Low dimensional approximation and generalization of multivariate functions on smooth manifolds using deep ReLU neural networks

Demetrio Labate[a,*], Ji Shi[a]

[a]*Department of Applied Mathematics, University of Houston, 651 Phillip G Hoffman, Houston, 77204-3008, Texas, USA*

Abstract

The expressive power of deep neural networks is manifested by their remarkable ability to approximate multivariate functions in a way that appears to overcome the curse of dimensionality. This ability is exemplified by their success in solving high-dimensional problems where traditional numerical solvers fail due to their limitations in accurately representing high-dimensional structures. To provide a theoretical framework for explaining this phenomenon, we analyze the approximation of Hölder functions defined on a d-dimensional smooth manifold $\mathcal{M}$ embedded in $\mathbb{R}^D$, with $d \ll D$, using deep neural networks. We prove that the uniform convergence estimates of the approximation and generalization errors by deep neural networks with ReLU activation functions do not depend on the ambient dimension D of the function but only on its lower manifold dimension d, in a precise sense. Our result improves existing results from the literature where approximation and generalization errors were shown to depend weakly on D.

Keywords: Approximation theory, approximation power, deep learning, deep neural network, generalization analysis, Hölder continuity, manifold learning
2010 MSC: 41A25, 41A30, 41A46, 60H25, 68T07

1. Introduction

One of the most striking properties of deep Neural Networks (NNs) is their remarkable ability to approximate high-dimensional functions in a way that appears to overcome the *curse of dimensionality* (COD). COD postulates that numerical approximation methods deteriorate exponentially fast with increasing dimension [1, 2] and poses a very significant challenge in many areas of applied mathematics. For instance, the computational cost of traditional discretization

[*] Corresponding author
Email addresses: dlabate@math.uh.edu (Demetrio Labate), jshi24@cougarnet.uh.edu (Ji Shi)

methods for Partial Differential Equations (PDEs), such as finite difference, finite element and spectral methods, scales with the dimension and becomes impractical as the dimension increases. By contrast, recent results have shown that deep NNs may perform very efficiently even in high dimensional numerical PDE problems where classical numerical solvers fail (e.g., [3]). These results have spurred a flurry of research activity aiming at integrating deep learning algorithms into traditional numerical methods.

Several arguments have been proposed to understand the approximation properties of NNs and explain the mechanisms by which deep NNs can avoid COD [4, 5]. A number of notable papers, for instance, have explored the role of compositionality by which deep ReLU NNs can outperform more traditional approximation methods [4, 6, 7]. Another group of works adopted the approach of framing the approximation problem within a target function space well suited to the approximation properties of deep ReLU NNs, such as the Barron spaces [8, 9, 10], or the Hölder-Zygmund spaces of mixed smoothness [11]. Yet another very appealing explanation for the ability of deep neural networks of avoiding COD is the so-called *manifold hypothesis*, a theoretical framework that is the foundation of manifold learning [12] and was already successfully exploited in nonlinear dimensionality reduction applications [13, 14]. Under the manifold hypothesis, high dimensional data are assumed to lie in the vicinity of a lower dimensional manifold. As a result, as we observe samples from an unknown function f defined on a compact subset of $\mathbb{R}^D$, we are not seeking to approximate f with respect to a norm on $\mathbb{R}^D$; rather, we consider a measure μ defined on a d-dimensional manifold $\mathcal{M}$, where $d < D$ (often $d \ll D$), and we estimate the error associated with the measure μ on $\mathcal{M}$. In this setting, D is often referred to as the *ambient dimension* as compared to the *manifold dimension* d. Under the manifold hypothesis, the ability of a neural network to avoid COD is explained as the ability of discovering appropriate local coordinate transformations, hence reducing the complexity of a high-dimensional problem to an underlying low-dimensional problem, given by the data distribution.

A seminal result in this direction is the work by Shaham et al. [15] proving that, up to some technical assumptions, *if f is a C^2 function with values on a d-dimensional smooth manifold $\mathcal{M} \subset \mathbb{R}^D$, then there exists a deep ReLU NN with W parameters and a finite number of layers computing a function f_W such that*

$$\|f - f_W\|_\infty \leq C\, W^{-\frac{2}{d}}, \tag{1}$$

for some C independent of W (but dependent on f, $\mathcal{M}$ and D). That is, the number of parameters W needed to achieve arbitrary approximation accuracy for f using a NN scales *essentially* with the manifold dimension d, and depends only weakly on the ambient dimension D, where this weak dependence is hidden non-explicitly in the constant C. Successive contributions from the literature extended and refined the approximation result stated above in various ways, by considering functions in a Hölder space with smoothness index $\beta \in (0, 1]$ and deriving, under the manifold hypothesis, estimates of the form

$$\|f - f_W\|_\infty \leq C\, W^{-\frac{\beta}{d}}, \tag{2}$$

where the constant C depends explicitly on the Hölder norm of f, $\mathcal{M}$ and D (cf. [16, 17, 18, 19]).

As the input layer necessarily depends on the input dimension D, the dependence of C on D in (1) and (2) cannot be avoided. The best one can hope for is to remove the dependence on D *up to the input layer.* In line with this observation, the major contribution of this paper is the derivation of a refined approximation estimate of the form

$$\|f - f_W\|_\infty \leq C \, \left(N_0^2 L_0^2 \log_3(N_0 + 2)\right)^{-\beta/d_e}, \tag{3}$$

where N_0 is the maximum width of all hidden layers, L_0 is the number of hidden layers of a deep ReLU NNs and d_e is an *effective dimension* closely related to d (see Theorem 1 and Remark 5 for the precise statement). We argue that controlling the approximation properties of an NN using the parameters of the hidden layers is practically and conceptually appropriate as only these parameters, unlike those in the input layer, are part of the NN design.

Remarkably, *our constant C in* (3) *does not depend on the ambient dimension D.* In our estimate, *the number of hidden parameters of the NN needed to achieve arbitrary approximation accuracy of f on $\mathcal{M}$ only depends on the manifold dimension* (through an appropriate effective dimension d_e dependent on d but not on D, as shown in our Theorem 1 below). Up to our knowledge, this is the first result of this type, as in all published results the constant C in (3) depends on D; further, no existing result of the form (2) can be converted directly to an estimate of the form (3) with C independent of D.

To achieve our improved approximation estimate, one of the novelties of our approach is the careful application of a version of the Johnson-Lindenstrauss Lemma on smooth manifolds (Theorem 4 below) that we use to map points from the ambient space nearly isometrically into a lower dimensional domain. An important implication of our new approximation theorem is an improved estimate of the generalization error using deep ReLU NNs. Our Theorem 2 proves that, *using samples taken from an unknown Hölder function satisfying the manifold hypothesis, not only the decay rate of the regression error but also the multiplicative constant are independent of the ambient dimension D; both quantities only depend on the manifold dimension.* This result significantly improves existing generalization estimates in the literature.

The rest of the paper is organized as follows. We introduce the relevant notation and definitions in Sec. 2. We next present our main results in Sec. 3, together with a discussion of the related literature. We finally present the proofs of our main theorems in Sec. 4.

2. Notation and Definitions

Throughout the paper, we denote as $\mathbb{N} = \{1, 2, \cdots\}$ the set of natural number and, for any $n \in \mathbb{N}$, we use the compact notation $[n] := \{1, 2, \cdots, n\}$.

For any a, $b \in \mathbb{R}$, we use the notation $a \vee b := \max\{a, b\}$. The floor of $a \in \mathbb{R}$, denoted as $\lfloor a \rfloor$, is the greatest integer less than or equal to a and the ceiling of a, denoted as $\lceil a \rceil$, is the least integer greater than or equal to a. We use boldface symbols to denote vectors and regular fonts with a subscript to denote vector coordinates, i.e., $\mathbf{x} \in \mathbb{R}^n$ and x_i is the i-th coordinate of $\mathbf{x}$. For a vector $\mathbf{x} \in \mathbb{R}^n$, we use the standard norm notation, $\|\mathbf{x}\|_p = \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}$ and $\|\mathbf{x}\|_\infty = \max_{1\leq i\leq n} |x_i|$. For a matrix $A \in \mathbb{R}^{m\times n}$, we use the norm notation $\|A\|_1 = \max_{1\leq j\leq n} \sum_{i=1}^m |a_{ij}|$ and $\|A\|_\infty = \max_{1\leq i\leq m} \sum_{j=1}^n |a_{ij}|$.

For a measure μ on a measurable set $E \subset \mathbb{R}^D$ and a measurable function $f : E \to \mathbb{R}$, the L^p norm of f, for $1 \leq p < \infty$, is the integral $\|f\|_{L^p(E,\mu)} := \left(\int_E |f|^p \, d\mu\right)^{1/p}$. For $p = \infty$, we denote as usual the supremum norm of f on E as $\|f\|_\infty := \operatorname{ess\,sup}_{\mathbf{x}\in E} |f(\mathbf{x})|$.

We will derive our approximation results on functions from the Hölder space.

Definition 1 (Hölder space). *Let $\beta > 0$ and $F \subseteq \mathbb{R}^D$ be a closed set. The Hölder space with degree of smoothness β on F is defined as*

$$\mathcal{H}(\beta, F) = \left\{ f \in C^{\lfloor \beta \rfloor}(F) \;\middle|\; ||f||_{\mathcal{H}(\beta,F)} < \infty \right\},$$

where, for $f : F \to \mathbb{R}$, the Hölder norm of f is defined by

$$||f||_{\mathcal{H}(\beta,F)} = \max\left\{ \max_{\alpha:\, ||\alpha||_1 \leq \lfloor\beta\rfloor} \sup_{\mathbf{x}\in F} |\partial^\alpha f(\mathbf{x})|, \max_{\alpha:\, ||\alpha||_1 = \lfloor\beta\rfloor} \sup_{\substack{\mathbf{x},\mathbf{y}\in F \\ \mathbf{x}\neq\mathbf{y}}} \frac{|\partial^\alpha f(\mathbf{x}) - \partial^\alpha f(\mathbf{y})|}{||\mathbf{x}-\mathbf{y}||_2^{\beta-\lfloor\beta\rfloor}} \right\}.$$

For $M > 0$, we denote the closed ball in $\mathcal{H}(\beta, F)$ with radius M as

$$\mathcal{H}(\beta, F, M) := \left\{ f \in C^{\lfloor \beta \rfloor}(F) \;\middle|\; ||f||_{\mathcal{H}(\beta,F)} \leq M \right\}.$$

We adopt the following definition of a NN and its realizations, similar to [20, 21, 22].

Definition 2 (Neural Network (NN)). *Given D, $L \in \mathbb{N}$, a NN with input dimension D and L layers is a sequence of matrix-vector tuples*

$$\Phi = ((A_1, b_1), (A_2, b_2), \cdots, (A_{L+1}, b_{L+1})),$$

where $w_0 = D$, $w_1, \cdots, w_{L+1} \in \mathbb{N}$, and where $A_l \in \mathbb{N}^{w_l \times w_{l-1}}$ and $b_l \in \mathbb{R}^{w_l}$ for $l = 1, \cdots, L+1$. We refer to w_{L+1} as the output dimension of Φ and to $w_1, \ldots, w_L$ as the width (or the number of neurons) of the inner layers. Also, we denote with the symbol $\mathcal{N}(\Phi)$ the maximum width for all **hidden** *layers (or width of Φ) with $\mathcal{L}(\Phi)$ the number of* **hidden** *layers of Φ and with $\mathcal{B}(\Phi) = \max_{l=1,\cdots,L+1}\{||Vec(A_l)||_\infty, ||b_l||_\infty\}$ the scale of the weights of Φ, where $Vec(A)$ is the vectorization of matrix A.*

For a NN Φ and an activation function $\varrho : \mathbb{R} \to \mathbb{R}$, the realization $R(\Phi)$ of Φ is the measurable map $R(\Phi) : \mathbb{R}^{w_0} \to \mathbb{R}^{w_{L+1}}$, where the output $\mathbf{x}_{L+1} = R(\Phi)(\mathbf{x}) \in \mathbb{R}^{w_{L+1}}$ is given by

$$\begin{aligned} \mathbf{x}_0 &:= \mathbf{x} \in \mathbb{R}^{w_0} \\ \mathbf{x}_l &:= \varrho(A_l \mathbf{x}_{l-1} + b_l), for\ l = 1, \cdots, L \\ \mathbf{x}_{L+1} &:= A_{L+1}\mathbf{x}_L + b_{L+1} \end{aligned} \tag{4}$$

and ϱ is understood to act component-wise.

Note: in the following, throughout the paper, we will assume that ϱ is the rectified linear unit (*ReLU*) activation function, defined by $\varrho(\cdot) = \max\{0, \cdot\}$.

Definition 3 (Neural Network Class). *For a tuple $(N_0, L_0, B_0) \in \mathbb{N} \times \mathbb{N} \times \mathbb{R}$, we define as $\mathcal{F}(W_0, L_0, B_0)$ the class of NNs*

$$\mathcal{F}(N_0, L_0, B_0) := \left\{ R(\Phi) : [0,1]^{w_0} \to \mathbb{R}^{w_{\mathcal{L}(\Phi)+1}} \,\middle|\, \mathcal{N}(\Phi) \leq N_0, \mathcal{L}(\Phi) \leq L_0, \mathcal{B}(\Phi) \leq B_0 \right\}.$$

Henceforth, the expression "a NN Φ with width N_0, depth L_0 and scale B_0" means that $R(\Phi) \in \mathcal{F}(N_0, L_0, B_0)$, i.e.,

- the maximum width of the NN for all **hidden** layers is no more than N_0;
- the number of **hidden** layers of the NN is no more than L_0;
- the scale of weights of the NN is no more than B_0.

For example, in the expression (4), Φ is a NN with width $\max_{l=1,\cdots,L}\{\omega_i\}$, depth L and scale $\max_{l=1,\cdots,L+1}\{||Vec(A_l)||_\infty, ||b_l||_\infty\}$.

We define concatenations and parallelizations of NNs following [20].

Definition 4 (Concatenation of NNs). *Given two NNs*

$$\Phi^1 = ((A_1^1, b_1^1), \cdots, (A_{L_1+1}^1, b_{L_1+1}^1)) \text{ and } \Phi^2 = ((A_1^2, b_1^2), \cdots, (A_{L_2+1}^2, b_{L_2+1}^2))$$

for L_1, $L_2 \in \mathbb{N}$ satisfying the condition that the input layer of Φ^2 has the same dimension as the output layer of Φ^1, which we assume to be w_{L_1+1}, the concatenation of Φ^1 and Φ^2 is

$$\Phi^2 \odot \Phi^1$$
$$= \left((A_1^1, b_1^1), \cdots, (A_{L_1}^1, b_{L_1}^1), \left(\begin{pmatrix} A_{L_1+1}^1 \\ -A_{L_1+1}^1 \end{pmatrix}, \begin{pmatrix} b_{L_1+1}^1 \\ -b_{L_1+1}^1 \end{pmatrix} \right), ([A_1^2; -A_1^2], b_1^2), (A_2^2, b_2^2), \cdots, (A_{L_2+1}^2, b_{L_2+1}^2) \right)$$

Remark 1. A straightforward calculation shows that

$$R(\Phi^2 \odot \Phi^1) = R(\Phi^2) \circ R(\Phi^1), \quad \mathcal{N}(\Phi^2 \odot \Phi^1) = max\{\mathcal{N}(\Phi^2), \mathcal{N}(\Phi^1), 2w_{L_1+1}\},$$
$$\mathcal{L}(\Phi^2 \odot \Phi^1) = 1 + \mathcal{L}(\Phi^2) + \mathcal{L}(\Phi^1), \quad \mathcal{B}(\Phi^2 \odot \Phi^1) = max\{\mathcal{B}(\Phi^2), \mathcal{B}(\Phi^1)\}.$$

By induction, $\Phi^m \odot \cdots \odot \Phi^1$ has the following properties

$$\begin{aligned}
&R(\Phi^m \odot \cdots \odot \Phi^1) = R(\Phi^m) \circ \cdots \circ R(\Phi^1), \\
&\mathcal{N}(\Phi^m \odot \cdots \odot \Phi^1) = \max_{1 \leq l \leq m} \mathcal{N}(\Phi^l) \vee 2 \max_{1 \leq l \leq m-1} w_{L_l+1}, \\
&\mathcal{L}(\Phi^m \odot \cdots \odot \Phi^1) = (m-1) + \sum_{l=1}^{m} \mathcal{L}(\Phi^l), \\
&\mathcal{B}(\Phi^m \odot \cdots \odot \Phi^1) = \max_{1 \leq l \leq m} \mathcal{B}(\Phi^l).
\end{aligned}$$

Definition 5 (Parallelization of NNs). *Let L, D_1, $D_2 \in \mathbb{N}$ and let*

$$\Phi^1 = ((A_1^1, b_1^1), \cdots, (A_{L+1}^1, b_{L+1}^1)),\ \Phi^2 = ((A_1^2, b_1^2), \cdots, (A_{L+1}^2, b_{L+1}^2))$$

be two NNs with L-layers and with D_1-dimensional and D_2-dimensional input, respectively. We define the parallelization with shared inputs of Φ^1 and Φ^2, denoted as $P(\Phi^1, \Phi^2)$, by

$$P(\Phi^1, \Phi^2) := ((\tilde{A}_1, \tilde{b}_1), (\tilde{A}_2, \tilde{b}_2), \cdots, (\tilde{A}_{L+1}, \tilde{b}_{L+1})), \quad \textit{if } D_1 = D_2,$$

and define the parallelization without shared inputs of Φ^1 and Φ^2, denoted as $FP(\Phi^1, \Phi^2)$, by

$$FP(\Phi^1, \Phi^2) := ((\tilde{A}_1, \tilde{b}_1), \cdots, (\tilde{A}_{L+1}, \tilde{b}_{L+1})),\ \textit{for any } D_1,\ D_2 \in \mathbb{N},$$

where, for $1 < l \leq L + 1$,

$$\tilde{A}_1 := \begin{pmatrix} A_1^1 \\ A_1^2 \end{pmatrix}, \tilde{b}_1 := \begin{pmatrix} b_1^1 \\ b_1^2 \end{pmatrix},\ \textit{and } \tilde{A}_l := \begin{pmatrix} A_l^1 & 0 \\ 0 & A_l^2 \end{pmatrix}, \hat{b}_l := \begin{pmatrix} b_l^1 \\ b_l^2 \end{pmatrix}.$$

Remark 2. A straightforward calculation gives the following relations

$$\begin{aligned} &\mathcal{N}(P(\Phi^1, \Phi^2)) = \mathcal{N}(FP(\Phi^1, \Phi^2)) = \mathcal{N}(\Phi^1) + \mathcal{N}(\Phi^2), \\ &\mathcal{L}(P(\Phi^1, \Phi^2)) = \mathcal{L}(FP(\Phi^1, \Phi^2)) = L, \\ &\text{and } \mathcal{B}(P(\Phi^1, \Phi^2)) = \mathcal{B}(FP(\Phi^1, \Phi^2)) = \max\{\mathcal{B}(\Phi^1), \mathcal{B}(\Phi^2)\}. \end{aligned}$$

By induction, we obtain the following properties about the generalization using L layers:

$$\begin{gathered} \mathcal{N}(P(\Phi^1, \cdots, \Phi^m)) = \mathcal{N}(FP(\Phi^1, \cdots, \Phi^m)) = \textstyle\sum_{l=1}^m \mathcal{N}(\Phi^l), \\ \mathcal{L}(P(\Phi^1, \cdots, \Phi^m)) = \mathcal{L}(FP(\Phi^1, \cdots, \Phi^m)) = L, \\ \mathcal{B}(P(\Phi^1, \cdots, \Phi^m)) = \mathcal{B}(FP(\Phi^1, \cdots, \Phi^m)) = \max_{1 \leq l \leq m} \mathcal{B}(\Phi^l). \end{gathered}$$

Definition 6 (Identity function). *For $D, L \in \mathbb{N}$, we define the following NN with depth L to approximate the identity function on $\mathbb{R}^D$:*

$$\Phi_{D,L}^{Id} := \left(\left(\begin{pmatrix} I_D \\ -I_D \end{pmatrix}, 0 \right), \underbrace{(I_{2D}, 0), \cdots, (I_{2D}, 0)}_{L-1 \textit{ times}}, ([I_D \mid -I_D], 0) \right),$$

where $I_D \in \mathbb{R}^{D \times D}$ is the identity matrix.

Remark 3. A direct calculation shows that $\Phi_{D,L}^{Id}$ is a L-layers NN of the identity function, i.e., $R(\Phi_{D,L}^{Id})(\mathbf{x}) = \mathbf{x}$ for $\mathbf{x} \in \mathbb{R}^D$. Also, $\mathcal{N}(\Phi_{D,L}^{Id}) = 2D$, $\mathcal{L}(\Phi_{D,L}^{Id}) = L$ and $\mathcal{B}(\Phi_{D,L}^{Id}) = 1$.

Definition 7 (Maximum function). *We define the following NN to emulate the maximum function*

$$\Phi^{max} := \left(\left(\left(\begin{array}{cc} 1 & 0 \\ -1 & 1 \end{array} \right), \left(\begin{array}{c} 0 \\ 0 \end{array} \right) \right), ((1,1),0) \right). \tag{5}$$

Remark 4. A direct observation shows that, if $R(\Phi^{max})(x_1, x_2) = \max\{x_1, x_2\}$, for all $(x_1, x_2) \in \mathbb{R}^2$, then

$$\mathcal{N}(\Phi^{max}) = 2, \quad \mathcal{L}(\Phi^{max}) = 1, \quad \mathcal{B}(\Phi^{max}) = 1.$$

3. Main Theorems

Our study considers functions in $\mathcal{H}(\beta, \mathcal{M}, M)$, where the domain $\mathcal{M} \subset [0,1]^D$ is a smooth compact d-dimensional manifold with $d \ll D$. Our first theorem proves that these functions can be approximated using a deep ReLU NN whose parameters, except for the input layer, depend on the manifold dimension d but not the ambient dimension D. Before presenting our main theorems, we briefly review the related literature.

Several works have recently studied the approximation and generalization capabilities of deep ReLU NNs under the assumption of low-dimensional data structures, i.e., for functions with domain in a low-dimensional embedded manifold. We can roughly divide these works into two groups that we describe as *constructive* and *non-constructive* approaches depending on whether the map between ambient space, where data are nominally defined, and a lower dimensional space is explicitly constructed or not.

Constructive-method papers typically define an explicit chart to map samples from the ambient space $\mathbb{R}^D$ into a d-dimensional manifold, with d smaller than D, and use the properties of the map to approximate functions on the manifold using NNs. Such papers include the work by Chen *et al.* [16], Schmidt-Hieber [18], Nakada and Imaizumi [19] and Cloninger and Klock [17], and derive uniform approximation estimates of the form (1) or (2), where the number of network parameters scales essentially with the dimension d. However, as we mentioned above, the constant C appearing in the estimates depends on the ambient dimension D.

By contrast, another group of papers - that we describe as non-constructive - use versions of the Johnson-Lindenstrauss lemma to claim the existence of a nearly isometric map between the ambient space $\mathbb{R}^D$ and a lower dimensional manifold. These papers include the work by Cai *et al.* [23] (2018), Shen *et al.* [21] (2020) and Jiao *et al.* [24] (2021). Since non-constructive approaches aim to preserve the global structure of the manifold, they usually require stronger regularity assumptions on the manifold as compared to constructive approaches which use local isometric charts (e.g., the constructive approach by Nakada and Imaizumi [19] does not require the manifold to be smooth). On the other hand, non-constructive methods generally yield better estimates with respect to the constant C appearing in the uniform estimate (2) or (3). To provide a more precise quantification of expressive power, some authors [21, 22, 25] recently proposed approximation estimates of the form (3). Using the maximum width of all hidden layers N_0 and the number of hidden layers L_0 is not only more practical as these numbers are design parameters of the NN, but also more general and precise; an estimate in terms of N_0, L_0 can be used to derive an estimate in terms of the total number of NN parameters W but not vice versa.

In this paper, we adopt the non-constructive point of view and apply a manifold version of the Johnson-Lindenstrauss lemma that is a variant of a result by Eftekhari and Wakin [26] and establishes the existence of a linear transformation mapping points in $\mathcal{M} \subseteq [0,1]^D$ nearly isometrically into a lower dimensional domain.

The application of this nearly isometric mapping requires some technical assumptions about the manifold. Namely, we assume that $\mathcal{M}$ is a compact d-dimensional Riemannian submanifold of $\mathbb{R}^D$ with a bounded condition number. We recall a Riemannian submanifold $\mathcal{M}$ of Riemannian manifold $\widetilde{\mathcal{M}}$ is a submanifold of $\widetilde{\mathcal{M}}$ equipped with the Riemannian metric inherited from $\widetilde{\mathcal{M}}$. We also recall that Riemannian manifolds extend the notion of Euclidean space into more general curved space. More precisely, a Riemannian manifold is a real, smooth manifold equipped with a positive-definite inner product g on the tangent space of the manifold at each point. The family g of inner products is called a Riemannian metric and allows one to define a distance so that a Riemannian manifold is also a metric space. The *condition number* of a manifold or a submanifold $\mathcal{M}$ is defined as $1/\tau$, where τ, called the *reach* of the manifold, is the largest number such that any point at distance less than τ from $\mathcal{M}$ has a unique nearest point on $\mathcal{M}$. The significance of the condition number is that it controls both local properties of the manifold, such as curvature (which is bounded by $1/\tau$), and global properties, such as self-avoidance. We refer to [27] for additional properties of the condition number. We also recall the definition of diameter of a Riemannian manifold $\mathcal{M}$:

$$\operatorname{diam}(\mathcal{M}) := \sup_{\mathbf{x},\mathbf{y}\in\mathcal{M}} \Delta(\mathbf{x},\mathbf{y}),$$

where $\Delta(\mathbf{x},\mathbf{y})$ is the geodesic distance between $\mathbf{x}$ and $\mathbf{y}$ on $\mathcal{M}$.

We can now state our first theorem, whose proof is given in Sec. 4.

Theorem 1. *For $D \in \mathbb{N}$, let $\mathcal{M}$ be a compact d-dimensional Riemannian submanifold contained in $[0,1]^D$, with $\mathbf{0} \in \mathcal{M}$, having condition number $1/\tau$ and volume $V_{\mathcal{M}}$ satisfying $\frac{V_{\mathcal{M}}}{\tau^d} \geq \left(\frac{21}{2\sqrt{d}}\right)^d$, and let $f_0 \in \mathcal{H}(\beta, \mathcal{M}, M)$, where $M > 0$ and $\beta \in (0,1]$. Let*

$$d_e = \left\lceil 828\left(24d + 2d\log\left(\frac{9\sqrt{d}}{\tau}\right) + \log(2V_{\mathcal{M}}^2)\right)\right\rceil. \tag{6}$$

Then, for any $N, L \in \mathbb{N}$, there exists a NN Φ^{f_0} with $R(\Phi^{f_0}) \in \mathcal{F}(N_0, L_0, B_0)$

where

$$\begin{aligned}
&N_0 = 2^{d_e/\beta+1}\,(6d_e+47)\,N,\\
&L_0 = (28d_e^2-15)L,\\
&B_0 = \tfrac{1}{4\,\mathrm{diam}\,(\mathcal{M})} \vee 3\left\lceil 2^{d_e/\beta}\left(N^2\log_3(N+2)\right)^{1/d_e}\right\rceil \left\lceil L^{2/d_e}\right\rceil \vee 64\,(\mathrm{diam}(\mathcal{M}))^{2\beta}\,{d_e}^{\beta}M^2,
\end{aligned}$$

so that

$$\begin{aligned}
&\max_{\mathbf{x}\in\mathcal{M}} |R(\Phi^{f_0})(\mathbf{x}) - f_0(\mathbf{x})|\\
&\leq \left(384\,(\mathrm{diam}(\mathcal{M}))^{2\beta}\,{d_e}^{\beta}M^2+6\right) d_e^{\beta/2}\left(N^2L^2\log_3(N+2)\right)^{-\beta/d_e}.
\end{aligned}$$

Remark 5. From Theorem 1, a direct calculation gives that there is a constant $C = C(\beta, d_e, M, \mathrm{diam}(\mathcal{M}))$ dependent on β, d_e, M and $\mathrm{diam}(\mathcal{M}))$ but not on D such that

$$\max_{\mathbf{x}\in\mathcal{M}} |R(\Phi^{f_0})(\mathbf{x}) - f_0(\mathbf{x})| \leq C\,\left(N_0^2L_0^2\log_3(N_0+2)\right)^{-\beta/d_e}.$$

This is the same decay rate found in [25], where it is also shown that this is the optimal decay rate, up to a constant [25, Thm 2.4]. As compared with the similar type of estimates in [21, 23, 24] where the multiplicative constant C depends on the ambient dimension D, the constant C in our estimate in Theorem 1 does not depend on D. Up to our knowledge, our approximation result is the first one of this form to avoid the dependence of the multiplicative constant C on D.

Remark 6. Theorem 1 and Remark 5 show that, under the manifold hypothesis, the approximation rate and the multiplicative constant C do not scale with the ambient dimension D but rather with an *effective dimension* d_e which depends on the manifold dimension d and the geometry of the manifold but not on D. The size of d_e depends on the complexity of the manifold in the sense that, as implied by (6), higher values of the volume $V_{\mathcal{M}}$ or condition number $1/\tau$ make d_e larger. To further illustrate the impact of the manifold geometry on the size of d_e let us consider, as an example, the case where $\mathcal{M}$ is a d-dimensional unit sphere. In this case, $\tau = 1$ and $V_{\mathcal{M}} = \frac{\pi^{d/2}}{\Gamma(d/2+1)} \propto d^{-d/2}$. It follows that $\log(V_{\mathcal{M}}^2) \propto -d\log d$ and this cancels the term $2d\log\sqrt{d} = d\log d$ on the right-hand side of (6), so that in this case d_e grows linearly on d, namely $d_e \approx 19{,}872d$). For a more general manifold, the volume also depends on the curvature so that, in general, d_e would grow like $d\log d$. In other words, a manifold with a larger volume $V_{\mathcal{M}}$ or a larger condition number $1/\tau$ (hence greater curvature) is more complex; hence the effective dimension d_e of the projection space into which manifold data are mapped nearly isometrically needs to be sufficiently large. Our observations show that the numerical value of d_e can be significantly larger than d and, thus, the estimate in terms of d_e is only useful when $D \gg d$. We believe that the multiplicative constant in (6) can be improved by refining the estimate of Theorem 4; this is however beyond the scope of this paper.

We next use our approximation result to derive an estimate of the form (2), where the approximation error is controlled by the total number of network parameters. Using the observation that a NN Φ^{f_0} with input space $[0,1]^D$ having L_0 inner layers and maximum width of the inner layers N_0 has at most $W = (Dd_e + d_e + N_0 d_e + N_0^2 + N_0)L_0 > (D+1)d_e + (L_0-2)N_0^2 + (L_0+d_\delta)N_0 + 1$ nonzero parameters, Theorem 1 implies the following Corollary.

Corollary 1. *Fix $D \in \mathbb{N}$, $M > 0$, $\beta \in (0,1]$. Let $\mathcal{M}$ be a compact d-dimensional Riemannian submanifold contained in $[0,1]^D$, with $\mathbf{0} \in \mathcal{M}$, having condition number $1/\tau$ and volume $V_{\mathcal{M}}$ satisfying $\frac{V_{\mathcal{M}}}{\tau^d} \geq \left(\frac{21}{2\sqrt{d}}\right)^d$. Let d_e be given by (6). Given $f_0 \in \mathcal{H}(\beta, \mathcal{M}, M)$, for any $L \in \mathbb{N}$, there exists a NN Φ^{f_0} with depth $L_0 = (28d_e^2 - 15)L$ and width $N_0 = 2^{d_e/\beta+1}(6d_e + 47)$ having at most $W := (Dd_e + d_e + N_0 d_e + N_0^2 + N_0)L_0$ nonzero parameters such that*

$$\max_{x \in \mathcal{M}} |R(\Phi^{f_0})(x) - f_0(x)| \leq C N_0^{-\beta/d_e} (Dd_e + d_e + N_0 d_e + N_0^2 + N_0)^{\beta/d_e} W^{-\beta/d_e},$$

where $C = C(\beta, d_e, M, \mathrm{diam}(\mathcal{M}))$ is independent of D.

Remark 7. By Corollary 1, our Remark 5 implies an estimate of the form (2) where the multiplicative constant C depends weakly on D, namely as D^{β/d_e}. For large ambient dimension D, our result improves upon existing estimates from the literature; note that the constant C depends as $(D \log D)^{\beta/d}$ in [16], as $(D^{3+D})^{\beta/d}$ in [18], as D^β in [19], as $\left(\frac{D}{\varepsilon^d} \log \frac{D}{\varepsilon^{3+d}}\right)^{\beta/d}$ in [17].

We observe that, while we have used our estimate of the form (3) to imply an estimate of the form (2), the converse implication does not follow by a direct argument.

Remark 8. The approach presented above through Theorem 1 and Corollary 1 addresses the issue of how well we can approximate a target function $f \in \mathcal{H}(\beta, \mathcal{M}, M)$ and is meant to provide an explanation of the approximation properties of DNNs observed in applications. Our result does not provide a practical method that could be implemented numerically. The construction of numerical approximation procedures raises other problems, most notably the stability of the approximation algorithm. We refer the reader to [7, Sec. 9] and [28] for more details about this topic.

We next analyze the generalization error of DNNs by considering a nonparametric regression problem associated with n observations $\{(X_i, Y_i)\}_{i=1}^n \subseteq \mathcal{M} \times \mathbb{R}$ from the model

$$Y_i = f_0(X_i) + \varepsilon_i, \quad i = 1, \cdots, n, \tag{7}$$

where $f_0 \in \mathcal{H}(\beta, \mathcal{M}, M)$, the covariates X_i marginally follow a probability measure μ and the errors ε_i are i.i.d normally distributed with mean 0 and variance σ^2, independently of the X_i.

As above, we consider the situation where f_0 is contained on a d-dimensional manifold $\mathcal{M}$ inside the compact set $[0,1]^D$. To find the solution of the regression problem, we introduce a NN class $\hat{\mathcal{F}}(N, L, B) := \{ g \circ \Psi \,|\, g \in \mathcal{F}(N, L, B)\}$, where $\Psi : \mathcal{M} \mapsto [0,1]^{d_e}$ is the affine transformation

$$\Psi(\mathbf{x}) := \frac{1}{4\,\mathrm{diam}(\mathcal{M})} A\,\mathbf{x} - \frac{1}{4\,\mathrm{diam}(\mathcal{M})} \mathbf{y}_0, \tag{8}$$

defined for $\mathbf{x} \in \mathcal{M}$, d_e is the effective dimension given by (6), A is a random $d_e \times D$ matrix populated with i.i.d. zero-mean Gaussian random variables with variance $1/d_e$ and $\mathbf{y}_0 \in \mathbb{R}^{d_e}$ is chosen to satisfy $A(\mathcal{M}) \subseteq \{4\,\mathrm{diam}(\mathcal{M})\mathbf{y} + \mathbf{y}_0 \mid \mathbf{y} \in [0,1]^{d_e}\}$. To estimate the function f_0, we then compute the least square estimator $\widehat{f} \in \hat{\mathcal{F}}(N, L, B)$ of f_0 associated with the following empirical risk minimization

$$\widehat{f} = \underset{f \in \hat{\mathcal{F}}(N,L,B)}{\mathrm{argmin}} \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - f(X_i)\right)^2 . \tag{9}$$

The following theorem, whose proof is given in Sec. 4, relies on the approximation properties of Theorem 1 to derive a bound for the generalization error. Due to the independence of our approximation estimate on the ambient dimension D, our generalization error bound depends on the effective manifold dimension d_e only and not on the ambient dimension D.

Theorem 2. *Fix $D \in \mathbb{N}$, $M > 0$, $\beta \in (0, 1]$. Let $\mathcal{M}$ be a compact d-dimensional Riemannian submanifold contained in $[0,1]^D$, with $\mathbf{0} \in \mathcal{M}$, having condition number $1/\tau$ and volume $V_{\mathcal{M}}$ satisfying $\frac{V_{\mathcal{M}}}{\tau^d} \geq \left(\frac{21}{2\sqrt{d}}\right)^d$ and let d_e be given by* (6). *Given $f_0 \in \mathcal{H}(\beta, \mathcal{M}, M)$, let $\widehat{f}$ be the solution of the minimization of empirical risk in* (9)*, where we choose the NN class $\hat{\mathcal{F}}(N_1, L_1, B_1)$ defined above with*

$$\begin{aligned} N_1 &= 2^{d_e/\beta+1}\,(6d_e + 47)\, n^{d_e/(4\beta+2d_e)}, \\ L_1 &= (28d_e^2 - 15)C_2^{d_e/2\beta}, \\ B_1 &= 3\left\lceil 2^{d_e/\beta} n^{1/(2\beta+d_e)} \left(\log_3(n^{d_e/(4\beta+2d_e)} + 2)\right)^{1/d_e} \right\rceil \lceil C_2^{1/\beta} \rceil, \end{aligned}$$

where $C_2 = d_e^{\beta/2}(384\,(\mathrm{diam}(\mathcal{M}))^{2\beta} {d_e}^{\beta} M^2 + 6)$. Then there exists a constant $C = C(\sigma, \beta, d_e, M, \mathrm{diam}(\mathcal{M}))$, independent of D, such that

$$\|\widehat{f} - f_0\|^2_{L^2(\mathcal{M},\mu)} \leq C n^{-2\beta/(2\beta+d_e)}(1 + \log n)^2$$

holds with probability at least $1 - 2\exp\left(-n^{d_e/(2\beta+d_e)}\right)$ for any $n \geq N$ with a sufficiently large N, where μ is the marginal distribution of X on $\mathcal{M}$.

Remark 9. The constant C controlling the generalization error in Theorem 2 depends only on the effective manifold dimension d_e and not on the ambient dimension D. For large values of D, our result improves existing estimates from the literature, including Nakada and Imaizumi [19], Chen *et al.* [16] where the constant C of the generalization bound depends on the ambient dimension D (polynomially in [19] and $D \log D$ in [16]). Schmidt-Hieber [18] derive the same decay rate in terms of the manifold dimension d rather than d_e; their multiplicative constant C is dependent on D.

4. Proofs of the theorems

4.1. Proof of Theorem 1

We first recall several estimates about the computational complexity of NNs approximating different types of elementary functions.

We start by recalling the approximation estimate of the function xy which is due to Yarotsky [29]. For consistency with our notation, we refer the reader to Lu *et al.* [22].

Lemma 1 (Lemma 4.2 in [22]). *For any $N, L \in \mathbb{N}$ and $a, b \in \mathbb{R}$ with $a < b$, there exists a NN Φ^{xy} with width $9N + 1$, depth L and scale $(b-a)^2$ such that*

$$|R(\Phi^{xy})(x, y) - xy| \leq 6(b-a)^2 N^{-L} \text{ for } x, y \in [a, b].$$

Remark 10. In particular, if $x = 0$ or $y = 0$, then $R(\Phi^{xy})(x, y) = 0$.

The lemma below constructs a NN to approximate the product function $f(x_1, \cdots, x_d) = \prod_{i=1}^d x_i$ on $[0, 1]^d$.

Lemma 2 (Lemma 5.3 in [22]). *For any $N, L, d \in \mathbb{N}$ with $d \geq 2$, there exists a NN Φ^{mult} with width $9(N+1) + d - 1$, depth $7d(d-1)L$ and scale 2 such that*

$$\left| R(\Phi^{mult})(\mathbf{x}) - \prod_{i=1}^{d} x_i \right| \leq 9(N+1)^{-7dL} \text{ for } \mathbf{x} = (x_1, x_2, \cdots, x_d) \in [0, 1]^d.$$

The following theorem constructs NNs approximating functions in the space $\mathcal{H}(\beta, [0, 1]^d, M)$, for $\beta \in (0, 1]$ and the statement is similar to [21, Theorem 1.1] and [25, Corollary 1.3]. However, our proof is more straightforward than the argument used in [21, 25] and adapts an idea from Yarotsky [29, Theorem 1] to derive an estimate in terms of N and L. The approximation rate is optimal up to a constant as shown by [25, Theorem 2.4].

Theorem 3. *For $d \in \mathbb{N}$, $\beta \in (0, 1]$, $M > 0$ with $d \geq 4$, take any $f_0 \in \mathcal{H}(\beta, [0, 1]^d, M)$. For any $N, L \in \mathbb{N}$, there is a NN Φ^{f_0} with width $2^{d/\beta+1}(6d + 47)N$, depth $(28d^2 - 16)L$ and scale $3\lceil 2^{d/\beta} \left(N^2 \log_3(N+2)\right)^{1/d} \rceil \lceil L^{2/d} \rceil \vee M^2$ such that*

$$\max_{\mathbf{x} \in [0,1]^d} |R(\Phi^{f_0})(\mathbf{x}) - f_0(\mathbf{x})| \leq 6(M^2 + 1) d^{\beta/2} \left(N^2 L^2 \log_3(N+2)\right)^{-\beta/d}.$$

Proof. The proof involves four steps. Firstly, we create a partition of unity and construct a NN to approximate it. Secondly, we build a NN to estimate the target function f_0. Thirdly, we estimate the approximation error. Lastly, we determine the size of the NN of the approximator of the target function f_0.

Let $K \in \mathbb{N}$. Consider a partition of unity formed by a grid of $(K+1)^d$ functions $\rho_{\mathbf{m}}$ on $[0,1]^d$:

$$\sum_{\mathbf{m}} \rho_{\mathbf{m}}(\mathbf{x}) \equiv 1, \quad \mathbf{x} \in [0,1]^d, \tag{10}$$

where $\mathbf{m} = (m_1, \cdots, m_d) \in \{0, 1, \cdots, K\}^d$, and the function $\rho_{\mathbf{m}}$ is defined as follows:

$$\rho_{\mathbf{m}}(\mathbf{x}) = \prod_{i=1}^{d} \psi\left(3K\left(x_i - \frac{m_i}{K}\right)\right), \tag{11}$$

where

$$\psi(x) = \begin{cases} 1, & |x| < 1 \\ 2 - |x|, & 1 \le |x| \le 2 \\ 0, & |x| > 2 \end{cases}. \tag{12}$$

We observe that

$$\|\rho_{\mathbf{m}}\|_\infty = 1 \tag{13}$$

and the support of $\rho_{\mathbf{m}}$ is contained in the set

$$\left\{\mathbf{x} : \left|x_i - \tfrac{m_i}{K}\right| \le \tfrac{2}{3K},\, i = 1, \cdots, d\right\} \subset \left\{\mathbf{x} : \left|x_i - \tfrac{m_i}{K}\right| \le \tfrac{1}{K},\, i = 1, \cdots, d\right\}. \tag{14}$$

A NN approximation of $\rho_{\mathbf{m}}$ in (11) is

$$\begin{aligned} \Phi^{\rho_{\mathbf{m}}} :=& \Phi^{\min} \odot P(((0,1)),((1,0))) \odot ((1, -9\left(2\lceil N^{(d-3)/d}\rceil + 1\right)^{-7d\lceil L^{(d-2)/d}\rceil}), (1,0), (1,0)) \\ &\odot \Phi^{\mathrm{mult}} \odot \underbrace{\Phi^{\psi} \odot ((3K, 0)) \odot \left(\left(I_{d\times d}, -\tfrac{\mathbf{m}}{K}\right)\right)}_{=:\Phi^{\mathbf{y}}}, \end{aligned} \tag{15}$$

where $I_{d\times d}$ is the identity matrix of size d; Φ^{mult} is defined in Lemma 2; $\Phi^{\min} := (-1, 0) \odot \Phi^{\max} \odot \left(\left(\begin{smallmatrix} -1 & 0 \\ 0 & -1 \end{smallmatrix}\right), \left(\begin{smallmatrix} 0 \\ 0 \end{smallmatrix}\right)\right)$ and $\Phi^{\max}$ is defined by (5); Φ^{ψ} is the NN realizing the function ψ in (12), that is,

$$\begin{aligned} \Phi^{\psi} := \quad & ((A, \mathbf{b})) \odot P(((1,2),(1,0),(1,0)),((1,1),(1,0),(1,0)), \\ & ((1,-1),(1,0),(1,0)),((1,-2),(1,0),(1,0))), \end{aligned} \tag{16}$$

where $A = \left(\begin{smallmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 \end{smallmatrix}\right)$ and $\mathbf{b} = (0,0,0,0)^T$.

By Lemma 2, there is a NN with width, depth and scale satisfying

$$\begin{aligned} &\mathcal{N}(\Phi^{\mathrm{mult}}) \le 9\left(2\lceil N^{(d-3)/d}\rceil + 1\right) + d - 1,\ \mathcal{L}(\Phi^{\mathrm{mult}}) \le 7d(d-1)\lceil L^{(d-2)/d}\rceil, \\ &\text{and } \mathcal{B}(\Phi^{\mathrm{mult}}) \le 2, \end{aligned} \tag{17}$$

such that

$$\max_{\mathbf{x}\in[0,1]^d} \left|R(\Phi^{\mathrm{mult}})(R(\Phi^{\mathbf{y}})(\mathbf{x})) - \rho_{\mathbf{m}}(\mathbf{x})\right| \le 9\left(2\lceil N^{(d-3)/d}\rceil + 1\right)^{-7d\lceil L^{(d-2)/d}\rceil}, \tag{18}$$

where $\Phi^{\mathbf{y}}$ is given by (15).

Applying (15) and (18), we have

$$R(\Phi^{\rho_{\mathbf{m}}})(\mathbf{x}) \in [0,1] \text{ for } \forall \mathbf{x} \in [0,1]^d; \tag{19}$$

$$\operatorname{supp}(R(\Phi^{\rho_{\mathbf{m}}})) \subseteq \left\{\mathbf{x} : \|\mathbf{x} - \tfrac{\mathbf{m}}{K}\|_\infty \le \tfrac{1}{K}\right\}; \tag{20}$$

$$\max_{\mathbf{x}\in[0,1]^d} |R(\Phi^{\rho_{\mathbf{m}}})(\mathbf{x}) - \rho_{\mathbf{m}}(\mathbf{x})| \le 9\left(2\lceil N^{(d-3)/d}\rceil + 1\right)^{-7d\lceil L^{(d-2)/d}\rceil}. \tag{21}$$

By Remark 1, Remark 2, Remark 4, (15) and (17), we have that

$$\mathcal{N}(\Phi^{\rho_{\mathbf{m}}}) \le 9\left(2\lceil N^{(d-3)/d}\rceil + 1\right) + 2d,\ \mathcal{L}(\Phi^{\rho_{\mathbf{m}}}) \le 7d(d-1)\lceil L^{(d-2)/d}\rceil + 15, \text{ and } \mathcal{B}(\Phi^{\rho_{\mathbf{m}}}) \le 3K. \tag{22}$$

Using the partition of unity (11), we decompose the target function $f_0 \in \mathcal{H}(\beta, [0,1]^d, M)$ as:

$$f_0 = \sum_{\mathbf{m}} \rho_{\mathbf{m}}\, f_0. \tag{23}$$

Next, for any $\mathbf{m} \in \{0, 1, \cdots, K\}^d$, we approximate the function $\rho_{\mathbf{m}} f_0$ using $\rho_{\mathbf{m}}\, f_0\left(\frac{\mathbf{m}}{K}\right)$. We show below that we can control the approximation error by choosing

$$K = \underbrace{\left\lceil 2^{d/\beta}\left(N^2 \log_3(N+2)\right)^{1/d}\right\rceil}_{=:\, n} \underbrace{\lceil L^{2/d}\rceil}_{=:\, l} - 1. \tag{24}$$

Observing that $d, N, L \in \mathbb{N}$ and $\beta \in (0,1]$, we get $K \in \mathbb{N}$, which satisfies the assumption of K made at the beginning of the proof.

Let φ denote the bijective map $\varphi : \{0, 1, \cdots, K\}^d \to \left[(K+1)^d\right]$. By (15), (23) and (24), we can approximate $\rho_{\mathbf{m}}\, f_0$, with $\varphi(\mathbf{m}) = i$ on $[0,1]^d$, using the NN

$$\Phi^i := \Phi^{xy} \odot P\left(\Phi^{Id}_{1,L^*} \odot \left(0, f_0\left(\tfrac{\varphi^{-1}(i)}{K}\right)\right), \Phi^{\rho_{\varphi^{-1}(i)}}\right). \tag{25}$$

where Φ^{Id}_{1,L^*}, with $L^* = 7d(d-1)\lceil L^{(d-2)/d}\rceil + 14$, is the identity function defined in Definition 6; by Lemma 1, there exists a NN Φ^{xy} with width, depth and scale satisfying

$$\mathcal{N}(\Phi^{xy}) \le 9\left(2\lceil N^{(d-3)/d}\rceil + 1\right) + 1, \quad \mathcal{L}(\Phi^{xy}) \le 2d\left(\lceil L^{(d-2)/d}\rceil + 1\right), \text{ and } \mathcal{B}(\Phi^{xy}) \le (M-1)^2 \tag{26}$$

such that

$$\begin{aligned}&\left|R(\Phi^{xy})(R(\Phi^{\rho_{\mathbf{m}}})(\mathbf{x}), f_0(\tfrac{\mathbf{m}}{K})) - R(\Phi^{\rho_{\mathbf{m}}})(\mathbf{x}) f_0(\tfrac{\mathbf{m}}{K})\right|\\ &\le 6(M-1)^2(2\lceil N^{(d-3)/d}\rceil + 1)^{-2d(\lceil L^{(d-2)/d}\rceil + 1)},\end{aligned} \tag{27}$$

where the inequality also uses by (19), i.e., $\max_{\mathbf{x}\in[0,1]^d} |R(\Phi^{\rho_{\mathbf{m}}})(\mathbf{x})| \le 1$ and the hypothesis $f_0 \in \mathcal{H}(\beta, [0,1]^d, M)$.

Using Remark 1, Remark 2, Remark 3, (25) and (26), we get

$$\begin{aligned}
&\mathcal{N}(\Phi^i) \le 9\left(2\lceil N^{(d-3)/d}\rceil + 1\right) + 2d + 2,\\
&\mathcal{L}(\Phi^i) \le (7d^2 - 5d)\lceil L^{(d-2)/d}\rceil + 2d + 16,\\
&\mathcal{B}(\Phi^i) \le 3K \vee (M-1)^2. \qquad (28)
\end{aligned}$$

For simplicity, we define $\Phi^{i\to j}$ with $i, j \in [(K+1)^d]$ and $i < j$ as follows:

$$\Phi^{i\to j} := (((\underbrace{1, \cdots, 1}_{j-i+1}), 0)) \odot P(\Phi^i, \Phi^{i+1}, \cdots, \Phi^j). \qquad (29)$$

Applying Remark 1, Remark 2, (28) and (29), we obtain that

$$\begin{aligned}
\mathcal{N}(\Phi^{i\to j}) &\le (j-i+1)\left(9\left(2\lceil N^{(d-3)/d}\rceil + 1\right) + 2d + 2\right),\\
\mathcal{L}(\Phi^{i\to j}) &\le (7d^2 - 5d)\lceil L^{(d-2)/d}\rceil + 2d + 17,\\
\mathcal{B}(\Phi^{i\to j}) &\le 3K \vee (M-1)^2. \qquad (30)
\end{aligned}$$

By (29), we construct a NN Φ^{simul} to approximate $\sum_{\mathbf{m}} \rho_{\mathbf{m}} f_0\left(\frac{\mathbf{m}}{K}\right)$ as illustrated in Figure 1, given by,

$$\begin{aligned}
\Phi^{\text{simul}} &= ((\mathbf{e}_{d+1}, 0))\\
&\odot\ FP(((I_{d\times d}, \mathbf{0}^T)), (((1,1),0))) \odot FP(P(\Phi^{Id}_{d,L_1}, \Phi^{(l-1)n+1\to ln}), \Phi^{Id}_{1,L_3})\\
&\ \vdots\\
&\odot\ FP(((I_{d\times d}, \mathbf{0}^T)), (((1,1),0))) \odot FP(P(\Phi^{Id}_{d,L_1}, \Phi^{2n+1\to 3n}), \Phi^{Id}_{1,L_3})\\
&\odot\ FP(((I_{d\times d}, \mathbf{0}^T)), (((1,1),0))) \odot FP(P(\Phi^{Id}_{d,L_1}, \Phi^{n+1\to 2n}), \Phi^{Id}_{1,L_3})\\
&\odot\ FP(((I_{d\times d}, \mathbf{0}^T)), (((1,1),0))) \odot P(\Phi^{Id}_{d,L_1}, \Phi^{1\to n}, \Phi^{Id}_{d,L_2} \odot ((\mathbf{0},0)))), \qquad (31)
\end{aligned}$$

where $\mathbf{0} = (\underbrace{0, \cdots, 0}_{d})$; $\mathbf{e}_{d+1} = (\underbrace{0, \cdots, 0}_{d}, 1)$; $I_{d\times d}$ is the identity matrix of size d; Φ^{Id}_{d,L_1} and Φ^{Id}_{1,L_3} are $\mathcal{L}(\Phi^{1\to n})$-layer identity function of the NN defined in Definition 6; Φ^{Id}_{d,L_2} is $(\mathcal{L}(\Phi^{1\to n}) - 1)$-layer identity function of the NN.

From Remark 1, Remark 2, Remark 3, (30) and (31), we have

$$\begin{aligned}
\mathcal{N}(\Phi^{\text{simul}}) &\le n\left(9\left(2\lceil N^{(d-3)/d}\rceil + 1\right) + 2d + 2\right) + 4d,\\
\mathcal{L}(\Phi^{\text{simul}}) &\le l\left((7d^2 - 5d)\lceil L^{(d-2)/d}\rceil + 2d + 19\right),\\
\mathcal{B}(\Phi^{\text{simul}}) &\le 3K \vee (M-1)^2 \vee M, \qquad (32)
\end{aligned}$$

where n and l are given by (24).

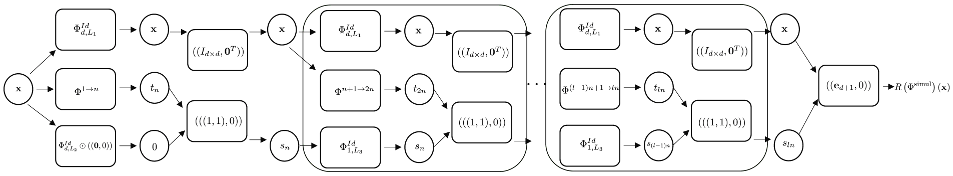

Figure 1: This NN computes Φ^{simul} where $s_m = \sum_{i=1}^m R(\Phi^i)(\mathbf{x})$ and $t_m = s_m - s_{m-1}$ for $m \in \mathbb{N}$.

We next define the NN $\Phi^{f_0} := \Phi^{\text{bound}} \odot \Phi^{\text{simul}}$ to approximate f_0, where

$$\Phi^{\text{bound}} := \Phi^{\min} \odot P(((0,M)),((1,0))) \odot \Phi^{\max} \odot P(((0,-M)),((1,0))),$$

$\Phi^{\max}$ is defined by (5), $\Phi^{\min}$ is defined in below (15). We also observe that $R(\Phi^{\text{bound}})(t) = \min\{\max\{-M,t\},M\}$ for all $t \in \mathbb{R}$ and

$$\mathcal{N}(\Phi^{\text{bound}}) = 4,\ \mathcal{L}(\Phi^{\text{bound}}) = 9,\ \mathcal{B}(\Phi^{\text{bound}}) = M. \tag{33}$$

We remark that the definition of Φ^{bound} leads to $\max_{\mathbf{x}\in[0,1]^d} \left|R(\Phi^{f_0})(\mathbf{x})\right| \leq M$ and we will use this property in the proof of Theorem 2.

The total approximation error is bounded by

$$\begin{aligned}
& \max_{\mathbf{x}\in[0,1]^d} \left|R(\Phi^{f_0})(\mathbf{x}) - f_0(\mathbf{x})\right| \\
= \quad & \max_{\mathbf{x}\in[0,1]^d} \left|R(\Phi^{\text{simul}})(\mathbf{x}) - f_0(\mathbf{x})\right| \\
= \quad & \max_{\mathbf{x}\in[0,1]^d} \left|\sum_{\mathbf{m}} \left(R(\Phi^{xy})\left(R(\Phi^{\rho_{\mathbf{m}}})(\mathbf{x}), f_0(\tfrac{\mathbf{m}}{K})\right) - \rho_{\mathbf{m}}(x) f_0(\mathbf{x})\right)\right| \\
\leq \quad & \max_{\mathbf{x}\in[0,1]^d} \sum_{\mathbf{m}} \left|R(\Phi^{xy})\left(R(\Phi^{\rho_{\mathbf{m}}})(\mathbf{x}), f_0(\tfrac{\mathbf{m}}{K})\right) - \rho_{\mathbf{m}}(x) f_0(\mathbf{x})\right| \\
\leq \quad & \underbrace{\max_{\mathbf{x}\in[0,1]^d} \sum_{\mathbf{m}} \left|R(\Phi^{xy})\left(R(\Phi^{\rho_{\mathbf{m}}})(\mathbf{x}), f_0(\tfrac{\mathbf{m}}{K})\right) - R(\Phi^{\rho_{\mathbf{m}}})(\mathbf{x}) f_0(\tfrac{\mathbf{m}}{K})\right|}_{\mathcal{E}_1} \\
+ \quad & \underbrace{\max_{\mathbf{x}\in[0,1]^d} \sum_{\mathbf{m}} \left|R(\Phi^{\rho_{\mathbf{m}}})(\mathbf{x}) f_0(\tfrac{\mathbf{m}}{K}) - \rho_{\mathbf{m}}(\mathbf{x}) f_0(\tfrac{\mathbf{m}}{K})\right|}_{\mathcal{E}_2} \\
+ \quad & \underbrace{\max_{\mathbf{x}\in[0,1]^d} \sum_{\mathbf{m}} \left|\rho_{\mathbf{m}}(\mathbf{x}) f_0(\tfrac{\mathbf{m}}{K}) - \rho_{\mathbf{m}}(\mathbf{x}) f_0(\mathbf{x})\right|}_{\mathcal{E}_3} \\
= \quad & \mathcal{E}_1 + \mathcal{E}_2 + \mathcal{E}_3.
\end{aligned} \tag{34}$$

Our next task is to estimate $\mathcal{E}_1$, $\mathcal{E}_2$ and $\mathcal{E}_3$. By (27), we have

$$\begin{aligned}
\mathcal{E}_1 \quad &:= \quad \max_{\mathbf{x}\in[0,1]^d} \sum_{\mathbf{m}} \left| R(\Phi^{xy}) \left(R(\Phi^{\rho_{\mathbf{m}}})(\mathbf{x}), f_0(\tfrac{\mathbf{m}}{K}) \right) - R(\Phi^{\rho_{\mathbf{m}}})(\mathbf{x}) f_0(\tfrac{\mathbf{m}}{K}) \right| \\
&\overset{(i)}{=} \quad \sum_{\left\{\mathbf{m}: \|\mathbf{x}-\frac{\mathbf{m}}{K}\|_\infty \le \frac{1}{K}\right\}} \left| R(\Phi^{xy}) \left(R(\Phi^{\rho_{\mathbf{m}}})(\mathbf{x}), f_0(\tfrac{\mathbf{m}}{K}) \right) - R(\Phi^{\rho_{\mathbf{m}}})(\mathbf{x}) f_0(\tfrac{\mathbf{m}}{K}) \right| \\
&\overset{(ii)}{\le} \quad 2^d \times 6(M-1)^2 \left(2\lceil N^{(d-3)/d} \rceil + 1 \right)^{-2d\left(\lceil L^{(d-2)/d} \rceil + 1\right)}, \qquad (35)
\end{aligned}$$

where (i) uses (20) and Remark 10; (ii) follows from the observation that, for $\mathbf{x} \in [0,1]^d$, $\#\left\{\mathbf{m} \in \{0,1,\cdots,K\}^d : \|\mathbf{x} - \frac{\mathbf{m}}{K}\|_\infty \le \frac{1}{K}\right\} \le 2^d$.

Similarly, we deduce that

$$\begin{aligned}
\mathcal{E}_2 \quad &:= \quad \max_{\mathbf{x}\in[0,1]^d} \sum_{\mathbf{m}} \left| R(\Phi^{\rho_{\mathbf{m}}})(\mathbf{x}) f_0(\tfrac{\mathbf{m}}{K}) - \rho_{\mathbf{m}}(\mathbf{x}) f_0(\tfrac{\mathbf{m}}{K}) \right| \\
&\overset{(i)}{=} \quad \sum_{\left\{\mathbf{m}: \|\mathbf{x}-\frac{\mathbf{m}}{K}\|_\infty \le \frac{1}{K}\right\}} \left| R(\Phi^{\rho_{\mathbf{m}}})(\mathbf{x}) f_0(\tfrac{\mathbf{m}}{K}) - \rho_{\mathbf{m}}(\mathbf{x}) f_0(\tfrac{\mathbf{m}}{K}) \right| \\
&= \quad \left| f_0(\tfrac{\mathbf{m}}{K}) \right| \sum_{\left\{\mathbf{m}: \|\mathbf{x}-\frac{\mathbf{m}}{K}\|_\infty \le \frac{1}{K}\right\}} \left| R(\Phi^{\rho_{\mathbf{m}}})(\mathbf{x}) - \rho_{\mathbf{m}}(\mathbf{x}) \right| \\
&\overset{(ii)}{\le} \quad 2^d \times 9M \left(2\lceil N^{(d-3)/d} \rceil + 1 \right)^{-7d\lceil L^{(d-2)/d} \rceil}, \qquad (36)
\end{aligned}$$

where (i) is obtained by (14) and (20); (ii) follows from (21) and the hypothesis $f_0 \in \mathcal{H}(\beta, [0,1]^d, M)$.

Using $\rho_{\mathbf{m}}(\mathbf{x}) \in [0,1]$ for all $\mathbf{x} \in [0,1]^d$ and the hypothesis $f_0 \in \mathcal{H}(\beta, [0,1]^d, M)$, we have

$$\begin{aligned}
\mathcal{E}_3 \quad &:= \quad \max_{\mathbf{x}\in[0,1]^d} \sum_{\mathbf{m}} \left| \rho_{\mathbf{m}}(\mathbf{x}) f_0(\tfrac{\mathbf{m}}{K}) - \rho_{\mathbf{m}}(\mathbf{x}) f_0(\mathbf{x}) \right| \\
&\overset{(i)}{=} \quad \sum_{\left\{\mathbf{m}: \|\mathbf{x}-\frac{\mathbf{m}}{K}\|_\infty \le \frac{1}{K}\right\}} \left| \rho_{\mathbf{m}}(\mathbf{x}) f_0(\tfrac{\mathbf{m}}{K}) - \rho_{\mathbf{m}}(\mathbf{x}) f_0(\mathbf{x}) \right| \\
&= \quad |\rho_{\mathbf{m}}(\mathbf{x})| \sum_{\left\{\mathbf{m}: \|\mathbf{x}-\frac{\mathbf{m}}{K}\|_\infty \le \frac{1}{K}\right\}} \left| f_0(\tfrac{\mathbf{m}}{K}) - f_0(\mathbf{x}) \right| \\
&\overset{(ii)}{\le} \quad 2^d \times M (\tfrac{\sqrt{d}}{K})^\beta \\
&\overset{(iii)}{\le} \quad 2^d \times 2M 2^{-d} d^{\beta/2} N^{-2\beta/d} \left(\log_3(N+2) \right)^{-\beta/d} L^{-2\beta/d}, \qquad (37)
\end{aligned}$$

where (i) uses (14); (ii) follows from $\|\rho_{\mathbf{m}}\|_\infty = 1$ in (13), the observation $\|\mathbf{x} - \frac{m}{K}\|_2 \le \frac{\sqrt{d}}{K}$ and the hypothesis $f_0 \in \mathcal{H}(\beta, [0,1]^d, M)$; (iii) follows by the definition of K in (24), which leads to $K = \lceil 2^{d/\beta} \left(N^2 \log_3(N+2) \right)^{1/d} \rceil \lceil L^{2/d} \rceil - 1 \ge \frac{1}{2} \lceil 2^{d/\beta} \left(N^2 \log_3(N+2) \right)^{1/d} \rceil \lceil L^{2/d} \rceil - 1 \ge \frac{2^{d/\beta}}{2} N^{2/d} \left(\log_3(N+2) \right)^{1/d} L^{2/d}$.

Recall that, for $N, L \in \mathbb{N}$, $d \geq 4$ and $\beta \in (0,1]$,

$$\begin{aligned}
& \left(2\lceil N^{(d-3)/d}\rceil + 1\right)^{-7d\lceil L^{(d-2)/d}\rceil} \\
\leq\ & \left(2\lceil N^{(d-3)/d}\rceil + 1\right)^{-2d\left(\lceil L^{(d-2)/d}\rceil + 1\right)} \\
\leq\ & \left(2\lceil N^{(d-3)/d}\rceil + 1\right)^{-2d} 2^{-2d\lceil L^{(d-2)/d}\rceil} \\
\leq\ & 2^{-2d}\lceil N^{(d-3)/d}\rceil^{-2d}\lceil L^{(d-2)/d}\rceil^{-2d} \\
\leq\ & 2^{-d} N^{-2\beta/d} \left(\log_3(N+2)\right)^{-\beta/d} L^{-2\beta/d}.
\end{aligned} \tag{38}$$

Combining (35), (36), (37) and (38), we have

$$\begin{aligned}
& \max_{\mathbf{x}\in[0,1]^d} \left|R(\Phi^{f_0})(\mathbf{x}) - f_0(\mathbf{x})\right| \\
\leq\ & \mathcal{E}_1 + \mathcal{E}_2 + \mathcal{E}_3 \\
\leq\ & 2^d \cdot 6(M-1)^2 \left(2\lceil N^{(d-3)/d}\rceil + 1\right)^{-2d\left(\lceil L^{(d-2)/d}\rceil + 1\right)} \\
+\ & 2^d \cdot 9M \left(2\lceil N^{(d-3)/d}\rceil + 1\right)^{-7d\lceil L^{(d-2)/d}\rceil} \\
+\ & 2^d \cdot 2M 2^{-d} d^{\beta/2} N^{-2\beta/d} \left(\log_3(N+2)\right)^{-\beta/d} L^{-2\beta/d} \\
\leq\ & \left(6(M-1)^2 + 9M + 2M\right) d^{\beta/2} N^{-2\beta/d} \left(\log_3(N+2)\right)^{-\beta/d} L^{-2\beta/d} \\
\leq\ & 6(M^2+1) d^{\beta/2} \left(N^2 L^2 \log_3(N+2)\right)^{-\beta/d}.
\end{aligned}$$

By Remark 1, (32), (33) and $\Phi^{f_0} := \Phi^{\text{bound}} \odot \Phi^{simul}$, we observe that

$$\begin{aligned}
\mathcal{N}(\Phi^{f_0}) &\leq n\left(9\left(2\lceil N^{(d-3)/d}\rceil + 1\right) + 2d + 2\right) + 4d \\
&= \lceil 2^{d/\beta}\left(N^2\log_3(N+2)\right)^{1/d}\rceil \left(18\lceil N^{(d-3)/d}\rceil + 2d + 11\right) + 4d \\
&\leq 2^{d/\beta+1}\left(6d+47\right) N, \\
\mathcal{L}(\Phi^{f_0}) &\leq l\left((7d^2 - 5d)\lceil L^{(d-2)/d}\rceil + 2d + 19\right) + 10 \\
&\overset{(i)}{\leq} (28d^2 - 16)L, \\
\mathcal{B}(\Phi^{f_0}) &\leq 9 \vee 3\lceil 2^{d/\beta}\left(N^2\log_3(N+2)\right)^{1/d}\rceil\lceil L^{2/d}\rceil \vee M^2 \\
&\overset{(ii)}{=} 3\lceil 2^{d/\beta}\left(N^2\log_3(N+2)\right)^{1/d}\rceil\lceil L^{2/d}\rceil \vee M^2,
\end{aligned}$$

where n and l are defined in (24), (i) comes from the hypothesis $d \geq 4$ and (ii) follows from the observation that - since $d_\delta \in \mathbb{N}$ and $\beta \in (0,1]$ - we have the inequality $3\lceil 2^{d_\delta/\beta} N^{2/d_\delta}\rceil\lceil(\log_3(N+2))^{1/d}\rceil\lceil L^{2/d_\delta}\rceil \geq 9$. □

As indicated above, our proof of Theorem 1 requires an extension of the celebrated Johnson-Lindenstrauss lemma to the manifold setting, that we apply to preserve pairwise ambient distances, up to a controllable distortion, when we project points from a manifold $\mathcal{M} \subset \mathbb{R}^D$ into a lower dimensional target space. Our theorem below is a slight modification of a result by Eftekhari and Wakin [26], providing additional control on the size of the entries of the random projection matrix as compared to the original result in [26]. We use this property in the proof of Theorem 1 to bound the scale of the approximating NN. We postpone the proof of Theorem 4 to Appendix A.

Theorem 4. *Let $\mathcal{M}$ be a compact K-dimensional Riemannian submanifold of $\mathbb{R}^D$ having condition number $1/\tau$ and volume $V_{\mathcal{M}}$, satisfying*

$$\frac{V_{\mathcal{M}}}{\tau^K} \geq \left(\frac{21}{2\sqrt{K}}\right)^K. \tag{39}$$

Fix $\epsilon \in (0, 1/3]$ and $\rho \in (0,1)$. Let A be a random $d_\epsilon \times D$ matrix populated with i.i.d. random variables $a_{i,j}$ where

$$a_{ij} = \begin{cases} +\frac{1}{\sqrt{d_\epsilon}} \text{ with probability } \frac{1}{2} \\ -\frac{1}{\sqrt{d_\epsilon}} \text{ with probability } \frac{1}{2} \end{cases}$$

and

$$d_\epsilon = \left\lceil 92\epsilon^{-2} \max\left\{24K + 2K\log\left(\frac{\sqrt{K}}{\tau\epsilon^2}\right) + \log(2V_{\mathcal{M}}^2),\, \log\left(\frac{20}{\rho}\right)\right\}\right\rceil. \tag{40}$$

Then with probability at least $1-\rho$ the following statement holds: for every pair of points $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{M}$,

$$(1-\epsilon)\|\mathbf{x}_1 - \mathbf{x}_2\|_2 \leq \|A\mathbf{x}_1 - A\mathbf{x}_2\|_2 \leq (1+\epsilon)\|\mathbf{x}_1 - \mathbf{x}_2\|_2.$$

Remark 11. Theorem 4, exactly as the original theorem by Eftekhari and Wakin [26], assumes inequality (39) which imposes a mild geometric condition on the reach. As observed in Remark 6, this condition is easily satisfied for the hyper-sphere. It is observed in [26] that one can relax inequality (39) and the result of the theorem would still hold even though with a possibly larger constants in (40).

We will also need the following extension result for function in Hölder spaces, which is similar to Lemma 4.1 in [21].

Lemma 3 (Extension Lemma). *Let $f \in \mathcal{H}(\beta, E, M)$, where $0 < \beta \leq 1$, $M > 0$ and $E \subseteq [0,1]^d$ is a closed set with $d \in \mathbb{N}$. Then there exists a function $g \in \mathcal{H}(\beta, [0,1]^d, 2d^{\beta/2}M)$ such that $g(\mathbf{x}) = f(\mathbf{x})$ for any $\mathbf{x} \in E$.*

Proof. For $\mathbf{x} \in [0.1]^d$, we define

$$g(\mathbf{x}) := \sup_{\mathbf{z}\in E} \left(f(\mathbf{z}) - M\|\mathbf{z} - \mathbf{x}\|_2^\beta \right). \tag{41}$$

Since $f \in \mathcal{H}(\beta, E, M)$, we have that $f(\mathbf{z}) - M\|\mathbf{z} - \mathbf{x}\|_2^\beta \le f(\mathbf{x})$ for any $\mathbf{x}, \mathbf{z} \in E$. This implies that $g(\mathbf{x}) \le f(\mathbf{x})$, for $\mathbf{x} \in E$. Together with the observation

$$g(\mathbf{x}) := \sup_{\mathbf{z}\in E} \left(f(\mathbf{z}) - M\|\mathbf{z} - \mathbf{x}\|_2^\beta \right) \ge f(\mathbf{x}) - M\|\mathbf{x} - \mathbf{x}\|_2^\beta = f(\mathbf{x}),$$

for any $\mathbf{x} \in E$, it follows that $f(\mathbf{x}) = g(\mathbf{x})$, for any $\mathbf{x} \in E$.

By the observation

$$\sup_{\mathbf{z}\in E} f_1(\mathbf{z}) - \sup_{\mathbf{z}\in E} f_2(\mathbf{z}) \le \sup_{\mathbf{z}\in E}(f_1(\mathbf{z}) - f_2(\mathbf{z})),$$

we have that

$$\begin{aligned} g(\mathbf{x}_1) - g(\mathbf{x}_2) &\le \sup_{\mathbf{z}\in E} \left(f(\mathbf{z}) - M\|\mathbf{z} - \mathbf{x}_1\|_2^\beta \right) - \sup_{\mathbf{z}\in E} \left(f(\mathbf{z}) - M\|\mathbf{z} - \mathbf{x}_2\|_2^\beta \right) \\ &\le \sup_{\mathbf{z}\in E} \left(M\|\mathbf{z} - \mathbf{x}_1\|_2^\beta - M\|\mathbf{z} - \mathbf{x}_2\|_2^\beta \right) \\ &\le M\|\mathbf{x}_1 - \mathbf{x}_2\|_2^\beta, \end{aligned}$$

for any $\mathbf{x}_1, \mathbf{x}_2 \in [0,1]^d$, where the last inequality comes from the fact that, for any $\beta \in (0,1]$ and a, $b \in \mathbb{R}$, we have the inequality $|a+b|^\beta \le |a|^\beta + |b|^\beta$.

Similarly, for any $\mathbf{x}_1, \mathbf{x}_2 \in [0,1]^d$, we have $g(\mathbf{x}_2) - g(\mathbf{x}_1) \le M\|\mathbf{x}_1 - \mathbf{x}_2\|_2^\beta$, which implies

$$|g(\mathbf{x}_1) - g(\mathbf{x}_2)| \le M\|\mathbf{x}_1 - \mathbf{x}_2\|_2^\beta.$$

From the definition of $g(\mathbf{x})$ in (41), using the assumption that $f \in \mathcal{H}(\beta, E, M)$ and the inequality $\|\mathbf{x} - \mathbf{y}\|_2 \le \sqrt{d}$ for any $\mathbf{x}, \mathbf{y} \in [0,1]^d$, we have that, for any $\mathbf{x} \in [0,1]^d$,

$$|g(\mathbf{x})| \le (d^{\beta/2} + 1)M \le 2d^{\beta/2}M.$$

□

With the above preparation, we can now prove Theorem 1.

Proof (Proof of Theorem 1). By Theorem 4, there exists a matrix $A \in \mathbb{R}^{d_\epsilon \times D}$, with d_ϵ given by (40), such that for every pair of points $\mathbf{x}_1$, $\mathbf{x}_2 \in \mathcal{M}$, we have

$$(1-\epsilon)\|\mathbf{x}_1 - \mathbf{x}_2\|_2 \le \|A\mathbf{x}_1 - A\mathbf{x}_2\|_2 \le (1+\epsilon)\|\mathbf{x}_1 - \mathbf{x}_2\|_2. \tag{42}$$

For any $\mathbf{x}_1$, $\mathbf{x}_2 \in \mathcal{M}$, using inequality (42), we have that

$$\begin{aligned} ||A\mathbf{x}_1 - A\mathbf{x}_2||_\infty &\overset{(i)}{\le} \|A\mathbf{x}_1 - A\mathbf{x}_2\|_2 \\ &\le (1+\epsilon)\|\mathbf{x}_1 - \mathbf{x}_2\|_2 \\ &\le (1+\epsilon)\,\mathrm{diam}(\mathcal{M}) \\ &\overset{(ii)}{\le} 2\ \mathrm{diam}(\mathcal{M}), \end{aligned} \tag{43}$$

where (i) follows from the observation that, for any $\mathbf{z} = (z_1, \cdots, z_{d_\epsilon})^T \in \mathbb{R}^{d_\epsilon}$, we have $|z_i| \leq \sqrt{z_1^2 + \cdots + z_{d_\epsilon}^2}$ for $i = 1, \cdots, d_\epsilon$; inequality (ii) follows from the assumption that $\epsilon \leq 1/3$.

Next, we use the matrix A to construct an affine transformation $\varPsi : \mathcal{M} \mapsto [0,1]^{d_\epsilon}$ as

$$\varPsi(\mathbf{x}) := \frac{1}{4 \ \mathrm{diam}(\mathcal{M})} A\,\mathbf{x} - \frac{1}{4 \ \mathrm{diam}(\mathcal{M})}\mathbf{y}_0, \tag{44}$$

where we choose $\mathbf{y}_0 \in \mathbb{R}^{d_\epsilon}$ such that $A(\mathcal{M}) \subseteq \{4 \ \mathrm{diam}(\mathcal{M})\mathbf{y} + \mathbf{y}_0 \mid \mathbf{y} \in [0,1]^{d_\epsilon}\}$ and, for any $\mathbf{x} \in \mathcal{M}$. By construction, we have that $\varPsi(\mathcal{M}) \subseteq [0,1]^{d_\epsilon}$.

We can define a NN $\hat{\varPsi}$ that realizes $\varPsi(\mathbf{x})$ exactly by setting

$$\hat{\varPsi} := \left(\left(\frac{1}{4 \ \mathrm{diam}(\mathcal{M})} A, -\frac{1}{4 \ \mathrm{diam}(\mathcal{M})}\mathbf{y}_0 \right) \right). \tag{45}$$

By (42) and (44), we have

$$\frac{1-\epsilon}{4 \ \mathrm{diam}(\mathcal{M})} \|\mathbf{x}_1 - \mathbf{x}_2\|_2 \leq \|\varPsi(\mathbf{x}_1) - \varPsi(\mathbf{x}_2)\|_2 \leq \frac{1+\epsilon}{4 \ \mathrm{diam}(\mathcal{M})} \|\mathbf{x}_1 - \mathbf{x}_2\|_2. \tag{46}$$

We will next define a unique low-dimensional function g_0, with values on $[0,1]^{d_\epsilon}$, to represent the function f_0 defined on $\mathcal{M}$. For any $\mathbf{y} \in \varPsi(\mathcal{M}) \subseteq [0,1]^{d_\epsilon}$, we define

$$g_0(\mathbf{y}) := f_0(\mathbf{x}_\mathbf{y}), \text{ where } \mathbf{x}_\mathbf{y} = \{\mathbf{x} \in \mathcal{M} \,\big|\, \varPsi(\mathbf{x}) = \mathbf{y}, \, \mathbf{y} \in \varPsi(\mathcal{M})\}. \tag{47}$$

Note that, by inequality (46), we have that the map $\varPsi$ is injective and, hence, it is bijective from $\mathcal{M}$ onto $\varPsi(\mathcal{M})$.

We claim that the function g_0 defined by (47) is a Hölder continuous function.

To show that this is the case, we first observe that, by the hypothesis of Theorem 1, the norm of f_0 in (47) is bounded by M. It follows that

$$|g(\mathbf{y})| \leq M, \text{ for any } \mathbf{y} \in \varPsi(\mathcal{M}). \tag{48}$$

Next, we observe that, for $\mathbf{y}_1, \mathbf{y}_2 \in \varPsi(\mathcal{M})$, there are $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{M}$ defined by $\mathbf{x}_i = \{\mathbf{x} \in \mathcal{M} \,|\, \varPsi(\mathbf{x}_i) = \mathbf{y}_i\}$, $i = 1, 2$. Using (46), it follows that

$$\begin{aligned} & |g_0(\mathbf{y}_1) - g_0(\mathbf{y}_2)| \\ = \ & |f_0(\mathbf{x}_1) - f_0(\mathbf{x}_2)| \\ \leq \ & M\|\mathbf{x}_1 - \mathbf{x}_2\|_2^\beta \\ \leq \ & M \left(\left(\tfrac{4 \ \mathrm{diam}(\mathcal{M})}{1-\epsilon} \right)^\beta \|\varPsi(\mathbf{x}_1) - \varPsi(\mathbf{x}_2)\|_2^\beta \right) \\ \leq \ & \left(\tfrac{4 \ \mathrm{diam}(\mathcal{M})}{1-\epsilon} \right)^\beta M\|\mathbf{y}_1 - \mathbf{y}_2\|_2^\beta. \end{aligned} \tag{49}$$

Combining (48) with (49), we conclude that $g_0 \in \mathcal{H}\left(\beta, \varPsi(\mathcal{M}), \left(\tfrac{4 \ \mathrm{diam}(\mathcal{M})}{1-\epsilon}\right)^\beta M\right)$.

Using Lemma 3, we now extend the function g_0, originally defined on $\Psi(\mathcal{M})$, to a Hölder function $\widetilde{g_0}$ defined on $[0,1]^{d_\epsilon}$. Note that the Hölder norm of $\widetilde{g_0}$ is bounded by $8\left(\frac{\operatorname{diam}(\mathcal{M})}{1-\epsilon}\right)^{\beta} {d_\epsilon}^{\beta/2} M$ and, thus, $\widetilde{g_0}$ belongs to the Hölder space $\mathcal{H}\left(\beta, [0,1]^{d_\epsilon}, 8\left(\frac{\operatorname{diam}(\mathcal{M})}{1-\epsilon}\right)^{\beta} {d_\epsilon}^{\beta/2} M\right)$.

We next show that we can approximate $\widetilde{g_0}$ using an appropriate NN. Using Theorem 3, there exists a NN $\Phi^{\widetilde{g_0}}$ with width, depth and scale satisfying

$$\begin{aligned} &\mathcal{N}(\Phi^{\widetilde{g_0}}) \le 2^{d_\epsilon/\beta+1}\,(6d_\epsilon+47)\,N, \quad \mathcal{L}(\Phi^{\widetilde{g_0}}) \le (28d_\epsilon^2-16)L, \\ &\mathcal{B}(\Phi^{\widetilde{g_0}}) \le 3\lceil 2^{d_\epsilon/\beta}\left(N^2\log_3(N+2)\right)^{1/d_\epsilon}\rceil\lceil L^{2/d_\epsilon}\rceil \vee 64\left(\frac{\operatorname{diam}(\mathcal{M})}{1-\epsilon}\right)^{2\beta} {d_\epsilon}^{\beta} M^2, \end{aligned} \tag{50}$$

so that

$$\begin{aligned} &\max_{\mathbf{y}\in[0,1]^{d_\epsilon}} |R(\Phi^{\widetilde{g_0}})(\mathbf{y}) - \widetilde{g_0}(\mathbf{y})| \\ &\le \left(384\left(\frac{\operatorname{diam}(\mathcal{M})}{1-\epsilon}\right)^{2\beta} {d_\epsilon}^{\beta} M^2 + 6\right) d_\epsilon^{\beta/2}\left(N^2L^2\log_3(N+2)\right)^{-\beta/d_\epsilon}. \end{aligned} \tag{51}$$

We can finally approximate f_0 using a NN. Namely, using the above NN $\Phi^{\widetilde{g_0}}$ and the NN $\hat{\Psi}$ given in (45), we define the NN $\Phi^{f_0} = \Phi^{\widetilde{g_0}} \odot \hat{\Psi}$.

For any $\mathbf{x} \in \mathcal{M}$, given any $\epsilon > 0$, we have that

$$\begin{aligned} &|f_0(\mathbf{x}) - R(\Phi^{f_0})(\mathbf{x})| \\ &\overset{(i)}{=} |g_0(\Psi(\mathbf{x})) - R(\Phi^{\widetilde{g_0}})(\Psi(\mathbf{x}))| \\ &\overset{(ii)}{=} |\widetilde{g_0}(\Psi(\mathbf{x})) - R(\Phi^{\widetilde{g_0}})(\Psi(\mathbf{x}))| \\ &\overset{(iii)}{\le} \left(384\left(\frac{\operatorname{diam}(\mathcal{M})}{1-\epsilon}\right)^{2\beta} {d_\epsilon}^{\beta} M^2 + 6\right) d_\epsilon^{\beta/2}\left(N^2L^2\log_3(N+2)\right)^{-\beta/d_\epsilon}, \end{aligned}$$

where equality (i) follows from the definition of g_0 and Φ^{f_0}; (ii) follows from the definition of $\widetilde{g_0}$; (iii) follows from (51).

By Remark 1, (50) and $\Phi^{f_0} = \Phi^{\widetilde{g_0}} \odot \hat{\Psi}$, we obtain

$$\begin{aligned} &\mathcal{N}(\Phi^{f_0}) \le 2^{d_\epsilon/\beta+1}\,(6d_\epsilon+47)\,N, \quad \mathcal{L}(\Phi^{f_0}) \le (28d_\epsilon^2-16)L + 1 \le (28d_\epsilon^2-15)L, \\ &\mathcal{B}(\Phi^{f_0}) \le C_0 \vee 3\lceil 2^{d_\epsilon/\beta}\left(N^2\log_3(N+2)\right)^{1/d_\epsilon}\rceil\lceil L^{2/d_\epsilon}\rceil \vee 64\left(\frac{\operatorname{diam}(\mathcal{M})}{1-\epsilon}\right)^{2\beta} {d_\epsilon}^{\beta} M^2, \end{aligned}$$

where we claim that C_0, the scale parameter associated with the NN $\hat{\Psi}$, given by (45), satisfies $C_0 = \max\{\frac{1}{2}, \frac{1}{4\operatorname{diam}(\mathcal{M})}\}$. In fact, by the definition of $\hat{\Psi}$, we have that

$$C_0 = \max\left\{\left\|Vec\left(\frac{1}{4\operatorname{diam}(\mathcal{M})}A\right)\right\|_\infty, \left\|\frac{1}{4\operatorname{diam}(\mathcal{M})}y_0\right\|_\infty\right\},$$

where y_0 is the displacement vector. Using Theorem 4, we can bound each entry of random matrix A by 1. By the hypothesis $\mathbf{0} \in \mathcal{M}$ and (43), we get $\frac{1}{4\ \mathrm{diam}(\mathcal{M})} A(\mathcal{M}) \subseteq [-1/2, 1/2]^{d_\epsilon}$. So we can take $y_0 \in [-2\,\mathrm{diam}(\mathcal{M}), 2\,\mathrm{diam}(\mathcal{M})]^{d_\epsilon}$ such that $A(\mathcal{M}) \subseteq \{4\ \mathrm{diam}(\mathcal{M})\mathbf{y} + \mathbf{y}_0 \,|\, \mathbf{y} \in [0,1]^{d_\epsilon}\}$. This shows that $C_0 = \max\{\frac{1}{2}, \frac{1}{4\,\mathrm{diam}(\mathcal{M})}\}$.

We finally show that, for an appropriate choice of ρ, we can express d_ϵ in terms of ϵ, d, $V_\mathcal{M}$ and τ. In fact, by taking

$$\rho = \frac{10\tau^{2d}\epsilon^{4d}}{(e^{24}d)^d V_\mathcal{M}^2}, \tag{52}$$

a direct calculation gives the following equality

$$\left(24d + 2d\log\left(\tfrac{\sqrt{d}}{\tau\epsilon^2}\right) + \log(2V_\mathcal{M}^2)\right) = \log\left(\tfrac{20}{\rho}\right),$$

so that, by (40), we have

$$d_\epsilon \quad = \quad \left\lceil 92\epsilon^{-2}\left(24d + 2d\log\left(\tfrac{\sqrt{d}}{\tau\epsilon^2}\right) + \log(2V_\mathcal{M}^2)\right)\right\rceil,$$

where there is no dependence on ρ. Additionally, by the condition $\frac{V_\mathcal{M}}{\tau^d} \geq \left(\frac{21}{2\sqrt{d}}\right)^d$ in the hypothesis, it follows that

$$\rho = \frac{10\tau^{2d}\epsilon^{4d}}{(e^{24}d)^d V_\mathcal{M}^2} \leq 10 \cdot \left(\frac{4\epsilon^4}{441e^{24}}\right)^d,$$

showing that ρ is very small. The proof is completed by choosing $\epsilon = 1/3$ and identifying $d_e = d_{1/3}$. □

4.2. Proof of Theorem 2

Our proof of Theorem 2 adapts ideas from [19, 30] in combination with classical techniques [31]. We present the entire argument for completeness.

We recall the definition of the image of a measure (cf. [32], Sec. 3.4).

Definition 8. *Let X and Y be two sets with σ-algebras $\mathcal{A}$ and $\mathcal{B}$ defined on X and Y, respectively, and let f be a $(\mathcal{A}, \mathcal{B})$-measurable mapping from X into Y. Then, for any bounded (or bounded from below) measure μ on $\mathcal{A}$, the formula*

$$f_\#\mu \ : \ B \mapsto \mu\left(f^{-1}(B)\right), \ \ B \in \mathcal{B},$$

defines a measure on $\mathcal{B}$ called the image of the measure μ *under the mapping f.*

We also need the following change of variables lemma (Theorem 3.6.1 in [32]).

Lemma 4. *Let X and Y be two sets with σ-algebras $\mathcal{A}$ and $\mathcal{B}$ defined on X and Y, respectively, and let μ be a non-negative measure on $\mathcal{A}$. A $\mathcal{B}$-measurable function g on Y is integrable with respect to the measure $f_\#\mu$ precisely when the function $g \circ f$ is integrable with respect to μ. In addition, we have*

$$\int_Y g \ d(f_\#\mu) = \int_X g \circ f \ d\mu.$$

We also recall the definition of covering number (cf. Definition 4.2.2 in [33] or [34], p.41).

Definition 9 (Covering number). *Given a metric or pseudo-metric space* $(T, dist)$ *and a set* $K \subset T$*, for any* $\epsilon > 0$*, the* covering number $n^*(\epsilon, K, dist)$ *is the smallest number of closed balls of radius* ϵ *needed to cover* K*. That is, denoting a closed balls of radius* ϵ *as* $\beta(t, \epsilon) := \{s \in T : dist(s,t) \leq \epsilon\}$*, we have*

$$n^*(\epsilon, K, dist) := \min \left\{ n : \text{ there exist } t_1, \cdots, t_n \in T \text{ such that } K \subseteq \bigcup_{i=1}^{n} \beta(t_i, \epsilon) \right\}.$$

Clearly, if a subset K of a metric space $(T, dist)$ is precompact, then we have that $n^*(\epsilon, K, dist) < \infty$ and the covering number is a measure of the compactness of K. From the above definition, specializing to the function class of NNs, we define the covering number for a NN class.

Definition 10 (Covering number of a NN class). *Given a NN class* $\mathcal{F}(N, L, B)$ *of mappings* $f : [0,1]^d \to \mathbb{R}$*, where* $N, L \in \mathbb{N}$ *and* $B \in \mathbb{R}$ *are fixed, its* covering number $n^*(\epsilon, \mathcal{F}(N, L, B), \|\cdot\|)$ *is the smallest number of* $\|\cdot\|$*-balls of radius* ϵ *that covers* $\mathcal{F}(N, L, B)$*, where* $\|\cdot\|$ *is a norm on* $\mathcal{F}(N, L, B)$*. That is, denoting a* $\|\cdot\|$*-ball of radius centered at* $f \in \mathcal{F}(N, L, B)$ *as* $\beta(f, \epsilon) := \{g \in \mathcal{F}(N, L, B) : \|f - g\| \leq \epsilon\}$*, then*

$$\begin{aligned} &n^*(\epsilon, \mathcal{F}(N, L, B), \|\cdot\|) \\ &:= \min \left\{ n \colon \mathcal{F}(N, L, B) \subseteq \bigcup_{i=1}^{n} \beta(f_i, \epsilon) \text{ for some } f_1, \cdots, f_n \in \mathcal{F}(N, L, B) \right\}, \end{aligned}$$

in which case we call the set $\{\beta(f_i, \epsilon)\}_{i=1}^{n}$ a minimal ϵ-cover *of* $\mathcal{F}(N, L, B)$.

In practice, the norm $\|\cdot\|$ in Definition 10 that we consider below for functions $f \in \mathcal{F}(N, L, B)$ is either the infinity-norm $\|\cdot\|_\infty$ or the *empirical norm* $\|\cdot\|_n$. We recall that, given n observations $\{X_i : i = 1, \ldots n\} \subseteq [0,1]^d$, the empirical norm of $f \in \mathcal{F}(N, L, B)$ is

$$\|f\|_n := \sqrt{\frac{1}{n} \sum_{i=1}^{n} f(X_i)^2}. \tag{53}$$

The following lemma gives an upper bound of $n^*(\epsilon, \mathcal{F}(N, L, B), \|\cdot\|_\infty)$. Similar results can be found in [16], [19], [31], [35].

Lemma 5. *Let* $\mathcal{F}(N, L, B)$ *be a NN class of mappings* $R(\Phi) : [0,1]^d \to \mathbb{R}$ *with* $B > 0$ *and* $N, L \in \mathbb{N}$ *satisfying* $BN > 2$*. For any* $\epsilon > 0$*, we have*

$$n^*(\epsilon, \mathcal{F}(N, L, B), \|\cdot\|_\infty) \leq \left(\frac{4(L+1)(B+2)B^{L+1}N^{L+1}}{\epsilon} \right)^W,$$

where $W := (d+1)N + (L-1)N^2 + LN + 1$ *is the maximum number of non-zero parameters of NN class* $\mathcal{F}(N, L, B)$*.*

Proof. We consider two NNs $\Phi, \Phi' \in \mathcal{F}(N, L, B)$ where $\Phi = ((A_1, b_1), \cdots, (A_{L+1}, b_{L+1}))$ and $\Phi' = ((A'_1, b'_1), \cdots, (A'_{L+1}, b'_{L+1}))$. Under the assumptions that all parameters of Φ and Φ' are at most h away from each other, we have

$$\begin{aligned}
& \sup_{\mathbf{x}\in[0,1]^d} |R(\Phi)(\mathbf{x}) - R(\Phi')(\mathbf{x})| \\
= \;& \|(A_{L+1} \cdot \varrho(A_L \cdots \varrho(A_1\mathbf{x} + b_1) \cdots + b_L) + b_{L+1}) \\
- \;& (A'_{L+1} \cdot \varrho(A'_L \cdots \varrho(A'_1\mathbf{x} + b'_1) \cdots + b'_L) + b'_{L+1})\|_\infty \\
\le \;& \|b_{L+1} - b'_{L+1}\|_\infty + \|A_{L+1} - A'_{L+1}\|_1 \, \|A_L \cdots \varrho(A_1\mathbf{x} + b_1) \cdots + b_L\|_\infty \\
+ \;& \|A'_{L+1}\|_1 \, \|(A_L \cdots \varrho(A_1\mathbf{x} + b_1) + b_L) - (A'_L \cdots \varrho(A'_1\mathbf{x} + b'_1) \cdots + b'_L)\|_\infty \\
\overset{(i)}{\le} \;& h + hN\|A_L \cdots \varrho(A_1\mathbf{x} + b_1) \cdots + b_L\|_\infty \\
+ \;& BN\|(A_L \cdots \varrho(A_1\mathbf{x} + b_1) \cdots + b_L) - (A'_L \cdots \varrho(A'_1\mathbf{x} + b'_1) \cdots + b'_L)\|_\infty
\end{aligned} \tag{54}$$

where inequality (i) follows from the observation that the width of their hidden layers and output layer is at most N. To bound the term $\|A_L \cdots \varrho(A_1\mathbf{x} + b_1) \cdots + b_L\|_\infty$, we observe that

$$\begin{aligned}
\|A_L \cdots \varrho(A_1\mathbf{x} + b_1) \cdots + b_{L-1}\|_\infty \;&\le\; \|A_L(\cdots \varrho(A_1\mathbf{x} + b_1) \cdots)\|_\infty + \|b_L\|_\infty \\
&\le\; \|A_L\|_1 \, \|A_{L-1} \cdots \varrho(A_1\mathbf{x} + b_1) \cdots + b_{L-1}\|_\infty + B \\
&\le\; BN\|A_{L-1} \cdots \varrho(A_1\mathbf{x} + b_1) \cdots + b_{L-1}\|_\infty + B \\
&\overset{(ii)}{\le}\; (BN)^L \cdot 1 + B\sum_{i=0}^{L-1}(BN)^i \\
&\le\; (BN)^L + B(BN)^L,
\end{aligned} \tag{55}$$

where inequality (ii) follows by induction and the observation that $\|\mathbf{x}\|_\infty \le 1$; the last inequality follows by observing that $BN > 2$ and $\sum_{i=0}^{L-1}(BN)^i \le \frac{1-(BN)^L}{1-BN} \le (BN)^L$.

Now, combining (54) and (55) yields that

$$\begin{aligned}
& \sup_{\mathbf{x}\in[0,1]^d} |R(\Phi)(\mathbf{x}) - R(\Phi')(\mathbf{x})| \\
\le \;& BN\|(A_L \cdots \varrho(A_1\mathbf{x} + b_1) \cdots + b_L) - (A'_L \cdots \varrho(A'_1\mathbf{x} + b'_1) \cdots + b'_L)\|_\infty \\
+ \;& h + hN\|A_L \cdots \varrho(A_1\mathbf{x} + b_1) \cdots + b_L\|_\infty \\
\le \;& BN\|(A_L \cdots \varrho(A_1\mathbf{x} + b_1) \cdots + b_L) - (A'_L \cdots \varrho(A'_1\mathbf{x} + b'_1) \cdots + b'_L)\|_\infty \\
+ \;& h + hN\left((BN)^L + B(BN)^L\right) \\
\le \;& BN\|(A_L \cdots \varrho(A_1\mathbf{x} + b_1) \cdots + b_L) - (A'_L \cdots \varrho(A'_1\mathbf{x} + b'_1) \cdots + b'_L)\|_\infty \\
+ \;& hN(B+2)(BN)^L \\
\overset{(iii)}{\le} \;& (BN)^L\|(A_1\mathbf{x} + b_1) - (A'_1\mathbf{x} + b'_1)\|_\infty + LhN(B+2)(BN)^L \\
\le \;& (L+1)hN(B+2)(BN)^L,
\end{aligned} \tag{56}$$

where (iii) is obtained by the induction.

Finally, we discretize the non-zero parameters of the NN using step-size $h = \epsilon/(2(L+1)N(B+2)(BN)^L)$, so that we have $2B/h$ discretization points (recall that parameters range in the interval $[-B, B]$). Observing that there are at most $(2B/h)^W$ possible assignments, where the number of the non-zero parameters is at most $W = (d+1)N + (L-1)N^2 + LN + 1$, we obtain the following bound for the minimal $\epsilon/2$-covering of $\mathcal{F}(N, L, B)$:

$$n^*(\epsilon, \mathcal{F}(N, L, B), \|\cdot\|_\infty) \leq \left(\frac{2B}{h}\right)^W = \left(\frac{4(L+1)(B+2)B^{L+1}N^{L+1}}{\epsilon}\right)^W .$$

□

The next result extends Lemma 5 to the situation where the covering number is defined in terms of the empirical norm (53) rather than the infinity-norm.

Lemma 6. *Let $\mathcal{F}(N, L, B)$ be a NN class of mappings $R(\Phi) : [0, 1]^d \to \mathbb{R}$ with $B > 0$ and $N, L \in \mathbb{N}$ satisfying $BN > 2$. For any $\epsilon > 0$, we have*

$$n^*(\epsilon, \mathcal{F}(N, L, B), \|\cdot\|_n) \leq \left(\frac{4(L+1)(B+2)B^{L+1}N^{L+1}}{\epsilon}\right)^W ,$$

where W, as in Lemma 5, is the maximum number of non-zero parameters of NN class $\mathcal{F}(N, L, B)$ and $\|\cdot\|_n$ is the empirical norm (53).

Proof. We consider two NNs $\Phi, \Phi' \in \mathcal{F}(N, L, B)$ as in the proof of Lemma 5. We also assume that all parameters of Φ and Φ' are at most h away from each other. By the definition of empirical norm (53), for $X_i \in [0, 1]^d$, $i = 1, \cdots, n$, we have that

$$\|R(\Phi) – R(\Phi')\|_n = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(R(\Phi)(X_i) - R(\Phi')(X_i)\right)^2} \leq \sup_{\mathbf{x}\in[0,1]^d} |R(\Phi)(\mathbf{x}) – R(\Phi')(\mathbf{x})|. \tag{57}$$

Combining inequality (56) with inequality (57), we have

$$\|R(\Phi) - R(\Phi')\|_n \leq (L+1)hN(B+2)(BW)^L.$$

The proof is completed by the same argument as the last part of the proof of Lemma 5. □

We will also need the following result [33, Theorem 2.8.4].

Theorem 5 (Bernstein's inequality for bounded distributions). *Consider n independent random variables $U_1, \ldots, U_n$ satisfying $E[U_i] = 0$ and $|U_i| \leq c$ for all $i = 1, \ldots, n$. Then, for any $t \geq 0$, we have*

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} U_i\right| \geq t\right) \leq 2\exp\left(-\frac{n^2t^2}{2\kappa^2 + \frac{2cnt}{3}}\right), \tag{58}$$

where $\kappa^2 = \sum_{i=1}^{n} E[U_i^2]$ is the variance of the sum.

We can now prove Theorem 2.

Proof (**Proof of Theorem 2).** The proof includes the following steps.

Step 1: we apply the affine transformation Ψ, given by (8), to map f_0 into a function g_0 defined in the lower dimensional space $\Psi(\mathcal{M}) \subseteq [0,1]^{d_e}$.

Step 2: we apply a bias-variance decomposition of the squared loss in $\Psi(\mathcal{M})$.

Step 3: we estimate the variance term using the empirical norm and then applying the Borell-Sudakov-Tsirelson concentration inequality.

Step 4: we estimate the bias term using Theorem 1.

We remark that Steps 2 and 3 adapt a similar structure to the proof of Theorem 7 in [19] and include some ideas from the proof of Theorem 1 in [30].

Step 1. Similar to the definition (47) in the proof of Theorem 1, we map the D-dimensional function f_0 into a lower dimensional space by defining $g_0(\mathbf{y}) := f_0(\mathbf{x}_\mathbf{y})$, where

$$\mathbf{x}_\mathbf{y} = \{\mathbf{x} \in \mathcal{M} \,\big|\, \Psi(\mathbf{x}) = \mathbf{y},\, \mathbf{y} \in \Psi(\mathcal{M})\}$$

holds for any $\mathbf{y} \in \Psi(\mathcal{M}) \subseteq [0,1]^{d_e}$ and Ψ is given by (8).

By the definition of $\widehat{f}$ in equation (9), there exists a function $\widehat{g} \in \mathcal{F}(N_1, L_1, B_1)$, for appropriate values of the parameters N_1, L_1 and B_1, such that $\widehat{f} = \widehat{g} \circ \Psi$. Since $\widehat{g} \in \mathcal{F}(N_1, L_1, B_1)$, $\widehat{g}$ is also a continuous function. Since Ψ is a bounded affine transformation, $\widehat{f}$ is continuous and $f_0 \in \mathcal{H}(\beta, \mathcal{M}, M)$, it follows that $(\widehat{f} - f_0)^2$ is continuous on the compact set $\mathcal{M}$ and, thus, $\max_{\mathbf{x}\in\mathcal{M}} \left(\widehat{f}(\mathbf{x}) - f_0(\mathbf{x})\right)^2 < \infty$. It also follows that

$$\begin{aligned}
\int_{\mathcal{M}} \left(\widehat{f}(\mathbf{x}) - f_0(\mathbf{x})\right)^2 d\mu(\mathbf{x}) &\leq \int_{\mathcal{M}} \max_{\mathbf{x}\in\mathcal{M}} \left(\widehat{f}(\mathbf{x}) - f_0(\mathbf{x})\right)^2 d\mu(\mathbf{x}) \\
&\leq \max_{\mathbf{x}\in\mathcal{M}} \left(\widehat{f}(\mathbf{x}) - f_0(\mathbf{x})\right)^2 \mu(\mathcal{M}) \\
&\overset{(i)}{\leq} \max_{\mathbf{x}\in\mathcal{M}} \left(\widehat{f}(\mathbf{x}) - f_0(\mathbf{x})\right)^2 < \infty,
\end{aligned}$$

where inequality (i) holds since μ is a probability measure. Thus, we have that $(\widehat{f} - f_0)^2$ is integrable with respect to μ and, by Lemma 4, we have that

$$\begin{aligned}
\|\widehat{f} - f_0\|^2_{L^2(\mathcal{M},\mu)} &= \|\widehat{g} \circ \Psi - g_0 \circ \Psi\|^2_{L^2(\mathcal{M},\mu)} \\
&= \int_{\mathcal{M}} (\widehat{g} \circ \Psi - g_0 \circ \Psi)^2 \, d\mu \\
&= \int_{\Psi(\mathcal{M})} (\widehat{g} - g_0)^2 \, d(\Psi_{\#}\mu) \\
&= \|\widehat{g} - g_0\|^2_{L^2(\Psi(\mathcal{M}),\Psi_{\#}\mu)}. \qquad (59)
\end{aligned}$$

Step 2. Our next task is to bound the norm $\|\widehat{g} - g_0\|^2_{L^2(\varPsi(\mathcal{M}), \varPsi_{\#}\mu)}$. Using Lemma 3, we extend the function g_0, originally defined on $\varPsi(\mathcal{M})$, to a Hölder function $\widetilde{g_0}$ defined on $[0,1]^{d_e}$. We denote

$$g^* := R(\Phi^{\widetilde{g_0}}) \in \mathcal{F}(N_1, L_1, B_1). \tag{60}$$

By Theorem 1 , the Hölder norm of $\widetilde{g}_0$ is bounded by

$$K(\beta, M) := 8\left(\frac{3\ \mathrm{diam}(\mathcal{M})}{2}\right)^{\beta} {d_e}^{\beta/2} M.$$

Hence, by the definition of $\Phi^{\widetilde{g_0}}$, $|g^*| \leq K(\beta, M)$. By the triangle inequality and the inequality $2ab \leq a^2 + b^2$, we have

$$\|\widehat{g} - g_0\|^2_{L^2(\varPsi(\mathcal{M}), \varPsi_{\#}\mu)} \leq 2\|\widehat{g} - g^*\|^2_{L^2(\varPsi(\mathcal{M}), \varPsi_{\#}\mu)} + 2\|g^* - g_0\|^2_{L^2(\varPsi(\mathcal{M}), \varPsi_{\#}\mu)},$$

so that, from (59) we obtain

$$\|\widehat{f} - f_0\|^2_{L^2(\mathcal{M}, \mu)} \leq 2\|\widehat{g} - g^*\|^2_{L^2(\varPsi(\mathcal{M}), \varPsi_{\#}\mu)} + 2\|g^* - g_0\|^2_{L^2(\varPsi(\mathcal{M}), \varPsi_{\#}\mu)}. \tag{61}$$

We identify the two terms on the left-hand side of the inequality above with a variance and bias term, respectively.

We start by estimating the variance term $\|\widehat{g} - g^*\|^2_{L^2(\varPsi(\mathcal{M}), \varPsi_{\#}\mu)}$.

Step 3. Given any $\tau > 0$, we select functions $\{g_1, \cdots, g_N\}$ that are centers of a minimal τ-cover of $\mathcal{F}(N_1, L_1, B_1)$ with $\|\cdot\|_\infty$ (cf. Definition 10), where $N = n^*(\tau, \mathcal{F}(N_1, L_1, B_1), \|\cdot\|_\infty)$. Accordingly, there exists a $g_j \in \{g_1, \cdots, g_N\}$ such that $\|\widehat{g} - g_j\|_\infty := \sup_{\mathbf{y} \in [0,1]^{d_e}} |\widehat{g} - g_j| \leq \tau$. Without loss of generality, we can assume that $|g_j| \leq K(\beta, M)$. By the triangle inequality and the inequality $2ab \leq a^2 + b^2$, we deduce that

$$\begin{aligned}
& \|\widehat{g} - g^*\|^2_{L^2(\varPsi(\mathcal{M}), \varPsi_{\#}\mu)} \\
\leq\ & 2\|\widehat{g} - g_j\|^2_{L^2(\varPsi(\mathcal{M}), \varPsi_{\#}\mu)} + 2\|g_j - g^*\|^2_{L^2(\varPsi(\mathcal{M}), \varPsi_{\#}\mu)} \\
\leq\ & 2\|\widehat{g} - g_j\|^2_\infty + 2\|g_j - g^*\|^2_{L^2(\varPsi(\mathcal{M}), \varPsi_{\#}\mu)} \\
\leq\ & 2\tau^2 + 2\|g_j - g^*\|^2_{L^2(\varPsi(\mathcal{M}), \varPsi_{\#}\mu)}.
\end{aligned} \tag{62}$$

Hence, to control $\|\widehat{g} - g^*\|^2_{L^2(\varPsi(\mathcal{M}), \varPsi_{\#}\mu)}$, we need to derive a bound for $\|g_j - g^*\|^2_{L^2(\varPsi(\mathcal{M}), \varPsi_{\#}\mu)}$.

Given the observations $\{(X_i, Y_i)\}_{i=1}^n \subseteq \mathcal{M} \times \mathbb{R}$, by setting $Z_i = \varPsi(X_i)$ for $i = 1, \cdots, n$, equation (7) can be reformulated as follows:

$$Y_i = g_0(Z_i) + \varepsilon_i, \quad Z_i \sim \varPsi_{\#}\mu. \tag{63}$$

Take any $g_j \in \{g_1, \cdots, g_N\}$. Given $\nu > 0$ (to be determined later), let

$$t = \max\left\{\nu,\ \frac{1}{2}\|g_j - g^*\|^2_{L^2(\varPsi(\mathcal{M}), \varPsi_{\#}\mu)}\right\} \text{ and } U_i = (g_j(Z_i) - g^*(Z_i))^2 - E[(g_j(Z_i) - g^*(Z_i))^2].$$

We note that $E[U_i]=0$. We also note that, by our observations above, we have that $|g_j| \leq K(\beta, M)$, $|g^*| \leq K(\beta, M)$ and, thus, $|U_i| \leq 8K(\beta, M)^2$. A direct calculation shows that

$$\begin{aligned}
E[U_i^2] &= E\left[\left((g_j(Z_i)-g^*(Z_i))^2 - E[(g_j(Z_i)-g^*(Z_i))^2]\right)^2\right] \\
&\overset{(i)}{\leq} E\left[(g_j(Z_i)-g^*(Z_i))^4\right] \\
&= E\left[(g_j(Z_i)-g^*(Z_i))^2(g_j(Z_i)-g^*(Z_i))^2\right] \\
&\leq 4K(\beta,M)^2\, E[(g_j(Z_i)-g^*(Z_i))^2] \\
&\leq 4K(\beta,M)^2\, \|g_j-g^*\|^2_{L^2(\Psi(\mathcal{M}),\Psi_{\#}\mu)} \\
&\overset{(ii)}{\leq} 8K(\beta,M)^2 t,
\end{aligned}$$

where (i) uses the fact that, for any random variable Z and $C \in \mathbb{R}$, $E[(Z-E[Z])^2] \leq E[(Z-C)^2]$; (ii) follows from our choice of t.

We can now apply Theorem 5 to $U_i = (g_j(Z_i)-g^*(Z_i))^2 - E[(g_j(Z_i)-g^*(Z_i))^2]$ with $c = 8K(\beta,M)^2$. We obtain that

$$P\left(\|g_j-g^*\|^2_{L^2(\Psi(\mathcal{M}),\Psi_{\#}\mu)} \geq \|g_j-g^*\|_n^2 + t\right) \leq \exp\left(-\frac{3n\nu}{64K(\beta,M)^2}\right), \quad (64)$$

where $\|\cdot\|_n$ is the empirical norm (53).

By our selection of t above, we have that $t \leq \nu + \frac{1}{2}\|g_j-g^*\|^2_{L^2(\Psi(\mathcal{M}),\Psi_{\#}\mu)}$ where $\nu > 0$. Thus, by setting $\nu = 64K(\beta,M)^2(n^{d_e/(2\beta+d_e)} + \log N)/(3n)$, we have

$$\begin{aligned}
\|g_j-g^*\|_n^2 + t &\leq \|g_j-g^*\|_n^2 + \frac{64K(\beta,M)^2 n^{-2\beta/(2\beta+d_e)}}{3} \\
&+ \frac{64K(\beta,M)^2\log N}{3n} + \frac{1}{2}\|g_j-g^*\|^2_{L^2(\Psi(\mathcal{M}),\Psi_{\#}\mu)}.
\end{aligned}$$

Combining the last inequality with expression (64) it follows that

$$\begin{aligned}
&\|g_j-g^*\|^2_{L^2(\Psi(\mathcal{M}),\Psi_{\#}\mu)} \\
&\leq 2\|g_j-g^*\|_n^2 + \frac{128K(\beta,M)^2 n^{-2\beta/(2\beta+d_e)}}{3} + \frac{128K(\beta,M)^2\log N}{3n}
\end{aligned}$$

holds with probability at least $1-\exp\left(-n^{d_e/(2\beta+d_e)}\right)$ uniformly for all $g_j \in \{g_1, \cdots, g_N\}$.

We can now go back to equation (62). Using the last inequality and setting

$\tau = n^{-\beta/(2\beta+d_e)}$, we deduce that

$$\begin{aligned}
& \|\widehat{g} - g^*\|^2_{L^2(\Psi(\mathcal{M}),\Psi_{\#}\mu)} \\
\leq\ & 2\tau^2 + 4\left(\|g_j - g^*\|_n^2 + \frac{64K(\beta,M)^2\tau^2}{3} + \frac{64K(\beta,M)^2\log N}{3n}\right) \\
\leq\ & 2\tau^2 + 4\left(2\|\widehat{g} - g_j\|_n^2 + 2\|\widehat{g} - g^*\|_n^2 + \frac{64K(\beta,M)^2\tau^2}{3} + \frac{64K(\beta,M)^2\log N}{3n}\right) \\
\leq\ & 2\tau^2 + 4\left(2\tau^2 + 2\|\widehat{g} - g^*\|_n^2 + \frac{64K(\beta,M)^2\tau^2}{3} + \frac{64K(\beta,M)^2\log N}{3n}\right) \\
\leq\ & 10n^{-2\beta/(2\beta+d_e)} + 8\left(\|\widehat{g} - g^*\|_n^2 + \tfrac{32K(\beta,M)^2 n^{-2\beta/(2\beta+d_e)}}{3} + \tfrac{32K(\beta,M)^2\log N}{3n}\right) \\
\leq\ & 10n^{-2\beta/(2\beta+d_e)} + 8\left(\|\widehat{g} - g^*\|_n^2 + \tfrac{32K(\beta,M)^2 n^{-2\beta/(2\beta+d_e)}}{3}\right. \\
+\ & \left.\tfrac{32K(\beta,M)^2 W}{3n}\log\tfrac{4(L_1+1)(B_1+2)B_1^{L_1+1}N_1^{L_1+1}}{n^{-\beta/(2\beta+d_e)}}\right) \qquad (65)
\end{aligned}$$

holds with probability at least $1 - \exp\left(-n^{d_e/(2\beta+d_e)}\right)$, where the last inequality follows by the definition of $N = n^*(\tau, \mathcal{F}(N_1, L_1, B_1), \|\cdot\|_\infty)$ and Lemma 5 and

$$W := (d_e+1)N_1 + (L_1-1)N_1^2 + L_1N_1 + 1. \qquad (66)$$

Our next task is to bound the empirical norm $\|\widehat{g} - g^*\|_n^2$.

With g^* given by (60), we denote the set resulting from the translation of $\mathcal{F}(N_1, L_1, B_1)$ by g^* as

$$\mathcal{G}(g^*) := \{(g - g^*) : g \in \mathcal{F}(N_1, L_1, B_1)\}.$$

Similarly, for any $\gamma > 0$, we define

$$\mathcal{G}_\gamma(g^*) := \{(g - g^*) : \|g - g^*\|_n \leq \gamma,\, g \in \mathcal{F}(N_1, L_1, B_1)\}.$$

Using the observation that $\mathcal{G}_\gamma(g^*) \subseteq \mathcal{G}(g^*)$ and Lemma 6, we have

$$\begin{aligned}
n^*(\epsilon, \mathcal{G}_\gamma(g^*), \|\cdot\|_n) &\leq n^*(\epsilon, \mathcal{G}(g^*), \|\cdot\|_n) \\
&= n^*(\epsilon, \mathcal{F}(N_1, L_1, B_1), \|\cdot\|_n) \\
&\leq \left(\tfrac{4(L_1+1)(B_1+2)B_1^{L_1+1}N_1^{L_1+1}}{\epsilon}\right)^W. \qquad (67)
\end{aligned}$$

Using the observations $Z_1, \cdots, Z_n$, given by (63), we define the Gaussian stochastic process $\frac{1}{n}\sum_{i=1}^n \varepsilon_i g(Z_i)$, where $g \in \mathcal{G}_\gamma(g^*)$ and the terms ε_i, given by (7), are normally distributed independent random variables with mean 0 and variance σ^2. In fact, this stochastic process is a sub-Gaussian process with respect to the $\|\cdot\|_n$, as shown by Lemma 5 in [30]. We refer to [34, Sec. 2.1.2] for definitions and basic results about sub-Gaussian stochastic processes.

Using the Borell-Sudakov-Tsirelson concentration inequality [34, Thm. 2.5.8], we have that for any $u \in \mathbb{R}$

$$P\left(\sup_{g\in\mathcal{G}_\gamma(g^*)} |\tfrac{1}{n}\sum_{i=1}^n \varepsilon_i g(Z_i)| \geq E\left[\sup_{g\in\mathcal{G}_\gamma(g^*)} |\tfrac{1}{n}\sum_{i=1}^n \varepsilon_i g(Z_i)|\right] + u\right) \leq \exp\left(\frac{-nu^2}{2\sigma^2\gamma^2}\right), \tag{68}$$

where σ^2 is the variance of the terms ε_i in (7). Next, using the chaining argument, cf. [34, Thm 2.3.6], and inequality (67), we deduce that

$$\begin{aligned}
& E\left[\sup_{g\in\mathcal{G}_\gamma(g^*)} \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i g(Z_i)\right|\right] \\
\leq\ & \frac{4\sqrt{2}\sigma}{\sqrt{n}} \int_0^\gamma \sqrt{\log 2n^*(\epsilon, \mathcal{G}_\gamma(g^*), \|\cdot\|_n)}\, d\epsilon \\
\leq\ & \frac{4\sqrt{2}\sigma}{\sqrt{n}} \int_0^\gamma \sqrt{\log 2\left(\frac{4(L_1+1)(B_1+2)B_1^{L_1+1}N_1^{L_1+1}}{\epsilon}\right)^W} d\epsilon \\
\leq\ & \frac{4\sqrt{2W}\sigma}{\sqrt{n}} \int_0^\gamma \sqrt{\log 2\frac{4(L_1+1)(B_1+2)B_1^{L_1+1}N_1^{L_1+1}}{\epsilon}}\, d\epsilon \\
\leq\ & \frac{4\sqrt{2W}\sigma}{\sqrt{n}} \int_0^\gamma \log \frac{8(L_1+1)(B_1+2)B_1^{L_1+1}N_1^{L_1+1}}{\epsilon}\, d\epsilon \\
=\ & \frac{4\sqrt{2W}\sigma\gamma}{\sqrt{n}} \left(\log \frac{8(L_1+1)(B_1+2)B_1^{L_1+1}N_1^{L_1+1}}{\gamma} + 1\right).
\end{aligned} \tag{69}$$

Using (68) and (69), we now have that for any $u \in \mathbb{R}$

$$\begin{aligned}
& \sup_{g\in\mathcal{G}_\gamma(g^*)} \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i g(Z_i)\right| \\
\leq\ & \frac{4\sqrt{2W}\sigma\gamma}{\sqrt{n}} \left(\log \frac{8(L_1+1)(B_1+2)B_1^{L_1+1}N_1^{L_1+1}}{\gamma} + 1\right) + u \\
\leq\ & \frac{1}{128}\gamma^2 + 2^{11}\sigma^2\frac{W}{n}\left(\log \frac{8(L_1+1)(B_1+2)B_1^{L_1+1}N_1^{L_1+1}}{\gamma} + 1\right)^2 + u,
\end{aligned}$$

holds with probability at least $1 - \exp\left(-nu^2/(2\sigma^2\gamma^2)\right)$, where the last inequality is a consequence of the algebraic inequality $ab \leq (1/32)a^2 + 16b^2$.

By setting $u = 2^{-7}\gamma^2$ in the last inequality, it follows that

$$\sup_{g\in\mathcal{G}_\gamma(g^*)} \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i g(Z_i)\right| \leq \frac{\gamma^2}{64} + 2^{11}\frac{\sigma^2 W}{n}\left(\log \frac{8(L_1+1)(B_1+2)B_1^{L_1+1}N_1^{L_1+1}}{\gamma} + 1\right)^2 \tag{70}$$

with probability at least $1 - \exp\left(-n\gamma^2/(2^{15}\sigma^2)\right)$.

We now claim that

$$\begin{aligned}\|\widehat{g}-g^*\|_n^2 &\le 9\|g^*-g_0\|_n^2 + 2^{14}\sigma^2\frac{W}{n}\left(\log\frac{(L_1+1)(B_1+2)B_1^{L_1+1}N_1^{L_1+1}}{16\sqrt{2}\sigma n^{-\beta/(2\beta+d_e)}}+1\right)^2 \\ &+ 2^{12}\sigma^2 n^{-2\beta/(2\beta+d_e)}. \qquad (71)\end{aligned}$$

To prove the claim, we set

$$\gamma = \max\left\{\sqrt{2^{15}\sigma^2 n^{-2\beta/(2\beta+d_e)}}, 2\|\widehat{g}-g_0\|_n\right\}. \qquad (72)$$

and we consider the two cases $\|\widehat{g}-g^*\|_n \le \gamma$ and $\|\widehat{g}-g^*\|_n \ge \gamma$ separately.

When $\|\widehat{g}-g^*\|_n \le \gamma$, we have

$$\begin{aligned}&\|\widehat{g}-g^*\|_n^2 \\ &\le 2\|\widehat{g}-g_0\|_n^2 + 2\|g_0-g^*\|_n^2 \\ &\overset{(i)}{\le} 4\|g_0-g^*\|_n^2 + 4\sup_{g\in\mathcal{G}_\gamma(g^*)}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i g(Z_i)\right| \\ &\overset{(ii)}{\le} 4\|g_0-g^*\|_n^2 + 2^{13}\sigma^2\frac{W}{n}\left(\log\frac{8(L_1+1)(B_1+2)B_1^{L_1+1}N_1^{L_1+1}}{\gamma}+1\right)^2 + \frac{\gamma^2}{16} \\ &\le 4\|g^*-g_0\|_n^2 + 2^{13}\sigma^2\frac{W}{n}\left(\log\frac{(L_1+1)(B_1+2)B_1^{L_1+1}N_1^{L_1+1}}{16\sqrt{2}\sigma n^{-\beta/(2\beta+d_e)}}+1\right)^2 \\ &+\frac{1}{16}\left(\sqrt{2^{15}\sigma^2 n^{-2\beta/(2\beta+d_e)}}\right)^2 + \frac{1}{16}\left(2\|\widehat{g}-g_0\|_n\right)^2 \\ &\le 4\|g^*-g_0\|_n^2 + 2^{13}\sigma^2\frac{W}{n}\left(\log\frac{(L_1+1)(B_1+2)B_1^{L_1+1}N_1^{L_1+1}}{16\sqrt{2}\sigma n^{-\beta/(2\beta+d_e)}}+1\right)^2 \\ &+2^{11}\sigma^2 n^{-2\beta/(2\beta+d_e)} + \frac{1}{2}\|\widehat{g}-g^*\|_n^2 + \frac{1}{2}\|g^*-g_0\|_n^2, \qquad (73)\end{aligned}$$

where (i) follows from

$$\|\widehat{g}-g_0\|_n^2 \le \|g^*-g_0\|_n^2 + \frac{2}{n}\sum_{i=1}^n \varepsilon_i\left(\widehat{g}(Z_i)-g^*(Z_i)\right), \qquad (74)$$

and (ii) follows from inequality (70). To justify inequality (74), we observe that, by the definition of $\widehat{g}$, $\|Y-\widehat{g}\|_n^2 \le \|Y-g\|_n^2$ for any $g\in\mathcal{F}(N_1,L_1,B_1)$, i.e.,

$$\frac{1}{n}\sum_{i=1}^n\left(Y_i-\widehat{g}(Z_i)\right)^2 \le \frac{1}{n}\sum_{i=1}^n\left(Y_i-g(Z_i)\right)^2. \qquad (75)$$

Substituting $Y_i = g_0(Z_i)+\varepsilon_i$, defined by (63), into (75), we obtain

$$\frac{1}{n}\sum_{i=1}^n\left(g_0(Z_i)+\varepsilon_i-\widehat{g}(Z_i)\right)^2 \le \frac{1}{n}\sum_{i=1}^n\left(g_0(Z_i)+\varepsilon_i-g(Z_i)\right)^2,$$

which, after a direct calculation, yields the following inequality

$$\|\widehat{g}-g_0\|_n^2 \le \|g-g_0\|_n^2 + \frac{2}{n}\sum_{i=1}^{n}\varepsilon_i\left(\widehat{g}(Z_i)-g(Z_i)\right).$$

Inequality (71) follows from simplifying inequality (73).

When $\|\widehat{g}-g^*\|_n \ge \gamma$, by the definition of γ in (72), it follows that $2\|\widehat{g}-g_0\|_n \le \|\widehat{g}-g^*\|_n$. Then we deduce that

$$\|\widehat{g}-g^*\|_n^2 \le 2\|\widehat{g}-g_0\|_n^2 + 2\|g^*-g_0\|_n^2 \le \frac{1}{2}\|\widehat{g}-g^*\|_n^2 + 2\|g^*-g_0\|_n^2.$$

By the above inequality, we easily get $\|\widehat{g}-g^*\|_n^2 \le 4\|g^*-g_0\|_n^2$. Thus, inequality (71) holds for $\|\widehat{g}-g^*\|_n \ge \gamma$.

Step 4. Using estimates (65) and (71) into (61), we have that

$$\begin{aligned}\|\widehat{f}-f_0\|_{L^2(\mathcal{M},\mu)}^2 &\le \left(20+2^{16}\sigma^2+\frac{2^9K(\beta,M)^2}{3}\right)n^{-2\beta/(2\beta+d_e)}\\ &+ 2^{18}\sigma^2\frac{W}{n}\left(\log\frac{(L_1+1)(B_1+2)B_1^{L_1+1}N_1^{L_1+1}}{16\sqrt{2}\sigma n^{-\beta/(2\beta+d_e)}}+1\right)^2\\ &+ \frac{2^9K(\beta,M)^2W}{3n}\log\frac{4(L_1+1)(B_1+2)B_1^{L_1+1}N_1^{L_1+1}}{n^{-\beta/(2\beta+d_e)}}\\ &+ 144\max_{\mathbf{y}\in\Psi(\mathcal{M})}|g^*(\mathbf{y})-g_0(\mathbf{y})| + 2\|g^*-g_0\|_{L^2(\Psi(\mathcal{M}),\Psi_{\#}\mu)}^2,\end{aligned} \tag{76}$$

with probability at least $1-2\exp\left(-n^{d_e/(2\beta+d_e)}\right)$.

By Theorem 1, we can find a NN $f^* \in \hat{\mathcal{F}}(N_1,L_1,B_1)$ such that

$$\sup_{\mathbf{y}\in\Psi(\mathcal{M})}|g^*(\mathbf{y})-g_0(\mathbf{y})| = \max_{\mathbf{x}\in\mathcal{M}}|f^*(\mathbf{x})-f_0(\mathbf{x})| \le n^{-\beta/(2\beta+d_e)}, \tag{77}$$

where $f^* = g^*\circ\Psi$ and $g^* \in \mathcal{F}(N_1,L_1,B_1)$.

Since $\|g^*-g_0\|_{L^2(\Psi(\mathcal{M}),\Psi_{\#}\mu)}^2 \le \max_{\mathbf{y}\in\Psi(\mathcal{M})}|g^*(\mathbf{y})-g_0(\mathbf{y})|$, it follows by applying (66) and (77) into (76) that there is a constant $C = C(\sigma,\beta,d_e,M,\mathrm{diam}(\mathcal{M}))$ such that

$$\|\widehat{f}-f_0\|_{L^2(\mathcal{M},\mu)}^2 \le Cn^{-2\beta/(2\beta+d_e)}(1+\log n)^2,$$

with probability at least $1-2\exp\left(-n^{d_e/(2\beta+d_e)}\right)$. □

Appendix A. Proof of Theorem 4.

Theorem 4 slightly modifies Theorem 2 in [26] by changing the family of random projections that are used to map points from a manifold $\mathcal{M}$ in $\mathbb{R}^D$ into a lower dimensional space. Namely, while the original random projections matrices were assumed to have entries with zero mean normal random variables, here we assume a different random distribution that allows us a better control on the size of the entries.

Since our proof follows closely the structure of the original proof in [26], we only report below the new elements and indicate how the original proof needs to be modified.

We start by recalling some useful definitions and observations, cf. [36].

Definition 11 (Sub-Gaussianity). *For any $\sigma > 0$, a zero-mean random variable X is σ^2-sub-Gaussian if, for all $\lambda \in \mathbb{R}$,*

$$\mathbb{E}[e^{\lambda X}] \leq \exp\left(\tfrac{\lambda^2\sigma^2}{2}\right).$$

Proposition 1 (Sums of sub-Gaussians). *For $i = 1, \dots, n$, let $a_i, \sigma_i \in \mathbb{R}, \sigma_i > 0$ and X_i be independent, mean zero σ_i^2-sub-Gaussian random variables. Then $\sum_{i=1}^n a_i X_i$ is $\sum_{i=1}^n a_i^2 \sigma_i^2$-sub-Gaussian.*

Definition 12 (Sub-exponential). *A mean-zero random variable X is (τ^2, b)-sub-exponential, with $\tau, b > 0$, if, for all $|\lambda| \leq \frac{1}{b}$,*

$$\mathbb{E}[e^{\lambda X}] \leq \exp\left(\tfrac{\lambda^2\tau^2}{2}\right).$$

Proposition 2 (Sums of sub-exponentials). *For $i = 1, \dots, n$, let $\tau_i \in \mathbb{R},$, $b_i > 0$ and X_i be independent (τ_i^2, b_i)-sub-exponential random variables. Then $\sum_{i=1}^n X_i$ is $(\sum_{i=1}^n \tau_i^2, b_*)$-sub-exponential, where $b_* = \max_i b_i$. In addition,*

$$P\left(\left|\frac{1}{n}\sum_{i=1}^n X_i - E[X_i]\right| \geq t\right) \leq 2\exp\left(-\min\left\{\frac{nt^2}{2\frac{1}{n}\sum_{i=1}^n \tau_i^2}, \frac{nt}{2b_*}\right\}\right).$$

Adapting the argument in [37, Appendix B], we have the following proposition.

Proposition 3. *If random variable X is σ^2-sub-Gaussian, then X^2 is $(18\sigma^4, 18\sigma^2)$-sub-exponential.*

Proof. For any integer $r \geq 1$, the moments of the σ^2-sub-Gaussian variable X are bounded by the inequality

$$E[|X|^r] \leq r\, 2^{r/2}\, \sigma^r\, \Gamma(r/2), \tag{A.1}$$

where Γ is the Gamma function.

Using a power series expansion, inequality (A.1) and the observation that

$\Gamma(r) = (r-1)!$ for an integer $r \geq 1$, we obtain that, for any $t \in \mathbb{R}$,

$$\begin{aligned}
E[e^{t(X^2-E[X^2])}] &= 1 + tE[X^2 - E[X^2]] + \sum_{r=2}^{\infty} \frac{t^r E[(X^2 - E[X^2])^r]}{r!} \\
&= 1 + \sum_{r=2}^{\infty} \frac{t^r E[(X^2 - E[X^2])^r]}{r!} \\
&\overset{(i)}{\leq} 1 + \sum_{r=2}^{\infty} \frac{|t|^r E[|X|^{2r}]}{r!} \\
&\leq 1 + \sum_{r=2}^{\infty} \frac{|t|^r 2r 2^r \sigma^{2r} \Gamma(r)}{r!} \\
&= 1 + \sum_{r=2}^{\infty} |t|^r 2^{r+1} \sigma^{2r} \\
&= 1 + \frac{8t^2\sigma^4}{1 - 2|t|\sigma^2}, \qquad \text{(A.2)}
\end{aligned}$$

where (i) uses the fact that, for any random variable X and $C \in \mathbb{R}$, $E[(X - E[X])^2] \leq E[(X - C)^2]$.

By choosing $|t| \leq 1/(18\sigma^2)$, we have that $8/(1 - 2|t|\sigma^2) \leq 9$. Finally, using this observation and the inequality $1 + \alpha \leq e^{\alpha}$, valid for any $\alpha \in \mathbb{R}$, from (A.2) we deduce that, for all $|t| \leq 1/(18\sigma^2)$, we have

$$E[e^{t(X^2-E[X^2])}] \leq e^{9\,t^2\sigma^4} = \exp(t^2 \frac{18\sigma^4}{2}).$$

□

The key element of the proof is the following lemma which is a modification of Lemma 17 in [26]. Once this lemma is proved, the proof of Theorem 4 then follows exactly as in the original proof of Theorem 2 in [26]. Hence the rest of the argument is omitted.

Lemma 7. *Let $A = (a_{i,j})$ be a $Q \times D$ matrix populated with i.i.d. random variables with entries*

$$a_{ij} = \begin{cases} +\frac{1}{\sqrt{Q}} & \text{with probability } \frac{1}{2} \\ -\frac{1}{\sqrt{Q}} & \text{with probability } \frac{1}{2}. \end{cases} \qquad \text{(A.3)}$$

Fix $0 \leq \lambda \leq 1/3$ and $\lambda' \geq 1$ or $\lambda' = 0$. Then, for fixed $y \in \mathbb{R}^D$, we have

$$\begin{aligned}
P\{|\|Ay\| - \|y\|| > \lambda\|y\|\} &\leq 2\exp\left(-\frac{Q\lambda^2}{36}\right) \qquad \text{(A.4)} \\
P\{\|Ay\| > (1 + \lambda')\|y\|\} &\leq 2\exp\left(-\frac{Q\lambda'}{36}\right). \qquad \text{(A.5)}
\end{aligned}$$

Proof. We claim that the random variable X defined in (A.3) is $\frac{1}{Q}$-sub-Gaussian. By taking expectations and using the power series expansion for the exponential, we obtain

$$\begin{aligned}
\mathbb{E}[e^{\lambda X}] &= \frac{1}{2}\left(e^{-\frac{\lambda}{\sqrt{Q}}} + e^{\frac{\lambda}{\sqrt{Q}}}\right) \\
&= \frac{1}{2}\left(\sum_{k=0}^{\infty}\frac{(-\lambda/\sqrt{Q})^k}{k!} + \sum_{k=0}^{\infty}\frac{(\lambda/\sqrt{Q})^k}{k!}\right) \\
&= \sum_{k=0}^{\infty}\frac{(\lambda/\sqrt{Q})^{2k}}{(2k)!} \\
&\leq 1 + \sum_{k=1}^{\infty}\frac{(\lambda/\sqrt{Q})^{2k}}{2^k k!} \\
&= \exp\left(\frac{\lambda^2/Q}{2}\right).
\end{aligned}$$

It is straightforward to verify that, for any $y = (y_j) \in \mathbb{R}^D$, we have $\mathbb{E}\|Ay\|^2 = \|y\|^2$. Without loss of generality, we assume that $\|y\| = 1$. Observe that, for $i = 1, \ldots, Q$, $(Ay)_i = \sum_{j=1}^{D} a_{ij}y_j$ and $\|y\|^2 = \sum_{j=1}^{D} y_j^2 = 1$. By Proposition 1, $(Ay)_i$ is $\frac{1}{Q}$-sub-Gaussian and hence, by Proposition 3, $(Ay)_i^2$ is $(\frac{18}{Q^2}, \frac{18}{Q})$-sub-exponential. By Proposition 2, we have that $\|Ay\|^2 = \sum_{i=1}^{Q}(Ay)_i^2$ is $(\frac{18}{Q}, \frac{18}{Q})$-sub-exponential and, for $k > 0$,

$$P\left(\left|\|Ay\|^2 - 1\right| \geq k\right) \leq 2\exp\left(-\min\left\{\frac{Qk^2}{36}, \frac{Qk}{36}\right\}\right). \qquad \text{(A.6)}$$

By inequality (A.6) and $0 \leq \lambda \leq 1$, we then can deduce that

$$\begin{aligned}
P\left(\|Ay\| - 1| \geq \lambda\right) &= P\left(\|Ay\| \geq 1 + \lambda\right) + P\left(\|Ay\| \leq 1 - \lambda\right) \\
&\leq P\left(\|Ay\|^2 \geq 1 + \lambda\right) + P\left(\|Ay\|^2 \leq 1 - \lambda\right) \\
&= P\left(\left|\|Ay\|^2 - 1\right| \geq \lambda\right) \\
&\leq 2\exp\left(-\frac{Q\lambda^2}{36}\right).
\end{aligned}$$

This establishes inequality (A.4). For inequality (A.5), using inequality (A.6) and $\lambda' \geq 1$, we obtain that

$$\begin{aligned}
P\left(\|Ay\| \geq 1 + \lambda'\right) &= P\left(\|Ay\|^2 \geq (1 + \lambda')^2\right) \\
&\leq P\left(\|Ay\|^2 \geq 1 + \lambda'\right) \\
&\leq P\left(\left|\|Ay\|^2 - 1\right| \geq \lambda\right) \\
&\leq 2\exp\left(-\frac{Q\lambda'}{36}\right).
\end{aligned}$$

□

References

[1] R. Bellman, On the theory of dynamic programming, Proceedings of the national Academy of Sciences 38 (1952) 716–719.

[2] E. Novak, H. Woźniakowski, Approximation of infinitely differentiable multivariate functions is intractable, Journal of Complexity 25 (2009) 398–404.

[3] P. Grohs, F. Hornung, A. Jentzen, P. Von Wurstemberger, A proof that artificial neural networks overcome the curse of dimensionality in the numerical approximation of black-scholes partial differential equations, Memoirs of the American Mathematical Society 284 (2023).

[4] T. Poggio, H. Mhaskar, L. Rosasco, B. Miranda, Q. Liao, Why and when can deep-but not shallow-networks avoid the curse of dimensionality: a review, International Journal of Automation and Computing 14 (2017) 503–519.

[5] J. Berner, P. Grohs, G. Kutyniok, P. Petersen, The Modern Mathematics of Deep Learning, Cambridge University Press, 2022, p. 1–111. doi:10.1017/9781009025096.002.

[6] I. Daubechies, R. DeVore, S. Foucart, B. Hanin, G. Petrova, Nonlinear approximation and (deep) ReLU networks, Constructive Approximation 55 (2022) 127–172.

[7] R. DeVore, B. Hanin, G. Petrova, Neural network approximation, Acta Numerica 30 (2021) 327–444.

[8] A. R. Barron, Universal approximation bounds for superpositions of a sigmoidal function, IEEE Transactions on Information theory 39 (1993) 930–945.

[9] J. W. Siegel, J. Xu, Characterization of the variation spaces corresponding to shallow neural networks, Constructive Approximation (2023) 1–24.

[10] S. Wojtowytsch, et al., Representation formulas and pointwise properties for barron functions, Calculus of Variations and Partial Differential Equations 61 (2022) 1–37.

[11] D. Dũng, et al., Deep relu neural networks in high-dimensional approximation, Neural Networks 142 (2021) 619–635.

[12] C. Fefferman, S. Mitter, H. Narayanan, Testing the manifold hypothesis, Journal of the American Mathematical Society 29 (2016) 983–1049.

[13] P. J. Bickel, B. Li, Local polynomial regression on unknown manifolds, Lecture Notes-Monograph Series (2007) 177–186.

[14] R. R. Coifman, M. Maggioni, Diffusion wavelets, Applied and computational harmonic analysis 21 (2006) 53–94.

[15] U. Shaham, A. Cloninger, R. R. Coifman, Provable approximation properties for deep neural networks, Applied and Computational Harmonic Analysis 44 (2018) 537–557.

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

[17] A. Cloninger, T. Klock, A deep network construction that adapts to intrinsic dimensionality beyond the domain, Neural Networks 141 (2021) 404–419.

[18] J. Schmidt-Hieber, Deep relu network approximation of functions on a manifold, arXiv preprint arXiv:1908.00695 (2019).

[19] R. Nakada, M. Imaizumi, Adaptive approximation and generalization of deep neural network with intrinsic dimensionality., J. Mach. Learn. Res. 21 (2020) 1–38.

[20] P. Petersen, F. Voigtlaender, Optimal approximation of piecewise smooth functions using deep relu neural networks, Neural Networks 108 (2018) 296–330.

[21] Z. Shen, H. Yang, S. Zhang, Deep network approximation characterized by number of neurons, Communications in Computational Physics 28 (2020) 1768–1811. doi:https://doi.org/10.4208/cicp.OA-2020-0149.

[22] J. Lu, Z. Shen, H. Yang, S. Zhang, Deep network approximation for smooth functions, SIAM Journal on Mathematical Analysis 53 (2021) 5465–5506.

[23] J.-F. Cai, D. Li, J. Sun, K. Wang, Enhanced expressive power and fast training of neural networks by random projections, CSIAM Transactions on Applied Mathematics 2 (2021) 532–550. doi:10.4208/csiam-am.SO-2020-0004.

[24] Y. Jiao, G. Shen, Y. Lin, J. Huang, Deep nonparametric regression on approximate manifolds: Nonasymptotic error bounds with polynomial prefactors, The Annals of Statistics 51 (2023) 691–716.

[25] Z. Shen, H. Yang, S. Zhang, Optimal approximation rate of relu networks in terms of width and depth, Journal de Mathématiques Pures et Appliquées 157 (2022) 101–135.

[26] A. Eftekhari, M. B. Wakin, New analysis of manifold embeddings and signal recovery from compressive measurements, Applied and Computational Harmonic Analysis 39 (2015) 67–109.

[27] P. Niyogi, S. Smale, S. Weinberger, Finding the homology of submanifolds with high confidence from random samples, Discrete & Computational Geometry 39 (2008) 419–441.

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

[29] D. Yarotsky, Error bounds for approximations with deep relu networks, Neural Networks 94 (2017) 103–114.

[30] T. Suzuki, Fast generalization error bound of deep learning from a kernel perspective, in: International Conference on Artificial Intelligence and Statistics, PMLR, 2018, pp. 1397–1406.

[31] M. Anthony, P. L. Bartlett, P. L. Bartlett, et al., Neural network learning: Theoretical foundations, volume 9, cambridge university press Cambridge, 1999.

[32] V. I. Bogachev, M. A. S. Ruas, Measure theory, volume 1, Springer, 2007.

[33] R. Vershynin, High-dimensional probability: An introduction with applications in data science, volume 47, Cambridge university press, 2018.

[34] E. Giné, R. Nickl, Mathematical foundations of infinite-dimensional statistical models, Cambridge university press, 2021.

[35] J. Schmidt-Hieber, Nonparametric regression using deep neural networks with relu activation function, The Annals of Statistics 48 (2020) 1875–1897.

[36] P. Rigollet, J.-C. Hütter, High-dimensional statistics, arXiv preprint arXiv:2310.19244 (2023).

[37] J. Honorio, T. Jaakkola, Tight bounds for the expected risk of linear classifiers and pac-bayes finite-sample guarantees, in: Artificial Intelligence and Statistics, PMLR, 2014, pp. 384–392.