Blessing of Dimensionality for Approximating Sobolev Classes on Manifolds

Hong Ye Tan[1,2], Subhadip Mukherjee[3], Junqi Tang[4], Carola-Bibiane Schönlieb[2]
[1]UCLA
[2]University of Cambridge
[3]IIT Kharagpur
[4]University of Birmingham
hyt35@math.ucla.edu, smukherjee@ece.iitkgp.ac.in, j.tang.2@bham.ac.uk, cbs31@cam.ac.uk

Abstract

The manifold hypothesis says that natural high-dimensional data lie on or around a low-dimensional manifold. The recent success of statistical and learning-based methods in very high dimensions empirically supports this hypothesis, suggesting that typical worst-case analysis does not provide practical guarantees. A natural step for analysis is thus to assume the manifold hypothesis and derive bounds that are independent of any ambient dimensions that the data may be embedded in. Theoretical implications in this direction have recently been explored in terms of generalization of ReLU networks and convergence of Langevin methods. In this work, we consider optimal uniform approximations with functions of finite statistical complexity. While upper bounds on uniform approximation exist in the literature using ReLU neural networks, we consider the opposite: lower bounds to quantify the fundamental difficulty of approximation on manifolds. In particular, we demonstrate that the statistical complexity required to approximate a class of bounded Sobolev functions on a compact manifold is bounded from below, and moreover that this bound is dependent only on the intrinsic properties of the manifold, such as curvature, volume, and injectivity radius.

1 Introduction

Data is ever growing, especially in the current era of machine learning. However, dimensionality is not always beneficial, and having too many features can confound simpler underlying truths. This is sometimes referred to as the curse of dimensionality (Altman and Krzywinski 2018). A classical example is manifold learning, which is known to scale exponentially in the intrinsic dimension (Narayanan and Niyogi 2009). In the current paradigm of increasing dimensionality, standard statistical tools and machine learning models continue to work, despite the high ambient dimensions arising in cases such as computational imaging (Wainwright 2019). One possible assumption to elucidate this phenomenon comes from the manifold hypothesis, also known as concentration of measure or the blessing of dimensionality (Bengio, Courville, and Vincent 2013). This states that real datasets are actually concentrated on or near low-dimensional manifolds, independently of the ambient dimension that the data is embedded in.

In this work, we explore the consequences of the manifold hypothesis through the lens of approximation theory and statistical complexity. For a class of functions with infinite statistical complexity, we consider a nonlinear width (Definition 2) in terms of how well it can be approximated in L^p with function classes of finite statistical complexity. We consider how difficult it is to optimally approximate classes of functions with functions of finite statistical complexity in terms of L^p distance. In particular, Theorem 2 demonstrates that on a Riemannian manifold, the optimal error incurred by approximating a bounded Sobolev class using function classes of finite statistical complexity can be lower bounded using only the *implicit* properties of the manifold.

1.1 Related Works

The manifold hypothesis is sometimes replaced with the "union of manifolds" hypothesis, where the component manifolds are allowed to have different intrinsic dimension (Vidal 2011; Brown et al. 2022). For estimating the intrinsic dimension, we refer to (Pope et al. 2021; Block et al. 2021; Levina and Bickel 2004; Fefferman, Mitter, and Narayanan 2016); for representing the manifold or dimension reduction, we refer to (Lee, Verleysen et al. 2007; Kingma and Welling 2013; Connor, Canal, and Rozell 2021; Tishby and Zaslavsky 2015; Shwartz-Ziv and Tishby 2017).

Intrinsic dimension estimation. Methods for empirically testing the manifold hypothesis typically involve assuming the samples follow some statistical process, where the dimension parameter is then estimated from samples using maximum likelihood estimation of distances between points (Pope et al. 2021; Block et al. 2021; Levina and Bickel 2004). (Fefferman, Mitter, and Narayanan 2016) provides a classical algorithm to test whether a set of points can be described by a manifold with sufficient regularity properties.

Learning the manifold/dimension reduction. In modern datasets, a reasonable proxy to the image manifold hypothesis is to have a representation of the low-dimensional structure, constructed from finitely many samples. (Fefferman, Mitter, and Narayanan 2016) provides a classical algorithm to test whether a set of points can be described by a manifold with sufficient regularity properties. Classical methods to find nonlinear low-dimensional manifolds include locally linear embedding (Roweis and Saul 2000), Isomap (Tenenbaum, Silva, and Langford 2000), eigenmaps (Belkin and

Niyogi 2001), and topological properties (Niyogi, Smale, and Weinberger 2008, 2011), with more comprehensive reviews given in (Lee, Verleysen et al. 2007). A more modern approach uses generative networks by using the latent space as the manifold, which bypasses using the intrinsic dimension itself as an algorithmic parameter (Wang, Yao, and Zhao 2016; DeMers and Cottrell 1992; Nakada and Imaizumi 2020).

Manifold-driven architectures. Driven by the manifold hypothesis, several machine learning approaches consider enforcing a network output to be low-dimensional. Common examples are variational auto-encoders, which consist of an "encoder" network mapping from the input to a latent space, and "decoder" network mapping from the latent space to an output (Kingma and Welling 2013; Connor, Canal, and Rozell 2021). By restricting the dimensionality of the latent space, the output manifold will automatically be restricted. Other methods include bottleneck layers in ResNets, which relate to the information bottleneck tradeoff between compression and prediction (Tishby and Zaslavsky 2015; Shwartz-Ziv and Tishby 2017).

Application: Langevin mixing times. For an isometrically embedded manifold, (Block et al. 2020) bounds a log-Sobolev constant for probability measures supported on the manifold that are absolutely continuous with respect to the volume measure, mollified with Gaussian densities in the ambient space. (Wang, Lei, and Panageas 2020) demonstrate linear convergence of the Kullback-Leibler divergence with rates depending only on the intrinsic dimension for the geodesic Langevin algorithm, which incorporates the Riemannian metric into the noise in the unadjusted Langevin algorithm.

(DeVore, Howard, and Micchelli 1989) show that for any continuously parameterized set of functions that can uniformly ε-approximate the unit ball of the Sobolev space $W^{k,\infty}(\mathbb{R}^d)$, the dimension of the parameter set must scale as $\Omega(\varepsilon^{-d/k})$. (Gao et al. 2019) shows that any class of functions that can robustly interpolate n samples has Vapnik-Chervonenkis (VC) dimension at least $\Omega(nD)$. (Bolcskei et al. 2019) show lower bounds for the sparsity of a deep neural network for approximating function classes in $L^2(\mathbb{R}^d)$. (Chen et al. 2019, 2022) provide approximation rates, empirical risk estimates and generalization bounds of ReLU networks for Hölder functions on manifolds, assuming isometric embedding in Euclidean space. (Labate and Shi 2023) consider generalization of the class of ReLU networks for Hölder functions on the manifold using the Johnson–Lindenstrauss lemma.

On the unit hypercube, (Yang, Yang, and Xiang 2024) addresses the complexity of approximating a Sobolev function *constructively* with ReLU DNNs by showing an upper bound on the VC dimension and pseudo-dimension of derivatives of neural networks based on the number of layers, input dimension, and maximum width. (Park et al. 2020; Kim, Min, and Park 2023; Hanin and Sellke 2017) consider lower bounds for the minimum width required for ReLU and ReLU-like networks to ε-approximate L^p functions on Euclidean space and the unit hypercube.

This work derives lower bounds (Theorems 2 and 3) on the *statistical complexity* in terms of the *nonlinear width*, c.f. Definition 2, required to approximate Sobolev functions on *compact Riemannian manifolds*. Sobolev functions define expressive classes of functions that can model many physical problems, while also having sufficient regularity properties allowing for functional analysis. In other words, we consider the difficulty of modelling physical problems over structured datasets with simple function classes.

2 Background

2.1 Pseudo-Dimension as Complexity

We consider a concept of statistical complexity called the pseudo-dimension (Pollard 2012; Anthony and Bartlett 1999). This extends the classical concept of Vapnik–Chervonenkis (VC) dimension from indicator-valued to real-valued functions.

Definition 1. *Let $\mathcal{H}$ be a class of real-valued functions with domain $\mathcal{X}$. Let $X_n = \{x_1, ..., x_n\} \subset \mathcal{X}$, and consider a collection of real numbers $s_1, ..., s_n \subset \mathbb{R}^n$. When evaluated at each x_i, a function $h \in \mathcal{H}$ will lie on one side[1] of the corresponding x_i, i.e.* $\operatorname{sign}(h(x_i) - s_i) = \pm 1$. *The vector of such sides $(\operatorname{sign}(h(x_i) - s_i))_{i=1}^n$ is thus an element of $\{\pm 1\}^n$.*

We say that $\mathcal{H}$ P-shatters *X_n if there exist real numbers $s_1, ..., s_n$ such that all possible sign combinations are obtained, i.e.,*

$$\{(\operatorname{sign}(h(x_i) - s_i))_i \mid h \in \mathcal{H}\} = \{\pm 1\}^n.$$

The pseudo-dimension $\dim_p(\mathcal{H})$ *is the cardinality of the largest set that is P-shattered:*

$$\dim_p(\mathcal{H}) = \sup\{n \in \mathbb{N} \mid \exists\{x_1, ..., x_n\} \subset \mathcal{X} \text{ that is P-shattered by } \mathcal{H}\}. \quad (1)$$

The VC dimension is defined similarly, but without the biases s_i and with $\mathcal{H}$ being a class of binary functions taking values in $\{\pm 1\}$. The pseudo-dimension satisfies similar properties as the VC dimension, such as coinciding with the standard notion of dimension for vector spaces of functions.

Proposition 1 (Anthony and Bartlett 1999, Thm. 11.4)**.** *If $\mathcal{H}'$ is an $\mathbb{R}$-vector space of real-valued functions, then $\dim_p(\mathcal{H}') = \dim(\mathcal{H}')$ as a vector space. In particular, if $\mathcal{H}$ is a subset of a vector space $\mathcal{H}'$ of real-valued functions, then $\dim_p(\mathcal{H}) \leq \dim(\mathcal{H}')$.*

Like the VC dimension and other statistical complexity measures such as the Rademacher or Gaussian complexity, low complexity leads to better generalization properties of empirical risk minimizers (Bartlett and Mendelson 2002). One example is as follows, where a precise definition of sample complexity can be found in Appendix A.

Proposition 2 (Anthony and Bartlett 1999, Thm. 19.2)**.** *Let $\mathcal{H}$ be a class of functions mapping from a domain X into*

[1]We adopt the notation of $\operatorname{sign}(0) = +1$ for well-definedness, but the other option is equally valid.

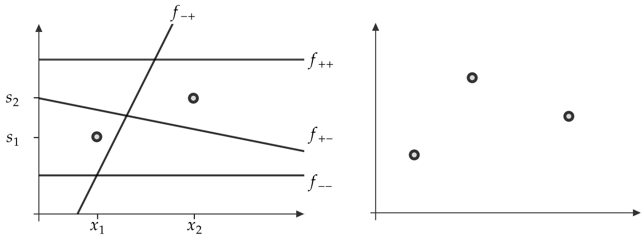

Figure 1: For the class of affine 1D functions $\{x \mapsto ax + b \mid a, b \in \mathbb{R}\}$, this choice of $\{x_1, x_2\} \subset \mathbb{R}$ and $s_1, s_2 \in \mathbb{R}$ on the left is P-shattered by the affine functions $f_{\pm\pm}$. However, there is no arrangement of three points that is P-shattered by affine functions, e.g. the arrangement on the right would not have a function that goes below the left and right points but above the middle point. Therefore, the pseudo-dimension of affine 1D functions is 2.

$[0, 1] \subset \mathbb{R}$*, and that* $\mathcal{H}$ *has finite pseudo-dimension. Then the* (ϵ, δ)*-sample complexity (Definition A.2) is bounded by*

$$m_L(\epsilon, \delta) \leq \frac{128}{\epsilon^2}\left(2\dim_p(\mathcal{H})\log\left(\frac{34}{\epsilon}\right) + \log\left(\frac{16}{\delta}\right)\right). \tag{2}$$

To compare the approximation of one function class by another, we consider a nonlinear width induced by a normed space.

Definition 2 (Nonlinear n-width)**.** *Let* $\mathcal{F}$ *be a normed space of functions. Given two subsets* $F_1, F_2 \subset \mathcal{F}$*, the* (asymmetric) Hausdorff distance *between the two subsets is the largest distance between elements of* F_1 *and their closest element in* F_2*:*

$$\operatorname{dist}(F_1, F_2; \mathcal{F}) = \sup_{f_1 \in F_1} \inf_{f_2 \in F_2} \|f_1 - f_2\|_{\mathcal{F}}. \tag{3}$$

For a subset $F \subset \mathcal{F}$*, the* nonlinear n-width *is given by the optimal (asymmetric) Hausdorff distance between* F *and* $\mathcal{H}^n$*, infimized over classes* $\mathcal{H}^n$ *in* $\mathcal{F}$ *with* $\dim_p(\mathcal{H}^n) \leq n$*:*

$$\rho_n(F, \mathcal{F}) \coloneqq \inf_{\mathcal{H}^n} \operatorname{dist}(F, \mathcal{H}^n; \mathcal{F}) = \inf_{\mathcal{H}^n} \sup_{f \in F} \inf_{h \in \mathcal{H}^n} \|f - h\|_{\mathcal{F}}. \tag{4}$$

This width measures the complexity in terms of how closely the entire function class can be approximated with another class of finite pseudo-dimension. This is useful in cases where F has infinite pseudo-dimension, and the nonlinear n-width acts as a surrogate measure of complexity, given by how well F can be approximated by classes of finite pseudo-dimension. In Section 3, we provide a lower bound on the nonlinear n-width of a bounded Sobolev class of functions. In terms of neural network approximation, these lower bounds complement existing approximation results of ReLU networks, which effectively provide an upper bound on the width by using the class of (bounded width, layers, and parameters) ReLU networks as the finite pseudo-dimension approximating class.

2.2 Riemannian Geometry

The manifold hypothesis can be readily expressed in terms of Riemannian geometry. A quick review and notation are given in Appendix B, and we further refer to (Bishop and Crittenden 2011; Gallot et al. 2004). Throughout, we will assume that our Riemannian manifold is finite-dimensional, compact, without boundary, and connected. We note that the connectedness assumption can be dropped by working instead on each connected component.

We state the celebrated Bishop–Gromov theorem (Petersen 2006; Bishop 1964). This is an essential volume-comparison theorem, used to tractably bound the volume of balls as they grow.

Theorem 1 (Bishop–Gromov)**.** *Let* (M, g) *be a complete* d*-dimensional Riemannian manifold whose Ricci curvature is bounded below by* $\operatorname{Ric} \geq (d-1)K$*, for some* $K \in \mathbb{R}$*. Let* M_K^d *be the complete* d*-dimensional simply connected space of constant sectional curvature* K*, i.e. a sphere, Euclidean space, or scaled hyperbolic space if* $K > 0$*,* $K = 0$*,* $K < 0$ *respectively. Then for any* $p \in M$ *and* $p_K \in M_K^d$*, we have that*

$$\phi(r) = \operatorname{vol}_M(B_r(p)) / \operatorname{vol}_{M_K^d}(B_r(p_K)) \tag{5}$$

is non-increasing on $(0, \infty)$*. In particular,* $\operatorname{vol}_M(B_r(p)) \leq \operatorname{vol}_{M_K^d}(B_r(p_K))$*.*

We note that in a space of constant sectional curvature M_K^d, the volume of a ball of radius r does not depend on the center. Without loss of generality, we write $\operatorname{vol}_{M_K^d}(B_r)$ to mean $\operatorname{vol}_{M_K^d}(B_r(p_K))$ for any point $p_K \in M_K^d$. Bishop–Gromov can be specialized as integrals in hyperbolic space.

Corollary 1 (Block et al. 2020; Ohta 2014)**.** *Let* (M, g) *be a complete* d*-dimensional Riemannian manifold such that* $\operatorname{Ric} \geq (d-1)K$ *for some* $K < 0$*. For any* $0 < r < R$*, we have*

$$\frac{\operatorname{vol}_M(B_R(x))}{\operatorname{vol}_M(B_r(x))} \leq \frac{\int_0^R s^{d-1}}{\int_0^r s^{d-1}}, \quad s(u) = \sinh(u\sqrt{|K|}). \tag{6}$$

We additionally need the following definitions, bounded in Proposition C.1 and Lemma C.2.

Definition 3 (Packing number)**.** *For a metric space* (M, d) *and radius* $\varepsilon > 0$*, the* packing number $N_\varepsilon(M)$ *is the maximum number of points* $x_1, ..., x_n \in M$ *such that the open balls* $B_\varepsilon(x_i)$ *are disjoint. The* ε-metric entropy $\mathcal{M}_\varepsilon$ *is the maximum number of points* $x_1, ..., x_m \in M$ *such that* $d(x_i, x_j) \geq \varepsilon$ *for* $i \neq j$*. Note that* $\mathcal{M}_{2\varepsilon} \leq N_\varepsilon \leq \mathcal{M}_\varepsilon$*.*

2.3 Sobolev Functions on Manifolds

We now define the bounded Sobolev ball on manifolds, which will be the subject of approximation in the next section. There are different ways to define the Sobolev spaces on manifolds due to the curvature differing only by constants, and we consider the variant presented in (Hebey 2000).

Definition 4 (Hebey 2000, Sec. 2.2)**.** *Let* (M, g) *be a smooth Riemannian manifold. For integer* k*,* $p \geq 1$*, and smooth* $u : M \to \mathbb{R}$*, define by* $\nabla^k u$ *the k'th covariant derivative of* u*, and* $|\nabla^k u|$ *its norm, defined in a local chart as*

$$|\nabla^k u|^2 = g^{i_1 j_1} ... g^{i_k j_k} (\nabla^k u)_{i_1 ... i_k} (\nabla^k u)_{j_1 ... j_k}, \tag{7}$$

using Einstein's summation convention where repeated indices are summed. Define the set of admissible test functions (with respect to the volume measure) as

$$\mathcal{C}^{k,p}(M) := \big\{u \in \mathcal{C}^\infty(M) \mid \forall j = 0, ..., k, \\ \int_M |\nabla^j u|^p \,\mathrm{dvol}_M < +\infty\big\}, \quad (8)$$

and for $u \in \mathcal{C}^{k,p}(M)$, the Sobolev $W^{k,p}$ norm as

$$\|u\|_{W^{k,p}} := \sum_{j=0}^{k} \left(\int_M |\nabla^j u|^p \,\mathrm{dvol}_M\right)^{1/p} = \sum_{j=0}^{k} \|\nabla^j u\|_p \quad (9)$$

The Sobolev space $W^{k,p}(M)$ is defined as the completion of $\mathcal{C}^{k,p}$ under $\|\cdot\|_{W^{k,p}}$.

It can be shown that Sobolev functions on a compact Riemannian manifold share similar embedding properties as in Euclidean space, and we refer to (Hebey 2000; Aubin 2012) for a more detailed treatment of such results. We additionally adopt the following definition of a bounded Sobolev ball, providing a compact space of functions to approximate.

Definition 5. *For constant $C \geq 0$, the bounded Sobolev ball $W^{k,p}(C; M)$ is given by the set of all functions with covariant derivatives bounded in $L^p = L^p(M, \mathrm{vol}_M)$ by C:*

$$W^{k,p}(C; M) = \big\{u \in W^{k,p}(M) \mid \forall l \leq k,\ \|\nabla^l u\|_p \leq C\big\} \quad (10)$$

We write $W^{k,p}(C)$ to mean $W^{k,p}(C; M)$.

3 Main Result

This section begins with a statement of the main approximation result, a lower bound on the nonlinear n-width (4) of bounded Sobolev balls $W^{1,p}(1)$ in L^q. This is followed by a high-level intuition behind the proof, then the proof in detail. The supporting lemmas are deferred to Appendix C.

Theorem 2. *Let (M, g) be a d-dimensional compact (separable) Riemannian manifold without boundary. From compactness, there exist real constants K, $\mathrm{inj}(M)$ such that:*

1. *The Ricci curvature satisfies $\mathrm{Ric} \geq (d-1)K$, where $K < 0$;*
2. *The injectivity radius is positive, $\mathrm{inj}(M) > 0$.*

For any $1 \leq p, q \leq +\infty$, the nonlinear width of $W^{1,p}(1)$ satisfies the lower bound for sufficiently large n:

$$\rho_n(W^{1,p}(1), L^q(M)) \\ \geq C(d, K, \mathrm{vol}(M), p, q)(n + \log n)^{-1/d}. \quad (11)$$

The (explicit) constant is independent of any ambient dimension that (M, g) may be embedded in.

Note that this statement does not refer to any ambient dimension or embeddings, and can be defined on abstract manifolds. This theorem should be contrasted with (Maiorov and Ratsaby 1999, Thm. 1), which exhibits a similar bound $n^{-1/d}$ for the bounded Sobolev space on the unit hypercube $[0, 1]^d$. The additional $\log n$ term is necessary due to the curvature of the space, but can otherwise be absorbed into the constant. We also note that it is possible to perform this analysis in the case of positive curvature and derive better bounds. For higher regularity functions, a similar result is available but without explicit constants. The proof of this extension is deferred to Appendix E.

Theorem 3. *Assume the setup of Theorem 2. For any $1 \leq p, q \leq +\infty$, the nonlinear width of $W^{k,p}(1)$ satisfies the lower bound for sufficiently large n:*

$$\rho_n(W^{k,p}(1), L^q(M)) \geq C(d, g, p, q, k)(n + \log n)^{-k/d}. \quad (12)$$

where the constant depends on the boundedness of the manifold geometry.

There are two major difficulties in converting the proof of (Maiorov and Ratsaby 1999) to the manifold setting, both arising from curvature. Firstly, the original proof uses a partition into hypercubes to construct the desired counterexample. As such hypercube partitions do not possess nice properties on manifolds, we instead consider a packing of geodesic balls, which does not fully cover the manifold and loosens the bound. The second major difference is the lack of global information, particularly for geodesic balls of the same radius which can have different volumes at different points, introducing additional constants into the final bound.

3.1 Proof Sketch

We can break down the proof into several steps.

1. We consider a class of simple functions, defined as sums of cutoff functions with disjoint supports. The class of simple functions is such that the L^1-norm of each component is large within the class of bounded $W^{1,p}(1)$ functions.
2. An appropriate subset of the simple functions is then taken, that is isometric to the hypercube graph $\{\pm 1\}^m$. Since we can find an ℓ_1-well separated subset of the hypercube graph, there exists an L^1-well separated subset of our simple functions.
3. The L^1 separation of the constructed set prevents approximation with classes of insufficiently large pseudo-dimension. This step uses an exponential lower bound on the metric entropy and a polynomial upper bound from Bishop–Gromov to derive a contradiction.
4. We conclude that the optimal approximation with function classes with bounded pseudo-dimension must incur an error, bounded from below as in the theorem statement. We conclude the proof by combining all the inequalities from the lemmas and Step 3.

3.2 Proof of Theorem 2

In the following, L^p spaces will be on (M, g) with respect to the underlying volume measure. Recall the definition of the bounded class of $W^{1,p}$ functions:

$$W^{1,p}(C) = \big\{u \in W^{1,p}(M) \mid \|u\|_{L^p}, \|\nabla u\|_{L^p} \leq C\big\}. \quad (13)$$

Step 1. Defining the base function class. Fix a radius $0 < r < \mathrm{inj}(M)$, which will be chosen appropriately later. Fix a maximal packing of geodesic r-balls, say with centers

$p_1, ..., p_{N_r}$, where $N_r = N_r^{\text{pack}}(M)$ is the packing number. By definition, $B_r(p_i)$ are disjoint for $i = 1, ..., N_r$. From Proposition C.1, the packing number satisfies the following where $D = \text{diam}(M)$:

$$\frac{\text{vol}(M)}{\text{vol}_{M_K^d}(B_{2r})} \le N_r^{\text{pack}} \le \frac{\text{vol}_{M_K^d}(B_D)}{\text{vol}_{M_K^d}(B_r)}. \tag{14}$$

For each ball $B_r(p_i)$, using (Azagra et al. 2007, Cor. 3), we can construct a $\mathcal{C}^\infty$ function $\phi_i' : M \to [0, r/4]$ with support $\text{supp}(\phi_i') \subset B_r(p_i)$ such that $|\nabla \phi_i'| \le 1$ pointwise and

$$\phi_i'(p) = \begin{cases} r/4, & d(p, p_i) \le r/2; \\ 0, & d(p, p_i) \ge r. \end{cases} \tag{15}$$

From (15), we have the L^1 lower bound

$$\|\phi_i'\|_1 \ge (r/4)\text{vol}_M(B_{r/2}(p_i)). \tag{16}$$

Moreover, we have the L^p bounds on ϕ_i' and $\nabla\phi_i'$,

$$\|\phi_i'\|_p \le (r/4)\text{vol}_M(B_r(p_i))^{1/p}, \tag{17}$$

$$\|\nabla\phi_i'\|_p \le \text{vol}_M(B_r(p_i))^{1/p}. \tag{18}$$

For $r < 4$, we have that $\phi_i' \in W^{1,p}(\text{vol}_M(B_r(p_i))^{1/p})$. Define a non-negative function $\phi_i \in W^{1,p}(1)$ with support in $B_r(p_i)$ satisfying:

$$\phi_i := \frac{\phi_i'}{\text{vol}_M(B_r(p_i))^{1/p}},\ \|\phi_i\|_1 \ge (r/4)\frac{\text{vol}_M(B_{r/2}(p_i))}{\text{vol}_M(B_r(p_i))^{1/p}}. \tag{19}$$

Moreover, $\phi_i = r/(4\text{vol}_M(B_r(p_i))^{1/p})$ on $B_{r/2}(p_i)$. We now consider the function class

$$F_r = \left\{ f_a = \frac{1}{N_r^{1/p}} \sum_{i=1}^{N_r} a_i \phi_i \,\middle|\, a_i \in \{\pm 1\},\, i = 1, ..., N_r \right\}. \tag{20}$$

Since the sum is over functions of disjoint support, we have that $\|f_a\|_p, \|\nabla f_a\|_p \le 1$, and thus each element of F_r also lies in $W^{1,p}(1)$. Moreover, every element $f_a \in F_r$ satisfies the L^1 lower bound using (19):

$$\|f_a\|_1 \ge \frac{r}{4N_r^{1/p}} \sum_{i=1}^{N_r} \frac{\text{vol}_M(B_{r/2}(p_i))}{\text{vol}_M(B_r(p_i))^{1/p}}, \quad \forall f_a \in F_r. \tag{21}$$

Step 2. L^1-well-separation of F_r. Consider the following lemma, which shows the existence of a large well-separated subset of ℓ_1^m.

Lemma 1 (Lorentz, von Golitschek, and Makovoz 1996, Lem. 2.2**).** *There exists a set $G \subset \{\pm 1\}^m$ of cardinality at least $2^{m/16}$ such that for any $v \ne v' \in G$, the distance $\|v - v'\|_{\ell_1^m} \ge m/2$. In particular, any two elements differ in at least $m/4$ entries.*

In particular, let $G \subset \{\pm 1\}^{N_r}$ be well separated by the above lemma. Denote by $F_r(G) = \{f_a \in F_r \mid a \in G\}$. the subset of F_r corresponding to these indices. For the specific choice of separated $G \subset \{\pm 1\}^{N_r}$ in the above lemma, we claim the following well-separation of $F_r(G)$, proved in Appendix D.

Claim 1. *There exists a constant $C_1(r) > 0$ such that for any $f \ne f' \in F_r(G)$, we have*

$$\|f - f'\|_1 \ge C_1(r) > 0. \tag{22}$$

Moreover, the following constant works:

$$C_1(r) = \frac{r N_r^{1-1/p} \int_0^{r/2} s^{d-1}}{8 \int_0^r s^{d-1}} \inf_{i \in [N_r]} \left[\text{vol}_M(B_r(p_i))^{1-1/p} \right]. \tag{23}$$

This shows L^1-well separation of the subset $F_r(G) \subset F_r \subset W^{1,p}(1)$, which consist of sums of disjoint cutoff functions. The key is to contrast this with the metric entropy bounds in Lemma C.2, by showing that $F_r(G)$ is difficult to approximate with function classes of low pseudo-dimension.

Step 3a. Construction of well-separated bounded set. Let $\mathcal{H}^n$ be a given set of vol_M-measurable functions with $\dim_p(\mathcal{H}^n) \le n$. Let $\varepsilon > 0$. Denote

$$\begin{aligned}\delta &:= \sup_{f \in F_r(G)} \inf_{h \in \mathcal{H}^n} \|f - h\|_1 + \varepsilon \\ &= \text{dist}(F_r(G), \mathcal{H}^n, L^1(M)) + \varepsilon. \end{aligned} \tag{24}$$

Define a projection operator $P : F_r(G) \to \mathcal{H}^n$, mapping $f \in F_r(G)$ to any element Pf in $\mathcal{H}^n$ such that $\|f - Pf\|_1 \le \delta$. We introduce a (measurable) clamping operator $\mathcal{C}$ for a function f:

$$\beta_i = r/(4\text{vol}_M(B_r(p_i))^{1/p} N_r^{1/p}), \quad i = 1, ..., N_r, \tag{25}$$

$$(\mathcal{C}f)(x) = \begin{cases} -\beta_i, & x \in B_r(p_i) \text{ and } f(x) < -\beta_i; \\ f(x), & x \in B_r(p_i) \text{ and } -\beta_i \le f(x) \le \beta_i; \\ \beta_i, & x \in B_r(p_i) \text{ and } f(x) > \beta_i; \\ 0, & \text{otherwise.} \end{cases} \tag{26}$$

Note that β_i are the bounds of $f_a \in F_r$ in the balls $B_r(p_i)$. Now consider the set of functions $\mathcal{S} := \mathcal{C}PF_r(G)$. Suppose $f \ne f' \in F_r(G)$. We show separation in $\mathcal{S}$ using triangle inequality:

$$\begin{aligned}\|\mathcal{C}Pf - \mathcal{C}Pf'\|_1 \ge \|f - f'\|_1 &- \|f - \mathcal{C}Pf\|_1 \\ &- \|f' - \mathcal{C}Pf'\|_1. \end{aligned} \tag{27}$$

For any $a \in G$, we have that $f_a \le \beta_i$ in $B_r(p_i)$, and both f_a and $\mathcal{C}Pf_a$ are zero on $M \setminus \bigsqcup_i B_r(p_i)$. We thus have that for any $x \in M$ and any $f_a \in F_r(G)$, $|f_a(x) - \mathcal{C}Pf_a(x)| \le |f_a(x) - Pf_a(x)|$. This inequality holds for $x \in B_r(p_i)$ since $\mathcal{C}$ clamps $Pf_a(x)$ towards $[-\beta_i, \beta_i]$, and holds trivially on $M \setminus \bigsqcup_i B_r(p_i)$. Integrating and by definition of P, we have that for any $f_a \in F_r(G)$,

$$\|f_a - \mathcal{C}Pf_a\|_1 \le \|f_a - Pf_a\|_1 \le \delta. \tag{28}$$

Using (22), (27) and (28), we thus have separation

$$\|\mathcal{C}Pf - \mathcal{C}Pf'\|_1 \ge \|f - f'\|_1 - 2\delta \ge C_1(r) - 2\delta. \tag{29}$$

Step 3b. Minimum distance by contradiction. Suppose for contradiction that $\delta \le C_1(r)/4$. Then from (29), we have

$$\|\mathcal{C}Pf - \mathcal{C}Pf'\|_1 \ge C_1(r)/2. \tag{30}$$

The separation implies that the $\mathcal{C}Pf$ are distinct for distinct $f \in F_r(G)$, thus $|\mathcal{S}| = |G| \geq 2^{N_r/16}$.

Define $\alpha = C_1(r)/2$. Consider the metric entropy in L^1, as given in Lemma C.2. By construction (29), $\mathcal{S}$ itself is an α-separated subset in L^1 as any two elements are L^1-separated by α, so

$$\mathcal{M}_\alpha(\mathcal{S}, L^1(\text{vol}_M)) \geq 2^{N_r/16}. \tag{31}$$

We now wish to obtain an upper bound on $\mathcal{M}_\alpha(\mathcal{S}, L^1)$ using Lemma C.2. From the definition of pseudo-dimension, we have $\dim_p(\mathcal{C}PF_r(G)) \leq \dim_p(PF_r(G))$, since any P-shattering set for $\mathcal{C}PF_r(G)$ will certainly P-shatter $PF_r(G)$. Since $PF_r(G) \subset \mathcal{H}^n$, we have $\dim_p(PF_r(G)) \leq \dim_p(\mathcal{H}^n) \leq n$. Thus $\dim_p(\mathcal{S}) = \dim_p(\mathcal{C}PF_r(G)) \leq n$. $\mathcal{S}$ is L^1-separated with distance at least α, and moreover consists of elements that are bounded by $\beta := \sup_i \beta_i$. Lemma C.2 now gives:

$$M_\alpha(\mathcal{S}, L^1(\text{vol}_M)) \leq e(n+1)\left(\frac{4e\beta\text{vol}(M)}{\alpha}\right)^n. \tag{32}$$

Intuitively, $N_r \sim r^{-d}$, so the lower bound (31) is exponential in r. Meanwhile, α and β are both polynomial in r, so the upper bound (32) is polynomial in r. So for sufficiently small r, we have a contradiction with the supposition that $\delta \leq C_1(r)/4$. We now show this formally. Recall:

$$\beta = \sup_{i\in[N_r]} \frac{r}{4\text{vol}_M(B_r(p_i))^{1/p} N_r^{1/p}},$$
$$\alpha = \frac{rN_r^{1-1/p}\int_0^{r/2} s^{d-1}}{16\int_0^r s^{d-1}} \inf_{i\in[N_r]}\left[\text{vol}_M(B_r(p_i))^{1-1/p}\right].$$

Note that the supremum in β and the infimum in α is attained by the same $i \in [N_r]$, namely, the p_i that has smallest $\text{vol}_M(B_r(p_i))$. Combining (31) and (32), where $s(u) = \sinh(u\sqrt{|K|})$,

$$\begin{aligned}
&\frac{1}{e(n+1)}2^{N_r/16} &(33)\\
&\leq \left(\frac{4e\beta\text{vol}(M)}{\alpha}\right)^n \\
&= \left(\frac{4e\text{vol}(M)\sup_{i\in[N_r]}[r/(4\text{vol}_M(B_r(p_i))^{1/p}N_r^{1/p})]}{\frac{rN_r^{1-1/p}\int_0^{r/2}s^{d-1}}{16\int_0^r s^{d-1}}\inf_{i\in[N_r]}\left[\text{vol}_M(B_r(p_i))^{1-1/p}\right]}\right)^n \\
&= \left(16e\frac{\text{vol}(M)\int_0^r s^{d-1}}{N_r\int_0^{r/2}s^{d-1}}\sup_{i,j\in[N_r]}\frac{\text{vol}_M(B_r(p_j))^{1/p-1}}{\text{vol}_M(B_r(p_i))^{1/p}}\right)^n \\
&= \left(16e\frac{\text{vol}(M)\int_0^r s^{d-1}}{N_r\int_0^{r/2}s^{d-1}}\sup_{i\in[N_r]}\left[\text{vol}_M(B_r(p_i))^{-1}\right]\right)^n &(34)
\end{aligned}$$

where the equalities come from definition of β and α and rearranging, and the last equality from noting the supremum is attained when $i = j \in [N_r]$ minimizes $\text{vol}_M(B_r(p_i))$. The following result lower-bounds the volume of small balls to control the supremum term.

Proposition 3 (Croke 1980, Prop. 14)**.** *For $r \leq \text{inj}(M)/2$, the volume of the ball $B_r(p)$ satisfies*

$$\text{vol}_M(B_r(p)) \geq C_2(d)r^d, \quad C_2(d) = \frac{2^{d-1}\text{vol}_{M_1^{d-1}}(B_1)^d}{d^d\text{vol}_{M_1^d}(B_1)^{d-1}}. \tag{35}$$

The volume of the d-dimensional hyperbolic sphere with sectional curvature K can be written in terms of the volume of the d-dimensional sphere[2]:

$$\text{vol}_{M_K^d}(B_\rho) = \text{vol}_{M_1^d}(B_1)\int_{t=0}^{\rho}\left(\frac{\sinh(\sqrt{|K|}t)}{\sqrt{|K|}}\right)^{d-1}\text{d}t. \tag{36}$$

Note that $x \leq \sinh(x) \leq 2x$ for $x \in [0, 2]$. Therefore, for $\rho \leq 2/\sqrt{|K|}$, we have $\sinh(\sqrt{|K|}t) \leq 2\sqrt{|K|}t$. We thus have that

$$\begin{aligned}
\text{vol}_{M_K^d}(B_\rho) &= \text{vol}_{M_1^d}(B_1)\int_{t=0}^{\rho}\left(\frac{\sinh(\sqrt{|K|}t)}{\sqrt{|K|}}\right)^{d-1}\text{d}t \\
&< \text{vol}_{M_1^d}(B_1)2^{d-1}\rho^d/d = C_3(d)\rho^d,
\end{aligned} \tag{37}$$

where $C_3(d) := \text{vol}_{M_1^d}(B_1)2^{d-1}/d$. Moreover, for $r < 1/\sqrt{K}$,

$$\frac{\int_0^r s^{d-1}}{\int_0^{r/2} s^{d-1}} = \frac{\int_0^r \sinh(\sqrt{|K|}u)^{d-1}\,\text{d}u}{\int_0^{r/2}\sinh(\sqrt{|K|}u)^{d-1}\,\text{d}u} \leq 2^d. \tag{38}$$

We continue the inequality (34) for $r < 1/\sqrt{|K|}$:

$$\begin{aligned}
&\frac{1}{e(n+1)}2^{N_r/16} &(39)\\
&\leq \left(16e\frac{\text{vol}(M)\int_0^r s^{d-1}}{N_r\int_0^{r/2}s^{d-1}}\sup_{i\in[N_r]}\left[\text{vol}_M(B_r(p_i))^{-1}\right]\right)^n &(40)\\
&\leq \left(16e\text{vol}_{M_K^d}(B_{2r})\frac{\int_0^r s^{d-1}}{\int_0^{r/2}s^{d-1}}C_2^{-1}r^{-d}\right)^n &(41)\\
&\leq \left(16eC_3(2r)^d\frac{\int_0^r s^{d-1}}{\int_0^{r/2}s^{d-1}}C_2^{-1}r^{-d}\right)^n &(42)\\
&\leq \left(2^{d+4}eC_3\frac{\int_0^r s^{d-1}}{\int_0^{r/2}s^{d-1}}C_2^{-1}\right)^n &(43)\\
&\leq \left(2^{2d+4}eC_3C_2^{-1}\right)^n = C_4(d)^n, &(44)
\end{aligned}$$

where $C_4 = C_4(d) := 2^{2d+4}eC_3C_2^{-1}$. We used (14), Prop. 3, (37), and (38) in the second, third, and final inequality respectively. We get a contradiction if

$$N_r > 16\left[n\log_2 C_4 + \log_2\left(e(n+1)\right)\right]. \tag{45}$$

[2]The volume of the d-dimensional sphere is $2\pi^{d/2}/\Gamma(d/2)$, where Γ is Euler's gamma function.

Recalling the lower bound (14) on N_r and using (38),

$$N_r \geq \frac{\mathrm{vol}(M)}{\mathrm{vol}_{M_K^d}(B_{2r})} > \frac{\mathrm{vol}(M)}{C_3(2r)^d}. \tag{46}$$

Take the following choice of r:

$$\min\left\{\frac{1}{2}\left(\frac{16C_3}{\mathrm{vol}(M)}\left[n\log_2 C_4 + \log_2(e(n+1))\right]\right)^{-1/d}, \frac{1}{\sqrt{|K|}}, \frac{\mathrm{inj}(M)}{2}, 4\right\}. \tag{47}$$

Using (46), this choice of r satisfies the contradiction condition (45). Note $r \sim (n+\log n)^{-1/d}$. The constants C_3, C_4 depend only on d.

Step 4. Concluding contradiction. This choice of r contradicts the assumption that $\delta \leq C_1(r)/4$. Therefore, we must have that $\delta > C_1/4$. Since the choice of r is independent of the choice of $\varepsilon > 0$ taken at the start of Step 3a, we have that

$$\mathrm{dist}(F_r(G), \mathcal{H}^n, L^1(\mathrm{vol}_M)) \geq C_1(r)/4, \tag{48}$$

where r is chosen as in (47). We obtain the chain of inequalities

$$\begin{aligned}
&\mathrm{dist}(W^{1,p}(1), \mathcal{H}^n, L^q) &(49)\\
&\geq \mathrm{dist}(W^{1,p}(1), \mathcal{H}^n, L^1)\mathrm{vol}(M)^{1/q-1}\\
&\geq \mathrm{dist}(F_r(G), \mathcal{H}^n, L^1)\mathrm{vol}(M)^{1/q-1}\\
&\geq C_1(r)\mathrm{vol}(M)^{1/q-1}/4\\
&= \frac{rN_r^{1-1/p}\mathrm{vol}(M)^{1/q-1}\int_0^{r/2} s^{d-1}}{32\int_0^r s^{d-1}}\\
&\quad \cdot \inf_{i\in[N_r]}\left[\mathrm{vol}_M(B_r(p_i))^{1-1/p}\right], &(50)
\end{aligned}$$

where the first inequality comes from Hölder's inequality $\|u\|_1 \leq \|u\|_q \mathrm{vol}(M)^{1-1/q}$, the second inequality from $F_r(G) \subset W^{1,p}(1)$, the third from (48) and the equality from definition (23) of $C_1(r)$. We conclude with recalling the bounds (38), (46), and Proposition 3. We have

$$\begin{aligned}
&\mathrm{dist}(W^{1,p}(1), \mathcal{H}^n, L^q)\\
&\geq \frac{r}{32}\underbrace{\left(\frac{\mathrm{vol}(M)}{C_3(2r)^d}\right)^{1-1/p}}_{(46)}\underbrace{2^{-d}}_{(38)}\underbrace{\left[C_2 r^d\right]^{1-1/p}}_{\text{Prop. 3}}\mathrm{vol}(M)^{1/q-1}\\
&= C_5(d, \mathrm{vol}(M), p, q)r,
\end{aligned}$$

where $C_5 = 2^{-2d-5+d/p}\mathrm{vol}(M)^{1/q-1/p}(C_2 C_3^{-1})^{1-1/p}$. Moreover, the constant C_5 and choice of r are independent of $\mathcal{H}^n$. Taking infimum over all choices of $\mathcal{H}^n$ with $\dim_p(\mathcal{H}^n) \leq n$ and using (47), we have

$$\rho_n(W^{1,p}(1), L^q) \geq C_5(d, \mathrm{vol}(M), p, q)r \gtrsim (n+\log n)^{-1/d}. \tag{51}$$

3.3 Relationship with Existing Bounds

(Yarotsky 2017) presents various results for approximating Sobolev balls on the unit hypercube. We note that the notion of complexity here is the number of neurons/weight.

Proposition 4 (Yarotsky 2017, Sec 3.2)**.** *Consider the Sobolev space $W^{k,\infty}([0,1]^d)$. Given $\varepsilon > 0$, there is a ReLU network architecture of complexity $\mathcal{O}(\varepsilon^{-d/k}\log(1/\varepsilon))$ that is able to approximate any function in the unit ball of $W^{k,\infty}([0,1]^d)$ with error ε.*

We note that the corresponding result of Theorem 3 in the case of $M = [0,1]^d$ in (Maiorov and Ratsaby 1999) gives the complexity lower bound $\Omega(\varepsilon^{-d/k})$ required to approximate the unit ball $W^{k,\infty}(1; [0,1]^d)$ with error ε. In particular, this states that ReLU neural networks with this architecture are asymptotically nearly optimal in achieving the best possible approximation.

In the context of manifolds, the following result gives the best available bounds in terms of ambient dimension dependence. The result considers using ReLU neural networks to approximate Hölder functions, which are closely related to Sobolev functions.

Proposition 5 (Chen et al. 2022, Thm 1)**.** *For a d-dimensional compact Riemannian manifold M without boundary isometrically embedded in $\mathbb{R}^D$, the unit ball of the Hölder space $C^{s,\alpha}(M)$ can be ε-approximated with a ReLU network architecture with complexity $\mathcal{O}(\varepsilon^{-\frac{d}{s+\alpha}}\log(\frac{1}{\varepsilon}) + D\log\frac{1}{\varepsilon} + D\log D)$. The constant depends on $d, s, \mathrm{vol}(M)$, the reach and diameter of the embedding, and the bounded geometry of M.*

We note that by the manifold version of Morrey's inequality, the Sobolev space of $W^{k,\infty}$ compactly and continuously embeds into $C^{k-1,1-\delta}$ for any $\delta \in (0,1)$ (Aubin 2012). This result thus gives a complexity upper bound of $\mathcal{O}(\varepsilon^{-\frac{d}{k-\delta}})$, nearly matching the lower bound of $\Omega(\varepsilon^{-d/k})$ in Theorem 3 up to log factors. However, our lower bound is independent of the ambient dimension D.

4 Conclusion

This work provides a theoretical motivation to further explore the manifold hypothesis. We show that the problem of approximating a bounded class of Sobolev functions depends only on the intrinsic properties of the supporting manifold. More precisely, the approximation error of the bounded $W^{k,p}$ space with respect to bounded pseudo-dimension classes is shown to be at least $(n + \log n)^{-k/d}$, where d is the intrinsic dimension of the underlying manifold. Since generalization error is linear in pseudo-dimension, this provides an ambient-dimension-free lower-bound on generalization error. This is in contrast to many works in the literature that provide constructive upper bounds on generalization error based on ReLU approximation properties that still depend on the embedding of the manifold in ambient Euclidean space. Followup work could consider Rademacher or Gaussian complexity, or alternative geometries such as Finsler manifolds (Busemann 2005).

Acknowledgments

HYT was supported by GSK.ai and the Masason Foundation. SM acknowledges support from the Faculty Start-up Research Grant (FSRG) provided by IIT Kharagpur (Project Code: RAI). CBS acknowledges support from the Philip Leverhulme Prize, the Royal Society Wolfson Fellowship, the EPSRC advanced career fellowship EP/V029428/1, EPSRC Grants EP/S026045/1 and EP/T003553/1, EP/N014588/1, EP/T017961/1, the Wellcome Innovator Awards 215733/Z/19/Z and 221633/Z/20/Z, the European Union Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Grant agreement No. 777826 NoMADS, the Cantab Capital Institute for the Mathematics of Information and the Alan Turing Institute.

References

Altman, N.; and Krzywinski, M. 2018. The curse (s) of dimensionality. *Nat Methods*, 15(6): 399–400.

Anthony, M.; and Bartlett, P. L. 1999. *Neural network learning: Theoretical foundations*, volume 9. Cambridge university press.

Aubin, T. 2012. *Nonlinear analysis on manifolds. Monge-Ampere equations*, volume 252. Springer Science & Business Media.

Azagra, D.; Ferrera, J.; López-Mesas, F.; and Rangel, Y. 2007. Smooth approximation of Lipschitz functions on Riemannian manifolds. *Journal of Mathematical Analysis and Applications*, 326(2): 1370–1378.

Bartlett, P. L.; and Mendelson, S. 2002. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov): 463–482.

Belkin, M.; and Niyogi, P. 2001. Laplacian eigenmaps and spectral techniques for embedding and clustering. *Advances in neural information processing systems*, 14.

Bengio, Y.; Courville, A.; and Vincent, P. 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8): 1798–1828.

Bishop, R. L. 1964. A relation between volume, mean curvature and diameter. In *Euclidean Quantum Gravity*, 161–161. World Scientific.

Bishop, R. L.; and Crittenden, R. J. 2011. *Geometry of Manifolds: Geometry of Manifolds*. Academic press.

Block, A.; Jia, Z.; Polyanskiy, Y.; and Rakhlin, A. 2021. Intrinsic dimension estimation using Wasserstein distances. *arXiv preprint arXiv:2106.04018*.

Block, A.; Mroueh, Y.; Rakhlin, A.; and Ross, J. 2020. Fast mixing of multi-scale Langevin dynamics under the manifold hypothesis. *arXiv preprint arXiv:2006.11166*.

Bolcskei, H.; Grohs, P.; Kutyniok, G.; and Petersen, P. 2019. Optimal approximation with sparsely connected deep neural networks. *SIAM Journal on Mathematics of Data Science*, 1(1): 8–45.

Brown, B. C.; Caterini, A. L.; Ross, B. L.; Cresswell, J. C.; and Loaiza-Ganem, G. 2022. Verifying the Union of Manifolds Hypothesis for Image Data. In *The Eleventh International Conference on Learning Representations*.

Busemann, H. 2005. *The geometry of geodesics*. Courier Corporation.

Chen, M.; Jiang, H.; Liao, W.; and Zhao, T. 2019. Efficient approximation of deep relu networks for functions on low dimensional manifolds. *Advances in neural information processing systems*, 32.

Chen, M.; Jiang, H.; Liao, W.; and Zhao, T. 2022. Nonparametric regression on low-dimensional manifolds using deep ReLU networks: Function approximation and statistical recovery. *Information and Inference: A Journal of the IMA*, 11(4): 1203–1253.

Connor, M.; Canal, G.; and Rozell, C. 2021. Variational autoencoder with learned latent structure. In *International conference on artificial intelligence and statistics*, 2359–2367. PMLR.

Croke, C. B. 1980. Some isoperimetric inequalities and eigenvalue estimates. In *Annales scientifiques de l'École normale supérieure*, volume 13, 419–435.

DeMers, D.; and Cottrell, G. 1992. Non-linear dimensionality reduction. *Advances in neural information processing systems*, 5.

DeVore, R. A.; Howard, R.; and Micchelli, C. 1989. Optimal nonlinear approximation. *Manuscripta mathematica*, 63: 469–478.

Fefferman, C.; Mitter, S.; and Narayanan, H. 2016. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4): 983–1049.

Gallot, S.; Hulin, D.; Lafontaine, J.; et al. 2004. *Riemannian geometry*, volume 3. Springer.

Gao, R.; Cai, T.; Li, H.; Hsieh, C.-J.; Wang, L.; and Lee, J. D. 2019. Convergence of adversarial training in overparametrized neural networks. *Advances in Neural Information Processing Systems*, 32.

Hanin, B.; and Sellke, M. 2017. Approximating continuous functions by relu nets of minimal width. *arXiv preprint arXiv:1710.11278*.

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

Kim, N.; Min, C.; and Park, S. 2023. Minimum width for universal approximation using ReLU networks on compact domain. *arXiv preprint arXiv:2309.10402*.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Labate, D.; and Shi, J. 2023. Low dimensional approximation and generalization of multivariate functions on smooth manifolds using deep neural networks. *Available at SSRN 4545106*.

Lee, J. A.; Verleysen, M.; et al. 2007. *Nonlinear dimensionality reduction*, volume 1. Springer.

Levina, E.; and Bickel, P. 2004. Maximum likelihood estimation of intrinsic dimension. *Advances in neural information processing systems*, 17.

Lorentz, G. G.; von Golitschek, M.; and Makovoz, Y. 1996. *Constructive approximation: advanced problems*, volume 304. Citeseer.

Maiorov, V.; and Ratsaby, J. 1999. On the degree of approximation by manifolds of finite pseudo-dimension. *Constructive approximation*, 15(2): 291–300.

Nakada, R.; and Imaizumi, M. 2020. Adaptive approximation and generalization of deep neural network with intrinsic dimensionality. *Journal of Machine Learning Research*, 21(174): 1–38.

Narayanan, H.; and Niyogi, P. 2009. On the Sample Complexity of Learning Smooth Cuts on a Manifold. In *COLT*.

Niyogi, P.; Smale, S.; and Weinberger, S. 2008. Finding the homology of submanifolds with high confidence from random samples. *Discrete & Computational Geometry*, 39: 419–441.

Niyogi, P.; Smale, S.; and Weinberger, S. 2011. A topological view of unsupervised learning from noisy data. *SIAM Journal on Computing*, 40(3): 646–663.

Ohta, S.-I. 2014. *Ricci curvature, entropy, and optimal transport*, 145–200. London Mathematical Society Lecture Note Series. Cambridge University Press.

Park, S.; Yun, C.; Lee, J.; and Shin, J. 2020. Minimum width for universal approximation. *arXiv preprint arXiv:2006.08859*.

Petersen, P. 2006. *Riemannian geometry*, volume 171. Springer.

Pollard, D. 2012. *Convergence of stochastic processes*. Springer Science & Business Media.

Pope, P.; Zhu, C.; Abdelkader, A.; Goldblum, M.; and Goldstein, T. 2021. The intrinsic dimension of images and its impact on learning. *arXiv preprint arXiv:2104.08894*.

Roweis, S. T.; and Saul, L. K. 2000. Nonlinear dimensionality reduction by locally linear embedding. *science*, 290(5500): 2323–2326.

Shwartz-Ziv, R.; and Tishby, N. 2017. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*.

Tenenbaum, J. B.; Silva, V. d.; and Langford, J. C. 2000. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500): 2319–2323.

Tishby, N.; and Zaslavsky, N. 2015. Deep learning and the information bottleneck principle. In *2015 ieee information theory workshop (itw)*, 1–5. IEEE.

Vidal, R. 2011. Subspace clustering. *IEEE Signal Processing Magazine*, 28(2): 52–68.

Wainwright, M. J. 2019. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press.

Wang, X.; Lei, Q.; and Panageas, I. 2020. Fast convergence of Langevin dynamics on manifold: Geodesics meet log-Sobolev. *Advances in Neural Information Processing Systems*, 33: 18894–18904.

Wang, Y.; Yao, H.; and Zhao, S. 2016. Auto-encoder based dimensionality reduction. *Neurocomputing*, 184: 232–242.

Yang, Y.; Yang, H.; and Xiang, Y. 2024. Nearly optimal VC-dimension and pseudo-dimension bounds for deep neural network derivatives. *Advances in Neural Information Processing Systems*, 36.

Yarotsky, D. 2017. Error bounds for approximations with deep ReLU networks. *Neural networks*, 94: 103–114.