Neural Networks 141 (2021) 404–419

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

A deep network construction that adapts to intrinsic dimensionality beyond the domain

Alexander Cloninger [a,b], Timo Klock [a,c,*]

[a] Department of Mathematics, University of California San Diego, 9500 Gilman Dr., La Jolla, CA 92093, United States
[b] Halıcığlu Data Science Institute, University of California San Diego, 10100 Hopkins Dr., La Jolla, CA 92093, United States
[c] Department of Numerical Analysis and Scientific Computing, Simula Research Laboratory, Martin Linges Vei 25, Fornebu 1364, Norway

ARTICLE INFO

Article history:
Received 6 August 2020
Received in revised form 26 April 2021
Accepted 3 June 2021
Available online 8 June 2021

Keywords:
Deep neural networks
Approximation theory
Curse of dimensionality
Composite functions
Noisy manifold models

ABSTRACT

We study the approximation of two-layer compositions $f(x) = g(\phi(x))$ via deep networks with ReLU activation, where ϕ is a geometrically intuitive, dimensionality reducing feature map. We focus on two intuitive and practically relevant choices for ϕ: the projection onto a low-dimensional embedded submanifold and a distance to a collection of low-dimensional sets. We achieve near optimal approximation rates, which depend only on the complexity of the dimensionality reducing map ϕ rather than the ambient dimension. Since ϕ encapsulates all nonlinear features that are material to the function f, this suggests that deep nets are faithful to an intrinsic dimension governed by f rather than the complexity of the domain of f. In particular, the prevalent assumption of approximating functions on low-dimensional manifolds can be significantly relaxed using functions of type $f(x) = g(\phi(x))$ with ϕ representing an orthogonal projection onto the same manifold.

1. Introduction

In the past decade neural networks emerged as powerful tools to construct state-of-the-art solutions for various different data analysis tasks. Much of this progress is of empirical nature and cannot be explained by current mathematical theory. This led to a re-emerging interest for developing a theoretical understanding of deep networks in recent years. In this work we contribute to the effort by studying the approximative capacity of deep networks with respect to practically motivated composite function classes in the high-dimensional regime.

Approximation properties of shallow and deep networks have been studied for over three decades and gained much traction during the rise of neural networks around the 80s and 90s (Cybenko, 1989; Hornik, Stinchcombe, & White, 1989; Leshno, Lin, Pinkus, & Schocken, 1993; Mhaskar, 1993, 1996). It is well-known that shallow networks (with non-polynomial activation) are universal approximators, which means they can approximate any continuous function on a compact subset of $\mathbb{R}^D$ arbitrarily well (Cybenko, 1989; Hornik et al., 1989; Leshno et al., 1993). Furthermore, it has been established that the number of required nonzero network parameters for uniformly approximating a $\mathcal{C}^\alpha$-function to accuracy ε on a compact subset of $\mathbb{R}^D$ is in $\mathcal{O}(\varepsilon^{-D/\alpha})$ (Mhaskar, 1996; Pinkus, 1999). Similar results hold for deep networks with the additional benefit that the approximation can be localized, contrary to approximation via shallow networks (Chui, Li, & Mhaskar, 1994, 1996; Mhaskar, 1993).

In modern networks differentiable sigmoidal activation functions are often replaced by the rectified linear unit activation (ReLU), because such networks do not suffer the vanishing gradient problem and can thus be more easily trained via backpropagation (Goodfellow, Bengio, Courville, & Bengio, 2016). Approximation properties of ReLU networks received much attention in recent years (Bölcskei, Grohs, Kutyniok, & Petersen, 2019; Grohs, Perekrestenko, Elbrächter, & Bölcskei, 2019; Petersen & Voigtlaender, 2018; Shaham, Cloninger, & Coifman, 2018; Shen, Yang, & Zhang, 2019; Telgarsky, 2017; Yarotsky, 2017, 2018). The bottom line is that ReLU networks are at least as expressive as networks with differentiable sigmoidal activation. Moreover, a series of recent works (Fang, Feng, Huang, & Zhou, 2020; Zhou, 2020a, 2020b) shows that this is also true for deep convolutional ReLU networks, which are significantly less flexible compared to fully-connected networks. To comply with modern neural network practice, we concentrate on the ReLU activation in this work, though we emphasize that we have no reason to believe our results are special to this choice.

Approximating functions either through differentiable sigmoidal networks or ReLU networks suffers from the curse of dimensionality, because the number of required parameters for approximating $f \in \mathcal{C}^\alpha$ on a compact subset of $\mathbb{R}^D$ is exponential

* Corresponding author at: Department of Numerical Analysis and Scientific Computing, Simula Research Laboratory, Martin Linges Vei 25, Fornebu 1364, Norway.
E-mail addresses: acloninger@ucsd.edu (A. Cloninger), timo@simula.no (T. Klock).

https://doi.org/10.1016/j.neunet.2021.06.004
0893-6080/

in D. Since high-dimensional problems are ubiquitous in applied areas, it is of great interest to identify narrower but sufficiently rich function classes that allow for faster approximation rates with at most polynomial dependency on D.

Three decades ago, the author of Barron (1993) showed that functions f, whose Fourier transform $\hat{f}$ satisfies

$$C_f = \int_{\mathbb{R}^D} \left|\omega \hat{f}(\omega)\right| d\omega < \infty,$$

can be approximated by a shallow network to accuracy ε using just $\mathcal{O}(\varepsilon^{-2})$ neurons. Functions satisfying such conditions are said to be of Barron-type and they are under continuous investigation ever since (Barron, 1994; Klusowski & Barron, 2016; Montanelli, Yang, & Du, 2019). Unfortunately, the constant involved in $\mathcal{O}(\varepsilon^{-2})$ depends on C_f, which in turn increases exponentially with the dimension D under standard regularity assumptions alone. Several works (Kurková & Sanguineti, 2001, 2002; Mhaskar, 2004) have subsequently investigated conditions on f that imply the growth of C_f is at most polynomial in D.

In Mhaskar, Liao, and Poggio (2016, 2017), Mhaskar and Poggio (2016), Poggio, Anselmi, and Rosasco (2015), Poggio, Mhaskar, Rosasco, Miranda, and Liao (2017) and Schmidt-Hieber (2020) the benefit of depth of networks has been analyzed by studying approximation properties of deep nets for compositional functions of the type $f(x) = g_L \circ \ldots \circ g_1(x)$. Intuitively, if all intermediate functions $g_\ell : \mathbb{R}^{\ell-1} \to \mathbb{R}^\ell$ are easier to approximate than the final target f, deep networks can approximate f more efficiently by mimicking the compositional structure of the function. This situation arises, for instance, if each component $g_{\ell,p} : \mathbb{R}^{\ell-1} \to \mathbb{R}$, $p = 1, \ldots, \ell$, depends on at most k of the $\ell - 1$ coordinates of the previous output, i.e., can be written as $\tilde{g}_{\ell,p}(I_{\ell,p}(x)) = g_{\ell,p}(x)$ for a map $I_{\ell,p} : \mathbb{R}^{\ell-1} \to \mathbb{R}^k$ that selects k coordinates, independently of x. In this case, assuming all components $g_{\ell,p}$, $p = 1, \ldots, \ell$, $\ell = 1, \ldots, L$ are α-Hölder, the function f can be approximated uniformly up to error ε using $\mathcal{O}(\varepsilon^{-k/\alpha})$ nonzero parameters (here, L is treated as a constant). The missing dependence on D in the exponent show that compositions pave a way for defining classes of functions that are narrow enough to avoid the curse of dimensionality (Mhaskar & Poggio, 2016, 2020; Poggio et al., 2017). This led to the notion of 'blessing of compositionality' as a cure to the curse of dimensionality.

Another line of research, which is motivated by the popularity of nonlinear dimension reduction methods, studies approximation of $f : \mathcal{M} \subseteq [0,1]^D \to \mathbb{R}$ on low-dimensional domains $\mathcal{M}$, such as a d-dimensional embedded submanifold. The authors of Shaham et al. (2018) established that uniform approximations to accuracy ε require just $\mathcal{O}(\varepsilon^{-d/\alpha})$ parameters, replacing the ambient dimension D with the intrinsic manifold dimension d. Similar results have been shown in Chen, Jiang, Liao, and Zhao (2019), Chui and Mhaskar (2018), Schmidt-Hieber (2019) and extended to more general notions of dimensionality or other types of neural networks (Mhaskar, 2020a, 2020b; Nakada & Imaizumi, 2019), including radial basis function networks and abstract generalizations thereof. Therefore, certain approximation systems, including deep networks, adapt to the intrinsic dimension of the domain of the target.

Approximation on low-dimensional domains is appealing because it is geometrically intuitive and can, to some extent, be checked in practice by analyzing local covariance matrices of a given data set. However, defining the complexity of an approximation task via the domain of the target has some significant drawbacks, which we highlight in the next section.

1.1. Drawbacks of measuring complexity by the target domain

Noisy manifold hypothesis. Many theoretical results that alleviate the curse of dimensionality are based either explicitly or implicitly on the exact manifold hypothesis, which states that data is supported on a low-dimensional manifold. In view of usually noisy real-world data, the exact manifold hypothesis seems overly stringent and in fact has been criticized for being rarely observable in practice (Hein & Maier, 2007a, 2007b). A more realistic alternative is to model real-world data as a sum of clean data, which is supported on a low-dimensional manifold $\mathcal{M}$ (think of the 'face manifold' consisting of images of faces He, Yan, Hu, Niyogi, & Zhang, 2005), plus noise, which generically pushes data points off the clean data manifold. If the noise is unstructured, we can simplistically assume that it concentrates in the local normal space of $\mathcal{M}$, and we may associate to $x \in \mathbb{R}^D$ the orthogonal projection $\pi_{\mathcal{M}}(x) = \operatorname{argmin}_{z\in\mathcal{M}} \|x - z\|_2$ as the clean data sample. We now aim for approximating functions $f(x) = g(\pi_{\mathcal{M}}(x))$, where $g : \mathcal{M} \to \mathbb{R}$ describes a function of interest defined on clean data. See Fig. 1a for an illustration of the setting.

Following results in Chen et al. (2019), Schmidt-Hieber (2019) and Shaham et al. (2018) about approximation over low-dimensional domains, we are tempted to think there is a significant difference between approximating a function $g : \mathcal{M} \to \mathbb{R}$ on $\mathcal{M}$ or a function $f(x) = g(\pi_{\mathcal{M}}(x))$ on a full-dimensional tubular domain around $\mathcal{M}$. We will prove that, in fact, both functions are approximable with similarly sized networks and by using the same amount of information about the target f.

We add that the stringency of the exact manifold hypothesis is often recognized and discussed in the literature. For instance, the authors of Chui and Mhaskar (2018) explain that their approximation results are robust to an inexact manifold hypothesis, because noise that spreads only in $s \ll D$ directions in the local normal space increases the dimensionality of the data manifold to just $d + s \ll D$. Furthermore, Mhaskar (2020b) proposes a Hermite polynomial based approximation scheme for functions on manifolds, which is robust to a degree of off-manifold noise. The theory in Cheng and Cloninger (2019) includes off-manifold noise under the assumption that the noise vanishes exponentially fast with increased distance from the manifold.

Adaptivity to function complexity. The same argument as in the previous paragraph can be made when approximating a function that just depends on a lower dimensional set of linear or nonlinear transformations of the input, as is common in the sufficient dimension reduction literature (Li, 2018). To give a simple example, we may consider the swiss role manifold $\mathcal{M}$ as in Figs. 1b–1c, where the colors indicate values of two different Lipschitz-continuous functions. Based on previously mentioned approximation results (Chen et al., 2019; Schmidt-Hieber, 2019; Shaham et al., 2018), both functions can be approximated using deep networks with $\mathcal{O}(\varepsilon^{-1/\dim(\mathcal{M})}) = \mathcal{O}(\varepsilon^{-1/2})$ parameters. However, the complexity of functions in 1b and 1c differs, because we can express f in 1b as $f(x) = g(\pi_\gamma(x))$, where γ is a one-dimensional manifold. In other words, there exists a submanifold $\gamma \subset \mathcal{M}$ with $\dim(\gamma) = 1$ that contains all material information for recovering the target function f.

Classification problems with class attractors. Another example, where the domain of the target is not a suitable measure of complexity, are classification problems with class attractors, see Fig. 1d. Here, we assume that the class label depends only on the proximity of the input to a low-dimensional attractor set, such as for instance a finite set of points. Hence, if we were aware of the attractor set, the target function is completely determined by evaluating the distance to the set, indicating that the complexity

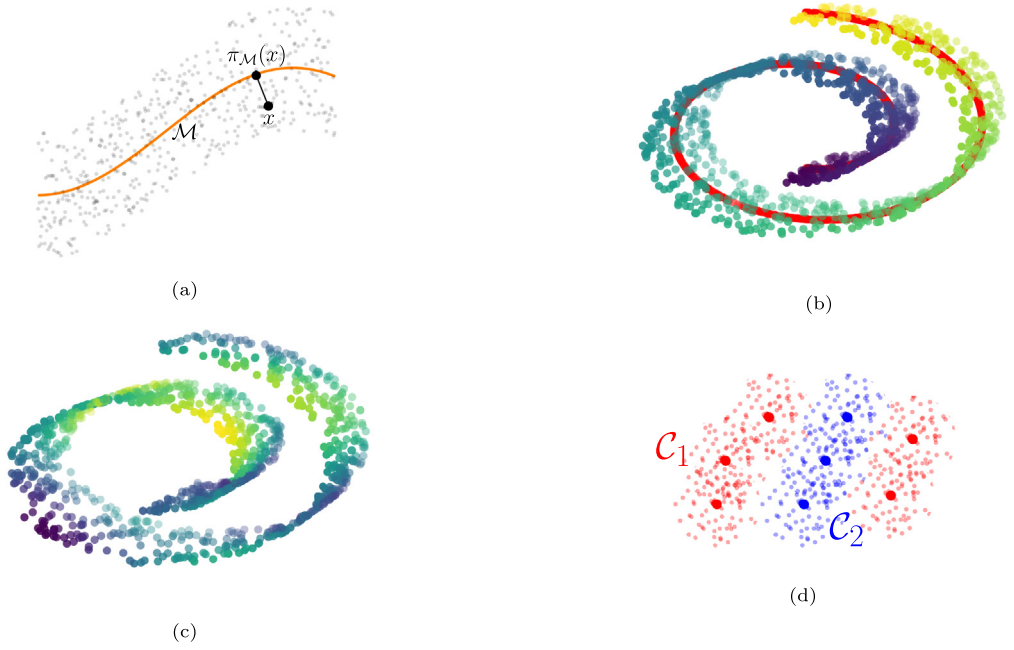

Fig. 1. Examples highlighting drawbacks of defining the approximation complexity via the target domain. In 1a the target function depends just on the projection of the input onto a low-dimensional manifold, yet the data is spread in a full-dimensional subset of $\mathbb{R}^D$. 1b–1c show two functions whose domain is the swiss role, but which are of different complexity because the function in 1b just depends on a single nonlinear transformation of the data (the red curve). 1d shows a classification problem where labels are assigned based on the proximity to a few class attractors (bold dots). In all three cases the dimensionality of the approximation domain is not a suitable measure for the difficulty of the approximation problem. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

of the target function is dictated by the complexity of the distance metric and the set of attractors, rather than the domain of the target. Classification problems, where a finite number of attractors exist, are the main object of study in few-shot learning, see for instance (Sung et al., 2018). In these problems the goal is to predict class labels after querying a tiny amount of samples, which are ideally points that serve as class attractors with respect to a, possibly prescribed, metric.

1.2. Contribution

Our main goal is to extend approximation guarantees of deep nets from functions defined on low-dimensional domains to functions that encode low-dimensionality in the joint input–output relation $x \mapsto f(x)$. We study two classes of functions, which resemble two layer composite functions $f(x) = g(\phi(x))$, where $\phi(x)$ takes the role of a geometrically intuitive, dimensionality reducing feature map. By resorting to such a function-driven notion of low-complexity, we alleviate the drawbacks raised in the previous section.

Functions of projections to low-dimensional sets. We first consider functions that model ϕ as an orthogonal projection onto a d-dimensional Riemannian submanifold $\mathcal{M} \subseteq [0,1]^D$. In this case we can write the target $f : \mathcal{A} \subseteq [0,1]^D \to \mathbb{R}$ as

$$f(x) = g(\pi_{\mathcal{M}}(x)) \quad \text{where} \quad \pi_{\mathcal{M}}(x) \in \operatorname*{argmin}_{z \in \mathcal{M}} \|x - z\|_2 \,, \tag{1}$$

and the approximation domain $\mathcal{A}$ is assumed to be contained in a tubular region around $\mathcal{M}$. The width of this region is constrained to guarantee that $\pi_{\mathcal{M}}(x)$ is Lipschitz-continuous, as described in detail in Section 2. We refer to the associated function class as Class 1 below.

Assumption (1) naturally includes the popular case $\mathcal{A} = \mathcal{M}$ and $\pi_{\mathcal{M}} = \mathrm{Id}$, which has been studied in Chen et al. (2019), Chui and Mhaskar (2018), Mhaskar (2020a, 2020b), Nakada and Imaizumi (2019), Schmidt-Hieber (2019) and Shaham et al. (2018). In the present case the approximation domain $\mathcal{A}$ does however not need to be low-dimensional. Rather, Eq. (1) imposes that f is locally constant in $D - d$ directions, corresponding to the local normal space of $\mathcal{M}$. If we were able to extract a subset of the approximation space $A \subseteq \mathcal{A}$, whose projection $\pi_{\mathcal{M}}(A)$ is supported on a small patch of the manifold $\mathcal{M}$ so that curvature effects of $\mathcal{M}$ are negligible, we can view $f|_A$ as a constant function with optimal regularity in $D - d$ directions corresponding to the local normal space, and regularity dictated by $g|_{\pi_{\mathcal{M}}(A)}$ in the remaining d directions. Following this intuition, our viewpoint is aligned with recent work on approximation of functions in anisotropic Besov spaces (Suzuki, 2018; Suzuki & Nitanda, 2019).

Contribution We achieve the same approximation guarantee that is achieved in Nakada and Imaizumi (2019), Schmidt-Hieber (2019) and Shaham et al. (2018) for the case $\mathcal{A} = \mathcal{M}$. Namely, if $\mathcal{M}$ is a d-dimensional manifold satisfying some common regularity assumptions and g is α-Hölder with respect to the geodesic metric on $\mathcal{M}$, functions of Class 1 can be approximated uniformly to accuracy ε using a deep ReLU network based on $\mathcal{O}(\varepsilon^{-d/\alpha})$ point queries of f and with $\mathcal{O}(\log(D)D\log^2(\varepsilon^{-1})\varepsilon^{-d/\alpha})$ nonzero parameters arranged in $\mathcal{O}(\log(D)\log^2(\varepsilon^{-1}))$ layers. The result is optimal in terms of the number of required function queries according to nonlinear width theory (DeVore, Howard, & Micchelli, 1989), and optimal (apart from logarithmic factors) in terms of the required network dimensions (Yarotsky, 2018, Theorem 1). We believe the result sheds a new light on the relevance of the manifold hypothesis, because we identify local invariances encoded in $x \mapsto f(x)$ as the key factor to simplify the approximation problem, as opposed to the complexity of the underlying data manifold.

Functions of distances to low-dimensional sets. Second, we study functions that depend only on distances to a collection of finite or low-dimensional sets $\mathcal{C}_1, \ldots, \mathcal{C}_M$. Mathematically, we assume $f : [0,1]^D \to \mathbb{R}$ can be written as

$$f(x) = \sum_{\ell=1}^{M} g_\ell \left(\min_{z \in \mathcal{C}_\ell} \mathrm{m}(x,z)^p \right), \tag{2}$$

where $\mathrm{m}(\cdot,\cdot)$ is a metric and $p \in \mathbb{N}$ can be an arbitrary scalar, which makes $\mathrm{m}(\cdot,\cdot)^p$ efficiently approximable by deep neural networks (think of $\mathrm{m}(\cdot,\cdot)^p = \|\cdot - \cdot\|_p^p$, which is a polynomial of degree p in the coordinates and thus efficiently approximable, see Lemma A.2). For functions satisfying (2), low-dimensionality will be encoded by assuming that packings of $\mathcal{C}_1, \dots, \mathcal{C}_M$ at scale ε with respect to $\mathrm{m}(\cdot,\cdot)$ have cardinality $\mathcal{O}(\varepsilon^{-d})$. This morally says $\mathcal{C}_1, \dots, \mathcal{C}_M$ are d-dimensional submanifolds, though we do not require any regularity about $\mathcal{C}_\ell$ and we also cover the case $d = 0$. The associated function class is referred to as Class 2 below.

Contribution For α-Hölder smooth $g_1, \dots, g_M$, we show that functions of type (2) can be uniformly approximated to accuracy ε with ReLU nets based on $\mathcal{O}(\varepsilon^{-\alpha})$ queries from each $g_1, \dots, g_M$ and with $\mathcal{O}\left(\log(\varepsilon^{-1})\varepsilon^{-\min\{1,d\}/\alpha} + \varepsilon^{-d/\alpha} P_{\mathrm{m}}(\varepsilon^{1/\alpha})\right)$ nonzero network parameters. Here, $P_{\mathrm{m}}(\varepsilon)$ describes the number of nonzero parameters required to uniformly approximate $\mathrm{m}(\cdot,\cdot)^p$ to accuracy ε. If the metric can be efficiently approximated by a deep net, e.g., by bounding $P_{\mathrm{m}}(\varepsilon) \in \mathcal{O}(D\log(D)\log(\varepsilon^{-1}))$ such as in the case $\mathrm{m}(\cdot,\cdot)^p = \|\cdot - \cdot\|_p^p$, we require in total $\mathcal{O}(D\log(D)\varepsilon^{-\min\{1,d\}/\alpha})$ parameters in the network. For $d \le 1$, which corresponds to the situation in Fig. 1d, the associated requirement $\mathcal{O}(D\log(D)\log(\varepsilon^{-1})\varepsilon^{-1/\alpha})$ is comparable to approximating a univariate function with a shallow or deep network (Mhaskar, 1996; Yarotsky, 2017, 2018). Similarly, the number of required function queries $\mathcal{O}(\varepsilon^{-\alpha})$ per g_i, $i = 1, \dots, M$, matches the minimal number of queries needed to approximate an arbitrary α-Hölder univariate functions according to nonlinear width theory (DeVore et al., 1989).

1.3. Organization of the paper

Section 2 rigorously introduces functions of type (1) and presents the corresponding approximation guarantee. Section 3 does the same for functions of type (2). Section 4 presents implications of our results to nonparametric estimation problems. Section 5 introduces preparatory material about ReLU calculus and Sections 6 and 7 present the proofs of our main results. We conclude in Section 8. Appendix contains some additional statements and proofs about differential geometry and ReLU approximation theory.

1.4. Notation

For $N \in \mathbb{N}$ we let $[N] := \{1, \dots, N\}$. $\mathrm{cl}(B)$ denotes the closure of a set B and $\mathrm{Im}(M)$ denotes the image of an operator M. $|A|$ denotes the absolute value if $A \in \mathbb{R}$, the length if A is an interval, and the cardinality if A is a finite set. We denote $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$. The ReLU activation function is denoted $(t)_+ = \max\{0, t\}$.

$\|\cdot\|_p$ denotes the standard Euclidean p-norm for vectors and $\|\cdot\|_2$ denotes the spectral norm for matrices. We denote $\mathrm{dist}(z; A) := \inf_{p\in A} \|z - p\|_2$ for $z \in \mathbb{R}^D$ and $A \subset \mathbb{R}^D$. $B_r(x)$ denotes the standard $\|\cdot\|_2$-ball of radius r around x, while $B_{\mathcal{M},r}(v)$ denotes the geodesic ball on a manifold $\mathcal{M}$ of radius r around v. $\|A\|_0$ counts the number of nonzero entries of a matrix A. $L_p(A)$ contains function with finite pth order Lebesgue norm.

We use $A \lesssim B$, respectively, $A \gtrsim B$, if there exists a uniform constant C such that $A \le CB$, respectively $A \ge CB$. Furthermore, we write $A \asymp B$ if $A \lesssim B$ and $A \gtrsim B$.

Finally, we define the ReLU activation function $(t)_+ = 0 \vee t = \max\{0, t\}$ and introduce the following definition of a deep ReLU network.

Definition 1.1 (*Grohs et al., 2019, Definition 2.1*)**.** Let $L \ge 2$ and $N_0, \dots, N_L \in \mathbb{N}_{>0}$. A map $\Phi : \mathbb{R}^{N_0} \to \mathbb{R}^{N_L}$ is called a ReLU network if there exist matrices $A_\ell \in \mathbb{R}^{N_\ell \times N_{\ell-1}}$ and vectors $b_\ell \in \mathbb{R}^{N_\ell}$ for $\ell \in [L]$ so that $\Phi(x) = W_L y_{L-1} + b_L$, where y_ℓ is recursively defined by $y_0 := x$ and

$$y_\ell := (A_\ell y_{\ell-1} + b_\ell)_+ \quad \text{for} \quad \ell \in [L-1].$$

Furthermore, we define $L(\Phi) := L$ as the number of layers, $W(\Phi) := \max_{\ell=0,\dots,L} N_\ell$ as the maximum width, $P(\Phi) := \sum_{\ell=1}^{L} \|A_\ell\|_0 + \|b_\ell\|_0$ as the number of nonzero parameters, and

$$B(\Phi) := \max\{|(b_\ell)_i|, |(A_\ell)_{ij}| : i \in N_\ell, j \in N_{\ell-1}, \ell \in [L]\}$$

as a bound for the absolute value over all parameters.

2. Main result: projection-based target functions

In this section we rigorously introduce projection-based functions as foreshadowed in (1) and we present the corresponding approximation guarantee. Before doing so, we introduce some well-known preparatory concepts from differential geometry. These are also summarized in Table 1.

Preparatory material from differential geometry. Let $\mathcal{M} \subseteq \mathbb{R}^D$ be a nonempty, connected, compact, d-dimensional Riemannian submanifold. A manifold $\mathcal{M}$ has an associated medial axis

$$\mathrm{Med}(\mathcal{M}) := \{x \in \mathbb{R}^D : \exists p \neq q \in \mathcal{M}, \ \|p - x\|_2 = \|q - x\|_2 = \mathrm{dist}(x; \mathcal{M})\}, \tag{3}$$

which contains all points $x \in \mathbb{R}^D$ with set-valued orthogonal projection $\pi_{\mathcal{M}}(x) = \mathrm{argmin}_{z\in\mathcal{M}} \|x - z\|_2$. The local reach (sometimes called local feature size Boissonnat & Ghosh, 2014) is defined by

$$\tau_{\mathcal{M}}(v) := \mathrm{dist}(v; \mathrm{Med}(\mathcal{M})) \tag{4}$$

and describes the minimum distance needed to travel from a point $v \in \mathcal{M}$ to the closure of the medial axis. The smallest local reach $\tau_{\mathcal{M}} := \inf_{v\in\mathcal{M}} \tau_{\mathcal{M}}(v)$ is called reach of $\mathcal{M}$.

Another important concept, which we use in the following, is the geodesic metric. Since compact Riemannian manifolds are geodesically complete by the Hopf–Rinow theorem, there exists a length-minimizing geodesic $\gamma : [t, t'] \to \mathcal{M}$ between any two points $\gamma(t) = v$ and $\gamma(t') = v'$, where the length is defined by $|\gamma| = \int_t^{t'} \|\dot\gamma(s)\|_2 \, ds$. The geodesic metric on $\mathcal{M}$ is defined as

$$d_{\mathcal{M}}(v, v') := \inf\{|\gamma| : \gamma \in \mathcal{C}^1([t, t']), \ \gamma : [t, t'] \to \mathcal{M}, \ \gamma(t) = v, \ \gamma(t') = v'\}. \tag{5}$$

We can extend $d_{\mathcal{M}}$ to tubular regions $T \supseteq \mathcal{M}$ around $\mathcal{M}$ by $d_T(x, x') := d_{\mathcal{M}}(\pi_{\mathcal{M}}(x), \pi_{\mathcal{M}}(x'))$, provided the orthogonal projection $\pi_{\mathcal{M}}$ is uniquely defined for $x, x' \in T$.

Main result. We are now interested in approximating functions of the type $f = g \circ \pi_{\mathcal{M}}$. To state the function class in rigorous terms, we define the set

$$\mathcal{M}(q) := \{x \in \mathbb{R}^D : x = v + u, \ v \in \mathcal{M}, \ u \in \ker(A(v)^\top), \ \|u\|_2 < q\tau_{\mathcal{M}}(v)\}, \tag{6}$$

where the columns of $A(v) \in \mathbb{R}^{D\times d}$ represent an orthonormal basis of the tangent space of $\mathcal{M}$ at v. The set $\mathcal{M}(q)$ represents a tubular region around the manifold $\mathcal{M}$ with local tube radius $q\tau_{\mathcal{M}}(v)$, where $\tau_{\mathcal{M}}(v)$ is the local reach as defined in (4). Since $\tau_{\mathcal{M}}(v) \ge \tau_{\mathcal{M}}$ for all $v \in \mathcal{M}$, $\mathcal{M}(q)$ contains, for instance, the tube of constant radius $q\tau_{\mathcal{M}}$ around $\mathcal{M}$. However, in regions where $\mathcal{M}$ has small curvature, the tube radius may also be significantly larger due to its scaling with the local reach.

The class of projection-based functions is defined as follows.

Table 1
Notations used for different geometrical concepts throughout the paper.

Symbol	Description
$\mathcal{M}$	A connected compact d-dimensional Riemannian submanifold of $\mathbb{R}^D$
d	Dimension of the manifold $\mathcal{M}$
$\pi_{\mathcal{M}}$	Orthogonal projection $\pi_{\mathcal{M}}(x) = \operatorname{argmin}_{z\in\mathcal{M}} \|x-z\|_2$
$\mathrm{Med}(\mathcal{M})$	Medial axis of $\mathcal{M}$, i.e. set with non-unique projections $\pi_{\mathcal{M}}(x)$
$A(v)$	$D \times d$ matrix containing columnwise orthonormal basis for the tangent space $\mathcal{M}$ at v
$\tau_{\mathcal{M}}(v)$	Local reach at $v \in \mathcal{M}$, i.e. distance to travel in $\mathrm{Im}(A(v))^{\perp}$ to reach $\mathrm{Med}(\mathcal{M})$
$\tau_{\mathcal{M}}$	Infimum over all local reaches, throughout assumed positive
$\mathcal{M}(q)$	Tube of radius $q \in [0,1)$ times local reach around $\mathcal{M}$, see (6)
$d_{\mathcal{M}}(v,v')$	Geodesic metric on $\mathcal{M}$
$d_T(v,v')$	Geodesic metric on $\mathcal{M}$ extended to $T \supseteq \mathcal{M}$ by $d_T(x,x') := d_{\mathcal{M}}(\pi_{\mathcal{M}}(x), \pi_{\mathcal{M}}(x'))$
$B_{\mathcal{M},r}(v)$	Geodesic ball of radius r around $v \in \mathcal{M}$
$\mathrm{Vol}(\mathcal{M})$	Volume of the manifold $\mathcal{M}$
$\mathcal{P}(\delta, C, \Delta)$	δ-packing number of a set C with respect to metric Δ

Class 1 The target $f : \mathcal{A} \subseteq [0,1]^D \to \mathbb{R}$ can be written as $f(x) = g(\pi_{\mathcal{M}}(x))$ for a connected, compact, nonempty, d-dimensional manifold $\mathcal{M}$ with $\tau_{\mathcal{M}} > 0$, $\mathcal{A} \subseteq \mathcal{M}(q) \subseteq [0,1]^D$ for some $q \in [0,1)$, and where $\pi_{\mathcal{M}}(x) := \operatorname{argmin}_{z\in\mathcal{M}} \|x-z\|_2$. The function $g : \mathcal{M} \to [0,1]$ is α-Hölder with Hölder constant L, i.e., satisfies for $\alpha \in (0,1]$ and $L \geq 0$

$$\left|g(v) - g(v')\right| \leq L d_{\mathcal{M}}^{\alpha}(v,v') \quad \text{for all} \quad v, v' \in \mathcal{M}. \tag{7}$$

The condition $\mathcal{A} \subseteq \mathcal{M}(q)$ for some $q < 1$ is important because it is a necessary for f to inherit smoothness properties from g. Namely, if $\mathcal{A}$ intersects the medial axis $\mathrm{Med}(\mathcal{M})$, see the definition in (3), the projection $\pi_{\mathcal{M}}$ is not uniquely defined over $\mathcal{A}$ and, as a consequence, f may not be well-defined as well. If $\mathcal{A} \cap \mathrm{Med}(\mathcal{M}) = \emptyset$ but $\mathrm{dist}(\mathcal{A}; \mathrm{Med}(\mathcal{M})) = 0$, f might be well-defined and continuous on $\mathcal{A}$, but we cannot expect f to be locally Hölder-continuous at points arbitrarily close to the medial axis. As shown in the following Lemma, enforcing $\mathcal{A} \subseteq \mathcal{M}(q)$ for some $q < 1$ solves these issues and implies that f inherits α-Hölder regularity of g with a Hölder constant equal to the product of the Hölder constant of g and $(1-q)^{-1}$.

Lemma 2.1. *Consider a connected, compact, d-dimensional Riemannian submanifold of $\mathcal{M} \subseteq \mathbb{R}^D$ with $\tau_{\mathcal{M}} > 0$ and let $q \in [0,1)$.*
(1) If $x \in \mathcal{M}(q)$ has decomposition $x = v + u$ for $v \in \mathcal{M}$ and $u \in \ker(A(v)^{\top})$ with $\|u\|_2 < q\tau_{\mathcal{M}}(v)$, then $\pi_{\mathcal{M}}(x)$ is uniquely determined by $\pi_{\mathcal{M}}(x) = v$.
(2) The projection $\pi_{\mathcal{M}}$ satisfies $\left\|\pi_{\mathcal{M}}(x) - \pi_{\mathcal{M}}(x')\right\|_2 \leq (1-q)^{-1}\left\|x - x'\right\|_2$ for all $x, x' \in \mathcal{M}(q)$.

Proof. The proof is deferred to Appendix A.1 in Appendix. □

We can now present our main approximation guarantee.

Theorem 2.2. *Let f be of Class 1 and define $C_{\mathcal{M}} := \mathrm{Vol}(\mathcal{M})d^{d/2}$, $C_q := C_d d^{d/2}(1-q)^{-2d}$, where C_d is the volume of the Euclidean unit ball in $\mathbb{R}^d$. For $\varepsilon \in (0, \tau_{\mathcal{M}}/2)$ there exists a ReLU network Φ, which uses $n \lesssim C_{\mathcal{M}}\varepsilon^{-d}$ point queries of f and has its dimensions bounded according to $B(\Phi) \lesssim \varepsilon^{-2}$, $W(\Phi) \lesssim DC_{\mathcal{M}}\varepsilon^{-d}$, and*

$$\begin{aligned} L(\Phi) &\lesssim C_q^4 \log^2\left(\frac{C_q}{\varepsilon^{\alpha}}\right) + \log\left(\frac{DC_qC_{\mathcal{M}}}{\tau_{\mathcal{M}}^2\varepsilon^{3+d}}\right), \\ P(\Phi) &\lesssim C_q^4 C_{\mathcal{M}}\varepsilon^{-d}\log^2\left(\frac{C_q}{\varepsilon^{\alpha}}\right) + D\varepsilon^{-d}\log\left(\frac{DC_qC_{\mathcal{M}}}{\tau_{\mathcal{M}}^2\varepsilon^{3+d}}\right), \end{aligned} \tag{8}$$

such that

$$\sup_{x\in\mathcal{A}} |f(x) - \Phi(x)| \lesssim \left(1 + \frac{L}{(1-q)^{2\alpha}}\right)\varepsilon^{\alpha}. \tag{9}$$

Alternatively, with access to $n \gtrsim (\tau_{\mathcal{M}}/2)^d C_{\mathcal{M}}$ point queries of f, we can construct a ReLU network Φ (with dimensions as in (8) and $\varepsilon \asymp (C_{\mathcal{M}}/n)^{1/d}$) that approximates f up to

$$\sup_{x\in\mathcal{A}} |f(x) - \Phi(x)| \lesssim \left(1 + \frac{L}{(1-q)^{2\alpha}}\right)\left(\frac{C_{\mathcal{M}}}{n}\right)^{\frac{\alpha}{d}}.$$

The same construction can be achieved with a network $\tilde{\Phi}$ with $L(\tilde{\Phi}) \lesssim \log(B(\Phi))L(\Phi)$, $W(\tilde{\Phi}) \lesssim (W(\Phi))^2$, $P(\tilde{\Phi}) \lesssim \log(B(\Phi))P(\Phi)$ and $B(\tilde{\Phi}) \leq 2$ according to Grohs et al. (2019, Proposition A.1).

Proof. A proof sketch and full proof details are given in Section 6. □

Theorem 2.2 shows that functions of Class 1 can be uniformly approximated to accuracy ε with a budget of $\mathcal{O}(\varepsilon^{-d/\alpha})$ queries of f and a network with $\mathcal{O}(\log^2(\varepsilon^{-1})\varepsilon^{-d/\alpha})$ nonzero parameters arranged in $\mathcal{O}(\log(\varepsilon^{-1}))$ layers. Since the problem class contains α-Hölder functions on $\mathbb{R}^d$, this result is optimal in terms of the number of needed function queries according to the theory of nonlinear width (DeVore et al., 1989). Moreover, apart from logarithmic factors, it is optimal in terms of the number of nonzero parameters in the network (Yarotsky, 2018, Theorem 1). A bound for the number of nonzero parameters can be used to control covering numbers of the associated ReLU function spaces (Schmidt-Hieber, 2020, Lemma 5). Bounds for covering numbers can then be combined with statistical learning theory to provide estimation guarantees for empirical risk minimization, see the details in Section 4. We also note that $W(\Phi)$ and $P(\Phi)$ have a mild log-linear dependency on the ambient dimension D, which is possibly not avoidable apart from cutting the log-factors.

The constant $C_{\mathcal{M}}$ is intrinsic to $\mathcal{M}$ and arises from bounding the cardinality of an ε-covering of $\mathcal{M}$ as in Lemma 6.1. The constant C_q and the factor $(1-q)^{-1}$ in (9) are extrinsic as they depend on the approximation domain $\mathcal{A}$ via $(1-q)^{-1}$. The factor $(1-q)^{-1}$ indicates that approximating f becomes increasingly challenging as $\mathrm{dist}(\mathcal{A}; \mathrm{Med}(\mathcal{M}))$ shrinks, i.e., as the approximation domain approaches the medial axis, where $\pi_{\mathcal{M}}$ is set-valued and f loses regularity.

The number of needed queries of f and the required dimension of the network in Theorem 2.2 are, apart from log-factors and constants, similar to the case $\mathcal{A} = \mathcal{M}$ and $\pi_{\mathcal{M}} = \mathrm{Id}$ (Nakada & Imaizumi, 2019; Schmidt-Hieber, 2019; Shaham et al., 2018). Hence, previously studied function classes can be significantly extended without compromising on the ability of deep networks to approximate them.

Remark 2.3. 1. Instead of defining $\mathcal{M}$ implicitly by the target f as in Class 1, we can also start with a fixed manifold $\mathcal{M}$, an associated approximation domain $\mathcal{M}(q)$ for $q \in [0,1)$, and ask how well all functions of the type $f(x) = g(\pi_{\mathcal{M}}(x))$ can be approximated over $\mathcal{M}(q)$. Theorem 2.2 applies to this case as

well. Furthermore, we note that all weights except for the last layer are used for the approximation of $\pi_{\mathcal{M}}$ in our construction. Therefore, if we approximate two functions $f(x) = g(\pi_{\mathcal{M}}(x))$ and $\widetilde{f}(x) = \widetilde{g}(\pi_{\mathcal{M}}(x))$ using the proposed construction, the associated networks differ only in the last layer.

2. As the proof in Section 6 will show, there is no significant advantage of the ReLU activation for the construction of the approximating network. Therefore, we believe that similar constructions are realizable with other common activation functions. We focus on the ReLU in this work simply because it is the most prominent choice in practice.

3. The results of Theorem 2.2 are achieved with networks that have duplicate weights, for the sake of an easier analysis. Removing duplicate weights only affects the constant factors in the bounds of Theorem 2.2.

As a corollary of Theorem 2.2, we can also derive an approximation guarantee for $\pi_{\mathcal{M}}$.

Corollary 2.4. *Let $q \in [0, 1)$ and let $\mathcal{M}$ be a nonempty, connected, compact d-dimensional manifold with $\tau_{\mathcal{M}} > 0$ and $\mathcal{M}(q) \subseteq [0, 1]^D$. For $\varepsilon \in (0, \tau_{\mathcal{M}}/2)$ there exists a ReLU network Φ with architecture constrained as in Theorem 2.2 and*

$$\sup_{x \in \mathcal{M}(q)} \|\pi_{\mathcal{M}}(x) - \Phi(x)\|_\infty \lesssim \varepsilon. \tag{10}$$

Proof. The proof is given at the end of Section 6. □

3. Main result: distance-based target functions

We now study distance-based target functions as foreshadowed by Eq. (2). The rigorous definition of the function class requires the well-known concept of packing numbers.

Definition 3.1 (*Vershynin, 2018, Section 4.2*)**.** Let $\mathcal{C}$ be a set endowed with a metric Δ and let $\delta > 0$. We say $\mathcal{Z} \subset \mathcal{C}$ is δ-separated if for any $z \neq z' \in \mathcal{Z}$ we have $\Delta(z, z') > \delta$. $\mathcal{Z}$ is maximal separated if adding any other point in $\mathcal{Z}$ destroys the separability property. The cardinality of the largest maximal separated set is called the packing number and denoted by $\mathcal{P}(\delta, \mathcal{Z}, \Delta)$.

Class 2 Let $\mathcal{C}_1, \dots, \mathcal{C}_M \subseteq [0, 1]^D$ be nonempty closed sets, let $\mathrm{m}(\cdot, \cdot) : [0, 1]^D \to [0, 1]$ be a continuous (normalized) metric, and assume there exists $\delta_0 > 0$ such that $\mathcal{P}(\delta, \mathcal{C}_\ell, \mathrm{m}) \lesssim \delta^{-d}$ for all $\delta < \delta_0$ and $\ell \in [M]$. Furthermore, assume there exists $p > 0$ so that m^p is ReLU-approximable in the sense that, for any fixed $z \in [0, 1]^D$ and $\varepsilon > 0$, there exists a ReLU net $\Psi_{z,\varepsilon}$ with $\sup_{x \in [0,1]^D} \left|\mathrm{m}(x, z)^p - \Psi_{z,\varepsilon}(x)\right| \leq \varepsilon$ and

$$L(\Psi_{z,\varepsilon}) \leq L_{\mathrm{m}}(\varepsilon), \quad W(\Psi_{z,\varepsilon}) \leq W_{\mathrm{m}}(\varepsilon), \quad P(\Psi_{z,\varepsilon}) \leq P_{\mathrm{m}}(\varepsilon), \quad B(\Psi_{z,\varepsilon}) \leq B_{\mathrm{m}}(\varepsilon). \tag{11}$$

We consider functions of the form $f(x) = \sum_{\ell=1}^{M} g_\ell \left(\min_{z \in \mathcal{C}_\ell} \mathrm{m}(x, z)^p\right)$, with $g_\ell : [0, 1] \to [0, 1]$ satisfying for some $\alpha \in (0, 1]$

$$\left|g_\ell(t) - g_\ell(t')\right| \leq L \left|t - t'\right|^\alpha \qquad \text{for all} \qquad t, t' \in [0, 1]. \tag{12}$$

The parameter $p \geq 1$ in Class 2 can be useful for making functions $\mathrm{m}(x, z)^p$ more easily approximable compared to $\mathrm{m}(x, z)$ (think of pth order Euclidean norms raised to the power p, which are degree p polynomials and can be easily approximated as shown in Lemma A.2). Furthermore, (11) should be seen as a definition of $L_{\mathrm{m}}(\varepsilon), W_{\mathrm{m}}(\varepsilon), P_{\mathrm{m}}(\varepsilon), B_{\mathrm{m}}(\varepsilon)$ rather than as an assumption, because it poses almost no restriction on the metric m in view of universal approximation theorems. However, if the approximation of m^p is responsible for an overwhelming majority of the required nonzero parameters in the network construction or scales exponentially in D, the corresponding metric m does not induce an interesting function class in the sense of reducing the original complexity of approximating f. We return to this point after stating the main result by discussing some practically relevant metrics m.

Theorem 3.2. *Let f be a function of Class 2. For any $\varepsilon \in (0, 2p\delta_0)$ there exists a ReLU network Φ, which uses $n \lesssim \varepsilon^{-1}$ point queries from each $g_1, \dots, g_M$ and has its dimensions bounded according to*

$$\begin{aligned} L(\Phi) &\lesssim d \log(p\varepsilon^{-1}) + L_{\mathrm{m}}(\varepsilon), \\ W(\Phi) &\lesssim M p^d \varepsilon^{-(1 \vee d)} W_{\mathrm{m}}(\varepsilon) \\ P(\Phi) &\lesssim M p^d d \log(p\varepsilon^{-1}) \varepsilon^{-(1 \vee d)} + M p^d \varepsilon^{-d} P_{\mathrm{m}}(\varepsilon) \end{aligned}$$

and $B(\Phi) \leq 1 \vee B_{\mathrm{m}}(\varepsilon)$, such that

$$\sup_{x \in [0,1]^D} |f(x) - \Phi(x)| \lesssim ML\varepsilon^\alpha. \tag{13}$$

Alternatively, with access to n point queries from each $g_1, \dots, g_M$, we can construct a ReLU network Φ (with dimensions as above and $\varepsilon \asymp n^{-1}$) that approximates f up to

$$\sup_{x \in \mathcal{A}} |f(x) - \Phi(x)| \lesssim \frac{ML}{n^\alpha}.$$

Proof. The proof is deferred to Section 7. □

As long as $L_{\mathrm{m}}(\varepsilon)$, $W_{\mathrm{m}}(\varepsilon)$, and $P_{\mathrm{m}}(\varepsilon)$ grow at most polylogarithmically in ε^{-1} and possibly polynomially in D, Theorem 3.2 shows that their contribution to the overall network complexity is negligible. Specifically, Theorem 3.2 then implies approximation of f to accuracy ε using $\mathcal{O}(\varepsilon^{-1})$ queries from each $g_1, \dots, g_M$ and $\mathcal{O}(\mathrm{polylog}(\varepsilon^{-1})\varepsilon^{-(1 \vee d)})$ nonzero parameters arranged in $\mathcal{O}(\mathrm{polylog}(\varepsilon^{-1}))$ layers. If $M = 1$, querying g_1 is similar to querying f and the result is optimal according to the theory of nonlinear width (DeVore et al., 1989). Moreover, if $d \leq 1$ and if we neglect logarithmic factors, the number of required nonzero parameters is optimal among all networks whose depth grows at most logarithmically in ε^{-1} (Yarotsky, 2018, Theorem 1). We remark that 2. and 3. of Remark 2.3 about the importance of the ReLU activation and the use of weight duplication apply to Theorem 3.2 as well.

Metrics induced by L_p-norms present a practical and versatile instance of metrics that can be efficiently approximated by deep networks. Specifically, we require $\mathcal{O}(D \log(D/\varepsilon))$ nonzero parameters, arranged in $\mathcal{O}(\log(D/\varepsilon))$ layers as shown in Lemma A.2, so that the overall number of nonzero parameters of the approximating network equals $\mathcal{O}(D \log(D/\varepsilon)\varepsilon^{-(1 \vee d)})$ (L_1 and L_∞ are actually exactly realizable with smaller networks, see also Remark A.7). We can also consider variations of L_p-norms, for instance by first transforming inputs through a sparsity inducing basis (e.g., a wavelet transformation operator) and then use an L_1-norm, or by considering weighted sums of multiple L_p-norms, where each L_p-norm measures the discrepancy of two points at different scales. To give a concrete example, we refer to the work Shirdhonkar and Jacobs (2008) and Leeb and Coifman (2016), who approximate the earth movers distance for histograms using a weighted sum of L_1-norms of wavelet coefficients of histogram differences.

We also note that Class 2 contains radial functions with $M = 1$, $d = 0$, and $\mathrm{m}(\cdot, \cdot) = \|\cdot - \cdot\|_2^2$. Chui, Lin, and Zhou (2019) prove

an approximation rate for radial functions similar to ours using smooth activation functions and McCane and Szymanski (2017) show dimension-free but sub-optimal rates for ReLU networks. Interestingly, Chui et al. (2019) also prove that shallow networks cannot achieve dimension-free rates, because they cannot leverage the compositional nature of f.

4. Implications on nonparametric estimation problems

In this section we briefly highlight some implications of our results on nonparametric estimation problems. We will focus on regression problems with X being a random input vector in $\mathbb{R}^D$, $Y = f(X) + \zeta$, and $\mathbb{E}[\zeta|X] = 0$. Furthermore, we assume f is of Class 1 or Class 2 (where the metric m is assumed to be as efficiently approximable as L_p-norms by a deep ReLU net).

Several very recent works (Bauer, Kohler, et al., 2019; Schmidt-Hieber, 2019, 2020) studied the performance of the empirical risk minimizer

$$\hat{\Phi} \in \operatorname*{argmin}_{\hat{\Psi}\in\mathcal{N}_N} \sum_{i=1}^{N} \left(\hat{\Psi}(X_i) - Y_i\right)^2, \tag{14}$$

where the hypothesis space $\mathcal{N}_N$ contains ReLU networks $\hat{\Psi}$ with complexity bounded by $L(\hat{\Psi}) \le L_N$, $W(\hat{\Psi}) \le W_N$, $P(\hat{\Psi}) \le P_N$, $B(\hat{\Psi}) \le B_N$, and L_N, W_N, P_N, B_N depend on the size of the training data $\{(X_i, Y_i) : i \in [N]\}$. The complexity of $\mathcal{N}_N$ can be controlled in terms of L_N, W_N, P_N and B_N (Schmidt-Hieber, 2020, Lemma 5), and a bias–variance tradeoff analysis allows for establishing estimation rates for (14), whenever the approximation error $\inf_{\Psi\in\mathcal{N}_N} \mathbb{E}\,(\Psi(X) - f(X))^2$ can be bounded in terms of L_N, W_N, P_N and B_N.

Following this strategy, Theorems 2.2 and 3.2 can be used to derive the estimation guarantees

$$\mathbb{E}\left(\hat{\Phi}(X) - f(X)\right)^2 \in \begin{cases} \tilde{\mathcal{O}}(N^{-\frac{2\alpha}{2\alpha+d}}), & \text{if f is of Class 1,} \\ \tilde{\mathcal{O}}(N^{-\frac{2\alpha}{2\alpha+(1\vee d)}}), & \text{if f is of Class 2,} \end{cases} \quad \text{as } N \to \infty, \tag{15}$$

where $\tilde{\mathcal{O}}$ absorbs log-factors in N. The corresponding relations between the architectural constraints and the size of the training data N are given by

$$\begin{aligned} L_N \in \tilde{\mathcal{O}}(1),\ P_N \in \tilde{\mathcal{O}}\left(N^{\frac{d}{2\alpha+d}}\right),\ W_N \in \tilde{\mathcal{O}}\left(N^{\frac{d}{2\alpha+d}}\right), \\ B_N \in \tilde{\mathcal{O}}\left(N^{\frac{2}{2\alpha+d}}\right),\ \text{for Class 1,} \\ \text{and}\ \ L_N \in \tilde{\mathcal{O}}(1),\ P_N \in \tilde{\mathcal{O}}\left(N^{\frac{1\vee d}{2\alpha+1\vee d}}\right),\ W_N \in \tilde{\mathcal{O}}\left(N^{\frac{1\vee d}{2\alpha+1\vee d}}\right), \\ B_N \in \tilde{\mathcal{O}}\left(1\right),\ \text{for Class 2.} \end{aligned} \tag{16}$$

The rates in (15) are statistically minimax optimal for Class 1 (even if X is supported exactly on a d-dimensional manifold) and minimax optimal for Class 2 if $d \le 1$ (Stone, 1982).

To the best of our knowledge, the literature does not provide algorithms for estimating f with the rate (15) under the assumptions imposed in Class 1 or Class 2. Focusing on Class 1, a few special cases have been considered in the literature. First, if X is supported exactly on $\mathcal{M}$, classical methods such as k nearest neighbors, piecewise polynomials, or kernel methods achieve the rate (15) (Bickel & Li, 2007; Kpotufe, 2011; Ye & Zhou, 2008). Second, if $\mathcal{M}$ is a linear subspace, methods from sufficient dimension reduction literature combined with traditional estimators achieve (15) under certain reasonable assumptions (Li, 2018; Ma & Zhu, 2013). Third, if $\dim(\mathcal{M}) = 1$ and g is strictly monotone along the manifold, Kereta, Klock, and Naumova (2020) achieve near-optimal rates in the case $\zeta \equiv 0$. Still, none of these approaches achieves (15) in the generality that is considered here, which indicates a gap between the performance of 'traditional estimators' and deep neural networks. We add though that computing the global minimizer (14) within small (polynomial) runtime is not well-understood, because networks $\hat{\Psi}_N \in \mathcal{N}_N$ are underparametrized by the choice $P(\hat{\Psi}_N) \ll N$.

Finally, we note that checking whether a function belongs to Class 1 or Class 2 is challenging in practice, because the input $\{X_i : i \in [N]\}$ does not reveal the compositional nature of $x \mapsto f(x)$ by itself. Instead, the compositional nature is only visible when jointly using $\{(X_i, Y_i) : i \in [N]\}$, for instance by inspecting derivative tensors of the function f. As an example, Hessian matrices of functions that belong to Class 1 have at most d nontrivial eigenvalues at any point $x \in \mathcal{A}$ and the nontrivial eigenspace corresponds to a subspace of the tangent space of $\mathcal{M}$. For functions of Class 2, derivative tensors also tend to have a specific shape, whose precise form depends on the distance m and the parameter M.

5. Preparatory material: a brief primer on ReLU calculus

ReLU calculus refers to a framework for developing ReLU network approximation guarantees based on successively approximating increasingly complex building blocks. Corresponding results have been developed in recent years (Bölcskei et al., 2019; Grohs et al., 2019; Petersen & Voigtlaender, 2018; Yarotsky, 2017, 2018), following the increased popularity of the ReLU activation in practice. This section gives an overview of some of the results, which we use in the remainder. Throughout, deep ReLU networks are defined as stated in Definition 1.1.

The first step towards developing approximation guarantees with ReLU nets is to endow the space of ReLU nets with two basic operations, namely compositions and linear combinations.

Lemma 5.1 (*Composition Grohs et al., 2019, Lemma 2.5*)**.** *Let $\Phi_1 : \mathbb{R}^{N_0} \to \mathbb{R}^{N_{L_1}}$ and $\Phi_2 : \mathbb{R}^{N_{L_1}} \to \mathbb{R}^{N_{L_2}}$ be two ReLU nets. There exists a ReLU net $\Psi : \mathbb{R}^{N_0} \to \mathbb{R}^{N_{L_2}}$ with $\Psi(x) = \Phi_2(\Phi_1(x))$ and $L(\Psi) = L(\Phi_1) + L(\Phi_2)$, $W(\Psi) = \max\{W(\Phi_1), W(\Phi_2), 2N_{L_1}\}$, $P(\Psi) = 2(P(\Phi_1) + P(\Phi_2))$, and $B(\Psi) \le B(\Phi_1) \vee B(\Phi_2)$.*

Lemma 5.2 (*Linear combination Grohs et al., 2019, Lemma 2.7*)**.** *Let $\{\Phi_i : i \in [N]\}$ be a set of ReLU networks with similar input dimension N_0. There exist ReLU networks Ψ_1 and Ψ_2 that realize the maps $\Psi_1(x) = (\alpha_1\Phi_1(x), \dots, \alpha_N\Phi_N(x))$ and $\Psi_2(x) = \sum_{i=1}^N \alpha_i\Phi_i(x)$. For $j \in \{1, 2\}$, they satisfy $L(\Psi_j) = \max_{i\in[N]} L(\Phi_i)$, $W(\Psi_j) \le \sum_{i=1}^N (2 \vee W(\Phi_i))$, $P(\Psi_j) = \sum_{i=1}^N (P(\Phi_i) + W(\Phi_i) + 2(L - L(\Phi_i)) + 1)$, and $B(\Psi_j) \le \max\{1, \max_{i\in[N]} B(\Phi_i) \vee \alpha_i\}$.*

Using compositions of ReLU nets, the next step is to approximate the square function $x \mapsto x^2$, for instance by using the so-called 'saw-tooth function' approximation (Yarotsky, 2017). Then, by using the identity

$$xy = \frac{1}{2}\left(x^2 + y^2 - (x - y)^2\right),$$

one can establish approximation guarantees for arbitrary multiplication and for multivariate polynomials of arbitrary degree. We exemplarily report the results of Grohs et al. (2019) in the next lemma.

Lemma 5.3 (*Grohs et al., 2019, Proposition 3.2, 3.4 and 3.6*)**.** *Let $\varepsilon \in (0, 1/2)$.*
(1) There exists a network with $L(\Phi) \lesssim \log(1/\varepsilon)$, $W(\Phi) = 3$, $P(\Phi) \lesssim \log(1/\varepsilon)$, and $B(\Phi) \le 1$ such that $\sup_{x\in[0,1]} \left|\Phi(x) - x^2\right| \le \varepsilon$.
(2) Let $R \ge 1$. There exists a network Φ with $L(\Phi) \lesssim \log(R/\varepsilon)$, $W(\Phi) \le 5$, $P(\Phi) \lesssim \log(R/\varepsilon)$ and $B(\Phi) \le 1$ so that $\sup_{(x,y)\in[-R,R]^2} |\Phi(x, y) - xy| \le \varepsilon$.

Table 2
Basic ReLU calculus results that are relevant to the manuscript. We use $x \in \mathbb{R}^D$ for vectors and $t \in \mathbb{R}$ for scalars. The approximation accuracy is ε in the respective metric and $\mathcal{O}(\cdot)$ means as $\varepsilon \to 0$. L, W, P, and B denote bounds on depth, width, number of parameters, and coefficient size of the network respectively.

Map	Metric	$L(\Phi)$	$W(\Phi)$	$P(\Phi)$	$B(\Phi)$	Reference
$x \mapsto \|x\|_p^p$	$L_\infty([-R,R]^D)$	$\mathcal{O}(p^2 \log(\lceil R \rceil D/\varepsilon))$	$9D$	$\mathcal{O}(DL(\Phi))$	1	Lemma A.2
$(x,t) \mapsto tx$	$L_\infty([-R,R]^{D+1})$	$\mathcal{O}(\log(R^2/\varepsilon))$	$5D$	$\mathcal{O}(DL(\Phi))$	1	Lemma A.3
$t \mapsto 1/t$	$L_\infty([R^{-1},R])$	$\mathcal{O}(R^4 \log^2(R/\varepsilon))$	9	$\mathcal{O}(L(\Phi))$	1	Lemma A.4
$x \mapsto x/\|x\|_1$	$L_\infty(\{x : R^{-1} \le \|x\|_1 \le R\})$	$\mathcal{O}(R^4 \log^2(R/\varepsilon))$	$\mathcal{O}(D)$	$\mathcal{O}(DL(\Phi))$	1	Lemma A.5
$x \mapsto \min_i x_i$	Exact on $\mathbb{R}^D$	$2\lceil \log_2(D) \rceil$	$3\lceil D/2 \rceil$	$11D\lceil \log_2(D) \rceil$	1	Lemma A.6

(3) Let $m \in \mathbb{N}$, $a \in \mathbb{R}^{m+1}$, $R \ge 1$. There exists a network Φ with $L(\Phi) \lesssim m\log(1/\varepsilon) + m^2\log(R) + m\log(\lceil \|a\|_\infty \rceil)$, $W(\Phi) \le 9$, $P(\Phi) \lesssim L(\Phi)$, $B(\Phi) \le 1$, and $\sup_{x\in[-R,R]} \left|\Phi(x) - \sum_{i=0}^m a_i x^i\right| \le \varepsilon$.

A natural next step is to study the approximation of functions with a certain degree of regularity. This can be done for instance by using local Taylor expansions and by approximating Taylor polynomials and indicator functions through deep networks. As a result, ReLU nets are able to uniformly approximate functions with a certain degree of regularity with an optimal number of function queries and nonzero parameters, see e.g. Schmidt-Hieber (2020) and Yarotsky (2017, 2018). As an example (and since it suffices for our purposes), we present a simplified version of Schmidt-Hieber (2020, Theorem 5) for α-Hölder univariate functions.

Theorem 5.4 (*Simplified Version of Schmidt-Hieber, 2020, Theorem 5*). *Let $L \ge 1$, $\alpha \in (0,1]$ and consider $f : [0,1] \to \mathbb{R}$ with $|f(t) - f(s)| \le L|t-s|^\alpha$ for all $t, s \in [0,1]$. For any $\varepsilon > 0$ there exists a ReLU network Φ that uses $n \lesssim \varepsilon^{-1}$ point queries of f and has complexity bounded by $L(\Phi) \lesssim \log(1/\varepsilon)$, $W(\Phi) \lesssim 1/\varepsilon$, $P(\Phi) \lesssim \log(1/\varepsilon)1/\varepsilon$, $B(\Phi) \le 1$ such that*

$$\sup_{x\in[0,1]^k} |f(x) - \Phi(x)| \le L\varepsilon^\alpha.$$

Proof. With $\alpha \in (0,1]$, Schmidt-Hieber (2020, Theorem 5) gives (using the same notation as in the reference)

$$\sup_{x\in[0,1]^k} |f(x) - \Phi(x)| \lesssim LN2^{-m} + LN^{-\alpha},$$

where N and m effectively describe width and depth of the approximating network Φ. By choosing $N \asymp 1/\varepsilon$ and $m \asymp \log_2(1/\varepsilon^{(1+\alpha)}))$ with suitable universal constants both summands are bounded by $L\varepsilon^\alpha/2$, giving the asserted approximation guarantee. The required network size can be read of from Schmidt-Hieber (2020, Theorem 5) by inserting N and m. For counting the number of required queries of f, we note that Φ approximates a piecewise constant approximation of f based on $\mathcal{O}(\varepsilon^{-1})$ subintervals. □

The aforementioned results present a small subset of existing approximation results for ReLU nets and give an idea how we can gradually approximate maps of increasing complexity. To facilitate the proofs for our results in the next two sections, we require some additional elementary approximations. These are listed in Table 2, with proofs deferred to Appendix A.2 in Appendix.

6. Proof of Theorem 2.2

We first give a proof sketch that outlines the strategy and additionally highlights the main challenges compared to the previously studied case $\mathcal{A} = \mathcal{M}$ and $\pi_\mathcal{M} = \mathrm{Id}$. Afterwards we present the proof details. Throughout we let $C_d, C_\mathcal{M}$ and C_q be the constants defined in Theorem 2.2.

6.1. Proof sketch and comparison with the case $\mathcal{A} = \mathcal{M}$

Our proof strategy shares some similarities with existing proof strategies for the case $\mathcal{A} = \mathcal{M}$, see for instance Nakada and Imaizumi (2019), Schmidt-Hieber (2019) and Shaham et al. (2018), but also differs in some aspects due to additional complications arising from the high-dimensional approximation domain. In both cases, we can start with a maximal separated δ-net $\{z_1, \dots, z_K\}$ of $\mathcal{M}$ (see Definition 3.1), which has cardinality bounded by $K \approx C_\mathcal{M}\delta^{-d}$ according to Lemma 6.1. Then, by defining U_i as geodesic balls $U_i := \{z \in \mathcal{M} : d_\mathcal{M}(z, z_i) \le \delta\}$, the subsets $U_1, \dots, U_K$ cover the manifold $\mathcal{M}$ and the preimages $\pi_\mathcal{M}^{-1}(U_1), \dots, \pi_\mathcal{M}^{-1}(U_K)$ cover the approximation domain $\mathcal{A} \subseteq \mathcal{M}(q)$. Hence, for any partition of unity $\eta_1, \dots, \eta_K$ subject to $\pi_\mathcal{M}^{-1}(U_1), \dots, \pi_\mathcal{M}^{-1}(U_K)$, we can express f by $f(x) = \sum_i f(x)\eta_i(x)$.

Let us now denote the orthoprojector onto the tangent space at $z_i \in \mathcal{M}$ by A_i. If we are in the case $\mathcal{A} = \mathcal{M}$, we naturally have $U_i = \pi_\mathcal{M}^{-1}(U_i) \cap \mathcal{A}$ and the sets U_i are isomorphic to $A_i(U_i)$ (provided $\delta < \tau_\mathcal{M}/2$, i.e., the covering of $\mathcal{M}$ is sufficiently fine Schmidt-Hieber, 2019; Shaham et al., 2018). Therefore, approximating $f\eta_i$ over U_i is morally like approximating a function on a compact subset of $\mathbb{R}^d$ and we can apply results from Mhaskar (1993, 1996) and Yarotsky (2017) to achieve approximation guarantees that depend exponentially on d instead of D. By linear combination of the resulting $C_\mathcal{M}\delta^{-d}$ approximants, we then obtain an approximation to f.

In the case $\mathcal{M} \subset \mathcal{A} \subseteq \mathcal{M}(q)$ the aforementioned strategy unfortunately cannot be used, because each η_i in the partition of unity is supported on a compact subset of $\mathbb{R}^D$, which is not isomorphic to a compact set in $\mathbb{R}^d$. Hence, naively using results from Mhaskar (1993, 1996) and Yarotsky (2017) to approximate an arbitrary partition of unity $\eta_1, \dots, \eta_K$ subject to $\pi_\mathcal{M}^{-1}(U_1), \dots, \pi_\mathcal{M}^{-1}(U_K)$ incurs the curse of dimensionality.

Instead, we will use a finer covering of $\mathcal{M}$ at the scale $\delta \approx \varepsilon$ (as opposed to $\delta \approx \tau_\mathcal{M}$ in the case $\mathcal{A} = \mathcal{M}$) and employ the piecewise constant approximation $f(x) = \sum_i f(x)\eta_i(x) \approx \sum_i g(z_i)\eta_i(x)$. If $\eta_1, \dots, \eta_K$ form a partition of unity with the localization property

$$\sup_{x\in\mathcal{M}(q):\eta_i(x)\neq 0} d_{\mathcal{M}(q)}(x, z_i) \lesssim \varepsilon, \tag{17}$$

the piecewise constant approximation $f(x) \approx \sum_i g(z_i)\eta_i(x)$ is accurate up to $\mathcal{O}(\varepsilon^\alpha)$ for α-Hölder g. We note however that we have to approximate $K \approx C_\mathcal{M}\varepsilon^{-d}$ functions $\eta_1, \dots, \eta_K$ by deep networks, which means that we can allocate at most $\mathcal{O}(\mathrm{polylog}(\varepsilon^{-1}))$ nonzero parameters for each individual approximation to match the overall result achieved in Theorem 2.2. Thus, η_i's have to satisfy (17), while also being approximable by relatively small networks.

Designing such η_i's is main difficulty of the proof. We first derive an auxiliary result to locally approximate the extended geodesic metric $d_{\mathcal{M}(q)}(x, z_i) = d_\mathcal{M}(\pi_\mathcal{M}(x), z_i)$ around z_i by basic features of the input vector x. Namely, Proposition 6.2 shows that, for any $p \in [q, 1)$, we have the local metric equivalence

$$\left\|A(z_i)^\top(x - z_i)\right\|_2 \lesssim d_{\mathcal{M}(q)}(x, z_i) \lesssim \frac{1}{1-p}\left\|A(z_i)^\top(x - z_i)\right\|_2 \tag{18}$$

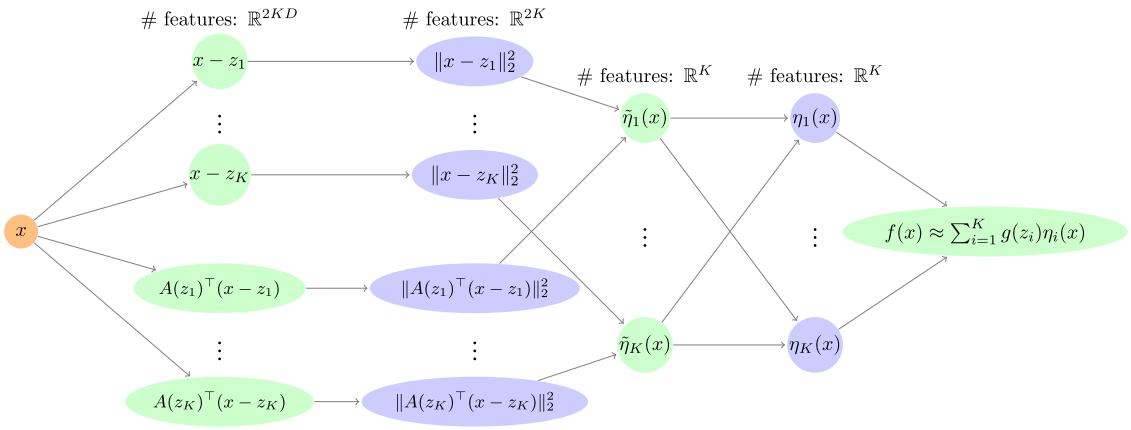

Fig. 2. Schematic ReLU construction used to approximate Class 1. At each node we illustrate the feature of x that is being approximated by the network. Green nodes can be exactly realized (assuming the previous layer is exact) with finite width layers, whereas blue nodes are approximated to accuracy $\mathcal{O}(\varepsilon)$ using $\mathcal{O}(\text{polylog}(\varepsilon^{-1}))$ layers. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

for every x contained in $B_{p\tau_{\mathcal{M}}(z_i)}(z_i)$ and with $\left\|A(z_i)^\top(x - z_i)\right\|_2 \lesssim (1-p)\tau_{\mathcal{M}}$. Intuitively, a point x satisfies $x \in B_{p\tau_{\mathcal{M}}(z_i)}(z_i)$ and $\left\|A(z_i)^\top(x - z_i)\right\|_2 \lesssim (1-p)\tau_{\mathcal{M}}$ if it is contained in an L_2-ball around z_i that extends up to $p\tau_{\mathcal{M}}(z_i)$ in normal direction and $(1-p)\tau_{\mathcal{M}}$ in tangential direction. Eq. (18) implies that we can approximate $d_{\mathcal{M}(q)}(x, z_i)$ on such balls using $\|A(z_i)^\top(x - z_i)\|_2$.

Crucially, $\|x - z_i\|_2$ and $\|A(z_i)^\top(x - z_i)\|_2$ are simple features of the input x, because (after taking squares) they are composed of a linear transformation followed by a polynomial of degree 2 of the input x. Hence, we can approximate these features efficiently using deep networks and construct a partition of unity function accordingly. The precise construction reads

$$\tilde{\eta}_i(x) := \left(1 - \left(\frac{\|x - z_i\|_2}{p\tau_{\mathcal{M}}(z_i)}\right)^2 - \left(\frac{\left\|A(z_i)^\top(x - z_i)\right\|_2}{h\varepsilon}\right)^2\right)_+$$

$$\text{and} \quad \eta_i(x) := \frac{\tilde{\eta}_i(x)}{\|\tilde{\eta}(x)\|_1},$$

where h is a bandwidth parameter that is suitably chosen as a function of q and $\tau_{\mathcal{M}}$. As shown in Proposition 6.3 and Lemma 6.4, η_i satisfies (17) and can be approximated to accuracy ε by a ReLU network Θ_i with $\mathcal{O}(\text{polylog}(\varepsilon^{-1}))$ nonzero parameters.

Finally, after recalling that linear combinations of ReLU networks are still ReLU networks, we approximate f by

$$\Phi(x) = \sum_{i=1}^{K} g(z_i)\Theta_i(x). \tag{19}$$

A schematic illustration of the complete approximating network is depicted in Fig. 2.

6.2. Proof details

Let us first collect some elementary facts from differential geometry that are required in the following.

Lemma 6.1. *Let $\mathcal{M}$ be a d-dimensional compact connected Riemannian manifold embedded in $[0,1]^D$ with reach $\tau_{\mathcal{M}} > 0$, Lebesgue volume Vol($\mathcal{M}$), and endowed with the Riemannian metric induced by $\mathbb{R}^D$. Let $v, z \in \mathcal{M}$.*
(1) If $\|v - z\|_2 \le \tau_{\mathcal{M}}/2$ then $d_{\mathcal{M}}(v,z) \le \tau_{\mathcal{M}}(1 - \sqrt{1 - 2\|v - z\|_2/\tau_{\mathcal{M}}})$.
(2) For any $r \in (0, \tau_{\mathcal{M}}/2)$ we have $Vol(B_{\mathcal{M},r}(v)) \le C_d(\tau_{\mathcal{M}}/(\tau_{\mathcal{M}} - 2r))^d r^d$.
(3) The tangent space orthoprojectors $A(v)A(v)^\top \in \mathbb{R}^{D\times D}$ satisfy perturbation bounds

$$\left\|A(v)A(v)^\top - A(z)A(z)^\top\right\|_2 \le \frac{1}{\tau_{\mathcal{M}}} d_{\mathcal{M}}(v,z). \tag{20}$$

(4) The local reach as defined in (4) satisfies the perturbation bound

$$|\tau_{\mathcal{M}}(v) - \tau_{\mathcal{M}}(z)| \le \|v - z\|_2 \le d_{\mathcal{M}}(v,z). \tag{21}$$

(5) We have $\mathcal{P}(\delta, \mathcal{M}, d_{\mathcal{M}}) \le 3^d Vol(\mathcal{M}) d^{d/2}\delta^{-d}$ for any $\delta \in (0, \frac{1}{2}\tau_{\mathcal{M}})$.
(6) Let $\mathcal{Z}$ be a maximal δ-separated set of $\mathcal{M}$ with respect to the geodesic metric. For any p with $p\delta \in (0, \tau_{\mathcal{M}}/4)$ we have $\left|\mathcal{Z} \cap B_{\mathcal{M},p\delta}(v)\right| \le C_d(5p\sqrt{d})^d$.

Proof. Property (1) can be found in Genovese, Perone-Pacifico, Verdinelli, and Wasserman (2012, Lemma 3) and (2) is derived in Chazal (2013, Proposition 1.1). (3) is similar to Boissonnat, Lieutier, and Wintraecken (2019, Corollary 3), after noticing that

$$\begin{aligned}\left\|A(v)A(v)^\top - A(z)A(z)^\top\right\|_2 &= \sin\angle\,(A(v), A(z)) \\ &\le 2\sin\left(\frac{\angle\,(A(v), A(z))}{2}\right),\end{aligned}$$

where $\angle\,(A(v), A(z))$ denotes the maximum principal angle between subspaces $\text{Im}(A(v))$ and $\text{Im}(A(z))$. For (4) we assume without loss of generality $\tau_{\mathcal{M}}(v) \ge \tau_{\mathcal{M}}(z)$. Then the result follows from

$$\begin{aligned}\tau_{\mathcal{M}}(v) - \tau_{\mathcal{M}}(z) &= \text{dist}(v; \text{Med}(\mathcal{M})) - \text{dist}(z; \text{Med}(\mathcal{M})) \\ &\le \|v - z\|_2 + \text{dist}(z; \text{Med}(\mathcal{M})) \\ &\quad - \text{dist}(z; \text{Med}(\mathcal{M})) = \|v - z\|_2 \le d_{\mathcal{M}}(v,z).\end{aligned}$$

Property (5) can be found in Baraniuk and Wakin (2009) and Niyogi, Smale, and Weinberger (2008). For (6) we first note that $\mathcal{Z} \cap B_{\mathcal{M},p\delta}(v)$ is still a δ-separated set of the geodesic ball $B_{\mathcal{M},p\delta}(v)$, which implies $\left|\mathcal{Z} \cap B_{\mathcal{M},p\delta}(v)\right| \le \mathcal{P}(\delta, B_{\mathcal{M},p\delta}(v), d_{\mathcal{M}})$. Since the reach of the geodesic ball $B_{\mathcal{M},p\delta}(v)$ is also bounded by $\tau_{\mathcal{M}}$, we can apply Property (2) and Property (5) to get

$$\begin{aligned}\mathcal{P}(\delta, B_{\mathcal{M},p\delta}(v), d_{\mathcal{M}}) &\le \frac{3^d \text{Vol}(B_{\mathcal{M},p\delta}(v)) d^{\frac{d}{2}}}{\delta^d} \le \frac{3^d 2^d C_d p^d \delta^d d^{\frac{d}{2}}}{\delta^d} \\ &= C_d(5p\sqrt{d})^d.\end{aligned}$$

□

The first step to prove Theorem 2.2 rigorously establishes the local metric equivalence (18) between the geodesic metric $d_{\mathcal{M}}(\pi_{\mathcal{M}}(x), z)$ and $\left\|A(z)^\top(x - z)\right\|_2$.

Proposition 6.2. *Let $\mathcal{M}$ be a connected compact d-dimensional Riemannian submanifold of $\mathbb{R}^D$ and let $q \in [0,1)$. For $x \in \mathcal{M}(q)$ with $v = \pi_{\mathcal{M}}(x)$ and arbitrary $z \in \mathcal{M}$ we have*

$$\left\|A(z)^\top (x-z)\right\|_2 \leq \left(1 + \frac{\operatorname{dist}(x;\mathcal{M})}{\tau_{\mathcal{M}} \vee (\tau_{\mathcal{M}}(v) - d_{\mathcal{M}}(v,z))}\right) d_{\mathcal{M}}(z,v). \tag{22}$$

Let now $p \in [q,1)$ arbitrary. Then for $x \in B_{p\tau_{\mathcal{M}}(z)}(z)$ with $\left\|A(z)^\top (x-z)\right\|_2 < \frac{1-p}{3}\tau_{\mathcal{M}}$, we have

$$d_{\mathcal{M}}(z,v) \leq \frac{3}{1-p}\left\|A(z)^\top (x-z)\right\|_2. \tag{23}$$

Proof. Throughout the proof we denote $P(z) = A(z)A(z)^\top$ as the orthoprojector onto the tangent space of $\mathcal{M}$ at $z \in \mathcal{M}$. For (22) we use $P(v)(x-v) = 0$ from part (1) of Lemma 2.1, $\|z-v\|_2 \leq d_{\mathcal{M}}(z,v)$, and the tangent perturbation bound (20) applied to the geodesic path $\gamma_{z\to v}$ from z to v with reach bound $\tau_{\gamma_{z\to v}} = \inf_{y\in \operatorname{Im}(\gamma_{z\to v})} \tau_{\mathcal{M}}(y)$ to compute

$$\begin{aligned}\|P(z)(x-z)\|_2 &\leq \|P(z)(v-z)\|_2 + \|P(z)(x-v)\|_2 \leq d_{\mathcal{M}}(v,z)\\ &\quad + \|P(z)-P(v)\|_2 \|x-v\|_2\\ &\leq d_{\mathcal{M}}(v,z) + \frac{\operatorname{dist}(x;\mathcal{M})}{\tau_{\gamma_{z\to v}}} d_{\mathcal{M}}(v,z).\end{aligned}$$

Furthermore, by the 1-Lipschitz property of the local reach, see (21), we have

$$\begin{aligned}\tau_{\gamma_{z\to v}} &= \inf_{y\in\operatorname{Im}(\gamma_{z\to v})} \tau_{\mathcal{M}}(y) \geq \tau_{\mathcal{M}}(v) - \sup_{y\in\operatorname{Im}(\gamma_{z\to v})} |\tau_{\mathcal{M}}(y) - \tau_{\mathcal{M}}(v)|\\ &\geq \tau_{\mathcal{M}}(v) - d_{\mathcal{M}}(v,z).\end{aligned}$$

Since the global bound $\tau_{\gamma_{z\to v}} \geq \tau_{\mathcal{M}}$ holds due to $\operatorname{Im}(\gamma_{z\to v}) \subset \mathcal{M}$, we obtain

$$\|P(z)(x-z)\|_2 \leq \left(1 + \frac{\operatorname{dist}(x;\mathcal{M})}{\tau_{\mathcal{M}} \vee (\tau_{\mathcal{M}}(v) - d_{\mathcal{M}}(v,z))}\right) d_{\mathcal{M}}(v,z).$$

For the opposite direction (23) we let $\omega := \|P(z)(x-z)\|_2$ and $\tilde{x} := z + Q(z)(x-z)$, where $Q(z) := \mathrm{Id} - P(z)$. By construction we have $P(z)(\tilde{x}-z) = 0$ and

$$\left\|x-\tilde{x}\right\|_2 = \|x - z - Q(z)(x-z)\|_2 = \|P(z)(x-z)\|_2 = \omega.$$

Furthermore, since $x \in B_{p\tau_{\mathcal{M}}(z)}(z)$, $\omega < \frac{1-p}{3}\tau_{\mathcal{M}}$, and $\tau_{\mathcal{M}} \leq \tau_{\mathcal{M}}(z)$, we can bound

$$\begin{aligned}\left\|\tilde{x}-z\right\|_2 &\leq \|x-z\|_2 + \left\|x-\tilde{x}\right\|_2 \leq p\tau_{\mathcal{M}}(z) + \omega\\ &< p\tau_{\mathcal{M}}(z) + \frac{1-p}{3}\tau_{\mathcal{M}} < \tilde{p}\tau_{\mathcal{M}}(z),\end{aligned}$$

for $\tilde{p} = \frac{1+2p}{3} < 1$. We thus have the decomposition $\tilde{x} = z + (\tilde{x}-z)$ for $z \in \mathcal{M}$, $\tilde{x}-z \perp \operatorname{Im}(P(z))$, and $\left\|\tilde{x}-z\right\|_2 < \tilde{p}\tau_{\mathcal{M}}(z)$. Part (1) in Lemma 2.1 implies $z = \pi_{\mathcal{M}}(\tilde{x})$ and $\tilde{x} \in \mathcal{M}(\tilde{p})$. Using the Lipschitz property of $\pi_{\mathcal{M}}$ in part (2) of Lemma 2.1 and $x \in \mathcal{M}(q) \subset \mathcal{M}(\tilde{p})$, $\tilde{x} \in \mathcal{M}(\tilde{p})$, we get

$$\|v-z\|_2 = \left\|\pi_{\mathcal{M}}(x) - \pi_{\mathcal{M}}(\tilde{x})\right\|_2 \leq \frac{1}{1-\tilde{p}}\left\|x-\tilde{x}\right\|_2 = \frac{3}{2(1-p)}\omega.$$

We further not that $\frac{3}{2(1-p)}\omega < \frac{1}{2}\tau_{\mathcal{M}}$, so that we can apply part (1) of Lemma 6.1 to get

$$\begin{aligned}d_{\mathcal{M}}(v,z) &\leq \tau_{\mathcal{M}} - \tau_{\mathcal{M}}\sqrt{1 - \frac{2\|v-z\|_2}{\tau_{\mathcal{M}}}} \leq \|v-z\|_2 + \frac{2\|v-z\|_2^2}{\tau_{\mathcal{M}}}\\ &\leq 2\|v-z\|_2. \quad \square\end{aligned}$$

We now introduce the partition of unity functions $\eta_1, \ldots, \eta_K$ and show that they satisfy the desired localization property.

Proposition 6.3. *Consider a connected compact d-dimensional Riemannian submanifold $\mathcal{M} \subseteq \mathbb{R}^D$ and let $q \in [0,1)$. Let $\mathcal{Z} = \{z_1, \ldots, z_{|\mathcal{Z}|}\} \subset \mathcal{M}$ be a maximal δ-separated set of $\mathcal{M}$ with respect to $d_{\mathcal{M}}$. Define bandwidth parameters $p := \frac{1}{2}(1+q)$ and $h := \frac{6}{1-qp^{-1}}$ and functions $\tilde{\eta}, \eta : \mathcal{M}(q) \to \mathbb{R}^{|\mathcal{Z}|}$ componentwise by*

$$\tilde{\eta}_i(x) = \left(1 - \left(\frac{\|x-z_i\|_2}{p\tau_{\mathcal{M}}(z_i)}\right)^2 - \left(\frac{\left\|A(z_i)^\top(x-z_i)\right\|_2}{h\delta}\right)^2\right)_+ \tag{24}$$

$$\text{and} \quad \eta_i(x) = \frac{\tilde{\eta}_i(x)}{\|\tilde{\eta}(x)\|_1}.$$

There exists a universal constant C such that if $\delta \in (0, C(1-q)^2\tau_{\mathcal{M}})$ we have

$$\sup_{x\in\mathcal{M}(q):\eta_i(x)\neq 0} d_{\mathcal{M}(q)}(x,z_i) \lesssim \frac{\delta}{(1-q)^2}, \tag{25}$$

$$(1-q) \lesssim \|\tilde{\eta}(x)\|_1 \lesssim C_q. \tag{26}$$

Proof. Denote $v = \pi_{\mathcal{M}}(x)$. We will a few times require in the following the bandwidth ratio

$$\frac{3h}{1-p} = \frac{36(q+1)}{(1-q)^2} \in \left[\frac{36}{(1-q)^2}, \frac{72}{(1-q)^2}\right).$$

By construction $\eta_i(x) \neq 0$ implies $x \in B_{p\tau_{\mathcal{M}}(z_i)}(z_i)$ and $\|A(z_i)^\top(x-z_i)\|_2 < h\delta$. Thus, as soon as $\delta < \frac{1-p}{3h}\tau_{\mathcal{M}}$, which is implied by $\delta < \frac{1}{36}(1-q)^2\tau_{\mathcal{M}}$, we have $\left\|A(z_i)^\top(x-z_i)\right\|_2 \leq \frac{1-p}{3}\tau_{\mathcal{M}}$. Applying Proposition 6.2 gives (25) by

$$d_{\mathcal{M}(q)}(x,z_i) = d_{\mathcal{M}}(v,z_i) \leq \frac{3h}{1-p}\delta \leq \frac{72}{(1-q)^2}\delta.$$

We now concentrate on the lower bound in (26). Denote $j \in \operatorname{argmin}_{i\in|\mathcal{Z}|} d_{\mathcal{M}(q)}(x,z_i)$. Since $\mathcal{Z}$ is a maximal δ-separated set of $\mathcal{M}$, we have $d_{\mathcal{M}(q)}(x,z_j) \leq \delta$. Eq. (22) in Proposition 6.2 implies

$$\begin{aligned}\left\|A(z_j)^\top(x-z_j)\right\|_2 &\leq \left(1 + \frac{\operatorname{dist}(x;\mathcal{M})}{\tau_{\mathcal{M}}(v) - \delta}\right)\delta \leq \left(1 + \frac{q\tau_{\mathcal{M}}(v)}{\tau_{\mathcal{M}}(v) - \delta}\right)\delta\\ &= \left(1 + q\frac{1}{1 - \frac{\delta}{\tau_{\mathcal{M}}(v)}}\right)\delta \leq (1+2q)\delta \leq 3\delta,\\ &\text{provided } \delta < \frac{1}{2}\tau_{\mathcal{M}}.\end{aligned}$$

Using the triangle inequality to get $\left\|x-z_j\right\|_2 \leq \delta + \|x-v\|_2$ and the 1-Lipschitz continuity of $\tau_{\mathcal{M}}(\cdot)$ in (21) to further bound $\|x-v\|_2 \leq q\tau_{\mathcal{M}}(v) \leq q(\delta + \tau_{\mathcal{M}}(z_j))$, it follows that

$$\frac{\left\|x-z_j\right\|_2}{p\tau_{\mathcal{M}}(z_j)} \leq \frac{\delta + q\delta + q\tau_{\mathcal{M}}(z_j)}{p\tau_{\mathcal{M}}(z_j)} \leq \frac{q}{p} + \frac{1+q}{p\tau_{\mathcal{M}}}\delta \leq \frac{q}{p} + \frac{4}{\tau_{\mathcal{M}}}\delta$$

Inserting the definition of the bandwidth parameter h, we thus obtain

$$\begin{aligned}1 - \frac{\left\|x-z_j\right\|_2}{p\tau_{\mathcal{M}}(z_j)} - \frac{\left\|A(z_j)^\top(x-z_j)\right\|_2}{h\delta} &\geq 1 - \frac{q}{p} - \frac{4}{\tau_{\mathcal{M}}}\delta - \frac{3}{h}\\ &\geq \frac{1}{2}\left(1 - \frac{q}{p}\right) - \frac{4}{\tau_{\mathcal{M}}}\delta.\end{aligned} \tag{27}$$

This is bounded from below by $\frac{1}{4}(1-qp^{-1})$ as soon as

$$\delta < \frac{\tau_{\mathcal{M}}}{16}\left(1 - \frac{q}{p}\right) = \frac{\tau_{\mathcal{M}}}{16}\frac{1-q}{1+q}, \qquad \text{which is implied by}$$

$$\delta < \frac{(1-q)\tau_{\mathcal{M}}}{16}.$$

Since squaring one of the subtracted terms in (27) reduces their size, we get the lower bound $\|\tilde{\eta}(x)\|_1 \geq \tilde{\eta}_i(x) \geq \frac{1}{4}(1-pq^{-1}) \geq 1/8(1-q)$.

For the upper bound on $\|\tilde{\eta}(x)\|_1$ we notice that $\eta_i(x) \neq 0$ implies by Proposition 6.2

$$h\delta > \left\|A(z_i)^\top (x - z_i)\right\|_2 \geq \frac{1-p}{3} d_{\mathcal{M}(q)}(z_i, x) \qquad \text{provided}$$
$$\delta < \frac{(1-q)^2}{36}\tau_{\mathcal{M}}.$$

Thus, $\tilde{\eta}_i(x) \neq 0$ implies $d_{\mathcal{M}(q)}(x, z_i) \leq 3h(1-p)^{-1}\delta$, i.e., all z_i's contributing to $\|\tilde{\eta}\|_1$ are contained within a geodesic ball of radius $3h(1-p)^{-1}\delta$ around v. As soon as $\frac{3h}{1-p}\delta < \frac{1}{4}\tau_{\mathcal{M}}$, which is implied by $\delta < 288(1-q)^2\tau_{\mathcal{M}}$, we can then use part (5) of Lemma 6.1 to bound

$$\left|\mathcal{Z} \cap B_{\mathcal{M}, \frac{3h}{1-p}\delta}(v)\right| \leq C_d \left(5\frac{3h}{1-p}\sqrt{d}\right)^d \leq C_d \left(5\frac{72}{(1-q)^2}\sqrt{d}\right)^d \lesssim C_q.$$

Since each $\tilde{\eta}_i(x)$ is individually bounded by 1, the upper bound on $\|\tilde{\eta}(x)\|_1$ in (26) follows. □

We next show that η can be uniformly approximated by a ReLU net of small complexity.

Lemma 6.4. *Assume the setting of Proposition 6.3 with $\delta < \tau_{\mathcal{M}}/2$ and $\mathcal{M}(q) \subseteq [0,1]^D$. For all $\varepsilon \in (0,1)$ there exists a ReLU-net Φ with complexity bounded as in (31) such that*

$$\sup_{x \in \mathcal{M}(q)} \|\eta(x) - \Phi(x)\|_1 \leq \varepsilon. \tag{28}$$

Proof. Recall that $\mathcal{Z} = \{z_1, \ldots, z_{|\mathcal{Z}|}\} \subset \mathcal{M}$ is a maximal δ-separated set of $\mathcal{M}$ with respect to $d_{\mathcal{M}}$ and that we have $C_q^{-1} \lesssim \|\tilde{\eta}(x)\|_1 \lesssim C_q$ (see right hand side in (26)). The proof is split into two parts. First, we describe how to approximate $\tilde{\eta}_i$ for some $i \in [|\mathcal{Z}|]$, and afterwards we describe how to combine the networks to approximate $\eta : \mathbb{R}^D \to \mathbb{R}^{|\mathcal{Z}|}$.

1. Approximating $\tilde{\eta}_i$: Let Θ be a ReLU net that approximates $\|\cdot\|_2^2$ over $[-1,1]^D$ to accuracy $\tilde{\varepsilon} > 0$ (existence is proven in Lemma A.2). Furthermore, let Ψ_i realize $x \mapsto x - z_i$, and Γ_i realize $x \mapsto A(z_i)^\top(x - z_i)$. For bandwidth parameters p and h as in Proposition 6.3, we then define a ReLU network

$$\tilde{\Phi}_i(x) := \left(1 - \frac{\Theta(\Psi_i(x))}{(p\tau_{\mathcal{M}}(z_i))^2} - \frac{\Theta(\Gamma_i(x))}{(h\delta)^2}\right)_+.$$

Comparing $\tilde{\Phi}_i$ with $\tilde{\eta}_i$ we obtain by 1-Lipschitzness of the ReLU and the triangle inequality

$$\begin{aligned}\sup_{x \in \mathcal{M}(q)} \left|\tilde{\Phi}_i(x) - \tilde{\eta}_i(x)\right| &\leq \left|\frac{\Theta(\Psi_i(x)) - \|x - z_i\|_2^2}{(p\tau_{\mathcal{M}}(z_i))^2}\right| \\ &\quad + \left|\frac{\Theta(\Gamma_i(x)) - \left\|A(z_i)^\top(x - z_i)\right\|_2^2}{(h\delta)^2}\right| \\ &\leq \frac{\tilde{\varepsilon}}{(p\tau_{\mathcal{M}}(z_i))^2} + \frac{\tilde{\varepsilon}}{(h\delta)^2} \\ &\leq \left(\frac{1}{(p\tau_{\mathcal{M}})^2} + \frac{1}{(h\delta)^2}\right) \leq \frac{5\tilde{\varepsilon}}{(\tau_{\mathcal{M}}\delta)^2}\end{aligned} \tag{29}$$

where we used $x - z_i \in [-1,1]^D$ since $x, z_i \in [0,1]^D$, $p \geq 1/2$, and $h > 1$. To compute the complexity of $\tilde{\Phi}_i$ we apply the rules of ReLU composition and linear combination in Lemmas 5.1 and 5.2, and the complexity bounds in Lemma A.2. We have $L(\Theta \circ \Psi_i) \leq L(\Theta) + L(\Psi_i) \lesssim \log(D/\tilde{\varepsilon})$, $W(\Theta \circ \Psi_i) \lesssim D$, $P(\Theta \circ \Psi_i) \lesssim D\log(D/\tilde{\varepsilon})$, and $B(\Theta \circ \Psi_i) \leq 1$, and the same bounds hold for $\Theta \circ \Gamma_i$. Thus, by the rules of ReLU linear combination in Lemma 5.2 (the additional ReLU activation in the last layer does not matter for the absolute bounds) we have

$$L(\tilde{\Phi}_i) \lesssim \log(D/\varepsilon^{-1}), \quad W(\tilde{\Phi}_i) \lesssim D, \quad P(\tilde{\Phi}_i) \lesssim D\log(D\tilde{\varepsilon}^{-1}),$$
$$B(\tilde{\Phi}_i) \leq 1/(p\tau_{\mathcal{M}})^2 \vee 1/(h\delta)^2.$$

2. Approximating η: Define now $\tilde{\Phi}(x) = (\tilde{\Phi}_1(x), \ldots, \tilde{\Phi}_{|\mathcal{Z}|}(x))$. Using (29) we note that

$$\begin{aligned}\left|\|\tilde{\Phi}(x)\|_1 - \|\tilde{\eta}(x)\|_1\right| &\leq \|\tilde{\Phi}(x) - \tilde{\eta}(x)\|_1 \\ &\leq \sum_{i=1}^{|\mathcal{Z}|} |\tilde{\Phi}_i(x) - \eta_i(x)| \leq \frac{5\,|\mathcal{Z}|\,\tilde{\varepsilon}}{(\tau_{\mathcal{M}}\delta)^2}\end{aligned} \tag{30}$$

Thus, with $C_q^{-1} \leq \|\tilde{\eta}(x)\|_1 \leq C_q$ we get $1/2C_q^{-1} \leq \|\tilde{\Phi}(x)\|_1 \leq 2C_q$ for $\tilde{\varepsilon} \leq (\tau_{\mathcal{M}}\delta)^2/(10C_q\,|\mathcal{Z}|)$. Now, let Λ be a network that approximates ℓ_1-normalization up to $\varepsilon/2$ for inputs u with $(2C_q)^{-1} \leq \|u\|_1 \leq 2C_q$ as in Lemma A.5. Setting $\Phi(x) := \Lambda(\Phi(x))$, the approximation error be decomposed into

$$\begin{aligned}\|\Phi(x) - \eta(x)\|_1 &= \left\|\Lambda(\tilde{\Phi}(x)) - \frac{\tilde{\Phi}(x)}{\|\tilde{\Phi}(x)\|_1}\right\|_1 + \left\|\frac{\tilde{\Phi}(x)}{\|\tilde{\Phi}(x)\|_1} - \eta(x)\right\|_1 \\ &\leq \frac{\varepsilon}{2} + \left\|\frac{\tilde{\Phi}(x)}{\|\tilde{\Phi}(x)\|_1} - \frac{\eta(x)}{\|\tilde{\eta}(x)\|_1}\right\|_1.\end{aligned}$$

For the second term, by twice applying triangle inequalities and reusing (30), we obtain

$$\begin{aligned}\left\|\frac{\tilde{\Phi}(x)}{\|\tilde{\Phi}(x)\|_1} - \frac{\tilde{\eta}(x)}{\|\tilde{\eta}(x)\|_1}\right\|_1 &\leq \frac{\|\tilde{\Phi}(x) - \tilde{\eta}(x)\|_1}{\|\tilde{\eta}(x)\|_1} \\ &\quad + \frac{|\|\tilde{\Phi}(x)\|_1 - \|\tilde{\eta}(x)\|_1|}{\|\tilde{\eta}(x)\|_1} \\ &\leq C_q \frac{10\,|\mathcal{Z}|\,\tilde{\varepsilon}}{(\tau_{\mathcal{M}}\delta)^2}.\end{aligned}$$

Combining both bounds yields and setting $\tilde{\varepsilon} = (\tau_{\mathcal{M}}\delta)^2\varepsilon/(20C_q\,|\mathcal{Z}|)$ yields the result.

Lastly, we bound the complexity of Φ. Following the rules of ReLU compositions and combinations in Lemmas 5.1 and 5.2, and using the cardinality bound $|\mathcal{Z}| \lesssim C_{\mathcal{M}}\delta^{-d}$ as in Lemma 6.1, we have

$$\begin{aligned}L(\Phi) &= L(\Lambda) + L(\tilde{\Phi}_1) \lesssim C_q^4 \log^2(C_q/\varepsilon) + \log\left(\frac{DC_q\,|\mathcal{Z}|}{(\tau_{\mathcal{M}}\delta)^2\varepsilon}\right) \\ &\leq C_q^4 \log^2\left(\frac{C_q}{\varepsilon}\right) + \log\left(\frac{DC_qC_{\mathcal{M}}}{\tau_{\mathcal{M}}^2\delta^{d+2}\varepsilon}\right) \\ W(\Phi) &\lesssim |\mathcal{Z}|\,W(\tilde{\Phi}_1) \lesssim DC_{\mathcal{M}}\delta^{-d}, \\ P(\Phi) &\lesssim P(\Lambda) + |\mathcal{Z}|\,P(\tilde{\Phi}_1) \lesssim C_q^4\,|\mathcal{Z}|\log^2\left(\frac{C_q}{\varepsilon}\right) \\ &\quad + |\mathcal{Z}|\,D\log\left(\frac{DC_q\,|\mathcal{Z}|}{\tau_{\mathcal{M}}^2\delta^2\varepsilon}\right) \\ &\lesssim C_q^4 C_{\mathcal{M}}\delta^{-d}\log^2\left(\frac{C_q}{\varepsilon}\right) + D\delta^{-d}\log\left(\frac{DC_qC_{\mathcal{M}}}{\tau_{\mathcal{M}}^2\delta^{2+d}\varepsilon}\right), \\ B(\Phi) &= B(\tilde{\Phi}_1) \lesssim \tau_{\mathcal{M}}^{-2} \vee \delta^{-2}. \quad \square\end{aligned} \tag{31}$$

Finally we combine Lemma 6.4 with the α-Hölder property of g to conclude the proof.

Proof of Theorem 2.2. Let $\mathcal{Z} := \{z_1, \ldots, z_K\}$ be a maximal separated ε-net of $\mathcal{M}$ with $K := |\mathcal{Z}| \lesssim C_{\mathcal{M}}\varepsilon^{-d}$ by Lemma 6.1 and let $g(Z) = (g(z_1), \ldots, g(z_K)) \in \mathbb{R}^K$. By Lemma 6.4, we can construct a network $\Theta : \mathbb{R}^D \to \mathbb{R}^K$, which approximates the partition of unity function $\eta(x)$ in (25) over $\mathcal{A} \subseteq \mathcal{M}(q)$ up to accuracy ε^α. To

approximate the target f we define the net

$$\Psi(x) := \sum_{i=1}^{K} (g(z_i)\Theta_i(x))_+ - (-g(z_i)\Theta_i(x))_+ = \langle g(Z), \Theta(x)\rangle. \quad (32)$$

Taking arbitrary $x \in \mathcal{A} \subseteq \mathcal{M}(q)$, we can first use triangle and Hölder inequalities to get

$$\begin{aligned}|f(x) - \Phi(x)| &= |g(\pi_{\mathcal{M}}(x)) - \langle g(Z), \Theta(x)\rangle| \\ &\le |g(\pi_{\mathcal{M}}(x)) - \langle g(Z), \eta(x)\rangle| \\ &\quad + |\langle g(Z), (\eta(x) - \Theta(x))\rangle| \\ &\le |\langle g(\pi_{\mathcal{M}}(x))\mathbb{1}_K - g(Z), \eta(x)\rangle| \\ &\quad + \|g(Z)\|_\infty \|\eta(x) - \Theta(x)\|_1 \\ &\le |\langle g(\pi_{\mathcal{M}}(x))\mathbb{1}_K - g(Z), \eta(x)\rangle| + \varepsilon^\alpha.\end{aligned}$$

where $\mathbb{1}_K = [1, \dots, 1] \in \mathbb{R}^K$ and where we used $\langle \mathbb{1}_K, \eta(x)\rangle = 1$, $\|g(Z)\|_\infty \le 1$. The first term can be bounded by Hölder's inequality and Proposition 6.3 according to

$$\begin{aligned}|\langle g(\pi_{\mathcal{M}}(x))\mathbb{1}_K - g(Z), \eta(x)\rangle| &\le \sup_{\substack{i\in[K]\\ \eta_i(x)\neq 0}} |g(\pi_{\mathcal{M}}(x)) - g(z_i)| \\ &\le L \sup_{\substack{i\in[K]\\ \eta_i(x)\neq 0}} d_{\mathcal{M}}^{\alpha}(\pi_{\mathcal{M}}(x), z_i) \\ &\lesssim L\left(\frac{1}{(1-q)^2}\varepsilon\right)^{\alpha},\end{aligned}$$

which shows the approximation error bound. To bound the complexity of Φ we note that the network is a composition of Θ with a two-layer network that has first layer weights $\pm g(Z) \in [-1, 1]^K$ and second layer weights ± 1. Therefore, the complexity of Φ is dominated by the complexity of Θ and can be read off from Lemma 6.4, respectively, from (31) in the proof. □

Proof of Corollary 2.4. For each $k \in [D]$ we can approximate $x \mapsto e_k^\top \pi_{\mathcal{M}}(x)$ via a ReLU net Φ_k using Theorem 2.2 for $g(\cdot) = e_k^\top(\cdot)$. By stacking these networks, we obtain the approximating network $\Phi(x) = (\Phi_1(x), \dots, \Phi_D(x))$, which achieves the asserted guarantee. Note that by construction (19) we have $\Phi_k(x) = \sum_{i=1}^K (z_i)_k \Theta_i(x)$. Since Θ_i is independent of k, corresponding weights can be shared in constructing Φ, so that the network architecture adheres to the bounds in Theorem 2.2. □

7. Proof of Theorem 3.2

Theorem 3.2 follows by the ReLU composition rule after approximating g and $x \mapsto \min_{z\in\mathcal{C}} \mathrm{m}(x, z)^p$. Let us separately prove the latter result now and then give the proof of Theorem 3.2.

Lemma 7.1. *Let $\mathcal{C} \subseteq [0, 1]^D$ be nonempty and closed, $\mathrm{m} : [0, 1]^D \times [0, 1]^D \to [0, 1]$ a metric satisfying (11) for some $p \ge 1$, and assume there exists $\delta_0 > 0$ so that $\mathcal{P}(\delta, \mathcal{C}, \mathrm{m}) \lesssim \delta^{-d}$ for all $\delta \in (0, \delta_0)$. For any $\varepsilon \in (0, 2p\delta_0)$ there exists a ReLU network Φ with $L(\Phi) \lesssim d\log(p\varepsilon^{-1}) + L_{\mathrm{m}}(\varepsilon)$, $W(\Phi) \lesssim (p/\varepsilon)^d W_{\mathrm{m}}(\varepsilon)$, $P(\Phi) \lesssim (p/\varepsilon)^d \left(d\log(p\varepsilon^{-1}) + P_{\mathrm{m}}(\varepsilon)\right)$, and $B(\Phi) \le 1 \vee B_{\mathrm{m}}(\varepsilon)$ satisfying*

$$\sup_{x\in[0,1]^D} \left|\min_{z\in\mathcal{C}} \mathrm{m}(x, z)^p - \Phi(x)\right| \le 2\varepsilon. \quad (33)$$

Proof. Let $\mathcal{Z} \subset \mathcal{C}$ be a maximal separated ε/p-net of $\mathcal{C}$, which has cardinality bounded according to $|\mathcal{Z}| \lesssim (p/\varepsilon)^d$ as soon as $\varepsilon < p\delta_0$. For each $z \in \mathcal{Z}$, let $\Psi_{z_i,\varepsilon}$ be a ReLU network that approximates $\mathrm{m}(x, z_i)^p$ up to accuracy ε and let $\Gamma : \mathbb{R}^{|\mathcal{Z}|} \to \mathbb{R}$ be a network that realizes $\Gamma(u) = \min_{i\in[|\mathcal{Z}|]} u_i$ (see Lemma A.6). We set $\Phi(x) = \Gamma(\Psi_{z_1,\varepsilon}, \dots, \Psi_{z_K,\varepsilon})$. Using the triangle inequality, we decompose

$$\begin{aligned}\left|\min_{z\in\mathcal{C}} \mathrm{m}(x, z)^p - \Phi(x)\right| &\le \left|\min_{z\in\mathcal{Z}} \mathrm{m}(x, z)^p - \Phi(x)\right| \\ &\quad + \left|\min_{z\in\mathcal{Z}} \mathrm{m}(x, z)^p - \min_{z\in\mathcal{C}} \mathrm{m}(x, z)^p\right| \\ &\le \left|\min_{z\in\mathcal{Z}} \mathrm{m}(x, z)^p - \min_{z\in\mathcal{Z}} \Psi_{z,\varepsilon}(x)\right| \\ &\quad + \left|\min_{z\in\mathcal{Z}} \mathrm{m}(x, z)^p - \min_{z\in\mathcal{C}} \mathrm{m}(x, z)^p\right| \quad (34)\end{aligned}$$

For the first term, we immediately have

$$\left|\min_{z\in\mathcal{Z}} \mathrm{m}(x, z)^p - \min_{z\in\mathcal{Z}} \Psi_{z,\varepsilon}(x)\right| \le \max_{z\in\mathcal{Z}} \left|\mathrm{m}(x, z)^p - \Psi_{z,\varepsilon}(x)\right| \le \varepsilon.$$

For the second term in (34) we note that there exists $v(x) \in \mathcal{C}$ satisfying $\mathrm{m}(x, v(x)) = \min_{z\in\mathcal{C}} \mathrm{m}(x, z)$ because $\mathcal{C}$ is closed and nonempty ($v(x)$ does need to be unique). Then, by $|a^p - b^p| \le p(a \vee b)^{(p-1)}|a - b|$, $\mathrm{m}(\cdot, \cdot) \in [0, 1]$, and the inverse triangle inequality we have

$$\begin{aligned}\left|\min_{z\in\mathcal{Z}} \mathrm{m}(x, z)^p - \min_{z\in\mathcal{C}} \mathrm{m}(x, z)^p\right| &= \left|\min_{z\in\mathcal{Z}} \mathrm{m}(x, z)^p - \mathrm{m}(x, v(x))^p\right| \\ &\le p \min_{z\in\mathcal{Z}} \mathrm{m}(z, v(x)) \le \varepsilon,\end{aligned}$$

with the last inequality following by the ε/p covering property of $\mathcal{Z}$. It remains to bound complexity of the network in terms of ε and $\mathcal{Z}$. Using the rules of compositions and linear combinations of networks in Lemmas 5.1 and 5.2 we have

$$\begin{aligned}L(\Phi) &= L(\Gamma) + L((\Psi_{z_1,\varepsilon}, \dots \Psi_{z_K,\varepsilon})) \lesssim \log(|\mathcal{Z}|) + L(\Psi_{z_1,\varepsilon/2}) \\ &\lesssim d\log(p\varepsilon^{-1}) + L_{\mathrm{m}}(\varepsilon) \\ W(\Phi) &= \max\{W(\Gamma), W((\Psi_{z_1,\varepsilon}, \dots \Psi_{z_K,\varepsilon})), 2|\mathcal{Z}|\} \lesssim |\mathcal{Z}|\, W(\Psi_{z_1,\varepsilon}) \\ &\lesssim (p/\varepsilon)^d W_{\mathrm{m}}(\varepsilon), \\ P(\Phi) &\lesssim P(\Gamma) + P((\Psi_{z_1,\varepsilon}, \dots \Psi_{z_K,\varepsilon})) \lesssim P(\Gamma) + |\mathcal{Z}|\, P(\Psi_{z_1,\varepsilon}) \\ &\lesssim |\mathcal{Z}| \log(|\mathcal{Z}|) + |\mathcal{Z}|\, P(\Psi_{z_1,\varepsilon}) \\ &\lesssim (p/\varepsilon)^d \left(d\log(p\varepsilon^{-1}) + P_{\mathrm{m}}(\varepsilon)\right), \\ B(\Phi) &\le B(\Gamma) \vee B((\Psi_{z_1,\varepsilon}, \dots \Psi_{z_K,\varepsilon})) \le 1 \vee B_{\mathrm{m}}(\varepsilon). \quad \square\end{aligned}$$

To prove Theorem 3.2 we now combine Lemma 7.1 with Theorem 5.4 in Section 5, which provides approximation bounds for univariate α-Hölder functions like $g_1, \dots, g_M$.

Proof of Theorem 3.2. Consider the case $M = 1$ first and let $g_1 = g$, $\mathcal{C}_1 = \mathcal{C}$. Let $\Psi : \mathbb{R}^D \to \mathbb{R}$ be the ReLU net approximating $x \mapsto \min_{z\in\mathcal{C}} \mathrm{m}(x, z)^p$ up to accuracy ε according to Lemma 7.1, with $\varepsilon < 2p\delta_0$, and let $\Theta : \mathbb{R} \to \mathbb{R}$ be a ReLU net that realizes $\Theta(t) = 1 \wedge t = 1 - (1 - t)_+$. Furthermore, by Theorem 5.4 there exists a ReLU network Ω that approximates g to accuracy $L\varepsilon^\alpha$ over $[0, 1]$. We define the overall approximation by $\Phi(x) := \Omega(\Theta(\Psi(x)))$ and compute

$$\begin{aligned}\left|g\left(\min_{z\in\mathcal{C}} \mathrm{m}(x, z)^p\right) - \Phi(x)\right| &= \left|g\left(\min_{z\in\mathcal{C}} \mathrm{m}(x, z)^p\right) - \Omega(\Theta(\Psi(x)))\right| \\ &\le \left|g\left(\min_{z\in\mathcal{C}} \mathrm{m}(x, z)^p\right) - g(\Theta(\Psi(x)))\right| \\ &\quad + |g(\Theta(\Psi(x))) - \Omega(\Theta(\Psi(x)))| \\ &\le \left|g\left(\min_{z\in\mathcal{C}} \mathrm{m}(x, z)^p\right) - g(\Theta(\Psi(x)))\right| + L\varepsilon^\alpha, \quad (35)\end{aligned}$$

where we used $\Theta(\Psi(x)) \in [0, 1]$ by construction and the approximation guarantees about Ω in the last step. For the first term in

(35), we use the α-Hölder property of g to get

$$\begin{aligned}\left|g\left(\min_{z\in C} \mathrm{m}(x,z)^p\right) - g(\Theta(\Psi(x)))\right| &\le L\left|\min_{z\in C} \mathrm{m}(x,z)^p - \Theta(\Psi(x))\right|^\alpha \\ &= L\left|\min_{z\in C} \mathrm{m}(x,z)^p - 1\wedge\Psi(x)\right|^\alpha \\ &\le L\left|\min_{z\in C} \mathrm{m}(x,z)^p - \Psi(x)\right|^\alpha \\ &\le L(2\varepsilon)^\alpha \lesssim L\varepsilon^\alpha,\end{aligned}$$

where the second to last inequality is an equality if $\Psi(x) < 1$, and follows from $\mathrm{m}(x,z) \le 1$ if $\Psi(x) \ge 1$. To bound the complexity of Φ we will use the rules of compositions according to Lemma 5.1. We have $B(\Phi) \le \max\{B(\Omega), B(\Theta), B(\Psi)\} \le 1 \vee B_{\mathrm{m}}(\varepsilon)$ and

$$\begin{aligned}L(\Phi) &\le L(\Omega) + L(\Theta) + L(\Psi) \lesssim \log(\varepsilon^{-1}) + d\log(p\varepsilon^{-1}) + L_{\mathrm{m}}(\varepsilon) \\ &\lesssim d\log(p\varepsilon^{-1}) + L_{\mathrm{m}}(\varepsilon), \\ W(\Phi) &\le \max\{W(\Omega), W(\Theta), W(\Psi)\} \lesssim \varepsilon^{-1} + p^d\varepsilon^{-d}W_{\mathrm{m}}(\varepsilon) \\ &\lesssim p^d\varepsilon^{-(1\vee d)}W_{\mathrm{m}}(\varepsilon), \\ P(\Phi) &\lesssim P(\Omega) + P(\Theta) + P(\Psi) \lesssim \log(\varepsilon^{-1})\varepsilon^{-1} \\ &\quad + p^d\varepsilon^{-d}\left(d\log(p\varepsilon^{-1}) + P_{\mathrm{m}}(\varepsilon)\right) \\ &\lesssim p^d d\log(p\varepsilon^{-1})\varepsilon^{-(1\vee d)} + p^d\varepsilon^{-d}P_{\mathrm{m}}(\varepsilon).\end{aligned}$$

For the case $M > 1$ we construct networks Φ_ℓ approximating $g_\ell(\mathrm{m}(x, C_\ell))$ to accuracy $L\varepsilon^\alpha$ each, and then use $x \mapsto \sum_{i=1}^M \Phi_\ell(x)$, which can be realized by a ReLU net according to Lemma 5.2. The error follows from the triangle inequality and the dimensions can be deduced from Lemma 5.2. □

8. Conclusion and future directions

In this work we study the uniform approximation of certain compositional functions by deep ReLU networks. The considered function classes are motivated by practical examples and generalize some frequently studied function classes, including functions defined on low-dimensional domains. We have proven uniform approximation guarantees with moderately deep networks, a near-optimal dependency on the number of nonzero network parameters, and optimal dependency on the number of required function queries. Our results suggest that local invariances encoded in the mapping $x \mapsto f(x)$ drive the approximation complexity rather than the complexity of the domain of the target.

We plan to extend our guarantees to projection-based functions $f(x) = g(\tilde{\pi}_{\mathcal{M}}(x))$ using projections $\tilde{\pi}_{\mathcal{M}}(x) = \mathrm{argmin}_{x\in\mathcal{M}} d(x,z)$ based on other metrics d and less regular sets $\mathcal{M}$. This allows for considering more general nonlinear reduction maps ϕ and thus further enhances our knowledge about the adaptivity of deep networks. Furthermore we plan to study the influence of the domain of the target (or more practically a given data set) on the training process of deep networks. While approximability is not crucially dependent on the data domain according to our results, training deep networks via backpropagation may still be affected by the domain of the data.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We thank the reviewers for their anonymous comments, which significantly improved earlier versions of this manuscript. AC is supported by National Science Foundation (Department of Mathematical Sciences) grants 1819222 and 2012266, and by Russell Sage Foundation grant 2196.

Appendix

A.1. Proof of Lemma 2.1

For the proof we recall simplified version of Federer (1959, Theorem 4.8) tailored to manifolds.

Theorem A.1 (*Federer, 1959, (6) and (7) in Theorem 4.8*)**.** *Let $\mathcal{M} \subset \mathbb{R}^D$ be a compact submanifold of $\mathbb{R}^D$.*

(1) Let $v \in \mathcal{M}$ and $x \in \mathbb{R}^D$ so that $\sup_{t\ge0}\{\pi_{\mathcal{M}}(v+tx) = v\} \in (0,\infty)$, then $v + rx \notin Int(Med(\mathcal{M})^C)$.

(2) Let $x \in Med(\mathcal{M})^C$ and $\tau_{\mathcal{M}}(\pi_{\mathcal{M}}(x)) > 0$. Then for any $z \in \mathcal{M}$ we have

$$\langle x - \pi_{\mathcal{M}}(x), \pi_{\mathcal{M}}(x) - z\rangle \ge -\frac{\|\pi_{\mathcal{M}}(x) - z\|_2^2\,\|x - \pi_{\mathcal{M}}(x)\|}{2\tau_{\mathcal{M}}(\pi_{\mathcal{M}}(x))}.$$

Proof of Lemma 2.1. Part 1: We first note that $\mathrm{dist}(x;\mathcal{M}) \le \|x - v\|_2 < q\tau_{\mathcal{M}}(v) \le \tau_{\mathcal{M}}(v)$, which implies $x \notin \mathrm{Med}(\mathcal{M})$, and thus there exists a unique projection $\pi_{\mathcal{M}}(x)$ according to the construction of $\mathrm{Med}(\mathcal{M})$. To show $\pi_{\mathcal{M}}(x) = v$, we consider a proof by contradiction. Assume $\pi_{\mathcal{M}}(x) \ne v$ and denote

$$l := \sup_{t\ge0}\left\{\pi_{\mathcal{M}}\left(v + t\frac{u}{\|u\|_2}\right) = v\right\}.$$

We have $l > 0$, since $u \perp \mathrm{Im}(A(v))$ and $\tau_{\mathcal{M}} > 0$ (see for instance Niyogi et al., 2008, Section 4), and $l < q\tau_{\mathcal{M}}(v)$, since $\pi_{\mathcal{M}}(x) \ne v$. By part (1) in Theorem A.1 we get $w := v + l\frac{u}{\|u\|_2} \notin \mathrm{Int}(\mathrm{Med}(\mathcal{M})^C)$. Therefore, for any $\varepsilon > 0$ there exists, with $B_\varepsilon(w)$ being a Euclidean ball of radius ε around w,

$$y \in B_\varepsilon(w) \cap \left(\mathrm{Int}(\mathrm{Med}(\mathcal{M})^C)\right)^C = B_\varepsilon(w) \cap \mathrm{cl}(\mathrm{Med}(\mathcal{M})).$$

Using the existence of such a y for every $\varepsilon > 0$, we get

$$\begin{aligned}\tau_{\mathcal{M}}(v) &\le \|v - y\|_2 \le \|v - w\|_2 + \|w - y\|_2 \\ &\le \|v - x\|_2 + \|w - y\|_2 < q\tau_{\mathcal{M}}(v) + \varepsilon.\end{aligned}$$

Letting $\varepsilon \to 0$ and recalling $q < 1$, this is a false statement.
Part 2: Using part (2) of Theorem A.1, we have for any $x \in \mathcal{M}(q)$ and $v \in \mathcal{M}$

$$\langle x - \pi_{\mathcal{M}}(x), \pi_{\mathcal{M}}(x) - v\rangle \ge -\frac{\|\pi_{\mathcal{M}}(x) - v\|_2^2\,\|x - \pi_{\mathcal{M}}(x)\|}{2\tau_{\mathcal{M}}(\pi_{\mathcal{M}}(x))}. \quad (36)$$

Taking arbitrary $x, x' \in \mathcal{M}(q)$ we obtain by the Cauchy–Schwartz inequality and (36)

$$\begin{aligned}\left\|x - x'\right\|_2\left\|\pi_{\mathcal{M}}(x) - \pi_{\mathcal{M}}(x')\right\|_2 &\ge \langle x - x', \pi_{\mathcal{M}}(x) - \pi_{\mathcal{M}}(x')\rangle \\ &= \langle x - \pi_{\mathcal{M}}(x) + \pi_{\mathcal{M}}(x) - \pi_{\mathcal{M}}(x') + \pi_{\mathcal{M}}(x') - x', \pi_{\mathcal{M}}(x) \\ &\quad - \pi_{\mathcal{M}}(x')\rangle \\ &\ge \left\|\pi_{\mathcal{M}}(x) - \pi_{\mathcal{M}}(x')\right\|_2^2 \\ &\quad \times \left(1 - \frac{1}{2}\frac{\|x - \pi_{\mathcal{M}}(x)\|_2}{\tau_{\mathcal{M}}(\pi_{\mathcal{M}}(x))} - \frac{1}{2}\frac{\left\|x' - \pi_{\mathcal{M}}(x')\right\|_2}{\tau_{\mathcal{M}}(\pi_{\mathcal{M}}(x'))}\right) \\ &= \left\|\pi_{\mathcal{M}}(x) - \pi_{\mathcal{M}}(x')\right\|_2^2(1 - q),\end{aligned}$$

where we used $x, x' \in \mathcal{M}(q)$ in the last inequality. □

A.2. Additional result from ReLU calculus

In this section we prove the approximation guarantees listed in Table 2.

Lemma A.2 (*pth Power of L_p-norm*). *Let $p \in \mathbb{N}$, $\varepsilon, R > 0$. There exists a ReLU network Φ with $L(\Phi) \lesssim p^2 \log(\lceil R \rceil D/\varepsilon)$, $W(\Phi) \leq 9D$, $P(\Phi) \lesssim Dp^2 \log(\lceil R \rceil D/\varepsilon)$ and $B(\Phi) \leq 1$ such that*

$$\sup_{x \in [-R,R]^D} \left| \|x\|_p^p - \Phi(x) \right| \leq \varepsilon.$$

Furthermore, $\|x\|_1$ can be realized exactly with $L(\Phi) = 2$, $W(\Phi) = 2D$, $P(\Phi) = 4D$ and $B(\Phi) = 1$.

Proof. Following part (3) in Lemma 5.3, there exists a ReLU network Γ that approximates $t \mapsto t^p$ to accuracy $\varepsilon D^{-1} > 0$ on $[-R, R]$. Set $\Phi(x) := \sum_{i=1}^D \Gamma(x_i)$. For arbitrary $x \in [-R, R]^D$ we have

$$\left| \|x\|_p^p - \Phi(x) \right| \leq \sum_{i=1}^D \left| x_i^p - \Theta(x_i) \right| \leq \varepsilon.$$

The complexity of Φ is bounded according to the rules in Lemmas 5.1, 5.2 and the bounds in Lemma 5.3 for the network Γ. We obtain $B(\Phi) \leq \max\{1, B(\Gamma)\} = 1$, $W(\Phi) = D(2 \vee W(\Gamma)) \leq 9D$, and

$$L(\Phi) = L(\Gamma) \lesssim p(\log(D/\varepsilon) + p \log(\lceil R \rceil)) \lesssim p^2 \log(\lceil R \rceil D/\varepsilon)$$
$$P(\Phi) = D(P(\Gamma) + W(\Gamma) + 1) \lesssim DL(\gamma) \leq Dp^2 \log(\lceil R \rceil D/\varepsilon).$$

For $p = 1$ we notice $\|x\|_1 = \sum_{i=1}^D |x_i| = \sum_{i=1}^D (x_i)_+ - (-x_i)_+$, which defines a shallow network with width $2D$ and $4D$ nonzero parameters. □

Lemma A.3 (*Multiplication*). *Let $\varepsilon \in (0, \frac{1}{2})$ and $a > 0$. There exists a ReLU network $\Phi : \mathbb{R}^D \times \mathbb{R} \to \mathbb{R}^D$ with $L(\Phi) \lesssim \log(a^2 \varepsilon^{-1})$, $W(\Phi) \leq 5D$, $P(\Phi) \lesssim D \log(a^2 \varepsilon^{-1})$ and $B(\Phi) \leq 1$ with*

$$\sup_{\|x\|_\infty \leq a,\ |y| \leq a} \|\Phi(x, y) - xy\|_\infty \leq \varepsilon.$$

Proof. By part b) of Lemma 5.3 there exists a ReLU net $\Psi : \mathbb{R}^2 \to \mathbb{R}$ approximating xy up to accuracy ε on $[-a, a]^2$. We set $\Phi(x, y) = (\Psi(x_1, y), \ldots, \Psi(x_D, y))$, which can be realized by a ReLU net (Lemma 5.2). Furthermore, using dimension bounds in Lemma 5.3, we get $L(\Phi) = L(\Psi) \lesssim \log(a^2 \varepsilon^{-1})$, $W(\Phi) \leq DW(\Psi) \leq 5D$, $P(\Phi) = D(P(\Psi) + W(\Psi) + 1) \lesssim D \log(a^2 \varepsilon^{-1})$ and $B(\Phi) \leq 1 \vee B(\Psi) \leq 1$. □

Lemma A.4 (*Division*). *Let $\varepsilon \in (0, 1)$ and $a \in \mathbb{R}_{\geq 1}$. There exists a network $\Phi : \mathbb{R} \to \mathbb{R}$ with $L(\Phi) \lesssim a^4 \log^2(a/\varepsilon)$, $W(\Phi) \leq 9$, $P(\Phi) \lesssim a^4 \log^2(a/\varepsilon)$ and $B(\Phi) \leq 1$, so that*

$$\sup_{t \in \left[\frac{1}{a}, a\right]} \left| \Phi(t) - \frac{1}{t} \right| \leq \varepsilon.$$

Proof. We follow the proof strategy of Telgarsky (2017, Lemma 3.6) but combine it with part c) of Lemma 5.3. Set $c = \frac{1}{a}$ and $r = \lceil a^2 \ln(\frac{2a}{\varepsilon}) \rceil$. First, we notice $t^{-1} = c \sum_{i=1}^\infty (1 - ct)^i$ so cutting the series at $i = r$ results in the approximation error

$$\left| \frac{1}{t} - c \sum_{i=1}^r (1 - ct)^i \right| = \left| c \sum_{i=r+1}^\infty (1 - ct)^i \right| \leq \frac{\varepsilon}{2}.$$

Now let $p(t) = c \sum_{i=1}^r z^i$ so that $p(1 - ct) = c \sum_{i=1}^r (1 - ct)^i$ and notice that $0 \leq 1 - ct \leq 1$ since $t \in [a^{-1}, a]$ and $c = a^{-1}$. Using part c) of Lemma 5.3, we can approximate p over $[0, 1]$ to accuracy $\frac{\varepsilon}{2}$ with a network Ψ adhering to the dimension bounds $L(\Psi) \lesssim r \log(1/\varepsilon) + r^2 + r \log(\lceil c \rceil)$, $W(\Psi) \leq 9$, $P(\Psi) \lesssim L(\Psi)$, and $B(\Psi) \leq 1$. Therefore, we get for any $t \in [a^{-1}, a]$

$$\left| \frac{1}{t} - \Psi(1 - ct) \right| \leq \left| \frac{1}{t} - p(1 - ct) \right| + |p(1 - ct) - \Psi(1 - ct)|$$
$$\leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} \leq \varepsilon.$$

We can simplify the bounds on $L(\Psi)$ and thus $P(\Psi)$ by recognizing that $r^2 \asymp a^4 \log^2(a/\varepsilon)$ dominates the terms $r \log(1/\varepsilon)$ and $r \log(\lceil c \rceil)$. □

Lemma A.5 (L_1-*normalization*). *Let $a \geq 1$, $\varepsilon \in (0, \frac{1}{2})$. There exists a ReLU network $\Phi : \mathbb{R}^D \to \mathbb{R}^D$ with $L(\Phi) \lesssim a^4 \log^2\left(\frac{a}{\varepsilon}\right)$, $W(\Phi) \lesssim D$, $P(\Phi) \lesssim a^4 D \log^2\left(\frac{a}{\varepsilon}\right)$, and $B(\Phi) \leq 1$ such that*

$$\sup_{\frac{1}{a} \leq \|x\|_1 \leq a} \left\| \Phi(x) - \frac{x}{\|x\|_1} \right\|_\infty \leq \varepsilon.$$

Proof. We combine four networks: a network realizing the identity, a network realizing the 1-norm, a network realizing approximate division based on Lemma A.4, and a network realizing approximate multiplication based on Lemma A.3. The identity map $\mathrm{Id}_D : \mathbb{R}^D \to \mathbb{R}^D$ can be realized by a two-layer net $\Psi(x) = (x)_+ - (-x)_+$ and $x \mapsto \|x\|_1$ can be realize by a two-layer ReLU net $\Theta(x) = \sum_{i=1}^D (x_i)_+ + (-x_i)_+$. Furthermore, let Γ denote a ReLU net approximating univariate division on $[a^{-1}, a]$ up to accuracy $\frac{\varepsilon}{2a}$, whose existence has been shown in Lemma A.4, and let Ω denote a ReLU net approximating $(x, y) \mapsto yx$ on $[-2a, 2a]^{D+1}$ to accuracy $\frac{\varepsilon}{2}$. Then we set $\Phi(x) = \Omega(\Psi(x), \Gamma(\Theta(x)))$, which satisfies

$$\left\| \Phi(x) - \frac{x}{\|x\|_1} \right\|_\infty \leq \|\Omega(x, \Gamma(\|x\|_1)) - x\Gamma(\|x\|_1)\|_\infty + \left\| x\Gamma(\|x\|_1) - \frac{x}{\|x\|_1} \right\|_\infty$$
$$\leq \frac{\varepsilon}{2} + \|x\|_\infty \frac{\varepsilon}{2a} \leq \frac{\varepsilon}{2} + \|x\|_1 \frac{\varepsilon}{2a} \leq \varepsilon,$$

where we used $\|x\|_1 \leq a$ in the last inequality. To compute the dimensions of Φ, first note that the composition rules in Lemma 5.1 imply $B(\Gamma \circ \Theta) = B(\Gamma) \vee B(\Theta) \leq 8 \vee a^{-1}$ and

$$L(\Gamma \circ \Theta) = L(\Theta) + L(\Gamma) \lesssim 2 + a^2 \log^2\left(\frac{a}{\varepsilon}\right) \lesssim a^2 \log^2\left(\frac{a}{\varepsilon}\right),$$
$$W(\Gamma \circ \Theta) = \max\{W(\Theta), W(\Gamma), 2\} \leq 2D \vee 16,$$
$$P(\Gamma \circ \Theta) = 2P(\Gamma) + 2P(\Theta) \lesssim a^2 \log^2\left(\frac{a}{\varepsilon}\right) + D.$$

Then, using linear combination and concatenation rules of ReLU nets in Lemmas 5.1, 5.2 we obtain

$$L(\Phi) = L(\Omega) + L((\Psi(x), \Gamma \circ \Theta)) \lesssim \log\left(a^2/\varepsilon\right) + 2 \vee a^4 \log^2(a/\varepsilon)$$
$$\lesssim a^4 \log^2(a/\varepsilon),$$
$$W(\Phi) = W(\Omega) \vee W((\Psi(x), \Gamma \circ \Theta))$$
$$\leq 5D \vee (4 + W(\Psi) + W(\Gamma \circ \Theta)) \lesssim D,$$
$$P(\Phi) \lesssim P(\Omega) + P((\Psi(x), \Gamma \circ \Theta)) \lesssim P(\Omega) + P(\Psi) + P(\Gamma \circ \Theta)$$
$$+ L(\Gamma \circ \Theta) + W(\Psi) + W(\Gamma \circ \Theta)$$
$$\lesssim D \log(a^2 \varepsilon^{-1}) + D + a^4 \log^2\left(\frac{a}{\varepsilon}\right) \lesssim a^4 D \log^2\left(\frac{a}{\varepsilon}\right),$$
$$B(\Phi) = B(\Omega) \vee B((\Psi, \Gamma \circ \Theta))$$
$$\leq \max\{1, B(\Psi), B(\Gamma), B(\Theta)\} \leq 1. \quad \square$$

Lemma A.6. *Let $K \geq 2$. There exists a ReLU network $\Phi_K : \mathbb{R}^K \to \mathbb{R}$ with $L(\Phi_K) \leq 2\lceil \log_2(K) \rceil$, $W(\Phi_K) \leq 3\lceil K/2 \rceil$, $P(\Phi_K) \leq 11K \lceil \log_2(K) \rceil$ and $B(\Phi_K) \leq 1$ such that $\Phi_K(x) = \min_{i \in [K]} x_i$.*

Proof. Without loss of generality we assume K is even as we can otherwise just replace x by repeating one of its arguments without changing the bounds on the dimension of the network. We proof the statement by induction. For $K = 2$ define a network

$$\Phi_2(x) = (x_1)_+ - (-x_1)_+ - (x_1 - x_2)_+ = x_1 - (x_1 - x_2)_+ = x_1 \wedge x_2.$$

Clearly, $L(\Phi_2) = 2$, $W(\Phi_2) = 3$, $P(\Phi_2) = 7$, and $B(\Phi_2) = 1$, which proves the induction start. For the induction step $(K - 1) \to K$ we assume the statement holds up to $K - 1$ and we set $\Phi_K = \Phi_{\frac{K}{2}}(\Phi_2(x_1, x_2), \ldots, \Phi_2(x_{K-1}, x_K))$, which realizes $\min_{x\in[K]} x_i$. To compute the network complexity we use composition and parallelization rules from Lemmas 5.1, 5.2. This gives $B(\Phi_K) \leq 1$ and

$$\begin{aligned}
L(\Phi_K) &= L\left(\Phi_{\frac{K}{2}}\right) + L(\Phi_2) = 2\left\lceil \log_2\left(\frac{K}{2}\right)\right\rceil + 2 \\
&= 2\left\lceil \log_2(K) - 1\right\rceil + 2 = 2\left\lceil \log_2(K)\right\rceil, \\
W(\Phi_K) &= \max\left\{W(\Phi_{\frac{K}{2}}, W(\Phi_2, \ldots, \Phi_2), K)\right\} \leq \frac{K}{2} W(\Phi_2) \leq 3\frac{K}{2}, \\
P(\Phi_K) &= 2P\left(\Phi_{\frac{K}{2}}\right) + 2P(\Phi_2, \ldots, \Phi_2) \\
&\leq 11K\left\lceil \log_2\left(\frac{K}{2}\right)\right\rceil + K(P(\Phi_2) + W(\Phi_2) + 1) \\
&\leq 11K\left\lceil \log_2(K)\right\rceil - 11K + 11K. \quad \square
\end{aligned}$$

Remark A.7. The L_∞-norm can be realized by a ReLU net due to Lemma A.6 and the identity

$$\|x\|_\infty = \max_{i\in[D]} |x_i| = \max_{i\in[D]} (x_i)_+ + (-x_i)_+ = -\min_{i\in[D]} -((x_i)_+ + (-x_i)_+).$$

References

Baraniuk, R. G., & Wakin, M. B. (2009). Random projections of smooth manifolds. *Foundations of Computational Mathematics*, *9*(1), 51–77.

Barron, A. R. (1993). Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, *39*(3), 930–945.

Barron, A. R. (1994). Approximation and estimation bounds for artificial neural networks. *Machine Learning*, *14*(1), 115–133.

Bauer, B., Kohler, M., et al. (2019). On deep learning as a remedy for the curse of dimensionality in nonparametric regression. *The Annals of Statistics*, *47*(4), 2261–2285.

Bickel, P. J., & Li, B. (2007). Local polynomial regression on unknown manifolds. In *Complex datasets and inverse problems* (pp. 177–186). Institute of Mathematical Statistics.

Boissonnat, J.-D., & Ghosh, A. (2014). Manifold reconstruction using tangential delaunay complexes. *Discrete & Computational Geometry*, *51*(1), 221–267.

Boissonnat, J.-D., Lieutier, A., & Wintraecken, M. (2019). The reach, metric distortion, geodesic convexity and the variation of tangent spaces. *Journal of Applied and Computational Topology*, *3*(1–2), 29–58.

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

Chazal, F. (2013). *An upper bound for the volume of geodesic balls in submanifolds of euclidean spaces*: *Technical note*, available at http://geometrica.saclay.inria.fr/team/Fred.Chazal/BallVolume2013.pdf.

Chen, M., Jiang, H., Liao, W., & Zhao, T. (2019). Efficient approximation of deep relu networks for functions on low dimensional manifolds. In *Advances in neural information processing systems* (pp. 8172–8182).

Cheng, X., & Cloninger, A. (2019). Classification logit two-sample testing by neural networks. arXiv preprint arXiv:1909.11298.

Chui, C., Li, X., & Mhaskar, H. N. (1994). Neural networks for localized approximation. *Mathematics of Computation*, *63*(208), 607–623.

Chui, C. K., Li, X., & Mhaskar, H. N. (1996). Limitations of the approximation capabilities of neural networks with one hidden layer. *Advances in Computational Mathematics*, *5*(1), 233–243.

Chui, C. K., Lin, S.-B., & Zhou, D.-X. (2019). Deep neural networks for rotation-invariance approximation and learning. *Analysis and Applications*, *17*(05), 737–772.

Chui, C. K., & Mhaskar, H. N. (2018). Deep nets for local manifold learning. *Frontiers in Applied Mathematics and Statistics*, *4*, 12.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems*, *2*(4), 303–314.

DeVore, R. A., Howard, R., & Micchelli, C. (1989). Optimal nonlinear approximation. *Manuscripta Mathematica*, *63*(4), 469–478.

Fang, Z., Feng, H., Huang, S., & Zhou, D.-X. (2020). Theory of deep convolutional neural networks II: Spherical analysis. *Neural Networks*, *131*, 154–162.

Federer, H. (1959). Curvature measures. *Transactions of the American Mathematical Society*, *93*(3), 418–491.

Genovese, C., Perone-Pacifico, M., Verdinelli, I., & Wasserman, L. (2012). Minimax manifold estimation. *Journal of Machine Learning Research*, *13*(May), 1263–1291.

Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). *Deep learning, Vol. 1*. MIT press Cambridge.

Grohs, P., Perekrestenko, D., Elbrächter, D., & Bölcskei, H. (2019). Deep neural network approximation theory. arXiv preprint arXiv:1901.02220.

He, X., Yan, S., Hu, Y., Niyogi, P., & Zhang, H.-J. (2005). Face recognition using laplacianfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(3), 328–340.

Hein, M., & Maier, M. (2007a). Manifold denoising. In *Advances in neural information processing systems* (pp. 561–568).

Hein, M., & Maier, M. (2007b). Manifold denoising as preprocessing for finding natural representations of data. In *AAAI* (pp. 1646–1649).

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, *2*(5), 359–366.

Kereta, Ž., Klock, T., & Naumova, V. (2020). Nonlinear generalization of the monotone single index model. *Information and Inference: A Journal of the IMA*, Eprint at https://academic.oup.com/imaiai/article-pdf/doi/10.1093/imaiai/iaaa013/33522904/iaaa013.pdf.

Klusowski, J. M., & Barron, A. R. (2016). Uniform approximation by neural networks activated by first and second order ridge splines. arXiv preprint arXiv:1607.07819.

Kpotufe, S. (2011). K-NN regression adapts to local intrinsic dimension. In *Advances in neural information processing systems* (pp. 729–737).

Kurková, V., & Sanguineti, M. (2001). Bounds on rates of variable-basis and neural-network approximation. *IEEE Transactions on Information Theory*, *47*(6), 2659–2665.

Kurková, V., & Sanguineti, M. (2002). Comparison of worst case errors in linear and neural network approximation. *IEEE Transactions on Information Theory*, *48*(1), 264–275.

Leeb, W., & Coifman, R. (2016). Hölder–Lipschitz norms and their duals on spaces with semigroups, with applications to earth mover's distance. *Journal of Fourier Analysis and Applications*, *22*(4), 910–953.

Leshno, M., Lin, V. Y., Pinkus, A., & Schocken, S. (1993). Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Networks*, *6*(6), 861–867.

Li, B. (2018). *Sufficient dimension reduction: methods and applications with r*. CRC Press.

Ma, Y., & Zhu, L. (2013). A review on dimension reduction. *International Statistical Review*, *81*(1), 134–150.

McCane, B., & Szymanski, L. (2017). Deep radial kernel networks: approximating radially symmetric functions with deep networks. arXiv preprint arXiv:1703.03470.

Mhaskar, H. N. (1993). Approximation properties of a multilayered feedforward artificial neural network. *Advances in Computational Mathematics*, *1*(1), 61–80.

Mhaskar, H. N. (1996). Neural networks for optimal approximation of smooth and analytic functions. *Neural Computation*, *8*(1), 164–177.

Mhaskar, H. N. (2004). On the tractability of multivariate integration and approximation by neural networks. *Journal of Complexity*, *20*(4), 561–590.

Mhaskar, H. N. (2020a). Dimension independent bounds for general shallow networks. *Neural Networks*, *123*, 142–152.

Mhaskar, H. (2020b). A direct approach for function approximation on data defined manifolds. *Neural Networks*, *132*, 253–268.

Mhaskar, H., Liao, Q., & Poggio, T. (2016). *Learning real and boolean functions: when is deep better than shallow*: *Technical report*, Center for Brains, Minds and Machines (CBMM), arXiv.

Mhaskar, H., Liao, Q., & Poggio, T. (2017). When and why are deep networks better than shallow ones?. In *Thirty-first AAAI conference on artificial intelligence*.

Mhaskar, H. N., & Poggio, T. (2016). Deep vs. shallow networks: An approximation theory perspective. *Analysis and Applications*, *14*(06), 829–848.

Mhaskar, H., & Poggio, T. (2020). Function approximation by deep networks. *Communications on Pure & Applied Analysis*, *19*(8).

Montanelli, H., Yang, H., & Du, Q. (2019). Deep relu networks overcome the curse of dimensionality for bandlimited functions. arXiv preprint arXiv:1903.00735.

Nakada, R., & Imaizumi, M. (2019). Adaptive approximation and estimation of deep neural network to intrinsic dimensionality. arXiv preprint arXiv:1907.02177.

Niyogi, P., Smale, S., & Weinberger, S. (2008). Finding the homology of submanifolds with high confidence from random samples. *Discrete & Computational Geometry*, *39*(1–3), 419–441.

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

Pinkus, A. (1999). Approximation theory of the MLP model in neural networks. *Acta Numerica*, *8*, 143–195.

Poggio, T., Anselmi, F., & Rosasco, L. (2015). *I-theory on depth vs width: hierarchical function composition*: *Technical report*, Center for Brains, Minds and Machines (CBMM).

Poggio, T., Mhaskar, H., Rosasco, L., Miranda, B., & Liao, Q. (2017). Why and when can deep-but not shallow-networks avoid the curse of dimensionality: a review. *International Journal of Automation and Computing*, *14*(5), 503–519.

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

Schmidt-Hieber, J. (2020). Nonparametric regression using deep neural networks with relu activation function. *The Annals of Statistics*, *48*(4), 1875–1897.

Shaham, U., Cloninger, A., & Coifman, R. R. (2018). Provable approximation properties for deep neural networks. *Applied and Computational Harmonic Analysis*, *44*(3), 537–557.

Shen, Z., Yang, H., & Zhang, S. (2019). Nonlinear approximation via compositions. *Neural Networks*, *119*, 74–84.

Shirdhonkar, S., & Jacobs, D. W. (2008). Approximate earth mover's distance in linear time. In *2008 IEEE conference on computer vision and pattern recognition* (pp. 1–8). IEEE.

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, 1040–1053.

Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P. H., & Hospedales, T. M. (2018). Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1199–1208).

Suzuki, T. (2018). Adaptivity of deep relu network for learning in besov and mixed smooth besov spaces: optimal rate and curse of dimensionality. arXiv preprint arXiv:1810.08033.

Suzuki, T., & Nitanda, A. (2019). Deep learning is adaptive to intrinsic dimensionality of model smoothness in anisotropic besov space. arXiv preprint arXiv:1910.12799.

Telgarsky, M. (2017). Neural networks and rational functions. In *Proceedings of machine learning research*: *Vol. 70*, *Proceedings of the 34th international conference on machine learning* (pp. 3387–3393). PMLR.

Vershynin, R. (2018). *High-dimensional probability: an introduction with applications in data science, Vol. 47*. Cambridge University Press.

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

Yarotsky, D. (2018). Optimal approximation of continuous functions by very deep relu networks. In *Conference on learning theory* (pp. 639–649).

Ye, G.-B., & Zhou, D.-X. (2008). Learning and approximation by Gaussians on Riemannian manifolds. *Advances in Computational Mathematics*, *29*(3), 291–310.

Zhou, D.-X. (2020a). Theory of deep convolutional neural networks: Downsampling. *Neural Networks*, *124*, 319–327.

Zhou, D.-X. (2020b). Universality of deep convolutional neural networks. *Applied and Computational Harmonic Analysis*, *48*(2), 787–794.