A margin-based multiclass generalization bound via geometric complexity

Michael Munn [*1] **Benoit Dherin** [*1] **Javier Gonzalvo** [1]

Abstract

There has been considerable effort to better understand the generalization capabilities of deep neural networks both as a means to unlock a theoretical understanding of their success as well as providing directions for further improvements. In this paper we investigate margin-based multiclass generalization bounds for neural networks which rely on a recent complexity measure, the geometric complexity, developed for neural networks and which measures the variability of the model function (Dherin et al., 2022).

We derive a new upper bound on the generalization error which scales with the margin-normalized geometric complexity of the network and which holds for a broad family of data distributions and model classes. Our generalization bound is empirically investigated for a ResNet18 model trained with SGD on the CIFAR10 and CIFAR100 datasets with both original and random labels.

1. Introduction

Within the field of machine learning, a model's ability to generalize well to unseen data is one of the key metrics of performance. Modern deep learning techniques have proven successful at this goal across multiple domains and applications. However, it is still not well understood how or why these neural network models exhibit such good generalization capabilities.

Classical statistical learning theory provides a theoretical framework to understand generalization and over the years various complexity measures have been proposed which aim to capture the relationship between generalization and complexity. In this context, the expectation is that lower model complexity should imply tighter generalization gaps. However, many of these well-known complexity measures, such as parameter count or parameter norms are not well-suited to the experimental results that are often observed when training neural networks (Jiang et al., 2019; Zhang et al., 2017). In particular, the phenomena of double descent (Nakkiran et al., 2021; Belkin et al., 2019; Belkin, 2021) clearly illustrates the drawbacks of relying on simple parameter count or model depth to measure model complexity and shows that as model size increases the model is able to perfectly fit the training data while also obtaining low test error. Other more theoretically motivated complexity measures may provide generalization error bounds which are vacuous or computationally intractable to compute. We refer the reader to (Jiang et al., 2019) which provides a well-written discussion of the landscape of current and commonly used complexity measures as well as an extensive empirical analysis of over 40 different generalization measures.

In this paper we focus on a complexity measure recently proposed in (Dherin et al., 2021; 2022). In those papers, the authors show through theoretical and empirical techniques that the geometric complexity is well-suited for the analysis of deep neural networks. Namely, they show that a large number of training and tuning heuristics in deep learning have the advantageous side-effect of decreasing the geometric complexity of the learned solution. These results hint that the geometric complexity can serve as a useful proxy for measuring network performance. Supporting this idea, thorough experimentation in (Novak et al., 2018) has shown empirically that a complexity measure similar to the geometric complexity correlates strongly with model generalization.

1.1. Contributions

Given these positive indicators, one natural question to explore is whether the geometric complexity can serve as an effective means for bounding the generalization error in neural networks. In this paper we show this to be true, and derive new upper bounds on the margin-based multi-class generalization error which scale with the margin-normalized geometric complexity of the network (Theorem 4.2).

This relationship can best be illustrated in Figure 1 which

[*]Equal contribution [1]Google Researh, USA. Correspondence to: Michael Munn <munn@google.com>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

plots the excess risk (test accuracy minus the training accuracy) across multiple epochs of training a ResNet18 model (He et al., 2016) on the CIFAR10 dataset (Krizhevsky et al., 2009); c.f. (Bartlett et al., 2017) which examines a similar behavior for the Lipschitz constant of the network (i.e., the product of the spectral norms of the weight matrices) of AlexNet also trained on CIFAR10. Similar to the behavior of the Lipschitz constant in that setting, note that the geometric complexity (see Definition 3.2 for the precise formulation) is correlated with the excess risk, both when training on original labels and random labels; see also Figure 2 in Appendix D which includes plots demonstrating similar behavior when training ResNet18 on the CIFAR100 dataset.

Our contributions can be summarized as follows:

- Theorem 1.1 below states the main generalization bound that is the basis of this work. It is important to note that this bound has no dependence on artifacts of the network architecture such as number or depth of layers and is multi-class. The full details of the proof can be found in the Appendix section C.
- Our proof relies on a novel covering number argument which ultimately follows from a consequence of an assumption on the underlying data distribution on which the model is trained. Namely, we require that the probability distribution from which data is sampled satisfies a Poincaré inequality (see Definition 3.1). This framework further highlights the data dependence of the complexity measure as well as geometric properties of the model function.
- The theorems we present here provide bounds on the generalization error which depend on the theoretical geometric complexity of the model function (see Definition 3.3). In Section 3.3.1 we show how the empirical geometric complexity (which is computed over a sample) compares to the theoretical geometric complexity (which is measured over the entire data distribution). Namely we show that for Lipschitz functions these two quantities are comparable; see Proposition 3.4.

The theorem below states that for a large class of data distributions (i.e., those that satisfy the Poincaré inequality which include the uniform distribution, Gaussian and mixtures of Gaussian distributions) the generalization error is bounded by the geometric complexity of the network:

Theorem 1.1. *Given a_1, a_2 be positive reals. Let $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ be i.i.d. input-output pairs in $\mathbb{R}^d \times \{1, \cdots, k\}$ and suppose the distribution μ of the x_i satisfies the Poincaré inequality with constant $\rho > 0$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, every margin $\gamma > 0$ and network $f : \mathbb{R}^d \to \mathbb{R}^k$ which satisfies $\mathrm{GC}(f, \mu) \leq a_1$ and $\|\mathbb{E}_\mu(f)\| \leq a_2$ satisfy*

$$\mathbb{P}\left[\arg\min_j f(x)_j \neq y\right] \leq \widehat{\mathcal{R}}_{S,\gamma}(f) + \frac{36\tilde{C}\sqrt{k\pi}}{\gamma m} + 3\sqrt{\frac{\log\frac{2}{\delta}}{2m}}$$

where $\widehat{\mathcal{R}}_{S,\gamma}(f) = m^{-1}\sum_i \mathbb{1}_{y_i f(x_i) \leq \gamma}$ and $\tilde{C} = a_2 + \sqrt{a_1\rho/\delta}$.

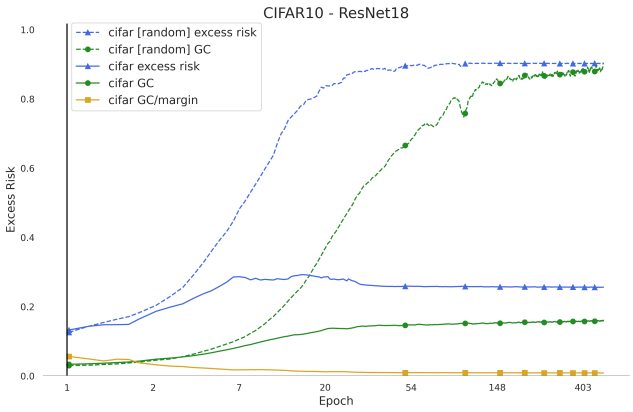

Figure 1. Analysis of ResNet18 (He et al., 2016) trained with SGD on CIFAR10 with both original and with random labels. The blue triangle-marked curves plot the excess risk across training epochs (on a log scale) while the green circle-marked curves track the geometric complexity (GC), normalized so that the two curves for random labels meet. Note that in both settings the GC is closely correlated with the excess risk. Furthermore, normalizing the GC by the margin (i.e., the square-marked curve) neutralizes growth across epochs. Similar plots for CIFAR100 can be found in Section D of the Appendix.

2. Related Work

There has been a considerable amount of work in this direction over the years. Here we survey just a few of the most relevant results as they relate to the current work.

Complexity measures for neural networks and double descent. An interesting aspect of the geometric complexity, as discussed in detail in (Dherin et al., 2022), is that it captures the double-descent phenomena (Belkin et al., 2019; Belkin, 2021; Nakkiran et al., 2021). Namely, when training multiple ResNet18 models on CIFAR10 with increasing layer width, both the test loss and the GC follow a double descent curve as the model width increases; see Figure 5 in

(Dherin et al., 2022). Recent alternative measures of complexity for neural networks, which also appear to correlate with generalization, appear to capture this phenomenon as well (Grant & Wu, 2022; Gamba et al., 2022; Achille & Soatto, 2018; Novak et al., 2018). This suggests interesting connections between the double-descent phenomena and generalization bounds for neural networks.

Implicit regularization and sharpness aware techniques. The generalization power of neural network in spite of their high expressivity has suggested that a form of implicit regularization keeps the model from overfitting (Neyshabur, 2017). Recent work (Barrett & Dherin, 2021; Smith et al., 2021; Ghosh et al., 2023; Ma & Ying, 2021) have shown that the optimization scheme implicitly regularizes the loss gradients favoring flat regions of the loss landscape. In turn these flatter regions with smaller gradients have been connected to smaller geometric complexity (Dherin et al., 2021; 2022). In (Zhang et al., 2023) the increased generalization power of the sharpness-aware optimizer constructed in (Foret et al., 2021) has been shown to rely of a similar mechanism of loss gradient regularization allowing the author to derive a generalization bound involving the loss gradients.

Generalization bounds for neural networks. At a high level this work can be seen as a natural companion of the spectrally-normalized margin bounds obtained for neural networks in (Bartlett et al., 2017). However, as their bound has clear dependence on the network architecture and norms of layer weights, ours is not directly comparable. While this lack of dependence on architecture can be viewed as an advantage of the GC as a complexity measure (primarily in its computation), our work should also be viewed in context of (Nagarajan & Kolter, 2019) which argues that such bounds alone cannot fully explain the generalization abilities of overparameterized neural networks. In particular, our bound does not reflect a width/depth dependence and, further, it remains to explore how our results perform in relation to the vacuity demonstrated by Nagarajan et al. We leave this detailed analysis for future work.

Finally, let us note that a number of other generalization bounds using different notions of complexity have been derived for neural networks; see for instance (Zhang et al., 2023; Ghosh et al., 2023; Foret et al., 2021; Sokolic et al., 2017; Chatterji et al., 2019; Long & Sedghi, 2020) as well as closely related results that approach generalization error bounds using a PAC-Bayes framework which has been shown to yield tighter bounds (Langford & Shawe-Taylor, 2002; McAllester, 1999; Dziugaite & Roy, 2017; Neyshabur et al., 2017).

3. Background and Notation

Let's introduce some notation and recall any relevant background material that is useful going forward.

3.1. Preliminaries

Let $(\Omega, \mathcal{F}, \mathbb{P})$ denote a probability space consisting of the sample space Ω, the σ-algebra $\mathcal{F}$ of subsets of Ω and the probability distribution $\mathbb{P}$ which maps sets of $\mathcal{F}$ to $[0, 1]$. A random variable is a function X from Ω to any set $\mathcal{S}$. In this paper, we are concerned with continuous random variables and thus will often take $\mathcal{S} = \mathbb{R}$ or $\mathbb{R}^d$ where $d > 1$. In this case, we require the function X to be measurable.

The distribution of a random variable $X : \Omega \to \mathcal{S}$ is a probability measure μ on $\mathcal{S}$ defined so that $\mu(A) = \mathbb{P}(X \in A)$ for any measurable set $A \subset \mathcal{S}$. We say the probability distribution μ is continuous provided μ is absolutely continuous with respect to the Lebesgue measure dx on $\mathbb{R}^d$; that is, there is a non-negative measurable function $u : \mathbb{R}^d \to [0, \infty]$ such that $d\mu = u(x)dx$.

For a real valued random variable X with continuous probability distribution μ, we denote the expected value of X and the variance of X with respect to the distribution μ as $\mathbb{E}_\mu[X]$ and $\mathrm{Var}_\mu(X)$, respectively.

Notation We use $\mathbb{1}_A$ to denote the indicator function on the set A. We use the $\|\cdot\|_p$ to denote the L^p-norm on $\mathbb{R}^n$, $\|\cdot\|_F$ denotes the Frobenius norm and $\|\cdot\|$ without a subscript denotes the usual Euclidean norm. $\mathbb{R}_+$ denote the positive reals. We will use S to denote a sample $\{(x_1, y_1), \ldots, (x_m, y_m)\}$ of m input-output pairs, at times just focusing on $S_X = \{x_1, \ldots, x_m\}$ the projection over the inputs, denoted as a dataset D. The sample S is drawn i.i.d. from some fixed but unknown data distribution $\mathscr{D}$ over the input space. In particular, we are concerned with probability data distributions which we denote by μ and their associated probability measure $\mathbb{P}$. For a set $A \subset \mathbb{R}^n$, $|A|$ denotes the Lebesgue measure of A and $\mu(A)$ denotes the measure of A for a measure μ. Given a point $p \in \mathbb{R}^n$ and a radius $r > 0$, let $B_p(r)$ denote the ball of radius r centered at p and we drop the subscript p to denote a ball centered at the origin; e.g. $B(r)$.

3.2. The Poincaré Inequality

Concentration inequalities are a cornerstone of convex geometry and have inspired many remarkable results and applications within numerous areas of theoretical computer science including randomized algorithms, Monte Carlo sampling methods, complexity approximation, and, of course, learning theory (Boucheron et al., 2004; Raginsky et al., 2013; Ledoux, 2001). Indeed, the concentration phenomena has been successfully adopted or adapted across different settings and appears when studying Gaussian space, Riemannian manifolds, discrete product spaces, and algebraic structures.

On a conceptual level, concentration inequalities quantify

the amount of random fluctuations of independent random variables by bounding the probability that such a function differs from its expected value by more than a certain amount (Boucheron et al., 2013; Ledoux, 2001). The Poincaré inequality can be seen as a form of concentration, since it implies that if the function is "spread out" over the domain (i.e., it has a large variance), then its gradient must be correspondingly large. Conversely, if the gradients are small, then the function must be concentrated around its mean.

More precisely, the classic Poincaré inequality provides L^p bounds on the oscillation of a function about its mean via L^p bounds on its derivative. For our purposes, we focus on the case when $p = 2$ though similar statement holds for all $1 \leq p \leq \infty$; see (Evans, 2022). In this case, the classic Poincaré inequality states that, given a connected, bounded domain $U \subset \mathbb{R}^n$ with Lipschitz boundary, there exists a constant $C := C(U) > 0$ such that for every function $u : U \subset \mathbb{R}^n \to \mathbb{R}$ with $\int u^2 dx < \infty$ and $\int \|\nabla u\|^2 dx < \infty$,

$$\|u - \bar{u}\|_{L^2(U)} \leq C \|\nabla u\|_{L^2(U)}, \tag{1}$$

where $\bar{u} = {}^1/_{|U|} \int_U u dx$ is the average value of u over the domain U and $|U|$ denotes the standard Lebesgue measure of $U \subset \mathbb{R}^n$. The optimal constant C in (1) is called the Poincaré constant and is related to the first eigenvalue of the negative Laplacian, computed via the Raleigh Quotient, and depends only on the geometry of U (Chavel, 1984).

The Poincaré inequality plays an important role in geometric and functional analysis and Poincaré type inequalities appear throughout the study of the geometry of metric measure spaces; i.e., metric spaces (X, d) equipped with a Borel measure. In fact, it is an active area of research to better understand which properties of a metric measure space support a Poincaré inequality and its applications in these more abstract settings. There is a long list of metric measure spaces supporting a (local and/or non-local) Poincaré inequality including $\mathbb{R}^n$ (Evans, 2022), Riemannian manifolds with non-negative Ricci curvature (Hebey, 2000) and their Gromov-Hausdorff limits (Lott & Villani, 2007), Carnot groups (Bruno et al., 2022), as well as other non-Riemannian metric measure spaces with fractional Hausdorff dimension (Heinonen et al., 2001). A more thorough overview of the literature would go out of the scope of the present paper, so we also refer the reader to (Heinonen et al., 2001) and the survey (Hajłasz & Koskela, 2000) and the references therein.

Returning to $\mathbb{R}^n$ with the standard Euclidean metric d, we can ask which measures μ on $(\mathbb{R}^n, d)$ satisfy a Poincaré type inequality as in (1) beyond the standard Lesbesgue measure dx. For example, when μ is the standard Gaussian measure on $\mathbb{R}^n$, then for any smooth function $u : \mathbb{R}^n \to \mathbb{R}$, then it is known that (see, for example (Bakry et al., 2014))

$$\mathrm{Var}_\mu(u) \leq \mathbb{E}_\mu \left[|\nabla u|^2 \right] \tag{2}$$

where $\mathrm{Var}_\mu(u) := \int |u - \int u d\mu|^2 d\mu$. This is known as the Gaussian Poincare Inequality and, in this setting, the Poincare constant $C = 1$ is optimal.

In fact, Poincaré type inequalities like (1) and (2) are known to hold for wide collection of measures μ on $\mathbb{R}^n$ and take a simple form, particularly when μ is absolutely continuous with respect to the Lebesgue measure, typically referred to as weighted Poincaré inequalities (Ferone et al., 2012). In (Bakry et al., 2008) the authors prove a Poincaré inequality for a large class of probability measures including log-concave probability measures on $\mathbb{R}^n$. In addition, (Schlichting, 2019) examines conditions on when a Poincaré inequality holds for mixtures of probability measures $\mu = t\mu_0 + (1-t)\mu_1$ for $t \in [0, 1]$ on $\mathbb{R}^n$ and derives explicit bounds on the Poincaré constant for a number of useful examples including mixtures of two Gaussian measures with equal covariance matrix, mixtures of a Gaussian and sub-Gaussian measure, mixtures of two centered Gaussians with different variance, and mixtures of uniform and Gaussian measures.

Whether or not a space supports a Poincaré inequality has deep connections to the geometry and analysis of the space and is closely related to the differentiability of Lipschitz functions in metric spaces (Cheeger, 1999). Following (Schlichting, 2019) and others, we define

Definition 3.1. For $n \geq 1$, a probability measure μ on $\mathbb{R}^n$ satisfies the Poincaré inequality with constant $\rho > 0$, if for all functions $u : \mathbb{R}^n \to \mathbb{R}$,

$$\mathrm{Var}_\mu(u) := \int \left| u - \int u d\mu \right|^2 d\mu \leq \rho \int |\nabla u|^2 d\mu.$$

In this case, we say μ satisfies $\mathrm{PI}(\rho)$.

While the condition that a probability measure satisfies a Poincaré inequality is fairly ubiquitous and can be assumed for a large class of data distributions, it is possible to devise examples that are pathological in the context of machine learning and which violate this property. For example, consider a discrete distribution μ concentrated on a finite number of points $\{x_1, \ldots, x_n\}$ for $x_i \in \mathbb{R}$; that is $\mu = \sum \delta_{x_i}$. Then, a function f which is constant around small neighborhoods of each x_i yet for which $f(x_i) \neq f(x_j)$ for $i \neq j$, will have have zero derivative on the support of μ but the variance of the function f is non-zero.

More generally, consider a distribution whose support consists of more than one connected components $U_1, \ldots, U_n$. Again, take f to be the sum of bump functions which are individually constant on a neighborhood of each U_i but for which $f|_{U_i} \neq f|_{U_j}$ when $i \neq j$. Then, the gradient of f

is zero on each component, which implies $\mathbb{E}_\mu[\|\nabla f|^2] = 0$; however, the variance of f is non-zero by construction. In particular, this points to a necessary condition for a distribution μ to satisfy $\mathrm{PI}(\rho)$. Namely, μ needs to have path connected support which is the case for most image distribution where one can morph one image smoothly into another and, more broadly, consistent with the data manifold hypothesis.

3.2.1. Poincaré and Isoperimetry

Lastly, let us note the connection to the Poincaré inequality and the isoperimetric inequality, particularly recent work on the robustness of model functions via isoperimetry (Bubeck & Sellke, 2023). Classically, in the field of geometric analysis, the two concepts are intimately related; in fact, the isoperimetric inequality can be proven directly from the Poincaré inequality. In the context of machine learning and measure concentration in high-dimensional geometry, it is known that if a probability measure μ on $\mathbb{R}^n$ satisfies $\mathrm{PI}(\rho)$ then the tails of μ are exponentially small (Bobkov & Ledoux, 1997); i.e., $\mu(|x| > t) \leq Ce^{-ct/\sqrt{\rho}}$, for some constants c, C. Essentially, the Poincaré inequality ensures random variables on $\mathbb{R}^n$ are $\sqrt{\rho}$-subgaussian.

As stated in (Bubeck & Sellke, 2023), a measure μ on $\mathbb{R}^n$ satisfies c-isoperimetry if for any bounded L-Lipschitz real-valued function satisfies

$$\mathbb{P}[|f - \mathbb{E}_\mu[f]| \geq t] \leq 2e^{\frac{-nt^2}{2cL^2}},$$

which implies that the output of any suitably rescaled Lipschitz function is in fact $O(1)$-subgaussian. In fact, as Bubeck et al. point out, their isoperimetry assumption is related to a log-Sobolov inequality, which is a strengthening of the regular Poincaré inequality. This suggests that our assumption that the distribution μ of the covariates x_i satisfies a Poincaré inequality as in our Theorem 1.1 is a comparable, or even a more primitive, requirement to that of isoperimetry.

3.3. Geometric Complexity

We now recall the definition of the geometric complexity introduced in (Dherin et al., 2022; 2021):

Definition 3.2 (Empirical Geometric Complexity)**.** Let $g : \mathbb{R}^d \to \mathbb{R}^k$ be a neural network. We can write $g(x) = \sigma(f(x))$ where σ denotes the last layer activation, and f its logit network. The **empirical geometric complexity** of the network over a dataset D is defined to be the discrete Dirichlet energy of its logit network over the dataset:

$$\mathrm{GC}(f, D) = \frac{1}{|D|} \sum_{x \in D} \|\nabla_x f(x)\|_F^2, \tag{3}$$

where $\|\nabla_x f(x)\|_F$ is the Frobenius norm of the network Jacobian.

In the case of simple linear transformations $f(x) = Ax + b$, the geometric complexity takes a very simple form; namely, $\mathrm{GC}(f, D) = \|A\|_F^2$. More generally, for Lipschitz functions, the geometric complexity is upper bounded by the square of Lipschitz constant. In fact, it is easy to construct examples for which this upper bound is strict. For example, consider the function such that $f|_{\{x \leq 0\}} = 0$, $f|_{\{x \geq 1\}} = 1$ and $f(x) = x$ otherwise. If the data set D is concentrated on the set $\{x < 0\}$ or $\{x > 1\}$ then $f'(x) = 0$ so $\mathrm{GC}(f, D) = 0$; however, the Lipschitz constant of f is 1. In the context of the generalization bound of this paper, this suggests that the comparable generalization bound proven in (Bartlett et al., 2017) should be looser compared to our Theorem 1.1, but we leave a more detailed comparison of the two bounds for later work.

Given a probability distribution μ on $\mathbb{R}^d$ and assuming the dataset D is drawn as an i.i.d. sample from μ, the geometric complexity in Definition 3.2 is an unbiased estimator of the following quantity

$$\mathbb{E}_\mu[\|\nabla f\|_F^2] = \int \|\nabla f(x)\|_F^2 d\mu(x). \tag{4}$$

We refer to the quantity defined in (4) as the theoretical geometric complexity as opposed to the empirical geometric complexity in Definition 3.2. (Note that (Novak et al., 2018) defines a similar notion of complexity using the full network rather than the logit network.)

Definition 3.3 (Theoretical Geometric Complexity)**.** Let μ denote a probability distribution on $\mathbb{R}^d$ and let $f : \mathbb{R}^d \to \mathbb{R}^k$ be the logit network of a neural network. The **theoretical geometric complexity** of f with respect to μ is the expectation of $\|\nabla_x f(x)\|_F^2$ over μ:

$$\mathrm{GC}(f, \mu) = \mathbb{E}_\mu[\|\nabla_x f\|_F^2]. \tag{5}$$

Both the empirical and theoretical geometric complexity are well-defined for any differentiable model, not only a neural network, and their definitions incorporate information about both the model function and the dataset over which the task is determined.

This being said, it is worth noting that many neural networks are not, rigorously speaking, differentiable everywhere. For example, the ReLU activation is not differentiable at the origin and thus even simple neural networks with ReLU activations are not differentiable. However, it is known by a theorem of Rademacher (see (Federer, 2014), Theorem 3.1.6) that any function $f : \mathbb{R}^m \to \mathbb{R}^n$ which is locally Lipschitz (i.e. the restriction of f to some neighborhood around any point is Lipschitz) is differentiable almost everywhere

with respect to Lebesgue measure. This means that f is differentiable everywhere except at most on a set of measure zero (with respect to Lebesgue). Since the theoretical geometric complexity is defined as an expectation over the distribution μ, which is absolutely continuous with respect to the Lebesgue meausure, any points of non-differentiability of the model function are essentially ignored when computing $\mathrm{GC}(f, \mu)$. Therefore, the study and application of the geometric complexity to complex architectures remains a worthwhile endeavor.

Furthermore, this condition of differentiability may also be relaxed when interpreting the Poincaré inequality as in Definition 3.1 as well. There has been considerable recent research focused on extending the theory of first order calculus and differentiable functions to the realm of general metric measure spaces; see, for example, the seminal work introduced by Cheeger in (Cheeger, 1999) and later (Cheeger & Kleiner, 2009). In (Cheeger, 1999), Cheeger generalizes Rademacher's theorem to a rich class of nontrivial metric measure spaces, named *Lipshitz differentiability spaces*, and shows that any doubling metric measure space satisfying a Poincarè inequality is a Lipschitz differentiable space. It was later shown (Bate & Li, 2018) that, in fact, the condition of satisfying a Poincaré type inequality is necessary; meaning that for any metric measure space (X, d, μ) which is Lipschitz differentiable, the measure μ satisfies a Poincaré type inequality.

3.3.1. Comparing the theoretical and empirical geometric complexity

One natural question that immediately arises is how the theoretical geometric complexity relates to (or can be bounded by) the empirical geometric complexity. In fact, we can show that for Lipschitz functions it is possible to obtain an explicit bound between the two.

Intuitively, the geometric complexity is closely related to the Lipschitz constant of a function though they are inherently different. Recall, the Lipschitz constant of a map $f : \mathbb{R}^d \to \mathbb{R}^k$ is the smallest $L > 0$ such that $\|f(x_1) - f(x_2)\| \leq L\|x_1 - x_2\|$ for all $x_1, x_2 \in \mathbb{R}^d$. Conceptually, the constant L measures the maximal amount of variation allowed by the function f over its entire domain provided the inputs change by a fixed amount. The geometric complexity measures intrinsic stretching of the function, more closely related to the Dirichlet energy, and is either averaged over a dataset (as in the empirical geometric complexity) or over the entire data distribution (as in the theoretical geometric complexity). We show

Proposition 3.4. *Let $f : \mathbb{R}^d \to \mathbb{R}^k$ be an L-Lipschitz map and $D = \{x_i\}_{i=1}^m$ a dataset of $m \geq 1$ points $x_i \in \mathbb{R}^d$ sampled from the continuous probability distribution μ. Then, for any $\delta \geq 0$, with probability at least $1 - \delta/2$, the following holds:*

$$\mathrm{GC}(f, \mu) \leq \mathrm{GC}(f, D) + L\sqrt{\frac{\log \frac{2}{\delta}}{2m}}.$$

Proof. See proof in Section A of the Appendix. □

3.4. Covering Numbers

The covering number for a class of functions plays an important role in Learning Theory and can used to estimate the probability or number of samples required to obtain a given confidence and error bound (Mendelson, 2001). In short, covering numbers provide a measure of the complexity of a class of functions; the larger the covering number the more complex the set of functions is.

In general, given a metric space (X, d) and a subset $U \subset X$, the **ϵ-covering number** of U, denoted $\mathcal{N}(U, \epsilon, d)$ is the minimal number of ϵ-balls (with respect to the metric d) needed to cover U. A collection $\mathcal{C}_\epsilon$ of elements $u_1, \ldots, u_k \in U$ is said to be an **ϵ-covering** of U if the union of balls $\cup_i B_{u_i}(\epsilon)$ contains U; that is, any $u \in U$, there exists $i \in [k]$ such that $d(u, u_i) \leq \epsilon$.

In this paper we are concerned with the class of differentiable functions $\mathcal{F} \ni f : \mathbb{R}^d \to \mathbb{R}^k$. Given a sample $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$, let $S_X = \{x_1, \ldots, x_m\}$ denote the projection over the inputs of S. We can also represent the input x_i's as a data matrix $X \in \mathbb{R}^{m \times d}$ by collecting the examples as rows of the matrix X. Denote by $\mu_{S_X} = m^{-1} \sum \delta_{\{x_i\}}$ the empirical measure supported on S_X and endow $\mathbb{R}^d$ with the Euclidean structure of $L^2(\mu_{S_X})$, which is isometric to ℓ^2, and let $\mathcal{F}|_S = \{\sum_i^m f(x_i)\mathbb{1}_{x_i} \mid f \in \mathcal{F}\} \subset L^2(\mu_{S_X})$.

For a given function $f \in \mathcal{F}$ we will consider the image $f(X) = \{f(x) \mid x \in X\} \subset \mathbb{R}^k$ of the set S_X and derive a covering number estimate on $\mathcal{F}|_S$; note that $\mathcal{F}|_S = \{f(X) \mid f \in \mathcal{F}\} = \{f(x) \mid f \in \mathcal{F}, x \in X\}$.

It is possible to show (see (Dudley, 1967; Tomczak-Jaegermann, 1989) and the discussion in (Mendelson, 2001)), that the empirical Rademacher complexity of $\mathcal{F}|_S$ can be upper bounded in terms in the covering numbers of $\mathcal{F}$ in $L^2(\mu_{S_X})$. One of the key components of our proof is an application of a variant of this Dudley entropy integral as utilized in (Bartlett et al., 2017).

4. Main Results

In this paper, we are concerned with a multiclass generalization bound for neural networks. We consider a collection of data points $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ sampled from a probability distribution over $\mathbb{R}^d \times \{1, \ldots, k\}$. Given a model function $f : \mathbb{R}^d \to \mathbb{R}^k$, for any input example $x \in \mathbb{R}^d$, the

network output $f(x) \in \mathbb{R}^k$ is converted to a class label $\{1, \ldots, k\}$. For the sake of clarity and simplicity, we will focus on the case when $k = 1$ and $k > 1$ separately though the argument follows the same idea; the details and notation requires just a bit more attention in the $k > 1$ setting.

In general, the argument to prove our generalization bound via geometric compleixty has three steps. **Step 1:** Assuming the Poincaré inequality for the input data distribution, we can show that the covering number of the class of functions with bounded GC is also bounded by a term involving the GC. **Step 2:** This bound on the covering number then allows us to derive a bound on the Rademacher complexity of that function class following a classical argument using the Dudley entropy integral. **Step 3:** We can at this point use the standard generalization bound in terms of the Rademacher complexity to derive our final generalization bound w.r.t. the GC.

4.1. The case $k = 1$

Let's work out carefully the binary classification case; i.e., when $k = 1$ and the neural network maps to $\mathbb{R}$.

Following the proof outline above, the following Lemma accomplishes Step 1. We include the full proof here to illustrate the main application of our concentration inequalities and how exactly they are exploiting given a bound on the geometric complexity of the functions in the class.

Lemma 4.1. *Let μ be a probability measure on $\mathbb{R}^d$ that satisfies a Poincaré inequality with constant $\rho > 0$. Denote by $X \in \mathbb{R}^{m\times d}$ a data matrix consisting of m points $x_i \in \mathbb{R}^d$ for $i \in [m]$. Let a_1, a_2, ϵ be positive reals and let $\mathcal{F}$ denote the class of differentiable maps*

$$\mathcal{F} := \{f : \mathbb{R}^d \to \mathbb{R} \mid \mathrm{GC}(f, \mu) \le a_1, |\mathbb{E}_\mu[f]| \le a_2\}.$$

Then, for any $\delta > 0$, with probability at least $1 - \delta$,

$$\mathcal{N}\left(\{f(X) \mid f \in \mathcal{F})\}, \epsilon, \|\cdot\|_2\right) \le \frac{1}{\epsilon}\left(a_2 + \sqrt{\frac{a_1\rho}{\delta}}\right).$$

Proof. Given $f \in \mathcal{F}$, define a new function $\tilde{f}(x) = f(x) - \mathbb{E}_\mu[f]$. Then $\mathbb{E}_\mu[\tilde{f}] = 0$ and $\mathrm{GC}(\tilde{f}, \mu) = \mathrm{GC}(f, \mu) \le a_1$. By Chebyshev's inequality and since μ satisfies $\mathrm{PI}(\rho)$, for any $t \in \mathbb{R}_+$, we have

$$\mathbb{P}\big(|\tilde{f}| \le t\big) \ge 1 - \frac{\mathrm{Var}_\mu(\tilde{f})}{t^2} \ge 1 - \frac{\rho\,\mathrm{GC}(\tilde{f}, \mu)}{t^2} \ge 1 - \frac{a_1\rho}{t^2}.$$

Taking $\delta = \frac{a_1\rho}{t^2}$ and solving for t we get $t = \sqrt{a_1\rho/\delta}$. Thus it follows that, for $\delta \in (0, 1)$,

$$\mathbb{P}\Big(|\tilde{f}| \le \sqrt{a_1\rho/\delta}\Big) \ge 1 - \delta.$$

Therefore,

$$\mathbb{P}\Big(|f - \mathbb{E}_\mu[f]| \le \sqrt{a_1\rho/\delta}\Big) \ge 1 - \delta,$$

and since $|\mathbb{E}_\mu[f]| \le a_2$, we have

$$\mathbb{P}\Big(|f| \le a_2 + \sqrt{a_1\rho/\delta}\Big) \ge 1 - \delta.$$

By the above argument, for any $f \in \mathcal{F}$ and for any $x \in \mathbb{R}^d$, with probability $1 - \delta$, $f(x)$ is contained inside the interval $\left[-a_2 - \sqrt{a_1\rho/\delta}, a_2 + \sqrt{a_1\rho/\delta}\right]$. Set $L := a_2 + \sqrt{a_1\rho/\delta}$; therefore, $\{f(x) \mid f \in \mathcal{F}, x \in X\} \subset [-L, L]$.

To find the size of an ϵ-covering of the interval $[-L, L]$, one can divide the interval into intervals of length 2ϵ. This bounds the size of the ϵ-covering of the convex hull of $f(X) = \{f(x) \mid x \in X\}$ of points $f(x_i) \in \mathbb{R}$ by L/ϵ. Thus, we get

$$\mathcal{N}(\{f(X) \mid f \in \mathcal{F}\}, \epsilon, \|\cdot\|_2) \le L/\epsilon = \frac{1}{\epsilon}\left(a_2 + \sqrt{\frac{a_1\rho}{\delta}}\right).$$

□

4.1.1. Confidence margin analysis

The confidence margin of a real-valued function f at a data point (x, y) is given by $yf(x)$ where we interpret the magnitude $|f(x)|$ as the confidence of the prediction when f classifies x correctly; i.e., $yf(x) > 0$. For any $\gamma > 0$, the γ-margin loss function, or ramp loss, penalizes f both when it misclassifies a point and when it correctly classifies x with confidence less than γ; i.e. $yf(x) \le \gamma$. We define the ramp loss $\ell_\gamma : \mathbb{R} \to \mathbb{R}_+$ as

$$\ell_\gamma(r) := \begin{cases} 0 & r < -\gamma, \\ 1 + r/\gamma & r \in [-\gamma, 0], \\ 1 & r > 0, \end{cases}$$

The parameter γ is the confidence margin demanded from a hypothesis f. Given a sample $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ the empirical margin loss is defined by

$$\widehat{\mathcal{R}}_{S,\gamma}(f) := m^{-1}\sum_{i=1}^{m} \ell_\gamma(-y_i f(x_i)).$$

We arrive at the following generalization bound:

Theorem 4.2. *Given a_1, a_2 be positive reals. Let $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ be i.i.d. input-output pairs in $\mathbb{R}^d \times \{\pm 1\}$ and suppose the distribution μ of the x_i satisfies the Poincaré inequality with constant $\rho > 0$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, every margin $\gamma > 0$*

and network $f : \mathbb{R}^d \to \mathbb{R}$ *which satisfies* $\mathrm{GC}(f, \mu) \leq a_1$ *and* $|\mathbb{E}_\mu(f)| \leq a_2$

$$\mathbb{P}\left[yf(x) \leq 0\right] \leq \widehat{\mathcal{R}}_{S,\gamma}(f) + \frac{12\tilde{C}\sqrt{\pi}}{\gamma m} + 3\sqrt{\frac{\log \frac{2}{\delta}}{2m}}$$

where $\widehat{\mathcal{R}}_{S,\gamma}(f) = m^{-1}\sum_i \mathbb{1}[y_i f(x_i) \leq \gamma]$ *and* $\tilde{C} = a_2 + \sqrt{a_1\rho/\delta}$.

See Appendix B for the complete proof of Theorem 4.2.

4.2. The case when $k > 1$

Most of the argument for the case $k > 1$ follows as above. However, in the multi-class setting the margin-based bound takes a slightly modified form c.f. (Bartlett et al., 2017; Mohri et al., 2018). In the multi-class setting the label associated to a data point x is the one resulting in the highest score from the model. This leads to the definition of the margin operator $\mathcal{M} : \mathbb{R}^k \times \{1, \ldots, k\} \to \mathbb{R}$ defined as $\mathcal{M}(v, y) := v_y - \max_{i \neq y} v_i$ where v_i denotes the i-th component of the vector v. Thus, given a funciton $f : \mathbb{R}^d \to \mathbb{R}^k$, the margin of the function at a labeled example (x, y) is

$$\mathcal{M}(f(x), y) = f(x)_y - \max_{i \neq y} f(x)_i$$

and f misclassifies the datapoint (x, y) iff $\mathcal{M}(f(x), y) \leq 0$. Similarly to the binary classification case $k = 1$, given a sample $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ and a confidence margin $\gamma > 0$, we can define the empirical margin loss of f for multi-class classification as

$$\widehat{\mathcal{R}}_{S,\gamma}(f) := m^{-1}\sum_{i=1}^{m} \ell_\gamma(-\mathcal{M}(f(x_i), y_i)).$$

Thus, the empirical margin loss $\widehat{\mathcal{R}}_{S,\gamma}(f)$ is upper bounded by the fraction of training examples that are either misclassified by f or that are correctly classified but with confidence below the threshold γ.

Similarly to Lemma 4.1, we have

Lemma 4.3. *Let* μ *be a probability measure on* $\mathbb{R}^d$ *that satisfies a multi-dimensional Poincaré inequality with constant* $\rho > 0$ *for maps* $f : \mathbb{R}^d \to \mathbb{R}^k$*, for* $k > 1$*. Denote by* $X \in \mathbb{R}^{m \times d}$ *a data matrix consisting of points* $x_i \in \mathbb{R}^d$ *for* $i \in [m]$*. Given positive reals* a_1, a_2*, and* ϵ *, let* $\mathcal{F}$ *denote the class of differentiable maps*

$$\mathcal{F} := \{f : \mathbb{R}^d \to \mathbb{R}^k \mid \mathrm{GC}(f, \mu) \leq a_1, \|\mathbb{E}_\mu[f]\| \leq a_2\}.$$

Then, for any $\delta > 0$*, with probability at least* $1 - \delta$*,*

$$\mathcal{N}(\{f(X) \mid f \in \mathcal{F}\}, \epsilon, \|\cdot\|_2) \leq \frac{3^k}{\epsilon^k}\left(a_2 + \sqrt{\frac{a_1\rho}{\delta}}\right)^k.$$

The proof follows a similar idea behind the proof of Lemma 4.1, taking care to adapt the argument to the setting of $\mathbb{R}^k$ and is contained in Appendix C.

With this Lemma we can now prove our multi-class margin based generalization bound, Theorem 1.1. The full proof is detailed in the Appendix C.

5. Discussion and Conclusion

In this paper we investigate margin-based multiclass generalization bounds for neural networks and show that the generalization error can be upper bounded by the geometric complexity of the learned model function. Indeed, we see that lower geometric complexity implies tighter generalization gaps. Furthermore, these upper bounds on the generalization error scale with the margin-normalized geometric complexity of the network.

In our experiments we train a ResNet18 model on the CIFAR10 and CIFAR100 datasets with original and random labels and see that indeed the excess risk during training is closely correlated with the geometric complexity of the model function in both settings, meaning they the two quantities tend to increase at a comparable rate during training

We also show how the empirical geometric complexity (computed over a data sample) compares to the theoretical geometric complexity (measured over the entire data distribution). Namely we show that for Lipschitz functions these two quantities are comparable.

Our results hold for a broad family of data distributions and model classes. Namely, we require that the distributions from which our data is sampled satisfies a Poincaré inequality. The Poincaré inequality plays an important role in geometric and functional analysis and Poincaré type inequalities appear throughout the study of the geometry of metric measure spaces. Furthermore, due to the nature of the definition of the geometric complexity, our bounds have no dependence on artifacts of the network architecture such as number or depth of layers.

We believe that this work gives a interesting and useful application of a recently proposed complexity measure for neural networks. Furthermore, our approach and proof techniques highlight the role and importance of the geometry of the underlying data distribution. We hope that our work will inspire further research in this area.

6. Acknowledgements

The authors would like to thank thank Scott Yak, Mehryar Mori and Peter Bartlett for many helpful discussions during the course of this work, as well as the reviewers for thoughtful and helpful feedback, which greatly improved the exposition.

References

Achille, A. and Soatto, S. Emergence of invariance and disentanglement in deep representations. In *2018 Information Theory and Applications Workshop (ITA)*, pp. 1–9, 2018. doi: 10.1109/ITA.2018.8503149.

Bakry, D., Barthe, F., Cattiaux, P., and Guillin, A. A simple proof of the poincaré inequality for a large class of probability measures. 2008.

Bakry, D., Gentil, I., Ledoux, M., et al. *Analysis and geometry of Markov diffusion operators*, volume 103. Springer, 2014.

Barrett, D. G. and Dherin, B. Implicit gradient regularization. In *International Conference on Learning Representations*, 2021.

Bartlett, P. L., Foster, D. J., and Telgarsky, M. J. Spectrally-normalized margin bounds for neural networks. *Advances in neural information processing systems*, 30, 2017.

Bate, D. and Li, S. Differentiability and poincaré-type inequalities in metric measure spaces. *Advances in Mathematics*, 333:868–930, 2018.

Belkin, M. Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation. 2021. URL https://arxiv.org/abs/2105.14368.

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

Bobkov, S. and Ledoux, M. Poincaré's inequalities and talagrand's concentration phenomenon for the exponential distribution. *Probability Theory and Related Fields*, 107: 383–400, 1997.

Boucheron, S., Lugosi, G., and Bousquet, O. Concentration inequalities. *Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia, February 2-14, 2003, Tübingen, Germany, August 4-16, 2003, Revised Lectures*, pp. 208–240, 2004.

Boucheron, S., Lugosi, G., and Massart, P. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

Bruno, T., Peloso, M. M., and Vallarino, M. Local and non-local poincaré inequalities on lie groups. *Bulletin of the London Mathematical Society*, 54(6):2162–2173, 2022.

Bubeck, S. and Sellke, M. A universal law of robustness via isoperimetry. *Journal of the ACM*, 70(2):1–18, 2023.

Chatterji, N. S., Neyshabur, B., and Sedghi, H. The intriguing role of module criticality in the generalization of deep networks. *ArXiv*, abs/1912.00528, 2019.

Chavel, I. *Eigenvalues in Riemannian geometry*. Academic press, 1984.

Cheeger, J. Differentiability of lipschitz functions on metric measure spaces. *Geometric & Functional Analysis GAFA*, 9:428–517, 1999.

Cheeger, J. and Kleiner, B. Differentiability of lipschitz maps from metric measure spaces to banach spaces with the radon–nikodym property. *Geometric and Functional Analysis*, 19(4):1017, 2009.

Dherin, B., Munn, M., and Barrett, D. G. The geometric occam's razor implicit in deep learning. *arXiv preprint arXiv:2111.15090*, 2021.

Dherin, B., Munn, M., Rosca, M., and Barrett, D. G. Why neural networks find simple solutions: the many regularizers of geometric complexity. *arXiv preprint arXiv:2209.13083*, 2022.

Dudley, R. M. The sizes of compact subsets of hilbert space and continuity of gaussian processes. *Journal of Functional Analysis*, 1(3):290–330, 1967.

Dziugaite, G. K. and Roy, D. M. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *arXiv preprint arXiv:1703.11008*, 2017.

Evans, L. C. *Partial differential equations*, volume 19. American Mathematical Society, 2022.

Federer, H. *Geometric measure theory*. Springer, 2014.

Ferone, V., Nitsch, C., and Trombetti, C. A remark on optimal weighted poincaré inequalities for convex domains. *Rendiconti Lincei*, 23(4):467–475, 2012.

Foret, P., Kleiner, A., Mobahi, H., and Neyshabur, B. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=6Tm1mposlrM.

Gamba, M., Englesson, E., Björkman, M., and Azizpour, H. Deep double descent via smooth interpolation. *arXiv preprint arXiv:2209.10080*, 2022.

Ghosh, A., Lyu, H., Zhang, X., and Wang, R. Implicit regularization in heavy-ball momentum accelerated stochastic gradient descent. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=ZzdBhtEH9yB.

Grant, E. and Wu, Y. Predicting generalization with degrees of freedom in neural networks. In *ICML 2022 2nd AI for Science Workshop*, 2022.

Hajłasz, P. and Koskela, P. *Sobolev met poincaré*, volume 688. American Mathematical Soc., 2000.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Hebey, E. *Nonlinear analysis on manifolds: Sobolev spaces and inequalities: Sobolev spaces and inequalities*, volume 5. American Mathematical Soc., 2000.

Heinonen, J. et al. *Lectures on analysis on metric spaces*. Springer Science & Business Media, 2001.

Jiang, Y., Neyshabur, B., Mobahi, H., Krishnan, D., and Bengio, S. Fantastic generalization measures and where to find them. *arXiv preprint arXiv:1912.02178*, 2019.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Langford, J. and Shawe-Taylor, J. Pac-bayes & margins. *Advances in neural information processing systems*, 15, 2002.

Ledoux, M. *The concentration of measure phenomenon*. Number 89. American Mathematical Soc., 2001.

Long, P. M. and Sedghi, H. Generalization bounds for deep convolutional neural networks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=r1e_FpNFDr.

Lott, J. and Villani, C. Weak curvature conditions and functional inequalities. *Journal of Functional Analysis*, 245(1):311–333, 2007.

Ma, C. and Ying, L. The sobolev regularization effect of stochastic gradient descent. 2021. URL https://arxiv.org/abs/2105.13462.

McAllester, D. A. Pac-bayesian model averaging. In *Proceedings of the twelfth annual conference on Computational learning theory*, pp. 164–170, 1999.

Mendelson, S. Geometric methods in the analysis of glivenko-cantelli classes. In *Computational Learning Theory: 14th Annual Conference on Computational Learning Theory, COLT 2001 and 5th European Conference on Computational Learning Theory, EuroCOLT 2001 Amsterdam, The Netherlands, July 16–19, 2001 Proceedings*, pp. 256–272. Springer, 2001.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. *Foundations of machine learning*. MIT press, 2018.

Nagarajan, V. and Kolter, J. Z. Uniform convergence may be unable to explain generalization in deep learning. *Advances in Neural Information Processing Systems*, 32, 2019.

Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B., and Sutskever, I. Deep double descent: Where bigger models and more data hurt. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124003, 2021.

Neyshabur, B. Implicit regularization in deep learning. *arXiv preprint arXiv:1709.01953*, 2017.

Neyshabur, B., Bhojanapalli, S., and Srebro, N. A pac-bayesian approach to spectrally-normalized margin bounds for neural networks. *arXiv preprint arXiv:1707.09564*, 2017.

Novak, R., Bahri, Y., Abolafia, D. A., Pennington, J., and Sohl-Dickstein, J. Sensitivity and generalization in neural networks: an empirical study. In *International Conference on Learning Representations*, 2018.

Raginsky, M., Sason, I., et al. Concentration of measure inequalities in information theory, communications, and coding. *Foundations and Trends® in Communications and Information Theory*, 10(1-2):1–246, 2013.

Schlichting, A. Poincaré and log–sobolev inequalities for mixtures. *Entropy*, 21(1):89, 2019.

Smith, S. L., Dherin, B., Barrett, D. G., and De, S. On the origin of implicit regularization in stochastic gradient descent. In *International Conference on Learning Representations*, 2021.

Sokolic, J., Giryes, R., Sapiro, G., and Rodrigues, M. R. D. Robust large margin deep neural networks. *IEEE Transactions on Signal Processing*, 65:4265–4280, 2017.

Tomczak-Jaegermann, N. *Banach-Mazur distances and finite-dimensional operator ideals*, volume 38. Longman Sc & Tech, 1989.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. „understanding deep learning requires rethinking generalization “, iclr 2017. *arXiv preprint arXiv:1611.03530*, 2017.

Zhang, T. Statistical analysis of some multi-category large margin classification methods. *Journal of Machine Learning Research*, 5(Oct):1225–1251, 2004.

Zhang, X., Xu, R., Yu, H., Zou, H., and Cui, P. Gradient norm regularizer seeks flat minima and improves generalization, 2023. URL https://openreview.net/forum?id=z4eslwuymzQ.

A. Comparing theoretical and geometric complexity

Proof of Proposition 3.4. For any dataset $D = \{x_i\}_{i=1}^m$ of $m \geq 1$ points $x_i \in \mathbb{R}^d$ drawn as i.i.d. samples from the continuous probability distribution μ over $\mathbb{R}^d$, the empirical geometric complexity over D is denoted by $\mathrm{GC}(f, D)$. We start by showing that

$$\mathbb{E}_{D\sim\mu^m}\left[\mathrm{GC}(f, D)\right] = \mathrm{GC}(f, \mu).$$

In fact, this follows by computation, keeping in mind that μ is a probability distribution and that the points are independently sampled. Note that,

$$\begin{aligned}
\mathbb{E}_{D\sim\mu^m}\left[\mathrm{GC}(f, D)\right] &= \mathbb{E}_{x_1,\ldots,x_m\sim\mu^m}\left[\frac{1}{m}\sum_{i=1}^m \|\nabla_x f(x_i)\|_F^2\right] \\
&= \frac{1}{m}\int_{\mathbb{R}^{m\times d}} \sum_{i=1}^m \|\nabla_x f(x_i)\|_F^2 d\mu^m(x_1, \ldots, x_m) \\
&= \frac{1}{m}\int_{\mathbb{R}^{m\times d}} \sum_{i=1}^m \|\nabla_x f(x_i)\|_F^2 u(x_1)\cdots u(x_m) dx_1 \cdots dx_m \\
&= \frac{1}{m}\sum_{i=1}^m \int_{\mathbb{R}^{(m-1)\times d}} \left[\int_{\mathbb{R}^d} \|\nabla_x f(x_i)\|_F^2 u(x_i) dx_i\right] u(x_1)\cdots \widehat{u(x_i)} \cdots u(x_m) dx_1 \cdots \widehat{dx_i} \cdots dx_m \\
&= \frac{1}{m}\sum_{i=1}^m \left[\int_{\mathbb{R}^d} \|\nabla_x f(x_i)\|_F^2 u(x_i) dx_i\right] \\
&= \frac{1}{m}\sum_{i=1}^m \left[\int_{\mathbb{R}^d} \|\nabla_x f(x_i)\|_F^2 d\mu(x_i)\right] \\
&= \frac{1}{m}\sum_{i=1}^m \mathrm{GC}(f, \mu) \\
&= \mathrm{GC}(f, \mu).
\end{aligned}$$

Let D and D' be two samples of size $m \geq 1$ which differ by exactly one point, say x_i in D and x_i' in D'. Then since the map f is L-Lipschitz we have

$$\mathrm{GC}(f, D) - \mathrm{GC}(f, D') = \frac{1}{m}\left(\|\nabla_x f(x_i)\|_F^2 - \|\nabla_x f(x_i')\|_F^2\right) \leq L^2/m,$$

and similarly, $\mathrm{GC}(f, D') - \mathrm{GC}(f, D) \leq L^2/m$. Thus, $|\,\mathrm{GC}(f, D) - \mathrm{GC}(f, D')| \leq L^2/m$ and by applying McDiarmind's inequality (e.g. (Mohri et al., 2018)), we have that for any $\epsilon > 0$,

$$\mathbb{P}\left[\mathrm{GC}(f, D) - \mathbb{E}_{D\sim\mu^m}[\mathrm{GC}(f, D)] \leq \epsilon\right] \geq 1 - \exp(-2m\epsilon^2/L^2). \tag{6}$$

Thus, since $\mathbb{E}_{D\sim\mu^m}[\mathrm{GC}(f, D)] = \mathrm{GC}(f, \mu)$ and setting $\delta/2 = \exp(-2m\epsilon^2/L^2)$ and substituting for ϵ in (6), we get that for any $\delta > 0$ with probability as least $1 - \delta/2$ the following holds:

$$\mathrm{GC}(f, \mu) \leq \mathrm{GC}(f, D) + L\sqrt{\frac{\log\frac{2}{\delta}}{2m}}.$$

This completes the proof. □

B. Proof of Theorem 4.2

Let us restate the theorem and provide the proof:

Theorem B.1. *Given a_1, a_2 be positive reals. Let $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ be i.i.d. input-output pairs in $\mathbb{R}^d \times \{\pm 1\}$ and suppose the distribution μ of the x_i satisfies the Poincaré inequality with constant $\rho > 0$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, every margin $\gamma > 0$ and network $f : \mathbb{R}^d \to \mathbb{R}$ which satisfies $\mathrm{GC}(f, \mu) \leq a_1$ and $|\mathbb{E}_\mu(f)| \leq a_2$*

$$\mathbb{P}\left[yf(x) \leq 0\right] \leq \widehat{\mathcal{R}}_{S,\gamma}(f) + \frac{12\tilde{C}\sqrt{\pi}}{\gamma m} + 3\sqrt{\frac{\log \frac{2}{\delta}}{2m}}$$

where $\widehat{\mathcal{R}}_{S,\gamma}(f) = m^{-1} \sum_i \mathbb{1}_{y_i f(x_i) \leq \gamma}$ and $\tilde{C} = a_2 + \sqrt{a_1 \rho / \delta}$.

The proof follows by combining fairly standard arguments in the literature. We include the full details here for completeness.

Proof. Let $\mathcal{F}$ denote the class of differentiable maps

$$\mathcal{F} := \{f : \mathbb{R}^d \to \mathbb{R} \mid \mathrm{GC}(f, \mu) \leq a_1, |\mathbb{E}_\mu[f]| \leq a_2\}.$$

and let $\widetilde{\mathcal{F}} = \{z = (x, y) \mapsto yf(x) \mid f \in \mathcal{F}\}$. For any $\gamma > 0$, define

$$\widetilde{\mathcal{F}}_\gamma := \{(x, y) \mapsto \ell_\gamma(-yf(x)) \mid f \in \mathcal{F}\}.$$

Since ℓ_γ has range $[0, 1]$, it follows classic generalization bounds based on the Rademacher complexity (see, for example Theorem 3.3 in (Mohri et al., 2018)) that, for any $\delta > 0$, with probability at least $1 - \delta$ over the draw of an i.i.d. sample S of size m, we have for all $f \in \widetilde{\mathcal{F}}_\gamma$:

$$\mathbb{E}[\ell_\gamma(-yf(x))] \leq \frac{1}{n} \sum_{i=1}^{n} \ell_\gamma(-y_i f(x_i)) + 2\widehat{\mathfrak{R}}_S(\widetilde{\mathcal{F}}_\gamma) + 3\sqrt{\frac{\log \frac{2}{\delta}}{2m}}. \tag{7}$$

We can further simplify the term $\widehat{\mathfrak{R}}_S(\widetilde{\mathcal{F}}_\gamma)$ here. Namely, $\widehat{\mathfrak{R}}_S(\widetilde{\mathcal{F}}_\gamma) = \widehat{\mathfrak{R}}_S(\ell_\gamma \circ \widetilde{\mathcal{F}})$ and, since the ramp loss ℓ_γ is $1/\gamma$-Lipschitz, by Talagrand's lemma (e.g. see Lemma 5.7 of (Mohri et al., 2018)), the empirical Rademacher complexity of $\ell_\gamma \circ \widetilde{\mathcal{F}}$ can be bounded in terms of the empirical Rademacher complexity of the original hypothesis set $\widetilde{\mathcal{F}}$; that is,

$$\widehat{\mathfrak{R}}_S(\ell_\gamma \circ \widetilde{\mathcal{F}}_\gamma) \leq \frac{1}{\gamma} \widehat{\mathfrak{R}}_S(\widetilde{\mathcal{F}}). \tag{8}$$

Since the $y_i \in \{\pm 1\}$, by computing the empirical Rademacher complexity of $\widetilde{\mathcal{F}}$ over the set S, we also have $\widehat{\mathfrak{R}}_S(\widetilde{\mathcal{F}}) = \widehat{\mathfrak{R}}_S(\mathcal{F})$. Therefore, and by recalling the definition of $\widehat{\mathcal{R}}_{S,\gamma}(f)$, (7) becomes

$$\mathbb{E}_\mu[\ell_\gamma(-yf(x))] \leq \widehat{\mathcal{R}}_{S,\gamma}(f) + \frac{2}{\gamma} \widehat{\mathfrak{R}}_S(\mathcal{F}) + 3\sqrt{\frac{\log \frac{2}{\delta}}{2m}}.$$

Focusing now on the left hand side of (7), note that by definition of the ramp loss, since $\mathbb{1}_{-yf(x) \geq 0} \leq \ell_\gamma(-yf(x))$, we have

$$\mathbb{E}[\mathbb{1}_{-yf(x) \geq 0}] \leq \mathbb{E}[\ell_\gamma(-yf(x))]$$

and $\mathbb{P}\left[yf(x) \leq 0\right] = \mathbb{E}[\mathbb{1}_{-yf(x) \geq 0}]$. Therefore,

$$\mathbb{P}\left[yf(x) \leq 0\right] \leq \widehat{\mathcal{R}}_{S,\gamma}(f) + \frac{2}{\gamma} \widehat{\mathfrak{R}}_S(\mathcal{F}) + 3\sqrt{\frac{\log \frac{2}{\delta}}{2m}}.$$

Furthermore, by definition of the ramp loss, we have that

$$\begin{aligned} \mathbb{P}\left[yf(x) \leq 0\right] &= \mu(-yf(x) \geq 0) \\ &= \mathbb{E}_\mu[\mathbb{1}_{-yf(x) \geq 0}] \\ &\leq \mathbb{E}_\mu[\ell_\gamma(-yf(x))]. \end{aligned}$$

Therefore,

$$\mathbb{P}\left[yf(x) \leq 0\right] \leq \widehat{\mathcal{R}}_{S,\gamma}(f) + \frac{2}{\gamma}\widehat{\mathfrak{R}}_S(\mathcal{F}) + 3\sqrt{\frac{\log\frac{2}{\delta}}{2m}}. \tag{9}$$

To complete the proof we can use a form of the Dudley entropy integral to deduce an upper bound on $\widehat{\mathfrak{R}}_S(\mathcal{F})$. The Dudley entropy integral lemma (see Lemma A.5 of (Bartlett et al., 2017)) states that

$$\widehat{\mathfrak{R}}_S(\mathcal{F}) \leq \inf_{\alpha>0}\left(\frac{4\alpha}{\sqrt{m}} + \frac{12}{m}\int_{\alpha}^{\sqrt{m}}\sqrt{\log\mathcal{N}\left(\mathcal{F}|_S, \epsilon, \|\cdot\|_2\right)}d\epsilon\right).$$

Examining the integral term above, note that $\mathcal{F}|_S = \{f(X) \mid f \in \mathcal{F}\} = \{f(x) \mid f \in \mathcal{F}, x \in X\}$ where $X = S_X$ is the projection of the sample S onto the inputs, so $\mathcal{N}(\mathcal{F}|_S, \epsilon, \|\cdot\|_2) = \mathcal{N}(\{f(X) \mid f \in \mathcal{F}\}, \epsilon, \|\cdot\|_2)$ and, as in Lemma 4.1 taking $\tilde{C} = \tilde{C}(a_1, a_2, \rho, \delta) := a_2 + \sqrt{a_1\rho/\delta}$, it follows that $\mathcal{N}(\{f(X) \mid f \in \mathcal{F}\}, \epsilon, \|\cdot\|_2) = 1$ for all $\epsilon \geq \tilde{C}$ since it requires only one ball of radius greater than or equal to $\tilde{C}$ to cover a ball of radius $\tilde{C}$. Thus, the integrand above is zero for any $\epsilon \geq \tilde{C}$. We can further upper bound this integral by swapping the integral limit $\sqrt{m}$ with $\tilde{C}$ since the integral of a positive function is no greater than the integral of that function over a potentially larger domain. Therefore, we get,

$$\begin{aligned}
\widehat{\mathfrak{R}}_S(\mathcal{F}) &\leq \inf_{\alpha>0}\left\{\frac{4\alpha}{\sqrt{m}} + \frac{12}{m}\int_{\alpha}^{\min(\sqrt{m},\tilde{C})}\sqrt{\log\mathcal{N}(\mathcal{F}|_{\mathcal{S}}, \epsilon, \|\cdot\|_2)}d\epsilon\right\} \\
&\leq \inf_{\alpha>0}\left\{\frac{4\alpha}{\sqrt{m}} + \frac{12}{m}\int_{\alpha}^{\tilde{C}}\sqrt{\log\mathcal{N}(\mathcal{F}|_{\mathcal{S}}, \epsilon, \|\cdot\|_2)}d\epsilon\right\}
\end{aligned}$$

To simplify this, let's first compute the integral term. By Lemma 4.1,

$$\begin{aligned}
\int_{\alpha}^{\tilde{C}}\sqrt{\log\mathcal{N}\left(\mathcal{F}|_S, \epsilon, \|\cdot\|_2\right)}d\epsilon &\leq \int_{\alpha}^{\tilde{C}}\sqrt{\log\left(\tilde{C}/\epsilon\right)}d\epsilon \\
&= \epsilon\sqrt{\log\left(\tilde{C}/\epsilon\right)}\Big|_{\epsilon=\alpha}^{\epsilon=\tilde{C}} - \frac{\tilde{C}\sqrt{\pi}}{2}\operatorname{erf}\left(\sqrt{\log\left(\tilde{C}/\epsilon\right)}\right)\Big|_{\epsilon=\alpha}^{\epsilon=\tilde{C}}
\end{aligned}$$

where erf denotes the error function

$$\operatorname{erf}(z) = \frac{2}{\sqrt{\pi}}\int_0^z e^{-t^2}dt.$$

Evaluating the right hand side fully, we get

$$\begin{aligned}
\int_{\alpha}^{\tilde{C}}\sqrt{\log\mathcal{N}\left(\mathcal{F}|_S, \epsilon, \|\cdot\|_2\right)}d\epsilon &\leq \epsilon\sqrt{\log\left(\tilde{C}/\epsilon\right)}\Big|_{\epsilon=\alpha}^{\epsilon=\tilde{C}} - \frac{\tilde{C}\sqrt{\pi}}{2}\operatorname{erf}\left(\sqrt{\log\left(\tilde{C}/\epsilon\right)}\right)\Big|_{\epsilon=\alpha}^{\epsilon=\tilde{C}} \\
&= -\alpha\sqrt{\log(\tilde{C}/\alpha)} + \frac{\tilde{C}\sqrt{\pi}}{2}\operatorname{erf}\left(\sqrt{\log\left(\tilde{C}/\alpha\right)}\right) \\
&\leq \frac{\tilde{C}\sqrt{\pi}}{2} - \alpha\sqrt{\log(\tilde{C}/\alpha)},
\end{aligned}$$

where in the last inequality we simply used the fact that for any $z > 0$ we have $\operatorname{erf}(z) \leq 1$. Therefore, substituting this back into the entropy bound for $\widehat{\mathfrak{R}}_S(\mathcal{F})$ above, and bounding the $\inf$ by taking the limit α goes to zero; we get,

$$\begin{aligned}
\widehat{\mathfrak{R}}_S(\mathcal{F}) &\leq \inf_{\alpha>0}\left\{\frac{4\alpha}{\sqrt{m}} + \frac{12}{m}\left(\frac{\tilde{C}\sqrt{\pi}}{2} - \alpha\sqrt{\log(\tilde{C}/\alpha)}\right)\right\} \\
&\leq \lim_{\alpha\to 0}\left\{\frac{4\alpha}{\sqrt{m}} + \frac{12}{m}\left(\frac{\tilde{C}\sqrt{\pi}}{2} - \alpha\sqrt{\log(\tilde{C}/\alpha)}\right)\right\} \\
&= \frac{6\tilde{C}\sqrt{\pi}}{m}
\end{aligned}$$

Note that in the inequalities above we are not finding the optimal or tightest upper bounds for $\widehat{\mathfrak{R}}_S(\mathcal{F})$ that are possible. However, given the nature of these expressions it is possible to determine bounds on how sharp these inequalities are. We simply note for the time being that, although these bounds are not sharp, they are not gross overestimates of the true infimum. Finally, substituting this bound on $\widehat{\mathfrak{R}}_S(\mathcal{F})$ into (9) we get.

$$\mathbb{P}\left[yf(x) \leq 0\right] \leq \widehat{\mathcal{R}}_{S,\gamma}(f) + \frac{12\tilde{C}\sqrt{\pi}}{\gamma m} + 3\sqrt{\frac{\log\frac{2}{\delta}}{2m}}$$

which completes the proof.

□

C. Proof of Theorem 1.1

One of the key components of our proof is the Poincaré inequality, originally stated for real-valued functions as in (Evans, 2022). Under similar assumption the Poincaré inequality naturally extends to vector valued maps. We include the proof here.

Let's now detail the proof of the main covering lemma behind Theorem 1.1. As mentioned previously, the proof follows the same logic as the idea as case $k = 1$ only here we need to be a bit more careful about multivariate norms. Note also, the final ball counting argument on the image in $\mathbb{R}^k$ incurs an additional cost resulting in an exponent k which ultimately incurs a cost of a factor $\sqrt{k}$ in our final bound; c.f. (Zhang, 2004).

Proof of Lemma 4.3. Given $f \in \mathcal{F}$, let f^i denote the component functions of f for $i \in [k]$ and define $\tilde{f} : \mathbb{R}^d \to \mathbb{R}^k$ by

$$\tilde{f} := (f^1 - \mathbb{E}_\mu[f^1], \ldots, f^k - \mathbb{E}_\mu[f^k]).$$

Thus, $\mathbb{E}_\mu[\tilde{f}] = 0 \in \mathbb{R}^k$ and $\mathrm{GC}(\tilde{f}, \mu) = \mathrm{GC}(f, \mu) \leq a_1$. Futhermore, by extending Chebyshev's inequality to this multivariate setting, we get that, for any $t \in \mathbb{R}_+$,

$$\mathbb{P}\left[\|\tilde{f}\| \leq t\right] \geq 1 - \frac{\sum_i \mathrm{Var}_\mu(\tilde{f}^i)}{t^2}.$$

Note that by the definition of $\tilde{f}$ and since μ satisfies $\mathrm{PI}(\rho)$, for each $i \in [k]$,

$$\mathrm{Var}_\mu(\tilde{f}^i) = \int |\tilde{f}^i|^2 d\mu \leq \rho \int \|\nabla \tilde{f}^i\|^2 d\mu = \rho\, \mathrm{GC}(\tilde{f}^i, \mu).$$

Furthermore, by the definition of $\mathrm{GC}(f, \mu)$ it follows that

$$\mathrm{GC}(\tilde{f}, \mu) = \int \|\nabla_x \tilde{f}\|_F^2 d\mu = \int \sum_{i,j} \left|\frac{\partial \tilde{f}^i}{\partial x_j}\right|^2 d\mu = \sum_i \mathrm{GC}(\tilde{f}^i, \mu).$$

Using this simplification of $\mathrm{GC}(\tilde{f}, \mu)$ and substituting for $\mathrm{Var}_\mu(\tilde{f}^i)$ in the application of Chebyshev's inequality above, we get

$$\mathbb{P}\left[\|\tilde{f}\| \leq t\right] \geq 1 - \frac{\rho\, \mathrm{GC}(\tilde{f}, \mu)}{t^2} \geq 1 - \frac{a_1 \rho}{t^2}.$$

As before, taking $\delta = a_1\rho/t^2$ and solving for t, we get $t = \sqrt{a_1\rho/\delta}$; thus, for any $\delta \in (0, 1)$, it follows that

$$\mathbb{P}\left[\|\tilde{f}\| \leq \sqrt{a_1\rho/\delta}\right] \geq 1 - \delta.$$

Therefore, since $\|\mathbb{E}_\mu[f]\| \leq a_2$, for any $f \in \mathcal{F}$ with high probability we can bound the image of f within a ball in $\mathbb{R}^k$; namely,

$$\mathbb{P}\left[\|f\| \leq a_2 + \sqrt{a_1\rho/\delta}\right] \geq 1 - \delta.$$

The rest of the argument follows from a standard ball counting argument in $\mathbb{R}^k$. Given $\epsilon > 0$, let $r := a_2 + \sqrt{a_1\rho/\delta}$ and take a maximal set of points $p_i \in B(r)$ such that $\mathrm{dist}(p_i, p_j) > \epsilon$ for $i \neq j$. It follows that $B_{p_i}(\epsilon/2) \cap B_{p_j}(\epsilon/2) = \emptyset$ and

$\cup_i B_{p_i}(\epsilon/2) \subset B(r(1+\epsilon/2))$. Thus, by construction and taking volumes on both sides, $\sum_i |B_{p_i}(\epsilon/2)| \leq |B(r(1+\epsilon/2))|$. Let N denote the number of points p_i and since $|B_{p_i}(\epsilon/2)| = |B(\epsilon/2)|$ for all $i \in [N]$, we get

$$N \leq \frac{|B(r(1+\epsilon/2))|}{|B(\epsilon/2)|} = r^k(1+2/\epsilon)^k.$$

Therefore, for small $\epsilon < 1$, $N \leq r^k(3/\epsilon)^k$ and thus

$$\mathcal{N}(\{f(X) \mid f \in \mathcal{F}\}, \epsilon, \|\cdot\|_2) \leq \frac{3^k}{\epsilon^k}\left(a_2 + \sqrt{\frac{a_1\rho}{\delta}}\right)^k.$$

This completes the proof. □

Using this covering lemma we can now prove our main Theorem 1.1. In fact, the argument follows the same lines as the case $k = 1$ with only slight modification to account for margin operator in the multi-class setting and the application of the Dudley entropy formula when bounding the empiricial Rademacher complexity.

Let us restate the theorem and provide the proof:

Theorem C.1. *Given a_1, a_2 be positive reals. Let $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ be i.i.d. input-output pairs in $\mathbb{R}^d \times \{1, \cdots, k\}$ and suppose the distribution μ of the x_i satisfies the Poincaré inequality with constant $\rho > 0$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, every margin $\gamma > 0$ and network $f : \mathbb{R}^d \to \mathbb{R}^k$ which satisfies $\mathrm{GC}(f, \mu) \leq a_1$ and $\|\mathbb{E}_\mu(f)\| \leq a_2$ satisfy*

$$\mathbb{P}\left[\arg\min_j f(x)_j \neq y\right] \leq \widehat{\mathcal{R}}_{S,\gamma}(f) + \frac{36\tilde{C}\sqrt{k\pi}}{\gamma m} + 3\sqrt{\frac{\log\frac{2}{\delta}}{2m}}$$

where $\widehat{\mathcal{R}}_{S,\gamma}(f) = m^{-1}\sum_i \mathbb{1}_{y_i f(x_i) \leq \gamma}$ and $\tilde{C} = a_2 + \sqrt{a_1\rho/\delta}$.

Proof of Theorem 1.1. Let $\mathcal{F}$ denote the class of differentiable maps

$$\mathcal{F} := \{f : \mathbb{R}^d \to \mathbb{R}^k \mid \mathrm{GC}(f, \mu) \leq a_1, \|\mathbb{E}_\mu[f]\| \leq a_2\}$$

and for any $\gamma > 0$ define

$$\widetilde{\mathcal{F}}_\gamma := \{(x, y) \mapsto \ell_\gamma(-\mathcal{M}(f(x), y) \mid f \in \mathcal{F}\}$$

where $\mathcal{M}(\cdot,\cdot)$ denotes the margin operator $\mathcal{M} : \mathbb{R}^k \times \{1, \ldots, k\} \to \mathbb{R}$ defined by $\mathcal{M}(v, y) = v_y - \max_{i \neq y} v_i$ and $\ell_\gamma : \mathbb{R} \to \mathbb{R}^+$ denotes the usual ramp loss.

Similar to the proof of Theorem 4.2, since ℓ_γ has range $[0, 1]$ and it follows from classic generalization bounds based on the Rademacher complexity (e.g., Theorem 3.3 in (Mohri et al., 2018)) that, for any $\delta > 0$, with probability at least $1 - \delta$ over the draw of an i.i.d. sample S of size m, we have for all $f \in \widetilde{\mathcal{F}}_\gamma$:

$$\mathbb{E}[\ell_\gamma(-\mathcal{M}(f(x), y))] \leq \widehat{\mathcal{R}}_{S,\gamma}(f) + 2\widehat{\mathfrak{R}}_S(\widetilde{\mathcal{F}}_\gamma) + 3\sqrt{\frac{\log\frac{2}{\delta}}{2m}} \tag{10}$$

where now $\widehat{\mathcal{R}}_{S,\gamma}(f) = m^{-1}\sum_i \ell_\gamma(-\mathcal{M}(f(x_i), y_i))$.

We can lower bound the left hand side of (10) (see Lemma A.4 of (Bartlett et al., 2017)) so that $\mathbb{P}\left[\arg\max_i f(x)_i \neq y\right] \leq \mathbb{E}[\ell_\gamma(-\mathcal{M}(f(x), y))]$ and, via Talagrand's lemma, we can also upper bound $\widehat{\mathfrak{R}}_S(\widetilde{\mathcal{F}}_\gamma)$ on the right hand side to get

$$\mathbb{P}\left[\arg\max_i f(x)_i \neq y\right] \leq \widehat{\mathcal{R}}_{S,\gamma}(f) + \frac{2}{\gamma}\widehat{\mathfrak{R}}_S(\mathcal{F}) + 3\sqrt{\frac{\log\frac{2}{\delta}}{2m}}$$

It remains to bound $\widehat{\mathfrak{R}}_S(\mathcal{F})$ which we can again accomplish through the Dudley entropy integral, as in the proof of Theorem 4.2, with only a very slight modification when using the covering number bound afforded by Lemma 4.3. Namely, taking as before $\tilde{C} = \tilde{C}(a_1, a_2, \rho, \delta) := a_2 + \sqrt{a_1\rho/\delta}$, then $\mathcal{N}\left(\mathcal{F}|_S, \epsilon, \|\cdot\|_2\right) \leq (3\tilde{C}/\epsilon)^k$. Following the same argument to

evaluate the integral we can obtain a comparable bound on the empirical Rademacher complexity of $\mathcal{F}$ over S, but now paying a cost of $\sqrt{k}$; i.e.,

$$\widehat{\mathfrak{R}}_S(\mathcal{F}) \leq \inf_{\alpha>0} \left\{ \frac{4\alpha}{\sqrt{m}} + \frac{12\sqrt{k}}{m} \left(\frac{3\tilde{C}\sqrt{\pi}}{2} - \alpha\sqrt{\log(3\tilde{C}/\alpha)} \right) \right\} \leq \frac{18\tilde{C}\sqrt{k}\sqrt{\pi}}{m}.$$

Thus, collecting terms we get

$$\mathbb{P}\left[\arg\min_j f(x)_j \neq y \right] \leq \widehat{\mathcal{R}}_{S,\gamma}(f) + \frac{36\tilde{C}\sqrt{k\pi}}{\gamma m} + 3\sqrt{\frac{\log\frac{2}{\delta}}{2m}}.$$

□

D. Experiment details

We trained a ResNet18 (He et al., 2016) with SGD on CIFAR10 and CIFAR100 with both original and random labels. During training we trained with batch size 256 for 100000 steps with learning rate 0.05. Here we plot the curves for the excess risk (test accuracy - train accuracy) and compare with the geometric complexity during training.

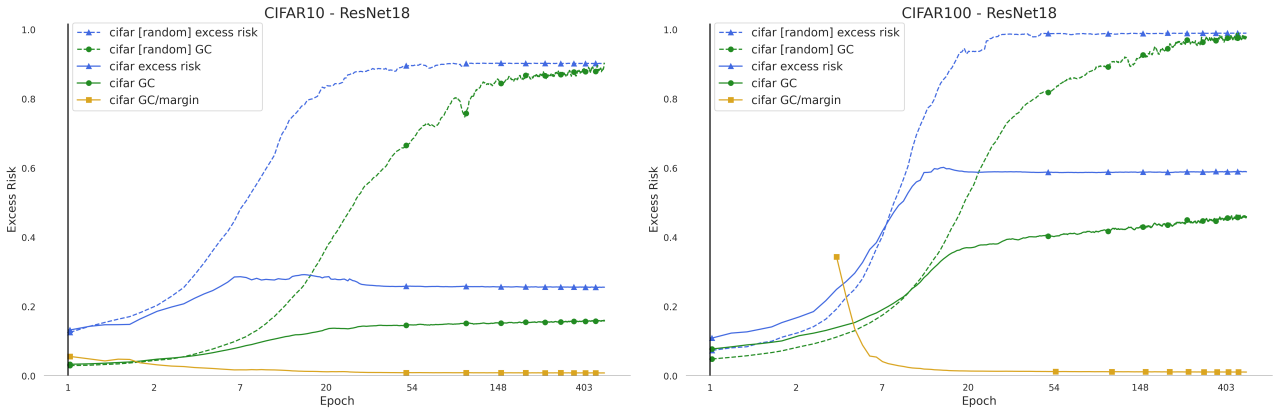

Figure 2. Analysis of ResNet18 (He et al., 2016) trained with SGD on CIFAR10 (left) and CIFAR100 (right) with both original and with random labels. The triangle-marked curves plot the excess risk across training epochs (on a log scale). Circle-marked curves track the geometric complexity (GC). Note that the GC is tightly correlated with the excess risk in both settings. Normalizing the GC by the margin neutralizes growth across epochs.