or pointwise) dimension of μ at x, and is denoted by $\dim_{\mathrm{loc}} \mu(x)$. The local dimensions describe the power-law behavior of $\mu(B(x, r))$ for small r [Fal97]. These dimensions are closely linked to the Hausdorff dimension.

S1.3 Existing Results

The following result from [Ram06] upper-bounds the Hausdorff dimension of the invariant measure of an IFS to the constituents of the IFS. We translate the result to our notation. Before we proceed, let us first introduce open set conditions from [Ram06]. The IFS satisfies the *open set condition* (OSC) if the IFS is contracting, and there exists an open set U such that $h_i(U) \subset U$ and $h_i(U) \cap h_j(U) = \emptyset$ for $i \neq j$. The IFS satisfies the *strong open set condition* (SOSC) if the IFS satisfies OSC for some open set U and if there exists some $R_1 > 0$ such that $\mathrm{dist}(h_i(U), h_j(U)) \geq R_1$. The IFS satisfies the *regular open set condition* (ROSC) if the IFS satisfies OSC for some open set U and in addition there exist some $R_2, R_3 > 0$ such that $\mathrm{vol}(B_r(x) \cap U) \geq R_3 r^d$ for any $r < R_2$, $x \in U$. We have the following result.

Theorem S2. *[Ram06, Theorem 2.1] Consider the IFS (6) and assume that conditions (7) and (8) are satisfied and h_i are continuously differentiable with derivatives J_{h_i} that are α-Hölder continuous for some $\alpha > 0$. The invariant measure $\mu_{W|\mathbf{S}_n}$ of the IFS satisfies*

$$\overline{\dim}_{\mathrm{H}} \mu_{W|\mathbf{S}_n} \leq s \quad \textit{where} \quad s := \frac{\mathcal{E}}{\sum_{i=1}^{m_b} p_i \int_{x \in \mathbb{R}^d} \log(\|J_{h_i}(w)\|) \mathrm{d}\mu_{W|\mathbf{S}_n}(w)},$$

where $\mathcal{E} := \sum_{i=1}^{m_b} p_i \log(p_i)$ is the (negative) entropy. Furthermore, if h_i are conformal and either SOSC or ROSC is satisfied, then we have

$$\dim_{\mathrm{H}}(\mu_{W|\mathbf{S}_n}) = \underline{\dim}_{\mathrm{H}}(\mu_{W|\mathbf{S}_n}) = s.$$

The next two results link the Hausdorff and Minkowski dimensions of a measure to its local dimension.

Proposition S6. *[Fal97, Propositions 10.3] For a finite Borel measure μ, the following identity holds:*

$$\overline{\dim}_{\mathrm{H}} \mu = \inf \left\{ s : \underline{\dim}_{\mathrm{loc}} \mu(x) \leq s \textit{ for } \mu\textit{-almost all } x \right\}. \tag{S4}$$

Theorem S3. *[Pes08, Theorem 7.1] Let μ be a finite Borel measure on $\mathbb{R}^d$. If $\overline{\dim}_{\mathrm{loc}} \mu(x) \leq \alpha$ for μ-almost every x, then $\overline{\dim}_{\mathrm{M}} \mu \leq \alpha$.*

The next theorem, called Egoroff's theorem, will be used in our proofs repeatedly. It provides a condition for measurable functions to be uniformly continuous in an almost full-measure set.

Theorem S4 (Egoroff's Theorem)**.** *[Bog07, Theorem 2.2.1] Let $(X, \mathcal{A}, \mu)$ be a space with a finite nonnegative measure μ and let μ-measurable functions f_n be such that μ-almost everywhere there is a finite limit $f(x) := \lim_{n \to \infty} f_n(x)$. Then, for every $\varepsilon > 0$, there exists a set $X_\varepsilon \in \mathcal{A}$ such that $\mu(X \backslash X_\varepsilon) < \varepsilon$ and the functions f_n converge to f uniformly on X_ε.*

S2 Additional Analytical Estimates for SGD

Support vector machines. Given the data points $z_i = (a_i, y_i)$ with the input data a_i and the output $y_i \in \{-1, 1\}$, consider support vector machines with smooth hinge loss:

$$\ell(w, z_i) := \ell_\sigma\left(y_i a_i^T w\right) + \lambda \|w\|^2 / 2, \tag{S5}$$

where $\sigma > 0$ is a smoothing parameter, $\lambda > 0$ is the regularization parameter and $\ell_\sigma(z) := 1 - z + \sigma \log(1 + e^{-\frac{1-z}{\sigma}})$. This loss function is a smooth version of the hinge loss that can be easier to optimize in some settings. In fact, it can be shown that as $\sigma \to 0$, this loss converges to the (non-smooth) hinge loss pointwise.

Proposition S7 (Support vector machines)**.** *Consider the support vector machines (S5). Assume the step-size $\eta < \frac{1}{\lambda + \|R\|^2/(4\rho)}$, where $R := \max_i \|a_i\|$ is finite. Then, we have:*

$$\overline{\dim}_{\mathrm{H}} \mu_{W|\mathbf{S}_n} \leq \frac{\log(n/b)}{\log(1/(1 - \eta\lambda))}. \tag{S6}$$

S3 Analytical Estimates for Preconditioned SGD

We consider the pre-conditioned SGD methods

$$w_k = w_{k-1} - \eta H^{-1} \nabla \tilde{R}_k(w_{k-1}), \tag{S7}$$

for a fixed square matrix H. Some choices of H includes a diagonal matrix, a block diagonal matrix or the Fisher-information matrix (see e.g. [ZMG19]). We assume that H is a positive definite matrix, and by Cholesky decomposition, we can write $H = SS^T$, where S is a real lower triangular matrix with positive diagonal entries. If we have $H = JJ^T$, where J is the Jacobian, then the corresponding least square problems is called the Gauss-Newton methods for least squares. Assume that H there exist some $m, M > 0$ such that:

$$0 \prec mI \preceq H \preceq MI. \tag{S8}$$

As illustrative examples; in the following, we will consider the setting where we divide $\mathbf{S}_n$ into $m_b = n/b$ batches with each batch having b elements, and then we discuss how analytical estimates on the (upper) Hausdorff dimension $\overline{\dim}_{\mathrm{H}} \mu_{W|\mathbf{S}_n}$ can be obtained for some particular problems including least squares, regularized logistic regression, support vector machines, and one hidden-layer network.

Least squares. We consider the least square problem with data points $z_i = (a_i, y_i)$ and the loss

$$\ell(w, z_i) := \frac{1}{2}\left(a_i^T w - y_i\right)^2 + \frac{\lambda}{2}\|w\|^2, \tag{S9}$$

where $\lambda > 0$ is a regularization parameter. If we apply preconditioned SGD this results in the recursion (6) with

$$\begin{aligned} h_i(w) &= M_i w + q_i \quad \text{with} \quad M_i := I - \eta\lambda H^{-1} - \eta H^{-1} H_i, \\ H_i &:= \frac{1}{b}\sum\nolimits_{j \in S_i} a_j a_j^T, \quad q_i := \frac{\eta}{b} H^{-1} \sum\nolimits_{j \in S_i} a_j y_j \,, \end{aligned} \tag{S10}$$

where $a_j \in \mathbb{R}^d$ are the input vectors, and y_j are the output variables, and $\{S_i\}_{i=1}^{m_b}$ is a partition of $\{1, 2, \ldots, n\}$ with $|S_i| = b$, where $i = 1, 2, \ldots, m_b$ with $m_b = n/b$. We have the following result.

Proposition S8 (Least squares)**.** *Consider the pre-conditioned SGD method* (S7) *for the least square problem* (S9)*. Assume that the step-size* $\eta < \frac{m}{R^2 + \lambda}$*, where* $R := \max_i \|a_i\|$ *is finite. Then, we have the following upper bound for the Hausdorff dimension:*

$$\overline{\dim}_{\mathrm{H}} \mu_{W|\mathbf{S}_n} \leq \frac{\log(n/b)}{\log(1/(1 - \eta M^{-1}\lambda))}. \tag{S11}$$

Regularized logistic regression. We consider the regularized logistic regression problem with the data points $z_i = (a_i, y_i)$ and the loss:

$$\ell(w, z_i) := \log\left(1 + \exp\left(-y_i a_i^T w\right)\right) + \frac{\lambda}{2}\|w\|^2, \tag{S12}$$

where $\lambda > 0$ is the regularization parameter.

Proposition S9 (Regularized logistic regression)**.** *Consider the pre-conditioned SGD method* (S7) *for regularized logistic regression* (S12)*. Assume that the step-size* $\eta < m/\lambda$ *and* $R := \max_i \|a_i\| < 2\sqrt{m\lambda/M}$*. Then, we have the following upper bound for the Hausdorff dimension:*

$$\overline{\dim}_{\mathrm{H}} \mu_{W|\mathbf{S}_n} \leq \frac{\log(n/b)}{\log(1/(1 - \eta M^{-1}\lambda + \frac{1}{4}\eta m^{-1} R^2))}. \tag{S13}$$

Next, we consider a non-convex formulation for robust regression. Consider the data points $z_i = (a_i, y_i)$ and the loss:

$$\ell(w, z_i) := \rho\left(y_i - \langle w, a_i \rangle\right) + \frac{\lambda}{2}\|w\|^2, \tag{S14}$$

where $\lambda > 0$ is a regularization parameter and ρ is a non-convex function. We have the following result.

Proposition S10 (Non-convex formulation for robust regression)**.** *Consider the pre-conditioned SGD method* (S7) *in the non-convex robust regression setting* (S14)*. Assume that the step-size* $\eta < \frac{m}{\lambda + R^2(2/t_0)}$ *and* $R := \max_i \|a_i\| < \sqrt{\lambda t_0 m/(2M)}$*. Then, we have the following upper bound for the Hausdorff dimension:*

$$\overline{\dim}_{\mathrm{H}} \mu_{W|\mathbf{S}_n} \leq \frac{\log(n/b)}{\log(1/(1 - \eta M^{-1}\lambda + \eta m^{-1} R^2 \frac{2}{t_0}))}. \tag{S15}$$

Support vector machines. We have the following result for pre-conditioned SGD when applied to the support vector machines problem (S5).

Proposition S11 (Support vector machines)**.** *Consider pre-conditioned SGD* (S7) *for support vector machines* (S5)*. Assume that the step-size* $\eta < \frac{m}{\lambda + \|R\|^2/(4\rho)}$ *where* $R := \max_i \|a_i\|$ *is finite. Then, we have the following upper bound for the Hausdorff dimension:*

$$\overline{\dim}_{\mathrm{H}} \mu_{W|\mathbf{S}_n} \leq \frac{\log(n/b)}{\log(1/(1 - \eta M^{-1}\lambda))}. \tag{S16}$$

One hidden-layered neural network. Consider the one-hidden-layer neural network setting as in Proposition 5, where the objective is to minimize the regularized squared loss with the loss function:

$$\ell(w, z_i) := \|y_i - \hat{y}_i\|^2 + \frac{\lambda}{2}\|w\|^2, \quad \hat{y}_i := \sum_{r=1}^{m} b_r \sigma\left(w_r^T a_i\right), \tag{S17}$$

where the non-linearity $\sigma : \mathbb{R} \to \mathbb{R}$ is smooth and $\lambda > 0$ is the regularization parameter.

Proposition S12 (One hidden-layer network)**.** *Consider the one-hidden-layer network* (S17)*. Assume that* $\eta < \frac{m}{C+\lambda}$ *and* $\lambda > \frac{M}{m}C$*, where* C *is defined in Corollary* 5*. Then, we have the following upper bound for the Hausdorff dimension:*

$$\overline{\dim}_{\mathrm{H}} \mu_{W|\mathbf{S}_n} \leq \frac{\log(n/b)}{\log(1/(1 - \eta(M^{-1}\lambda - m^{-1}C)))}. \tag{S18}$$

S4 Analytical Estimates for Stochastic Newton

We consider the stochastic Newton method

$$w_k = w_{k-1} - \eta \left[\tilde{H}_k(w_{k-1})\right]^{-1} \nabla \tilde{\mathcal{R}}_k(w_{k-1}), \quad \text{where} \quad \tilde{H}_k(w) := (1/b) \sum\nolimits_{i \in \Omega_k} \nabla^2 \ell(w, z_i),$$

see e.g. [RKM16], where $\Omega_k = S_i$ with probability p_i with $i = 1, 2, \ldots, m_b$, where $m_b = n/b$.

For simplicity, we focus on the least square problem, with the data points $z_i = (a_i, y_i)$ and the loss:

$$\ell(w, z_i) := \frac{1}{2}\left(a_i^T w - y_i\right)^2 + \frac{\lambda}{2}\|w\|^2, \tag{S19}$$

where $\lambda > 0$ is a regularization parameter. If we apply stochastic Newton this results in the recursion (6) with

$$h_i(w) = M_i w + q_i \quad \text{with} \quad M_i := (1 - \eta) I, \tag{S20}$$
$$\tilde{H}_i := \frac{1}{b} \sum\nolimits_{j \in S_i} a_j a_j^T + \lambda I, \quad q_i := \frac{\eta}{b} \tilde{H}_i^{-1} \sum\nolimits_{j \in S_i} a_j y_j \,,$$

where $a_j \in \mathbb{R}^d$ are the input vector, and y_j are the output variable, and $\{S_i\}_{i=1}^{m_b}$ is a partition of $\{1, 2, \ldots, n\}$ with $|S_i| = b$, where $i = 1, 2, \ldots, m_b$ with $m_b = n/b$. Therefore, $J_{h_i}(w) = (1 - \eta) I$. By following the similar argument as in the proof of Proposition S8, we conclude that for any $\eta \in (0, 1)$,

$$\overline{\dim}_{\mathrm{H}} \mu_{W|\mathbf{S}_n} \leq \frac{\log(n/b)}{\log(1/(1 - \eta))}, \tag{S21}$$

where the upper bound is decreasing in step-size η and batch-size b.

S5 Illustration of the Conditions (7) and (8)

In this section, we will demonstrate how to validate the conditions (7) and (8) on the least squares problem:

$$\hat{\mathcal{R}}(w) := \hat{\mathcal{R}}(w, \mathbf{S}_n) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - x_i^T w\right)^2,$$

where $x_i \in \mathbb{R}^d$'s are isotropic sub-Gaussian random vectors.

Consider a uniform subset $S \subset \{1, ..., n\}$ of size b. By the properties of sub-Gaussian random vectors (e.g. see [Ver18, Theorem 4.7.1]), we have $\mathbb{E}[\|I - \frac{1}{b}\sum_{i\in S} x_i x_i^T\|] \leq CK^2\sqrt{\frac{d}{b}}$ where C is a positive constant, K is the sub-Gaussian norm of x_i, which is true for a sufficiently large batch size.

Now, consider the SGD update with mini-batch size b:

$$w_{k+1} = w_k - \frac{\eta}{b}\sum_{i\in S} x_i(x_i^T w - y_i) =: h_S(w_k, S).$$

We look at the Lipschitz constant of this function system: For any S,

$$\|h_S(w, S) - h_S(w', S)\| = \left\| w - w' - \frac{\eta}{b}\sum_{i\in S} x_i x_i^T(w - w')\right\| \leq \left\| I - \frac{\eta}{b}\sum_{i\in S} x_i x_i^T \right\| \|w - w'\|.$$

Notice that the quantity $\|I - \frac{\eta}{b}\sum_{i\in S} x_i x_i^T\|$ is an upper bound on the Lipschitz constant L_S of h_S. Investigating the condition (8), we have $\mathbb{E}[\log(L_S)] \leq \log\left(\eta CK^2\sqrt{\frac{d}{b}} + (1-\eta)\right)$, where the first step follows from Jensen's inequality and the second follows from sub-Gaussian property. The right hand size of the above bound can be made smaller than 0 for a sufficiently small step size choice η.

S6 Estimating the Complexity R for SGD

Estimating R, as detailed in Equation (21), requires drawing N_W samples from the invariant measure and N_U batches of from the training data.

As mentioned in the main text, to approximate the summation over N_W samples from the invariant measure, assuming (8) is ergodic [DF99], we treat the iterates w_k as i.i.d. samples from $\mu_{W|S}$ for large k, hence, the norm of the Jacobian $\log(\|J_{h_{I_j}}(W_i)\|)$ can be efficiently computed on these iterates. Thus, first we train a neural-network to convergence, whereby convergence is defined as the model reaching some accuracy level (if the dataset is a classification task) *and* achieving a loss below some threshold on training data. We assume that after convergence the SGD iterates will be drawn from the invariant measure. As such we run the training algorithm for another 200 iterates, saving a snapshot of the model parameters at each step, such that $N_W = 200$ in Equation (21). For each of these snapshots we estimate the spectral norm $\|J_{h_I}(W)\|$ using a simple modification of the power iteration algorithm of [YGKM20], detailed in Section S6.1 below. This modified algorithm is scalable to neural networks with millions of parameters and we apply it to 50 of the batches used during training, such that $N_U = 50$ in (21).

S6.1 Power Iteration Algorithm for $\|J_{h_i}(w)\|$

We re-purpose the power iteration algorithm of [YGKM20] adding a small modification that allows for the estimation of the spectral norm $\|J_{h_i}(w)\|$. We first note that

$$J_{h_i}(w) = I - \eta\nabla^2\tilde{\mathcal{R}}_i(w) \tag{S22}$$

where $\nabla^2\tilde{\mathcal{R}}_i(w)$ is the Hessian for the i^{th} batch. As such our power iteration algorithm needs to estimate the operator norm of the matrix $I - \eta\nabla^2\tilde{\mathcal{R}}_i(w)$ and not just that of the Hessian of the network. To do this we just need to change the 'vector-product' step of the power-iteration algorithm of [YGKM20]. Our modified method has the same convergence guarantees, namely that the method will converge to the 'true' top eigenvalue if this eigenvalue is 'dominant', in that it dominates all other eigenvalues in absolute value, i.e if λ_1 is the top eigenvalue then we must have that:

$$|\lambda_1| > |\lambda_2| \geq \ldots |\lambda_n|$$

to guarantee convergence.

```
Algorithm 1: Power Iteration for Top Eigenvalue Computation of $J_{h_i}(w)$
Input: Network Parameters: $w$, Loss function: $f$, Learning rate: $\eta$
Compute the gradient of $f$ by backpropagation, i.e., compute $g_w = \frac{df}{dw}$.
Draw a random vector $v$ from $N(0,1)$ (same dimension as $w$).
Normalize $v$, $v = \frac{v}{\|v\|_2}$
for $i = 1, 2, \ldots$ do                  // Power Iteration
    Compute $gv = g_w^T v$                                  // Inner product
    Compute $Hv$ by backpropagation, $Hv = \frac{d(gv)}{dw}$      // Get Hessian vector product
    Compute $J_{h_i}v$, $J_{h_i}v = (I - \eta H)v = v - \eta Hv$        // Get $J_{h_i}$ vector product
    Normalize and reset $v$, $v = \frac{J_{h_i}v}{\|J_{h_i}v\|_2}$
end
```

S7 Experiment Hyperparameters

Training Parameters: All models in Figures 3 and 4 were trained using SGD with batch sizes of 50 or 100 and were considered to have converged for CIFAR10 and SVHN if they reached 100% accuracy and less than 0.0005 loss on the training set. For BHP convergence was considered to have been achieved after 100000 training steps. For all models except VGG16 in Figures 3 and 4 we use learning rates in

$$\{0.0075, 0.02, 0.025, 0.03, 0.04, 0.06, 0.07, 0.075, 0.08, 0.09, 0.1, 0.11, 0.12,$$
$$0.13, 0.14, 0.15, 0.16, 0.17, 0.18, 0.19, 0.194, 0.2, 0.22, 0.24, 0.25, 0.26\}.$$

VGG16 models were trained with learning rates in $\{0.0075, 0.02, 0.03, 0.06, 0.07, 0.08\}$.

Network Architectures: BHP FCN had 2 hidden layers and were 10 neurons wide. Similarly CIFAR10 FCN were 5 and 7 layers deep with 2048 neurons per layer. 9-layer CONV networks were VGG11 networks with the final 2 layers removed. 16-layer CONV networks were simply the standard implementation of VGG16 networks.

Run-time: The full battery of fully connected models split over two *GeForce GTX 1080* GPUs took two days to train to convergence and the subsequent power iterations took less than a day. Similarly the full gamut of VGG11 models took a day to train to convergence over four *GeForce GTX 1080* GPUs and the subsequent power iterations took less than a day to converge. The VGG16 models took a day to train over four *GeForce GTX 1080* GPUs but the power iterations **for each model** took roughly 24 hours on a single *GeForce GTX 1080* GPU.

S8 Postponed Proofs

S8.1 Proof of Proposition 1

Proof. Denote $\alpha := \overline{\dim}_{\mathrm{H}}\mu_{W|\mathbf{S}_n}$. By Assumption **H**1, we have

$$\underline{\dim}_{\mathrm{loc}}\mu_{W|\mathbf{S}_n}(w) = \overline{\dim}_{\mathrm{loc}}\mu_{W|\mathbf{S}_n}(w),$$

for $\mu_{W|\mathbf{S}_n}$-a.e. w, and by Proposition S6 we have

$$\overline{\dim}_{\mathrm{loc}}\mu_{W|\mathbf{S}_n}(w) \leq \alpha + \epsilon, \tag{S23}$$

for all $\epsilon > 0$ and for $\mu_{W|\mathbf{S}_n}$-a.e. w. By invoking Theorem S3, we obtain

$$\overline{\dim}_{\mathrm{M}} \mu_{W|\mathbf{S}_n} \leq \alpha + \epsilon. \tag{S24}$$

Since this holds for any ϵ, $\overline{\dim}_{\mathrm{M}} \mu_{W|\mathbf{S}_n} \leq \alpha$. By definition, we have for almost all $\mathbf{S}_n$:

$$\overline{\dim}_{\mathrm{M}} \mu_{W|\mathbf{S}_n} = \lim_{\delta \to 0} \inf \left\{ \overline{\dim}_{\mathrm{M}} A : \mu_{W|\mathbf{S}_n}(A) \geq 1 - \delta \right\}. \tag{S25}$$

Hence, given a sequence $(\delta_k)_{k \geq 1}$ such that $\delta_k \downarrow 0$, and $\mathbf{S}_n$, and any $\epsilon > 0$, there is a $k_0 = k_0(\epsilon)$ such that $k \geq k_0$ implies

$$\begin{aligned} \inf \left\{ \overline{\dim}_{\mathrm{M}} A : \mu_{W|\mathbf{S}_n}(A) \geq 1 - \delta_k \right\} \leq & \overline{\dim}_{\mathrm{M}} \mu_{W|\mathbf{S}_n} + \epsilon & \text{(S26)} \\ \leq & \alpha + \epsilon. & \text{(S27)} \end{aligned}$$

Hence, for any $\epsilon_1 > 0$ and $k \geq k_0$, we can find a bounded Borel set $A_{\mathbf{S}_n,k}$, such that $\mu_{W|\mathbf{S}_n}(A_{\mathbf{S}_n,k}) \geq 1 - \delta_k$, and

$$\overline{\dim}_{\mathrm{M}} A_{\mathbf{S}_n,k} \leq \alpha + \epsilon + \epsilon_1. \tag{S28}$$

Note that the boundedness of $A_{\mathbf{S}_n,k}$ follows from the fact that its upper-Minkowski dimension is finite. By choosing $\epsilon = \epsilon_1 = \frac{\varepsilon}{2}$, it yields the desired result. This completes the proof. □

S8.2 Proof of Theorem 1

Proof. We begin similarly to the proof of Proposition 1. Denote

$$\alpha(\mathbf{S}, n) := \overline{\dim}_{\mathrm{H}} \mu_{W|\mathbf{S}_n}.$$

By Assumption **H**1, we have $\underline{\dim}_{\mathrm{loc}} \mu_{W|\mathbf{S}_n}(w) = \overline{\dim}_{\mathrm{loc}} \mu_{W|\mathbf{S}_n}(w)$ for $\mu_{W|\mathbf{S}_n}$-almost every w, and by Proposition S6 we have

$$\overline{\dim}_{\mathrm{loc}} \mu_{W|\mathbf{S}_n}(w) \leq \alpha(\mathbf{S}, n) + \epsilon, \tag{S29}$$

for all $\epsilon > 0$ and for $\mu_{W|\mathbf{S}_n}$-a.e. w. By invoking Theorem S3, we obtain

$$\overline{\dim}_{\mathrm{M}} \mu_{W|\mathbf{S}_n} \leq \alpha(\mathbf{S}, n) + \epsilon. \tag{S30}$$

Since this holds for any $\epsilon > 0$, $\overline{\dim}_{\mathrm{M}} \mu_{W|\mathbf{S}_n} \leq \alpha(\mathbf{S}, n)$. By definition, we have for all $\mathbf{S}$ and n:

$$\overline{\dim}_{\mathrm{M}} \mu_{W|\mathbf{S}_n} = \lim_{\delta \to 0} \inf \left\{ \overline{\dim}_{\mathrm{M}} A : \mu_{W|\mathbf{S}_n}(A) \geq 1 - \delta \right\}. \tag{S31}$$

Hence, for a countable sequence of $\delta \downarrow 0$ and each n, there exists a set Ω_n of full measure such that

$$f_\delta^n(\mathbf{S}) := \inf \left\{ \overline{\dim}_{\mathrm{M}} A : \mu_{W|\mathbf{S}_n}(A) \geq 1 - \delta \right\} \to \overline{\dim}_{\mathrm{M}} \mu_{W|\mathbf{S}_n}, \tag{S32}$$

for all $\mathbf{S} \in \Omega_n$. Let $\Omega^* := \cap_n \Omega_n$. Then for $\mathbf{S} \in \Omega^*$ we have that for all n

$$f_\delta^n(\mathbf{S}) \to \overline{\dim}_{\mathrm{M}} \mu_{W|\mathbf{S}_n}, \tag{S33}$$

and therefore, on this set we also have

$$\sup_n \frac{1}{\xi_n} \min \left\{ 1, \left| f_\delta^n(\mathbf{S}) - \overline{\dim}_{\mathrm{M}} \mu_{W|\mathbf{S}_n} \right| \right\} \to 0,$$

where ξ_n is a monotone increasing sequence such that $\xi_n \geq 1$ and $\xi_n \to \infty$.

By applying Theorem S4 to the collection of random variables:

$$F_\delta(\mathbf{S}) := \sup_n \frac{1}{\xi_n} \min \left\{ 1, \left| f_\delta^n(\mathbf{S}) - \overline{\dim}_{\mathrm{M}} \mu_{W|\mathbf{S}_n} \right| \right\}, \tag{S34}$$

for any $\zeta > 0$, we can find a subset $\mathfrak{Z} \subset \mathcal{Z}^\infty$, with probability at least $1 - \zeta$ under π^∞, such that on $\mathfrak{Z}$ the convergence is uniform, that is

$$\sup_{\mathbf{S} \in \mathfrak{Z}} \sup_n \frac{1}{\xi_n} \min \left\{ 1, \left| f_\delta^n(\mathbf{S}) - \overline{\dim}_{\mathrm{M}} \mu_{W|\mathbf{S}_n} \right| \right\} \leq c(\delta), \tag{S35}$$

where for any ζ, $c(\delta) := c(\delta; \zeta) \to 0$ as $\delta \to 0$. Hence, for any δ, $\mathbf{S} \in \mathfrak{Z}$, and n, we have

$$f_\delta^n(\mathbf{S}) \leq \overline{\dim}_{\mathrm{M}} \mu_{W|\mathbf{S}_n} + \xi_n c(\delta) \tag{S36}$$

$$\leq \alpha(\mathbf{S}, n) + \xi_n c(\delta). \tag{S37}$$

Consider a sequence $(\delta_k)_{k \geq 1}$ such that $\delta_k \downarrow 0$ and $\delta_k \in \mathbb{Q}_{>0}$. Then, for any $\mathbf{S} \in \mathfrak{Z}$ and $\epsilon > 0$, we can find a bounded Borel set $A_{\mathbf{S}_n,k}$, such that $\mu_{W|\mathbf{S}_n}(A_{\mathbf{S}_n,k}) \geq 1 - \delta_k$, and

$$\overline{\dim}_{\mathrm{M}} A_{\mathbf{S}_n,k} \leq \alpha(\mathbf{S}, n) + \xi_n c(\delta_k) + \epsilon. \tag{S38}$$

Define the set

$$\mathcal{W}_{n,\delta_k} := \bigcup_{\mathbf{S} \in \mathcal{Z}^\infty} A_{\mathbf{S}_n,k}. \tag{S39}$$

By using $\mathcal{G}(w) := |\mathcal{R}(w) - \hat{\mathcal{R}}(w, \mathbf{S}_n)|$, under the joint distribution of $(W, \mathbf{S}_n)$, such that $\mathbf{S} \sim \pi^\infty$ and $W \sim \mu_{W|\mathbf{S}_n}$, we have:

$$\mathbb{P}\left(\mathcal{G}(W) > \varepsilon\right) \leq \zeta + \mathbb{P}\left(\{\mathcal{G}(W) > \varepsilon\} \cap \{\mathbf{S} \in \mathfrak{Z}\}\right) \tag{S40}$$

$$\leq \zeta + \delta_k + \mathbb{P}\left(\{\mathcal{G}(W) > \varepsilon\} \cap \{W \in A_{\mathbf{S}_n,k}\} \cap \{\mathbf{S} \in \mathfrak{Z}\}\right) \tag{S41}$$

$$\leq \zeta + \delta_k + \mathbb{P}\left(\left\{\sup_{w \in A_{\mathbf{S}_n,k}} \mathcal{G}(w) > \varepsilon\right\} \cap \{\mathbf{S} \in \mathfrak{Z}\}\right). \tag{S42}$$

Now, let us focus on the last term of the above equation. First we observe that as ℓ is L-Lipschitz, so are $\mathcal{R}$ and $\hat{\mathcal{R}}$. Hence, by considering the particular forms of the β-covers in **H**2, for any $w' \in \mathbb{R}^d$ we have:

$$\mathcal{G}(w) \leq \mathcal{G}(w') + 2L \left\| w - w' \right\|, \tag{S43}$$

which implies

$$\sup_{w \in A_{\mathbf{S}_n,k}} \mathcal{G}(w) \leq \max_{w \in N_{\beta_n}(A_{\mathbf{S}_n,k})} \mathcal{G}(w) + 2L\beta_n. \tag{S44}$$

Now, notice that the β-covers of **H**2 still yield the same Minkowski dimension in (5) [ŞSDE20]. Then by definition, we have for all $\mathbf{S}$ and n:

$$\limsup_{\beta \to 0} \frac{\log |N_\beta(A_{\mathbf{S}_n,k})|}{\log(1/\beta)} = \lim_{\beta \to 0} \sup_{r < \beta} \frac{\log |N_r(A_{\mathbf{S}_n,k})|}{\log(1/r)} = \overline{\dim}_{\mathrm{M}} A_{\mathbf{S}_n,k} := d_{\mathrm{M}}(\mathbf{S}, n, k). \tag{S45}$$

Hence for each n

$$g_\beta^{n,k}(\mathbf{S}) := \sup_{\mathbb{Q} \ni r < \delta} \frac{\log |N_r(A_{\mathbf{S}_n,k})|}{\log(1/r)} \to d_{\mathrm{M}}(\mathbf{S}, n, k), \tag{S46}$$

almost surely. By using the same reasoning in (S32), we have, for each n, there exists a set Ω'_n of full measure such that

$$g_\beta^{n,k}(\mathbf{S}) = \sup_{\mathbb{Q} \ni r < \beta} \frac{\log |N_r(A_{\mathbf{S}_n,k})|}{\log(1/r)} \to d_{\mathrm{M}}(\mathbf{S}, n, k), \tag{S47}$$

for all $\mathbf{S} \in \Omega'_n$. Define $\Omega^{**} := (\cap_n \Omega'_n) \cap \Omega^*$. Hence, on Ω^{**} we have:

$$G_\beta^k(\mathbf{S}) := \sup_n \frac{1}{\xi_n} \min\left\{1, \left| g_\beta^{n,k}(\mathbf{S}) - d_{\mathrm{M}}(\mathbf{S}, n, k) \right| \right\} \to 0, \tag{S48}$$

By applying Theorem S4 to the collection $\{G_\beta^k(\mathbf{S})\}_\beta$, for any $\zeta_1 > 0$ we can find a subset $\mathfrak{Z}_1 \subset \mathcal{Z}^\infty$, with probability at least $1 - \zeta_1$ under π^∞, such that on $\mathfrak{Z}_1$ the convergence is uniform, that is

$$\sup_{\mathbf{S} \in \mathfrak{Z}_1} \sup_n \frac{1}{\xi_n} \min\{1, |g_\beta^{n,k}(\mathbf{S}) - d_{\mathrm{M}}(\mathbf{S}, n, k)|\} \leq c'(\beta), \tag{S49}$$

where for any ζ_1, $c'(\beta) := c'(\beta; \zeta_1, \delta_k) \to 0$ as $\beta \to 0$. Hence, denoting $\mathfrak{Z}^* := \mathfrak{Z} \cap \mathfrak{Z}_1$ by using (S38) we have:

$$\{\mathbf{S} \in \mathfrak{Z}^*\} \subseteq \bigcap_n \left\{ \left| N_\beta\left(A_{\mathbf{S}_n,k}\right) \right| \leq \left(\frac{1}{\beta}\right)^{\alpha(\mathbf{S},n) + \xi_n c(\delta_k) + \xi_n c'(\beta) + \epsilon} \right\}.$$

Let $(\beta_n)_{n\geq 0}$ be a decreasing sequence such that $\beta_n \in \mathbb{Q}$ for all n and $\beta_n \to 0$. We then have

$$\mathbb{P}\Bigg(\{\mathbf{S} \in \mathfrak{Z}\} \cap \Big\{\max_{w \in N_{\beta_n}(A_{\mathbf{S}_n,k})} \mathcal{G}_n(w) \geq \varepsilon\Big\}\Bigg) \leq \mathbb{P}\Bigg(\{\mathbf{S} \in \mathfrak{Z}^*\} \cap \Big\{\max_{w \in N_{\beta_n}(A_{\mathbf{S}_n,k})} \mathcal{G}_n(w) \geq \varepsilon\Big\}\Bigg) + \zeta_2.$$

For $\rho > 0$ and $m \in \mathbb{N}_+$ let us define the interval $J_m(\rho) := (m\rho, (m+1)\rho]$. Furthermore, for any $t > 0$ define

$$\varepsilon(t) := \sqrt{\frac{2\nu^2}{n}\Big[\log(1/\beta_n)\left(t + \xi_n c(\delta_k) + \xi_n c'(\beta_n) + \epsilon\right) + \log(M/\zeta_2)\Big]}. \tag{S50}$$

For notational simplicity, denote $N_{\beta_n,k} := N_{\beta_n}(\mathcal{W}_{n,\delta_k})$, where $\mathcal{W}_{n,\delta_k}$ is defined in (S39) and

$$\tilde{\alpha}(\mathbf{S}, n, k, \epsilon) := \alpha(\mathbf{S}, n) + \xi_n c(\delta_k) + \xi_n c'(\beta_n) + \epsilon. \tag{S51}$$

Let d^* be the smallest real number such that $\alpha(\mathbf{S}, n) \leq d^*$ almost surely[4], we therefore have:

$$\begin{aligned}
&\mathbb{P}\Bigg(\{\mathbf{S} \in \mathfrak{Z}\} \cap \Big\{\max_{w \in N_{\beta_n}(A_{\mathbf{S}_n,k})} \mathcal{G}_n(w) \geq \varepsilon(\alpha(\mathbf{S}, n))\Big\}\Bigg) \\
&\leq \zeta_2 + \mathbb{P}\Bigg(\Bigg\{|N_{\beta_n}(A_{\mathbf{S}_n,k})| \leq \left(\frac{1}{\beta_n}\right)^{\tilde{\alpha}(\mathbf{S},n,k,\epsilon)}\Bigg\} \\
&\qquad\qquad \cap \Big\{\max_{w \in N_{\beta_n}(A_{\mathbf{S}_n},k)} |\hat{\mathcal{R}}_n(w) - \mathcal{R}(w)| \geq \varepsilon(\alpha(\mathbf{S}, n))\Big\}\Bigg) \\
&= \zeta_2 + \sum_{m=0}^{\lceil \frac{d^*}{\rho} \rceil} \mathbb{P}\Bigg(\Bigg\{|N_{\beta_n}(A_{\mathbf{S}_n,k})| \leq \left(\frac{1}{\beta_n}\right)^{\tilde{\alpha}(\mathbf{S},n,k,\epsilon)}\Bigg\} \\
&\qquad\qquad \cap \Big\{\max_{w \in N_{\beta_n}(A_{\mathbf{S}_n,k})} \mathcal{G}_n(w) \geq \varepsilon(\alpha(\mathbf{S}, n))\Big\} \cap \{\alpha(\mathbf{S}, n) \in J_m(\rho)\}\Bigg) \\
&= \zeta_2 + \sum_{m=0}^{\lceil \frac{d^*}{\rho} \rceil} \mathbb{P}\Bigg(\Bigg\{|N_{\beta_n}(A_{\mathbf{S}_n,k})| \leq \left(\frac{1}{\beta_n}\right)^{\tilde{\alpha}(\mathbf{S},n,k,\epsilon)}\Bigg\} \cap \{\alpha(\mathbf{S}, n) \in J_m(\rho)\} \\
&\qquad\qquad \cap \bigcup_{w \in N(\beta_n)} \Big(\{w \in N_{\beta_n}(A_{\mathbf{S}_n,k})\} \cap \{\mathcal{G}_n(w) \geq \varepsilon(\alpha(\mathbf{S}, n))\}\Big)\Bigg) \\
&\leq \zeta_2 + \sum_{m=0}^{\lceil \frac{d^*}{\rho} \rceil} \sum_{w \in N_{\beta_n,k}} \mathbb{P}\Bigg(\Big\{\mathcal{G}_n(w) \geq \varepsilon(m\rho)\Big\} \\
&\qquad \cap \Big\{w \in N_{\delta_n}(A_{\mathbf{S}_n,k})\Big\} \cap \Bigg\{|N_{\beta_n}(A_{\mathbf{S}_n,k})| \leq \left(\frac{1}{\beta_n}\right)^{\tilde{\alpha}(\mathbf{S},n,k,\epsilon)}\Bigg\} \cap \{\alpha(\mathbf{S}, n) \in J_m(\rho)\}\Bigg),
\end{aligned}$$

where we used the fact that on the event $\alpha(\mathbf{S}, n) \in J_m(\rho)$, $\varepsilon(\alpha(\mathbf{S}, n)) \geq \varepsilon(m\rho)$.

Notice that the events

$$\Big\{w \in N_{\beta_n}(A_{\mathbf{S}_n,k})\Big\}, \Big\{|N_{\beta_n}(A_{\mathbf{S}_n,k})| \leq (1/\beta_n)^{\tilde{\alpha}(\mathbf{S},n,k,\epsilon)}\Big\}, \{\alpha(\mathbf{S}, n) \in J_m(\rho)\}$$

are in $\mathfrak{G}$. On the other hand, the event $\{\mathcal{G}_n(w) \geq \varepsilon(m\rho)\}$ is clearly in $\mathfrak{F}$ (see **H**2 for definitions).

[4] Notice that we trivially have $d^* \leq d$; yet, d^* can be much smaller than d.

Therefore, we have

$$\begin{aligned}
&\mathbb{P}\left(\{\mathbf{S} \in \mathfrak{Z}\} \cap \left\{\max_{w \in N_{\beta_n}(A_{\mathbf{S}_n,k})} \mathcal{G}_n(w) \geq \varepsilon(\alpha(\mathbf{S},n))\right\}\right) \\
&\leq \zeta_2 + M \sum_{m=0}^{\lceil \frac{d^*}{\rho} \rceil} \sum_{w \in N_{\beta_n,k}} \mathbb{P}\left(\left\{\mathcal{G}_n(w) \geq \varepsilon(m\rho)\right\}\right) \\
&\qquad \times \mathbb{P}\left(\left\{w \in N_{\beta_n}(A_{\mathbf{S}_n,k})\right\} \cap \left\{|N_{\beta_n}(A_{\mathbf{S}_n,k})| \leq \left(\frac{1}{\beta_n}\right)^{\tilde{\alpha}(\mathbf{S},n,k,\epsilon)}\right\} \cap \{\alpha(\mathbf{S},n) \in J_m(\rho)\}\right),
\end{aligned}$$

Recall that $\mathcal{G}_n(w) = \frac{1}{n}\sum_{i=1}^n [\ell(w,z_i) - \mathbb{E}_{z\sim\pi}\ell(w,z)]$. Since the $(z_i)_i$ are i.i.d. by Assumption 3 it follows that $\mathcal{G}_n(w)$ is $(\nu/\sqrt{n}, \kappa/n)$-sub-exponential and from [Wai19, Proposition 2.9] we have that

$$\mathbb{P}\left(\left\{\mathcal{G}_n(w) \geq \varepsilon(m\rho)\right\}\right) \leq 2\exp\left(-\frac{n\varepsilon(m\rho)^2}{2\nu^2}\right),$$

as long as $\varepsilon(m\rho) \leq \nu^2/\kappa$. For n large enough we may assume that $\varepsilon(d^*) \leq \nu^2/\kappa$, and thus

$$\begin{aligned}
&\mathbb{P}\left(\{\mathbf{S} \in \mathfrak{Z}\} \cap \left\{\max_{w \in N_{\beta_n}(A_{\mathbf{S}_n,k})} \mathcal{G}_n(w) \geq \varepsilon(\alpha(\mathbf{S},n))\right\}\right) \\
&\leq 2M \sum_{m=0}^{\lceil \frac{d}{\rho} \rceil} \mathrm{e}^{-\frac{2n\varepsilon^2(m\rho)}{B^2}} \sum_{w \in N_{\beta_n,k}} \mathbb{P}\left(\left\{w \in N_{\beta_n}(A_{\mathbf{S}_n,k})\right\} \cap \left\{|N_{\beta_n}(A_{\mathbf{S}_n,k})| \leq \left(\frac{1}{\beta_n}\right)^{\tilde{\alpha}(\mathbf{S},n,k,\epsilon)}\right\}\right. \\
&\qquad\qquad \left. \cap \{\alpha(\mathbf{S},n) \in J_m(\rho)\}\right) + \zeta_2 \\
&\leq 2M \sum_{m=0}^{\lceil \frac{d}{\rho} \rceil} \mathrm{e}^{-\frac{n\varepsilon^2(m\rho)}{2\nu^2}} \sum_{w \in N_{\beta_n,k}} \mathbb{E}\left[\mathbb{1}\left\{w \in N_{\beta_n}(A_{\mathbf{S}_n,k})\right\}\right. \\
&\qquad \left. \times \mathbb{1}\left\{|N_{\beta_n}(A_{\mathbf{S}_n,k})| \leq \left(\frac{1}{\beta_n}\right)^{\tilde{\alpha}(\mathbf{S},n,k,\epsilon)}\right\} \times \mathbb{1}\left\{\alpha(\mathbf{S},n) \in J_m(\rho)\right\}\right] + \zeta_2 \\
&\leq 2M \sum_{m=0}^{\lceil \frac{d}{\rho} \rceil} \mathrm{e}^{-\frac{n\varepsilon^2(m\rho)}{2\nu^2}} \mathbb{E}\left[\sum_{w \in N_{\beta_n,k}} \mathbb{1}\left\{w \in N_{\beta_n}(A_{\mathbf{S}_n,k})\right\}\right. \\
&\qquad \left. \times \mathbb{1}\left\{|N_{\beta_n}(A_{\mathbf{S}_n,k})| \leq \left(\frac{1}{\beta_n}\right)^{\tilde{\alpha}(\mathbf{S},n,k,\epsilon)}\right\} \times \mathbb{1}\left\{\alpha(\mathbf{S},n) \in J_m(\rho)\right\}\right] + \zeta_2 \qquad \text{(S52)} \\
&\leq 2M \sum_{m=0}^{\lceil \frac{d}{\rho} \rceil} \mathrm{e}^{-\frac{n\varepsilon^2(m\rho)}{2\nu^2}} \mathbb{E}\left[|N_{\beta_n}(A_{\mathbf{S}_n,k})| \times \mathbb{1}\left\{|N_{\beta_n}(A_{\mathbf{S}_n,k})| \leq \left(\frac{1}{\beta_n}\right)^{\tilde{\alpha}(\mathbf{S},n,k,\epsilon)}\right\}\right. \\
&\qquad\qquad \left. \times \mathbb{1}\left\{\alpha(\mathbf{S},n) \in J_m(\rho)\right\}\right] + \zeta_2 \\
&= \zeta_2 + 2M \sum_{m=0}^{\lceil \frac{d}{\rho} \rceil} \mathrm{e}^{-\frac{n\varepsilon^2(m\rho)}{2\nu^2}} \mathbb{E}\left[\left[\frac{1}{\beta_n}\right]^{\tilde{\alpha}(\mathbf{S},n,k,\epsilon)} \times \mathbb{1}\left\{\alpha(\mathbf{S},n) \in J_m(\rho)\right\}\right],
\end{aligned}$$

where (S52) follows from Fubini's theorem.

Now, notice that the mapping $t \mapsto \varepsilon^2(t)$ is linear with derivative bounded by

$$\frac{2\nu^2}{n}\log(1/\beta_n).$$

Therefore, on the event $\{\alpha(\mathbf{S},n)\in J_m(\rho)\}$ we have

$$\begin{aligned}\varepsilon^2(\alpha(\mathbf{S},n))-\varepsilon^2(m\rho)\leq&(\alpha(\mathbf{S},n)-m\rho)\frac{2\nu^2}{n}\log(1/\beta_n) &\text{(S53)}\\ \leq&\rho\frac{2\nu^2}{n}\log(1/\beta_n). &\text{(S54)}\end{aligned}$$

By choosing $\rho=\rho_n=1/\log(1/\beta_n)$, we have

$$\varepsilon^2(m\rho_n)\geq\varepsilon^2\left(\alpha(\mathbf{S},n)\right)-\frac{2\nu^2}{n}.$$

Therefore, we have

$$\begin{aligned}&\mathbb{P}\Bigg(\{\mathbf{S}\in\mathfrak{Z}\}\cap\Big\{\max_{w\in N_{\beta_n}(A_{\mathbf{S}_n,k})}\mathcal{G}_n(w)\geq\varepsilon(\alpha(\mathbf{S},n))\Big\}\Bigg)\\ &\leq\zeta_2+2M\mathbb{E}\left[\sum_{m=0}^{\lceil\frac{d}{\rho_n}\rceil}\mathrm{e}^{-\frac{n\varepsilon^2(m\rho_n)}{2\nu^2}}\left[\frac{1}{\beta_n}\right]^{\tilde{\alpha}(\mathbf{S},n,k,\epsilon)}\times\mathbb{1}\left\{\alpha(\mathbf{S},n)\in J_m(\rho_n)\right\}\right]\\ &\leq\zeta_2+2M\mathbb{E}\left[\sum_{m=0}^{\lceil\frac{d}{\rho_n}\rceil}\mathrm{e}^{-\frac{n\varepsilon^2(\alpha(\mathbf{S},n))}{2\nu^2}+1}\left[\frac{1}{\beta_n}\right]^{\tilde{\alpha}(\mathbf{S},n,k,\epsilon)}\times\mathbb{1}\left\{\alpha(\mathbf{S},n)\in J_m(\rho_n)\right\}\right]\\ &=\zeta_2+2M\mathbb{E}\left[\mathrm{e}^{-\frac{n\varepsilon^2(\alpha(\mathbf{S},n))}{2\nu^2}+1}\left[\frac{1}{\beta_n}\right]^{\tilde{\alpha}(\mathbf{S},n,k,\epsilon)}\right].\end{aligned}$$

By the definition of $\varepsilon(t)$, for any $\mathbf{S}$ and n we have that:

$$2M\mathrm{e}^{-\frac{n\varepsilon^2(\alpha(\mathbf{S},n))}{2\nu^2}+1}\left[\frac{1}{\beta_n}\right]^{\tilde{\alpha}(\mathbf{S},n,k,\epsilon)}=2\mathrm{e}\zeta_2.$$

Therefore,

$$\mathbb{P}\Bigg(\{\mathbf{S}\in\mathfrak{Z}\}\cap\Big\{\max_{w\in N_{\beta_n}(A_{\mathbf{S}_n,k})}\mathcal{G}_n(w)\geq\varepsilon(\alpha(\mathbf{S},n))\Big\}\Bigg)\leq(1+2\mathrm{e})\zeta_2.$$

Therefore, by using the definition of $\varepsilon(t)$, (S42), and (S44), with probability at least $1-\zeta-\delta_k-(1+2\mathrm{e})\zeta_2$, we have

$$\begin{aligned}|\hat{\mathcal{R}}(W,\mathbf{S}_n)-\mathcal{R}(W)|\leq2\sqrt{\frac{2\nu^2}{n}\left[\log\left(\frac{1}{\beta_n}\right)\Big(\alpha(\mathbf{S},n)+\xi_nc(\delta_k)+\xi_nc'(\beta_n)+\epsilon\Big)+\log\left(\frac{M}{\zeta_2}\right)\right]}\\ +2L\beta_n.\end{aligned}$$

Choose k such that $\delta_k\leq\zeta/2$, $\zeta_2=\zeta/(2+4\mathrm{e})$, $\xi_n=\log\log(n)$, $\epsilon=\alpha(\mathbf{S},n)$, and $\beta_n=\sqrt{2\nu^2/L^2n}$. Then, with probability at least $1-2\zeta$, we have

$$\begin{aligned}&|\hat{\mathcal{R}}(W,\mathbf{S}_n)-\mathcal{R}(W)| &\text{(S55)}\\ &\leq4\sqrt{\frac{4\nu^2}{n}\left[\frac{1}{2}\log\left(nL^2\right)\Big(2\alpha(\mathbf{S},n)+c(\delta_k)\log\log(n)+o(\log\log(n))\Big)+\log\left(\frac{13M}{\zeta}\right)\right]}. &\text{(S56)}\end{aligned}$$

Finally, as we have $\alpha(\mathbf{S},n)\log(n)=\omega(\log\log(n))$, for n large enough, we obtain

$$|\hat{\mathcal{R}}(W,\mathbf{S}_n)-\mathcal{R}(W)|\leq8\nu\sqrt{\frac{\alpha(\mathbf{S},n)\log^2\left(nL^2\right)}{n}+\frac{\log\left(13M/\zeta\right)}{n}}. \tag{S57}$$

This completes the proof. □

S8.3 Proof of Proposition 2

Proof. If we apply SGD this results in the recursion (6) with

$$h_i(w) = M_i w + q_i \quad \text{with} \quad M_i := (1-\eta\lambda)I - \eta H_i, \tag{S58}$$
$$H_i := \frac{1}{b}\sum\nolimits_{j\in S_i} a_j a_j^T, \quad q_i := (\eta/b)\sum\nolimits_{j\in S_i} a_j y_j \,,$$

where $a_j \in \mathbb{R}^d$ are the input vector, and y_j are the output variable, and $\{S_i\}_{i=1}^{m_b}$ is a partition of $\{1,2,\dots,n\}$ with $|S_i| = b$ with $i = 1,2,\dots,m_b$ and $m_b = n/b$. Let L_i be the Lipschitz constant of $\nabla\ell(w,z_i)$. It can be seen that $\nabla\ell(w,z_i)$ is Lipschitz with constant $L_i = R_i^2 + \lambda$, where $R_i = \max_{j\in S_i}\|a_j\|$. We assume $\eta < 2/L = 2/(R^2+\lambda)$, where $R = \max_i R_i$, otherwise the expectation of the iterates can diverge from some initializations and for some choices of the batch-size. We have

$$h_i(u) - h_i(v) = M_i(u-v),$$

where

$$0 \preceq \left(1-\eta\lambda-\eta R_i^2\right) I \preceq M_i \preceq (1-\eta\lambda)I.$$

Hence, h_i is bi-Lipschitz in the sense of [Anc16] where

$$\gamma_i\|u-v\| \le \|h_i(u)-h_i(v)\| \le \Gamma_i\|u-v\|,$$

with

$$\gamma_i = \min\left(\left|1-\eta\lambda-\eta R_i^2\right|, |1-\eta\lambda|\right), \tag{S59}$$
$$\Gamma_i = \max\left(\left|1-\eta\lambda-\eta R_i^2\right|, |1-\eta\lambda|\right) < 1, \tag{S60}$$

as long as $\gamma_i > 0$. For simplicity of the presentation, we assume $\eta < \frac{1}{R^2+\lambda}$ in which case the expressions for γ_i and Γ_i simplify to:

$$\gamma_i = 1-\eta\lambda-\eta R_i^2, \quad \Gamma_i = 1-\eta\lambda.$$

In this case, it is easy to see that

$$0 < \gamma_i \le \|J_{h_i}(w)\| \le \Gamma_i < 1,$$

and it follows from Theorem S2 that

$$\overline{\dim}_{\mathrm{H}}\mu_{W|\mathbf{S}_n} \le \frac{\mathcal{E}}{\sum_{i=1}^{m_b} p_i \log(\Gamma_i)} = \frac{-\mathcal{E}}{\sum_{i=1}^{m_b} p_i \log(1/\Gamma_i)}. \tag{S61}$$

By Jensen's inequality, we have

$$-\mathcal{E} = \sum_{i=1}^{m_b} p_i \log\left(\frac{1}{p_i}\right) \le \log\left(\sum_{i=1}^{m_b} p_i \cdot \frac{1}{p_i}\right) = \log(m_b), \tag{S62}$$

where $m_b = n/b$. When $\eta < \frac{1}{R^2+\lambda}$, we recall that $\gamma_i = 1-\eta\lambda-\eta R_i^2$ and $\Gamma_i = 1-\eta\lambda$. Therefore,

$$\overline{\dim}_{\mathrm{H}}\mu_{W|\mathbf{S}_n} \le \frac{-\mathcal{E}}{\sum_{i=1}^{m_b} p_i \log(1/\Gamma_i)} \le \frac{\log(m_b)}{\log(1/(1-\eta\lambda))} = \frac{\log(n/b)}{\log(1/(1-\eta\lambda))}. \tag{S63}$$

The proof is complete. □

S8.4 Proof of Proposition 3

Proof. When the batch-size is equal to b, we can compute that the Jacobian is given by

$$J_{h_i}(w) = \frac{1}{b}\sum_{j\in S_i}\left(1-\eta\lambda+\eta y_j^2\left[\frac{e^{-y_j a_j^T w}}{(1+e^{-y_j a_j^T w})^2}\right] a_j a_j^T\right), \tag{S64}$$

where $\{S_i\}_{i=1}^{m_b}$ is a partition of $\{1,2,\dots,n\}$ with $|S_i| = b$, where $i = 1,2,\dots,m_b$ and $m_b = n/b$. Note that the input data is bounded, i.e. $R_i := \max_{j\in S_i}\|a_j\| < \infty$, and $R := \max_i R_i < 2\sqrt{\lambda}$.

Recall that the step-size is sufficiently small, i.e. $\eta < 1/\lambda$. One can provide the upper bound on $J_{h_i}(w)$:

$$\|J_{h_i}(w)\| \leq \Gamma_i := 1 - \eta\lambda + \frac{1}{4}\eta R_i^2 \leq 1 - \eta\lambda + \frac{1}{4}\eta R^2, \tag{S65}$$

so that

$$\begin{aligned}
\overline{\dim}_{\mathrm{H}}\mu_{W|\mathbf{S}_n} &\leq \frac{\mathcal{E}}{\sum_{i=1}^{m_b} p_i \log(\Gamma_i)} = \frac{-\mathcal{E}}{\sum_{i=1}^{m_b} p_i \log(1/(1-\eta\lambda+\frac{1}{4}\eta R_i^2))} \\
&\leq \frac{\log m_b}{\log(1/(1-\eta\lambda+\frac{1}{4}\eta R^2))} \qquad \text{(S66)} \\
&= \frac{\log(n/b)}{\log(1/(1-\eta\lambda+\frac{1}{4}\eta R^2))}, \qquad \text{(S67)}
\end{aligned}$$

where we used (S65) and (S62) in (S66). The proof is complete. □

S8.5 Proof of Proposition 4

Proof. We can compute that

$$\nabla\ell(w, z_i) = -a_i\rho'\left(y_i - \langle w, a_i\rangle\right) + \lambda w, \tag{S68}$$

$$h_i(w) = \frac{1}{b}\sum_{j\in S_i}(1-\eta\lambda)w + \eta a_j\rho'\left(y_j - \langle w, a_j\rangle\right), \tag{S69}$$

$$J_{h_i}(w) = \frac{1}{b}\sum_{j\in S_i}(1-\eta\lambda)I - \eta a_j a_j^T\rho''\left(y_j - \langle w, a_j\rangle\right), \tag{S70}$$

where $\{S_i\}_{i=1}^{m_b}$ is a partition of $\{1,2,\ldots,n\}$ with $|S_i| = b$, where $i = 1,2,\ldots,m_b$ with $m_b = n/b$. Furthermore, $\|\rho''_{\text{exp}}\|_\infty = \rho''_{\text{exp}}(0) = \frac{2}{t_0}$. Therefore, for $\eta \in (0, \frac{1}{\lambda+R^2(2/t_0)})$,

$$0 < (1-\eta\lambda) - \eta R^2\frac{2}{t_0} \leq \|J_{h_i}(w)\| \leq (1-\eta\lambda) + \eta R^2\frac{2}{t_0}, \tag{S71}$$

where $R = \max_i \|a_i\| < \sqrt{\lambda t_0/2}$. We have

$$\overline{\dim}_{\mathrm{H}}\mu_{W|\mathbf{S}_n} \leq \frac{\log m_b}{\log(1/(1-\eta\lambda+\eta R^2\frac{2}{t_0}))} = \frac{\log(n/b)}{\log(1/(1-\eta\lambda+\eta R^2\frac{2}{t_0}))}, \tag{S72}$$

where we used (S71) and (S62). The proof is complete. □

S8.6 Proof of Proposition S7

Proof. We can compute that

$$\begin{aligned}
\nabla\ell(w, z_i) &= y_i\ell'_\sigma\left(y_i a_i^T w\right)a_i + \lambda w, \\
\nabla^2\ell(w, z_i) &= y_i^2\ell''_\sigma\left(y_i a_i^T w\right)a_i a_i^T + \lambda, \\
h_i(w) &= w - \frac{\eta}{b}\sum_{j\in S_i}\nabla\ell(w, z_j),
\end{aligned}$$

where $\{S_i\}_{i=1}^{m_b}$ is a partition of $\{1,2,\ldots,n\}$ with $|S_i| = b$, where $i = 1,2,\ldots,m_b$ with $m_b = n/b$, so that

$$J_{h_i}(w) = I - \frac{\eta}{b}\sum_{j\in S_i}\nabla^2\ell(w, z_j) = (1-\eta\lambda)I - \frac{\eta}{b}\sum_{j\in S_i}y_j^2\ell''_\sigma\left(y_j a_j^T w\right)a_j a_j^T,$$

with

$$\ell''_\sigma(z) = \frac{1}{\sigma}\frac{e^{-(1-z)/\sigma}}{(1+e^{-(1-z)/\sigma})^2} \geq 0, \quad \|\ell''_\sigma\|_\infty = \ell''_\sigma(1) = \frac{1}{4\rho}.$$

Therefore, if $\eta \in (0, \frac{1}{\lambda+\|R\|^2/(4\rho)})$ and $R := \max_i \|a_i\|$, then

$$1-\eta\lambda-\eta\frac{1}{4\rho}R^2 \le \|J_{h_i}(w)\| \le 1-\eta\lambda.$$

This implies that

$$\overline{\dim}_{\mathrm{H}}\mu_{W|\mathbf{S}_n} \le \frac{\log m_b}{\log(1/(1-\eta\lambda))} = \frac{\log(n/b)}{\log(1/(1-\eta\lambda))}, \tag{S73}$$

where we used (S62). The proof is complete. □

S8.7 Proof of Proposition 5

Proof. We recall that the loss is given by:

$$\ell(w,z_i) := \|y_i-\hat{y}_i\|^2 + \lambda\|w\|^2/2, \quad \hat{y}_i := \sum\nolimits_{r=1}^{m} b_r\sigma\left(w_r^T a_i\right), \tag{S74}$$

where the non-linearity $\sigma : \mathbb{R} \to \mathbb{R}$ is smooth and $\lambda > 0$ is a regularization parameter. Note that we can re-write (S74) as $\ell(w,z_i) = \left\|y_i - b^T\sigma\left(w_r^T a_i\right)\right\|^2 + \lambda\|w\|^2/2$. We can compute that

$$\frac{\partial\ell(w,z_i)}{\partial w_r} = -(y_i-\hat{y}_i)\frac{\partial\hat{y}_i}{\partial w_r} + \lambda w_r = -(y_i-\hat{y}_i)b_r\sigma'(w_r^T a_i)a_i + \lambda w_r. \tag{S75}$$

Therefore,

$$\nabla\ell(w,z_i) = -(y_i-\hat{y}_i)v_i + \lambda w, \quad \text{where} \quad v_i := \begin{bmatrix} b_1\sigma'(w_1^T a_i)a_i \\ b_2\sigma'(w_2^T a_i)a_i \\ \cdots \\ b_m\sigma'(w_m^T a_i)a_i \end{bmatrix},$$

with

$$h_i(w) = w - \frac{\eta}{b}\sum_{j\in S_i}\nabla\ell(w,z_i),$$

and

$$\begin{aligned} J_{h_i}(w) &= (1-\eta\lambda)I - \frac{\eta}{b}\sum_{j\in S_i} v_j \otimes v_j^T \\ &\quad + \frac{\eta}{b}\sum_{j\in S_i}(y_j-\hat{y}_j)\begin{bmatrix} b_1\sigma''(w_1^T a_j)a_j a_j^T & 0_d & \cdots & 0_d \\ 0_d & b_2\sigma''(w_2^T a_j)a_j a_j^T & \cdots & 0_d \\ \vdots & \vdots & \ddots & \vdots \\ 0_d & 0_d & \cdots & b_m\sigma''(w_m^T a_j)a_j a_j^T \end{bmatrix} \\ &= (1-\eta\lambda)I - \frac{\eta}{b}\sum_{j\in S_i}\mathrm{diag}(\{B_r^{(j)}\}_{r=1}^m), \end{aligned} \tag{S76}$$

where $\{S_i\}_{i=1}^{m_b}$ is a partition of $\{1,2,\ldots,n\}$ with $|S_i| = b$, where $i = 1,2,\ldots,m_b$ with $m_b = n/b$, and

$$B_r^{(i)} := b_r\left[-(y_i-\hat{y}_i)\sigma''(w_r^T a_i) + (\sigma'(w_r^T a_i))^2\right]a_i a_i^T, \tag{S77}$$

and 0_d is a $d\times d$ zero matrix and $\mathrm{diag}(\{B_r^{(i)}\}_{r=1}^m)$ denotes a block diagonal matrix with the matrices $B_r^{(i)}$ on the diagonal. We assume the output y_i and the activation function σ and its second derivative σ'' is bounded.[5] This would for instance clearly hold for classification problems where y_i can take integer values on a compact set with a sigmoid or hyperbolic tangent activation function. Then, under this assumption, there exists a constant $M_y > 0$ such that $\max_i \|y_i - \hat{y}_i\| \le M_y$. Then for $\eta \in (0, \frac{1}{2\lambda})$ and $\lambda > C$ where $C := M_y\|b\|_\infty\|\sigma''\|_\infty R^2 + (\max_j \|v_j\|_\infty)^2$, we get

$$1-\eta(C+\lambda) \le \|J_{h_i}(w)\| \le 1-\eta(\lambda-C).$$

This implies that

$$\overline{\dim}_{\mathrm{H}}\mu_{W|\mathbf{S}_n} \le \frac{\log m_b}{\log(1/(1-\eta(\lambda-C)))} = \frac{\log(n/b)}{\log(1/(1-\eta(\lambda-C)))}, \tag{S78}$$

where we used (S62). The proof is complete. □

[5] Since the final layer is fixed at initialization, the output is bounded.

S8.8 Proof of Proposition S8

Proof. Recall that H is positive-definite and there exist some $m, M > 0$:

$$0 \prec mI \preceq H \preceq MI. \tag{S79}$$

We have

$$h_i(u) - h_i(v) = M_i(u - v),$$

where

$$0 \preceq \left(1 - \eta\lambda m^{-1} - \eta m^{-1} R_i^2\right) I \preceq M_i \preceq \left(1 - \eta\lambda M^{-1}\right) I, \tag{S80}$$

where $R_i := \max_{j\in S_i} \|a_j\|$, and we recall the assumption that $\eta < \frac{m}{R^2+\lambda}$, with $R := \max_i R_i$. Hence, h_i is bi-Lipschitz in the sense of [Anc16] where

$$\gamma_i(u - v) \leq \|h_i(u) - h_i(v)\| \leq \Gamma_i(u - v),$$

with

$$\gamma_i = \min\left(\left|1 - \eta\lambda m^{-1} - \eta m^{-1} R_i^2\right|, \left|1 - \eta M^{-1}\lambda\right|\right), \tag{S81}$$

$$\Gamma_i = \max\left(\left|1 - \eta\lambda m^{-1} - \eta m^{-1} R_i^2\right|, \left|1 - \eta M^{-1}\lambda\right|\right) < 1, \tag{S82}$$

as long as $\gamma_i > 0$. We recall the assumption $\eta < \frac{m}{R^2+\lambda}$, where $R := \max_i R_i$, in which case the expressions for γ_i and Γ_i simplify to:

$$\gamma_i = 1 - \eta m^{-1}\lambda - \eta m^{-1} R_i^2, \quad \Gamma_i = 1 - \eta M^{-1}\lambda.$$

In this case, it is easy to see that

$$0 < \gamma_i \leq \|J_{h_i}(w)\| \leq \Gamma_i < 1,$$

and it follows from Theorem S2 that

$$\overline{\dim}_{\mathrm{H}}\mu_{W|\mathbf{S}_n} \leq \frac{\mathcal{E}}{\sum_{i=1}^{m_b} p_i \log(\Gamma_i)} \leq \frac{\log m_b}{\log(1/(1 - \eta M^{-1}\lambda))} = \frac{\log(n/b)}{\log(1/(1 - \eta M^{-1}\lambda))}, \tag{S83}$$

where we used (S62). The proof is complete. □

S8.9 Proof of Proposition S9

Proof. Similar as in the proof of Proposition 3, we can compute that the Jacobian is given by

$$J_{h_i}(w) = \frac{1}{b}\sum_{j\in S_i}\left(1 - \eta H^{-1}\lambda + \eta H^{-1} y_j^2 \left[\frac{e^{-y_j a_j^T w}}{(1 + e^{-y_j a_j^T w})^2}\right] a_j a_j^T\right), \tag{S84}$$

where $\{S_i\}_{i=1}^{m_b}$ is a partition of $\{1, 2, \ldots, n\}$ with $|S_i| = b$, where $i = 1, 2, \ldots, m_b$ with $m_b = n/b$, and H is a positive-definite matrix with $0 \prec mI \preceq H \preceq MI$. recall that the input data is bounded, i.e. $\max_{j\in S_i} \|a_j\| \leq R_i$ for some R_i, and $R := \max_i R_i$ satisfying $R < 2\sqrt{m\lambda/M}$. Also recall the step-size is sufficiently small, i.e. $\eta < m/\lambda$. One can provide upper bounds and lower bounds on $J_{h_i}(w)$:

$$\|J_{h_i}(w)\| \leq \Gamma_i := 1 - \eta M^{-1}\lambda + \frac{1}{4}\eta m^{-1} R_i^2, \tag{S85}$$

$$\|J_{h_i}(w)\| \geq \gamma_i := 1 - \eta m^{-1}\lambda, \tag{S86}$$

so that

$$\overline{\dim}_{\mathrm{H}}\mu_{W|\mathbf{S}_n} \leq \frac{-\mathcal{E}}{\sum_{i=1}^{m_b} p_i \log(1/(1 - \eta M^{-1}\lambda + \frac{1}{4}\eta m^{-1} R_i^2))}$$

$$\leq \frac{\log m_b}{\log(1/(1 - \eta M^{-1}\lambda + \frac{1}{4}\eta m^{-1} R^2))} \tag{S87}$$

$$= \frac{b\log(n/b)}{\log(1/(1 - \eta M^{-1}\lambda + \frac{1}{4}\eta m^{-1} R^2))}, \tag{S88}$$

where we used (S62) in (S87). The proof is complete. □

S8.10 Proof of Proposition S10

Proof. Similar as in the proof of Proposition 4, we can compute that

$$J_{h_i}(w) = \frac{1}{b}\sum_{j\in S_i}\left(I - \eta H^{-1}\lambda\right) - \eta H^{-1} a_j a_j^T \rho''\left(y_j - \langle w, a_j\rangle\right), \tag{S89}$$

where $\{S_i\}_{i=1}^{m_b}$ is a partition of $\{1, 2, \ldots, n\}$ with $|S_i| = b$, where $i = 1, 2, \ldots, m_b$ with $m_b = n/b$, and H is a positive-definite matrix with $0 \prec mI \preceq H \preceq MI$. For the function ρ, a standard choice is exponential squared loss: $\rho_{\exp}(t) = 1 - e^{-|t|^2/t_0}$, where $t_0 > 0$ is a tuning parameter. We can compute that $\|\rho''_{\exp}\|_\infty = \rho''_{\exp}(0) = \frac{2}{t_0}$. Therefore, for $\eta \in (0, \frac{m}{\lambda + R^2(2/t_0)})$,

$$0 < 1 - \eta m^{-1}\lambda - \eta m^{-1}R^2\frac{2}{t_0} \leq \|J_{h_i}(w)\| \leq 1 - \eta M^{-1}\lambda + \eta m^{-1}R^2\frac{2}{t_0}, \tag{S90}$$

where $R = \max_i \|a_i\|$ and we recall that $R < \sqrt{\lambda t_0 m/(2M)}$. We have

$$\overline{\dim}_{\mathrm{H}}\mu_{W|\mathbf{S}_n} \leq \frac{\log m_b}{\log(1/(1 - \eta M^{-1}\lambda + \eta m^{-1}R^2\frac{2}{t_0}))} = \frac{\log(n/b)}{\log(1/(1 - \eta M^{-1}\lambda + \eta m^{-1}R^2\frac{2}{t_0}))}, \tag{S91}$$

where we used (S62). The proof is complete. □

S8.11 Proof of Proposition S11

Proof. Similar as in the proof of Proposition S7, we can compute that

$$J_{h_i}(w) = I - \frac{\eta}{b}H^{-1}\sum_{j\in S_i}\nabla^2\ell(w, z_j) = (1 - \eta\lambda H^{-1})I - \frac{\eta}{b}H^{-1}\sum_{j\in S_i} y_j^2 \ell''_\sigma\left(y_j a_j^T w\right) a_j a_j^T,$$

where $\{S_i\}_{i=1}^{m_b}$ is a partition of $\{1, 2, \ldots, n\}$ with $|S_i| = b$, where $i = 1, 2, \ldots, m_b$ with $m_b = n/b$, and H is a positive-definite matrix with $0 \prec mI \preceq H \preceq MI$, and

$$\ell''_\sigma(z) = \frac{1}{\sigma}\frac{e^{-(1-z)/\sigma}}{(1 + e^{-(1-z)/\sigma})^2} \geq 0, \quad \|\ell''_\sigma\|_\infty = \ell''_\sigma(1) = \frac{1}{4\rho}.$$

Therefore, if $\eta \in (0, \frac{m}{\lambda + \|R\|^2/(4\rho)})$ where $R := \max_i \|a_i\|$, then

$$1 - \eta m^{-1}\lambda - \eta m^{-1}\frac{1}{4\rho}R^2 \leq \|J_{h_i}(w)\| \leq 1 - \eta M^{-1}\lambda.$$

This implies that

$$\overline{\dim}_{\mathrm{H}}\mu_{W|\mathbf{S}_n} \leq \frac{\log m_b}{\log(1/(1 - \eta M^{-1}\lambda))} = \frac{\log(n/b)}{\log(1/(1 - \eta M^{-1}\lambda))}, \tag{S92}$$

where we use (S62). The proof is complete. □

S8.12 Proof of Proposition S12

Proof. By following the similar derivations as in Proposition 5, we obtain

$$J_{h_i}(w) = \left(1 - \eta\lambda H^{-1}\right) I - \frac{\eta}{b}H^{-1}\sum_{j\in S_i}\mathrm{diag}\left(\left\{B_r^{(j)}\right\}_{r=1}^{m}\right), \tag{S93}$$

where $\{S_i\}_{i=1}^{m_b}$ is a partition of $\{1, 2, \ldots, n\}$ with $|S_i| = b$, where $i = 1, 2, \ldots, m_b$, with $m_b = n/b$, and H is a positive-definite matrix and $0 \prec mI \preceq H \preceq MI$ for some $m, M > 0$, and $\mathrm{diag}(\{B_r^{(i)}\}_{r=1}^m)$ denotes a block diagonal matrix with the matrices $B_r^{(i)}$ on the diagonal defined in Proposition 5. As in Proposition 5, there exists a constant $M_y > 0$ such that $\max_i \|y_i - \hat{y}_i\| \leq M_y$. Then for $\eta \in (0, \frac{m}{C+\lambda})$ and $\lambda > \frac{M}{m}C$ where $C := M_y\|b\|_\infty\|\sigma''\|R^2 + (\max_j \|v_j\|_\infty)^2$, we get

$$1 - \eta m^{-1}(C + \lambda) \leq \|J_{h_i}(w)\| \leq 1 - \eta\left(M^{-1}\lambda - m^{-1}C\right).$$

This implies that

$$\overline{\dim}_{\mathrm{H}}\mu_{W|\mathbf{S}_n} \leq \frac{\log m_b}{\log(1/(1 - \eta(M^{-1}\lambda - m^{-1}C)))} = \frac{\log(n/b)}{\log(1/(1 - \eta(M^{-1}\lambda - m^{-1}C)))}, \tag{S94}$$

where we used (S62). The proof is complete. □