Efficient Approximation of Deep ReLU Networks for Functions on Low Dimensional Manifolds

Minshuo Chen **Haoming Jiang** **Wenjing Liao** **Tuo Zhao**
Georgia Institute of Technology
`{mchen393, jianghm, wliao60, tourzhao}@gatech.edu`

Abstract

Deep neural networks have revolutionized many real world applications, due to their flexibility in data fitting and accurate predictions for unseen data. A line of research reveals that neural networks can approximate certain classes of functions with an arbitrary accuracy, while the size of the network scales exponentially with respect to the data dimension. Empirical results, however, suggest that networks of moderate size already yield appealing performance. To explain such a gap, a common belief is that many data sets exhibit low dimensional structures, and can be modeled as samples near a low dimensional manifold. In this paper, we prove that neural networks can efficiently approximate functions supported on low dimensional manifolds. The network size scales exponentially in the approximation error, with an exponent depending on the intrinsic dimension of the data and the smoothness of the function. Our result shows that exploiting low dimensional data structures can greatly enhance the efficiency in function approximation by neural networks. We also implement a sub-network that assigns input data to their corresponding local neighborhoods, which may be of independent interest.

1 Introduction

In the past decade, neural networks have made astonishing breakthroughs in many real world applications, such as computer vision (Krizhevsky et al., 2012; Goodfellow et al., 2014; Long et al., 2015), natural language processing (Graves et al., 2013; Bahdanau et al., 2014; Young et al., 2018), healthcare (Miotto et al., 2017; Jiang et al., 2017), robotics (Gu et al., 2017), etc.

Although data sets in these applications are highly complex, neural networks have achieved overwhelming successes. For image classification, the winner of the 2017 ImageNet challenge retained a top-5 error rate of 2.25% (Hu et al., 2018), while the data set consists of about 1.2 million labeled high resolution images in 1000 categories. For speech recognition, Amodei et al. (2016) reported that deep neural networks outperformed humans with a 5.15% word error rate on the LibriSpeech corpus constructed from audio books (Panayotov et al., 2015). Such a data set consists of approximately 1000 hours of 16kHz read English speech from 8000 audio books. These empirical results suggest that neural networks can well approximate complex distributions and functions on data.

A line of research attempts to explain the success of neural networks through the lens of expressivity — neural networks can effectively approximate various classes of functions. Among existing works, the most well-known results are the universal approximation theorems, see Irie and Miyake (1988); Funahashi (1989); Cybenko (1989); Hornik (1991); Chui and Li (1992); Leshno et al. (1993). Specifically, Cybenko (1989) showed that neural networks with one single hidden layer and continuous sigmoidal[1] activations can approximate continuous functions in a unit cube with arbitrary accuracy. Later, Hornik (1991) extended the universal approximation theorem to general feed-forward networks

---

[1] A function $\sigma(x)$ is sigmoidal, if $\sigma(x) \to 0$ as $x \to -\infty$, and $\sigma(x) \to 1$ as $x \to \infty$.

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

with a single hidden layer, while the width of the network has to be exponentially large. Specific approximation rates of shallow networks (with one hidden layer) with smooth activation functions were given in Barron (1993) and Mhaskar (1996). Recently, Lu et al. (2017) proved the universal approximation theorem for width-bounded deep neural networks, and Hanin (2017) improved the result with ReLU (Rectified Linear Units) activations, i.e. $\mathrm{ReLU}(x) = \max\{0, x\}$. Yarotsky (2017) further showed that deep ReLU networks can uniformly approximate functions in Sobolev spaces, while the network size scales exponentially in the approximation error with an exponent depending on the data dimension. Moreover, the network size in Yarotsky (2017) matches its lower bound.

The network size considered in applications, however, is significantly smaller than what is predicted by the theory above. In the ImageNet challenge, data are RGB images with a resolution of 224×224. The theory above suggests that, to achieve a ϵ uniform approximation error, the number of neurons has to scale as $\epsilon^{-224\times224\times3/2}$ (Barron, 1993). Setting $\epsilon = 0.1$ already gives rise to $10^{224\times224\times3/2}$ neurons. However, the AlexNet (Krizhevsky et al., 2012) only consists of 650000 neurons and 60 million parameters to beat the state-of-the-art. To boost the performance on the ImageNet, several more sophisticated network structures were proposed later, such as VGG16 (Simonyan and Zisserman, 2014) which consists of about 138 million parameters. The size of both networks remains extremely small compared to $10^{224\times224\times3/2}$. Why is there a tremendous gap between theory and practice?

A common belief is that real world data sets often exhibit low dimensional structures. Many images consist of projections of 3-dimensional objects followed by some transformations, such as rotation, translation, and skeleton. Such a generating mechanism induces a small number of intrinsic parameters. Speech data are composed of words and sentences following the grammar, and therefore have a small degree of freedom. More broadly, visual, acoustic, textual, and many other types of data all have low dimensional structures due to rich local regularities, global symmetries, repetitive patterns, or redundant sampling. It is plausible to model these data as samples near a low dimensional manifold (Tenenbaum et al., 2000; Roweis and Saul, 2000). Then a natural question is:

Can deep neural networks efficiently approximate functions supported on low dimensional manifolds?

Function approximation on manifolds has been well studied using local polynomials (Bickel et al., 2007) and wavelets (Coifman and Maggioni, 2006). However, studies using neural networks are very limited. Two noticeable works are Chui and Mhaskar (2016) and Shaham et al. (2018). In Chui and Mhaskar (2016), high order differentiable functions on manifolds are approximated by neural networks with smooth activations, e.g., sigmoid activations and rectified quadratic unit functions ($\sigma(x) = (\max\{0, x\})^2$). These smooth activations, however, are rarely used in the mainstream applications such as computer vision (Krizhevsky et al., 2012; Long et al., 2015; Hu et al., 2018). In Shaham et al. (2018), a 4-layer network with ReLU activations was proposed to approximate C^2 functions on low dimensional manifolds that have absolutely summable wavelet coefficients. However, this theory does not cover arbitrarily smooth functions, and the analysis is built upon a restrictive assumption — there exists a linear transformation that maps the input data to sparse coordinates, but such transformation is not explicitly given.

In this paper, we propose a framework to construct deep neural networks with nonsmooth activations to approximate functions supported on a d-dimensional smooth manifold isometrically embedded in $\mathbb{R}^D$. We prove that, in order to achieve a fixed approximation error, the network size scales exponentially with respect to the intrinsic dimension d, instead of the ambient dimension D. Our framework is flexible: **1**). It applies to nonsmooth activations, e.g., ReLU and leaky ReLU activations; **2**). It applies to a wide class of functions, such as Sobolev and Hölder classes which are typical examples in nonparametric statistics (Györfi et al., 2006); **3**). It exploits high order smoothness of functions for making the approximation as efficient as possible.

Theorem (informal)**.** Let $\mathcal{M}$ be a d-dimensional compact Riemannian manifold isometrically embedded in $\mathbb{R}^D$ with $d \ll D$. Assume $\mathcal{M}$ satisfies some mild regularity conditions. Given any $\epsilon \in (0, 1)$, there exists a ReLU neural network structure such that, for any C^n ($n \geq 1$) function $f : \mathcal{M} \to \mathbb{R}$, if the weight parameters are properly chosen, the network yields a function $\widehat{f}$ satisfying $\|f - \widehat{f}\|_\infty \leq \epsilon$. Such a network has no more than $c_1 \left(\log \frac{1}{\epsilon} + \log D\right)$ layers, and at most $c_2 \left(\epsilon^{-d/n} \log \frac{1}{\epsilon} + D \log \frac{1}{\epsilon} + D \log D\right)$ neurons and weight parameters, where c_1, c_2 depend on d, n, f, and $\mathcal{M}$.

Our network size scales like $\epsilon^{-d/n}$ and only weakly depends on the ambient dimension D. This is consistent with empirical observations, and partially justifies why the networks of moderate size have achieved a great success on aforementioned learning tasks. Moreover, we show that our network size matches its lower bound up to a logarithmic factor (see Theorem 2).

Our theory applies to general C^n functions and leverages the benefits of exploiting high order smoothness. Our result improves Shaham et al. (2018) for C^n functions with $n > 2$. In this case, our network size scales like $\epsilon^{-d/n}$, which is significantly smaller than the one in Shaham et al. (2018) in the order of $\epsilon^{-d/2}$.

Here we state the theorem for C^n functions for simplicity, and similar results hold for Hölder functions (see Theorem 1). Our framework can be easily applied to leaky ReLU activations, since leaky ReLU can be implemented by the difference of two ReLU functions.

The high level idea of our framework is to partition the low dimensional manifold into a collection of open sets, and then use Taylor expansions to approximate the function in each neighborhood. A new technique is developed to implement a sub-network that assigns the input to its corresponding neighborhood on the manifold, which may be of independent interest.

Notations: We use bold-faced letters to denote vectors, and normal font letters with a subscript to denote its coordinate, e.g., $\mathbf{x} \in \mathbb{R}^d$ and x_k being the k-th coordinate of $\mathbf{x}$. Given a vector $\mathbf{n} = [n_1, \dots, n_d]^\top \in \mathbb{N}^d$, we define $\mathbf{n}! = \prod_{i=1}^d n_i!$ and $|\mathbf{n}| = \sum_{i=1}^d n_i$. We define $\mathbf{x}^{\mathbf{n}} = \prod_{i=1}^d x_i^{n_i}$. Given a function $f : \mathbb{R}^d \mapsto \mathbb{R}$, we denote its derivative as $D^{\mathbf{n}} f = \frac{\partial^{|\mathbf{n}|} f}{\partial x_1^{n_1} \dots \partial x_d^{n_d}}$, and its ℓ_∞ norm as $\|f\|_\infty = \max_{\mathbf{x}} |f(\mathbf{x})|$. We use $\circ$ to denote the composition operator.

2 Preliminaries

We briefly review manifolds, partition of unity, and function spaces defined on smooth manifolds. Details can be found in Tu (2010) and Lee (2006).

Let $\mathcal{M}$ be a d-dimensional Riemannian manifold isometrically embedded in $\mathbb{R}^D$.

Definition 1 (Chart)**.** A chart for $\mathcal{M}$ is a pair (U, ϕ) such that $U \subset \mathcal{M}$ is open and $\phi : U \mapsto \mathbb{R}^d$, where ϕ is a homeomorphism (i.e., bijective, ϕ and ϕ^{-1} are both continuous).

The open set U is called a coordinate neighborhood, and ϕ is called a coordinate system on U. A chart essentially defines a local coordinate system on $\mathcal{M}$. We say two charts (U, ϕ) and (V, ψ) on $\mathcal{M}$ are C^k compatible if and only if the transition functions, $\phi \circ \psi^{-1} : \psi(U \cap V) \mapsto \phi(U \cap V)$ and $\psi \circ \phi^{-1} : \phi(U \cap V) \mapsto \psi(U \cap V)$ are both C^k. Then we give the definition of an atlas.

Definition 2 (C^k Atlas)**.** An atlas for $\mathcal{M}$ is a collection $\{(U_\alpha, \phi_\alpha)\}_{\alpha \in \mathcal{A}}$ of pairwise C^k compatible charts such that $\bigcup_{\alpha \in \mathcal{A}} U_\alpha = \mathcal{M}$.

Definition 3 (Smooth Manifold)**.** A smooth manifold is a manifold $\mathcal{M}$ together with a C^∞ atlas.

Classical examples of smooth manifolds are the Euclidean space $\mathbb{R}^D$, the torus, and the unit sphere. The existence of an atlas on $\mathcal{M}$ allows us to define differentiable functions.

Definition 4 (C^n Functions on $\mathcal{M}$)**.** Let $\mathcal{M}$ be a smooth manifold in $\mathbb{R}^D$. A function $f : \mathcal{M} \mapsto \mathbb{R}$ is C^n if for any chart (U, ϕ), the composition $f \circ \phi^{-1} : \phi(U) \mapsto \mathbb{R}$ is continuously differentiable up to order n.

Remark 1. The definition of C^n functions is independent of the choice of the chart (U, ϕ). Suppose (V, ψ) is another chart and $V \bigcap U \neq \emptyset$. Then we have $f \circ \psi^{-1} = (f \circ \phi^{-1}) \circ (\phi \circ \psi^{-1})$. Since $\mathcal{M}$ is a smooth manifold, (U, ϕ) and (V, ψ) are C^∞ compatible. Thus, $f \circ \phi^{-1}$ is C^n and $\phi \circ \psi^{-1}$ is C^∞, and their composition is C^n.

We next introduce partition of unity, which plays a crucial role in our construction of neural networks.

Definition 5 (Partition of Unity)**.** A C^∞ partition of unity on a manifold $\mathcal{M}$ is a collection of nonnegative C^∞ functions $\rho_\alpha : \mathcal{M} \mapsto \mathbb{R}_+$ for $\alpha \in \mathcal{A}$ such that **1)**. the collection of supports, $\{\text{supp}(\rho_\alpha)\}_{\alpha \in \mathcal{A}}$ is locally finite[2]; **2)**. $\sum \rho_\alpha = 1$.

For a smooth manifold, a C^∞ partition of unity always exisits.

---

[2] A collection $\{A_\alpha\}$ is locally finite if every point has a neighborhood that meets only finitely many of A_α's.

Proposition 1 (Existence of a C^∞ partition of unity)**.** Let $\{U_\alpha\}_{\alpha \in \mathcal{A}}$ be an open cover of a smooth manifold $\mathcal{M}$. Then there is a C^∞ partition of unity $\{\rho_i\}_{i=1}^\infty$ with every ρ_i having a compact support such that $\mathrm{supp}(\rho_i) \subset U_\alpha$ for some $\alpha \in \mathcal{A}$.

Proposition 1 gives rise to the decomposition $f = \sum_{i=1}^\infty f_i$ with $f_i = f\rho_i$. Note that the f_i's have the same regularity as f, since $f_i \circ \phi^{-1} = (f \circ \phi^{-1}) \times (\rho_i \circ \phi^{-1})$ for a chart (U, ϕ). This decomposition has the advantage that every f_i is only supported in a single chart. Then the approximation of f boils down to the approximations of the f_i's, which are localized and have the same regularity as f.

To characterize the curvature of a manifold, we adopt the following geometric concept.

Definition 6 (Reach, *Definition 2.1* in Aamari et al. (2019))**.** Denote $\mathcal{C}(\mathcal{M}) = \{\mathbf{x} \in \mathbb{R}^D : \exists \mathbf{p} \neq \mathbf{q} \in \mathcal{M}, \|\mathbf{p} - \mathbf{x}\|_2 = \|\mathbf{q} - \mathbf{x}\|_2 = \inf_{\mathbf{y} \in \mathcal{M}} \|\mathbf{y} - \mathbf{x}\|_2\}$ as the set of points that have at least two nearest neighbors on $\mathcal{M}$. Then the reach $\tau > 0$ is defined as $\tau = \inf_{\mathbf{x} \in \mathcal{M}, \mathbf{y} \in \mathcal{C}(\mathcal{M})} \|\mathbf{x} - \mathbf{y}\|_2$.

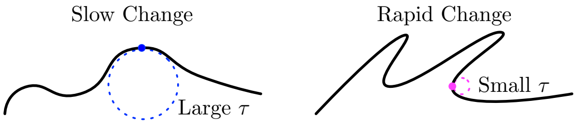

Figure 1: Manifolds with large and small reach.

Reach has a straightforward geometrical interpretation: At each point $\mathbf{x} \in \mathcal{M}$, the radius of the osculating circle is greater or equal to τ. A large reach for $\mathcal{M}$ essentially requires the manifold $\mathcal{M}$ not to change "rapidly" as shown in Figure 1.

Reach determines a proper choice of an atlas for $\mathcal{M}$. In Section 4, we choose each chart U_α contained in a ball of radius less than $\tau/2$. For smooth manifolds with a small τ, we need a large number of charts. Therefore, the reach of a smooth manifold reflects the difficulty of function approximations on $\mathcal{M}$.

3 Main Result

We next present how to construct a ReLU network to approximate $f : \mathcal{M} \mapsto \mathbb{R}$ with error ϵ, under certain assumptions on $\mathcal{M}$ and f.

Assumption 1. $\mathcal{M}$ is a d-dimensional compact Riemannian manifold isometrically embedded in $\mathbb{R}^D$. There exists a constant B such that for any point $\mathbf{x} \in \mathcal{M}$, we have $|x_i| \leq B$ for all $i = 1, \ldots, D$.

Assumption 2. The reach of $\mathcal{M}$ is $\tau > 0$.

Assumption 3. $f : \mathcal{M} \mapsto \mathbb{R}$ belongs to the Hölder space $H^{n,\alpha}$ with a positive integer n and $\alpha \in (0, 1]$, in the sense that $f \in C^{n-1}$ and for any chart (U, ϕ) and $|\mathbf{n}| = n$, we have

$$\left| D^{\mathbf{n}}(f \circ \phi^{-1})\big|_{\phi(\mathbf{x}_1)} - D^{\mathbf{n}}(f \circ \phi^{-1})\big|_{\phi(\mathbf{x}_2)} \right| \leq \|\phi(\mathbf{x}_1) - \phi(\mathbf{x}_2)\|_2^\alpha, \quad \forall \mathbf{x}_1, \mathbf{x}_2 \in U. \tag{1}$$

Assumption 3 says that all n-th order derivatives of $f \circ \phi^{-1}$ are Hölder continuous. Here Hölder functions are defined on manifolds. We recover the standard Hölder class on Euclidean spaces by taking ϕ as the identity map. We also note that Assumption 3 does not depend on the choice of charts.

We now formally state our main result. Extensions to functions in Sobolev spaces are straightforward.

Theorem 1. Suppose Assumptions 1 and 2 hold. Given any $\epsilon \in (0, 1)$, there exists a ReLU network structure such that, for any $f : \mathcal{M} \to \mathbb{R}$ satisfying Assumption 3, if the weight parameters are properly chosen, the network yields a function $\widehat{f}$ satisfying $\|\widehat{f} - f\|_\infty \leq \epsilon$. Such a network has no more than $c_1(\log \frac{1}{\epsilon} + \log D)$ layers, and at most $c_2(\epsilon^{-\frac{d}{n+\alpha}} \log \frac{1}{\epsilon} + D \log \frac{1}{\epsilon} + D \log D)$ neurons and weight parameters, where c_1, c_2 depend on d, n, f, τ, and the surface area of $\mathcal{M}$.

The network structure identified by Theorem 1 consists of three sub-networks as shown in Figure 2:

- *Chart determination sub-network*, which assigns the input to its corresponding neighborhoods;
- *Taylor approximation sub-network*, which approximates f by polynomials in each neighborhood;

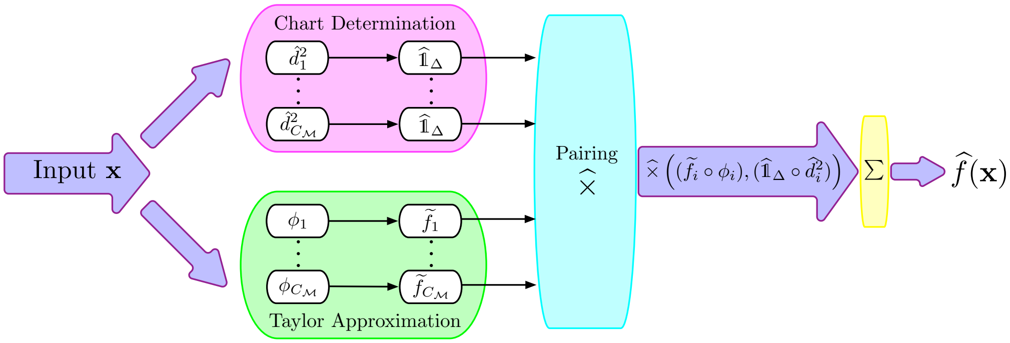
Chart Determination
$\widehat{d}_1^2$
$\widehat{\mathbb{1}}_\Delta$
$\widehat{d}_{C_\mathcal{M}}^2$
$\widehat{\mathbb{1}}_\Delta$
Input $\mathbf{x}$
Pairing
$\widehat{\times}$
$\widehat{\times}\left((\widetilde{f}_i \circ \phi_i), (\widehat{\mathbb{1}}_\Delta \circ \widehat{d}_i^2)\right)$
Σ
$\widehat{f}(\mathbf{x})$
ϕ_1
$\widetilde{f}_1$
$\phi_{C_\mathcal{M}}$
$\widetilde{f}_{C_\mathcal{M}}$
Taylor Approximation

Figure 2: The ReLU network identified by Theorem 1.

- *Pairing sub-network*, which yields multiplications of the proper pairs of outputs from the chart determination and the Taylor approximation sub-networks.

Specifically, we partition the manifold as $\mathcal{M} = \bigcup_{i=1}^{C_\mathcal{M}} U_i$, where the U_i's are open sets contained in a Euclidean ball of radius less than $\tau/2$. $C_\mathcal{M}$ depends on the reach τ, the surface area of $\mathcal{M}$, and the dimension d (see Section 4 for an explicit characterization). For each chart, the chart determination sub-network computes an approximation of the indicator function on U_i. The Taylor approximation sub-network provides a local polynomial approximation of f on U_i. Then the pairing sub-network approximates the product for the proper pairs of outputs in the previous two sub-networks. Finally, $\widehat{f}$ is obtained by taking a sum over $C_\mathcal{M}$ outputs from the pairing sub-network.

The size of our ReLU network matches its lower bound up to a logarithmic factor for the approximation of functions in Hölder spaces. Denote $F^{n,d}$ as functions defined on $[0,1]^d$ in the Hölder space $H^{n-1,1}$. We state a lower bound due to DeVore et al. (1989).

Theorem 2. Fix d and n. Let W be a positive integer and $\kappa : \mathbb{R}^W \mapsto C([0,1]^d)$ be any mapping. Suppose there is a continuous map $\Theta : F^{n,d} \mapsto \mathbb{R}^W$ such that $\|f - \kappa(\Theta(f))\|_\infty \leq \epsilon$ for any $f \in F^{n,d}$. Then $W \geq c\epsilon^{-\frac{d}{n}}$ with c depending on n only.

We take $\mathbb{R}^W$ as the parameter space of a ReLU network, and κ as the network structure. Then to approximate any $f \in F^{n,d}$, the ReLU network has at least $c\epsilon^{-\frac{d}{n}}$ weight parameters. Although Theorem 2 holds for functions on $[0,1]^d$, our network size remains in the same order up to a logarithmic factor even when the function is supported on a manifold of dimension d.

4 Proof of the Main Result

Due to limited space, we present a sketch of the proof for Theorem 1. Before we proceed, we show how to approximate the multiplication operation using ReLU networks. This operation is heavily used in the Taylor approximation sub-network, since Taylor polynomials involve sum of products. We first show ReLU networks can approximate quadratic functions.

Lemma 1 (Proposition 2 in Yarotsky (2017))**.** The function $f(x) = x^2$ with $x \in [0,1]$ can be approximated by a ReLU network with any error $\epsilon > 0$. The network has depth and the number of neurons and weight parameters no more than $c\log(1/\epsilon)$ with an absolute constant c.

This lemma is proved in Appendix A.1. The idea is to approximate quadratic functions using a weighted sum of a series of sawtooth functions. Those sawtooth functions are obtained by compositing the triangular function

$$g(x) = 2\mathrm{ReLU}(x) - 4\mathrm{ReLU}(x - 1/2) + 2\mathrm{ReLU}(x - 1),$$

which can be implemented by a single layer ReLU network.

We then approximate the multiplication operation by invoking the identity $ab = \frac{1}{4}((a+b)^2 - (a-b)^2)$ where the two squares can be approximated by ReLU networks in Lemma 1.

Corollary 1 (Proposition 3 in Yarotsky (2017))**.** Given a constant $C > 0$ and $\epsilon \in (0, C^2)$, there is a ReLU network which implements a function $\widehat{\times} : \mathbb{R}^2 \mapsto \mathbb{R}$ such that: **1)**. For all inputs x and

y satisfying $|x| \leq C$ and $|y| \leq C$, we have $|\widehat{\times}(x, y) - xy| \leq \epsilon$; **2)**. The depth and the weight parameters of the network is no more than $c \log \frac{C^2}{\epsilon}$ with an absolute constant c.

The ReLU network in Theorem 1 is constructed in the following 5 steps.

Step 1. Construction of an atlas. Denote the open Euclidean ball with center $\mathbf{c}$ and radius r in $\mathbb{R}^D$ by $\mathcal{B}(\mathbf{c}, r)$. For any r, the collection $\{\mathcal{B}(\mathbf{x}, r)\}_{\mathbf{x} \in \mathcal{M}}$ is an open cover of $\mathcal{M}$. Since $\mathcal{M}$ is compact, there exists a finite collection of points $\mathbf{c}_i$ for $i = 1, \dots, C_{\mathcal{M}}$ such that $\mathcal{M} \subset \bigcup_i \mathcal{B}(\mathbf{c}_i, r)$.

Now we pick the radius $r < \tau/2$ so that $U_i = \mathcal{M} \cap \mathcal{B}(\mathbf{c}_i, r)$ is diffeomorphic[3] to a ball in $\mathbb{R}^d$ (Niyogi et al., 2008). Let $\{(U_i, \phi_i)\}_{i=1}^{C_{\mathcal{M}}}$ be an atlas on $\mathcal{M}$, where ϕ_i is to be defined in **Step 2**. The number of charts $C_{\mathcal{M}}$ is upper bounded by

$$C_{\mathcal{M}} \leq \left\lceil \frac{SA(\mathcal{M})}{r^d} T_d \right\rceil,$$

where $SA(M)$ is the surface area of $\mathcal{M}$, and T_d is the thickness[4] of the U_i's.

Remark 2. The thickness T_d scales approximately linear in d. As shown in Conway et al. (1987), there exists covering with $\frac{d}{e\sqrt{e}} \lesssim T_d \leq d \log d + d \log \log d + 5d$.

Step 2. Projection with rescaling and translation. We denote the tangent space at $\mathbf{c}_i$ as $T_{\mathbf{c}_i}(\mathcal{M}) = \mathrm{span}(\mathbf{v}_{i1}, \dots, \mathbf{v}_{id})$, where $\{\mathbf{v}_{i1}, \dots, \mathbf{v}_{id}\}$ form an orthonormal basis. We obtain the matrix $V_i = [\mathbf{v}_{i1}, \dots, \mathbf{v}_{id}] \in \mathbb{R}^{D \times d}$ by concatenating $\mathbf{v}_{ij}$'s as column vectors.

Define $\phi_i(\mathbf{x}) = b_i(V_i^\top(\mathbf{x} - \mathbf{c}_i) + \mathbf{s}_i) \in [0, 1]^d$ for any $\mathbf{x} \in U_i$, where $b_i \in (0, 1]$ is a scaling factor and $\mathbf{s}_i$ is a translation vector. Since U_i is bounded, we can choose proper b_i and $\mathbf{s}_i$ to guarantee $\phi_i(\mathbf{x}) \in [0, 1]^d$. We rescale and translate the projection to ease the notation for the development of local Taylor approximations in **Step 4**. We also remark that each ϕ_i is a linear function, and can be realized by a single-layer linear network.

Step 3. Chart determination. This step is to locate the charts that a given input $\mathbf{x}$ belongs to. This avoids projecting $\mathbf{x}$ using unmatched charts (i.e., $\mathbf{x} \notin U_j$ for some j) as illustrated in Figure 3.

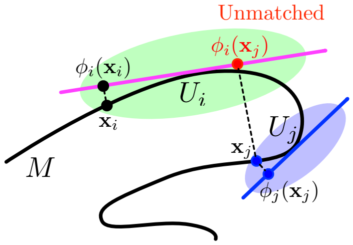

Figure 3: Projecting $\mathbf{x}_j$ using a matched chart (blue) (U_j, ϕ_j), and an unmatched chart (green) (U_i, ϕ_i).

Proper charts[5] can be determined by compositing an indicator function and the squared Euclidean distance $d_i^2(\mathbf{x}) = \|\mathbf{x} - \mathbf{c}_i\|_2^2 = \sum_{j=1}^D (x_j - c_{i,j})^2$ for $i = 1, \dots, C_{\mathcal{M}}$. The squared distance $d_i^2(\mathbf{x})$ is a sum of univariate quadratic functions, thus, we can apply Lemma 1 to approximate $d_i^2(\mathbf{x})$ by ReLU networks. Denote $\widehat{h}_{\mathrm{sq}}$ as an approximation of the quadratic function x^2 on $[0, 1]$ with an approximation error ν. Then we define

$$\widehat{d_i^2}(\mathbf{x}) = 4B^2 \sum_{j=1}^D \widehat{h}_{\mathrm{sq}} \left(\left| \frac{x_j - c_{i,j}}{2B} \right| \right).$$

as an approximation of $d_i^2(\mathbf{x})$. The approximation error is $\|\widehat{d_i^2} - d_i^2\|_\infty \leq 4B^2 D\nu$, by the triangle inequality. We next consider an approximation of the indicator function of an interval as in Figure 4:

$$\widehat{\mathbb{1}}_\Delta(a) = \begin{cases} 1 & a \leq r^2 - \Delta + 4B^2 m\nu \\ -\frac{1}{\Delta - 8B^2 m\nu} a + \frac{r^2 - 4B^2 m\nu}{\Delta - 8B^2 m\nu} & a \in [r^2 - \Delta + 4B^2 m\nu, r^2 - 4B^2 m\nu] \\ 0 & a > r^2 - 4B^2 m\nu \end{cases}, \tag{2}$$

where Δ ($\Delta \geq 8B^2 m\nu$) will be chosen later according to the accuracy ϵ. Note that $\widehat{\mathbb{1}}_\Delta$ can be implemented exactly by a single layer ReLU network: $\widehat{\mathbb{1}}_\Delta(a) = \frac{1}{\Delta - 8B^2 m\nu} \mathrm{ReLU}(-a + r^2 -$

---

[3] P is diffeomorphic to Q if there is a mapping $\Gamma : P \mapsto Q$ bijective, C^∞, and its inverse also being C^∞.

[4] Thickness is the average number of U_i's that contain a point on $\mathcal{M}$ (Conway et al., 1987).

[5] Note that an input $\mathbf{x}$ can belong to multiple charts. Accordingly, the chart determination sub-network determines all these charts.

$4B^2m\nu) - \frac{1}{\Delta - 8B^2m\nu}\mathrm{ReLU}(-a + r^2 - \Delta + 4B^2m\nu)$. We use $\widehat{\mathbb{1}}_\Delta \circ \widehat{d}_i^2$ to approximate the indicator function on U_i: if $\mathbf{x} \notin U_i$, i.e., $d_i^2(\mathbf{x}) \geq r^2$, we have $\widehat{\mathbb{1}}_\Delta \circ \widehat{d}_i^2(\mathbf{x}) = 0$; if $\mathbf{x} \in U_i$ and $d_i^2(\mathbf{x}) \leq r^2 - \Delta$, we have $\widehat{\mathbb{1}}_\Delta \circ \widehat{d}_i^2(\mathbf{x}) = 1$.

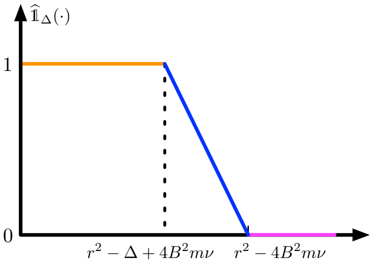

Figure 4: The approximation of the indicator function $\widehat{\mathbb{1}}_\Delta$ in (2).

Step 4. Taylor approximation. In each chart (U_i, ϕ_i), we locally approximate f using Taylor polynomials of order n. Specifically, we decompose f as $f = \sum_{i=1}^{C_\mathcal{M}} f_i$ with $f_i = f\rho_i$ where ρ_i is an element in a C^∞ partition of unity on $\mathcal{M}$ which is supported inside U_i. The existence of such a partition of unity is guaranteed by Proposition 1. Since $\mathcal{M}$ is compact and ρ_i is C^∞, f_i preserves the regularity (smoothness) of f such that $f_i \in H^{n,\alpha}$ for $i = 1, \dots, C_\mathcal{M}$.

Lemma 2. Suppose Assumption 3 holds. For $i = 1, \dots, C_\mathcal{M}$, the function f_i belongs to $H^{n,\alpha}$: there exists a Hölder coefficient L_i depending on d, f_i, and ϕ_i such that for any $|\mathbf{n}| = n$, we have

$$\left| D^\mathbf{n}(f_i \circ \phi_i^{-1})\big|_{\phi_i(\mathbf{x}_1)} - D^\mathbf{n}(f_i \circ \phi_i^{-1})\big|_{\phi_i(\mathbf{x}_2)} \right| \leq L_i \left\| \phi_i(\mathbf{x}_1) - \phi_i(\mathbf{x}_2) \right\|_2^\alpha, \quad \forall \mathbf{x}_1, \mathbf{x}_2 \in U_i.$$

Proof Sketch. We provide a sketch here. Details can be found in Appendix B.1. Denote $g_1 = f \circ \phi_i^{-1}$ and $g_2 = \rho_i \circ \phi_i^{-1}$. We have $D^\mathbf{n}(f_i \circ \phi_i^{-1}) = D^\mathbf{n}(g_1 \times g_2) = \sum_{|\mathbf{p}|+|\mathbf{q}|=n} \binom{n}{|\mathbf{p}|} D^\mathbf{p} g_1 D^\mathbf{q} g_2$, by the Leibniz rule. Consider each term in the sum: for any $\mathbf{x}_1, \mathbf{x}_2 \in U_i$,

$$\begin{aligned}
&\left| D^\mathbf{p} g_1 D^\mathbf{q} g_2 |_{\phi_i(\mathbf{x}_1)} - D^\mathbf{p} g_1 D^\mathbf{q} g_2 |_{\phi_i(\mathbf{x}_2)} \right| \\
\leq & \left| D^\mathbf{p} g_1(\phi_i(\mathbf{x}_1)) \right| \left| D^\mathbf{q} g_2 |_{\phi_i(\mathbf{x}_1)} - D^\mathbf{q} g_2 |_{\phi_i(\mathbf{x}_2)} \right| + \left| D^\mathbf{q} g_2(\phi_i(\mathbf{x}_2)) \right| \left| D^\mathbf{p} g_1 |_{\phi_i(\mathbf{x}_1)} - D^\mathbf{p} g_1 |_{\phi_i(\mathbf{x}_2)} \right| \\
\leq & \lambda_i \theta_{i,\alpha} \left\| \phi_i(\mathbf{x}_1) - \phi_i(\mathbf{x}_2) \right\|_2^\alpha + \mu_i \beta_{i,\alpha} \left\| \phi_i(\mathbf{x}_1) - \phi_i(\mathbf{x}_2) \right\|_2^\alpha .
\end{aligned}$$

Here λ_i and μ_i are uniform upper bounds on the derivatives of g_1 and g_2 with order up to n, respectively. The last inequality above is derived as follows: by the mean value theorem, we have

$$\begin{aligned}
&\left| D^\mathbf{q} g_2 |_{\phi_i(\mathbf{x}_1)} - D^\mathbf{q} g_2 |_{\phi_i(\mathbf{x}_2)} \right| \leq \sqrt{d} \mu_i \left\| \phi_i(\mathbf{x}_1) - \phi_i(\mathbf{x}_2) \right\|_2 \\
= & \sqrt{d} \mu_i \left\| \phi_i(\mathbf{x}_1) - \phi_i(\mathbf{x}_2) \right\|_2^{1-\alpha} \left\| \phi_i(\mathbf{x}_1) - \phi_i(\mathbf{x}_2) \right\|_2^\alpha \leq \sqrt{d} \mu_i (2r)^{1-\alpha} \left\| \phi_i(\mathbf{x}_1) - \phi_i(\mathbf{x}_2) \right\|_2^\alpha ,
\end{aligned}$$

where the last inequality is due to the fact that $\left\| \phi_i(\mathbf{x}_1) - \phi_i(\mathbf{x}_2) \right\|_2 \leq b_i \left\| V_i \right\| \left\| \mathbf{x}_1 - \mathbf{x}_2 \right\|_2 \leq 2r$. Then we set $\theta_{i,\alpha} = \sqrt{d} \mu_i (2r)^{1-\alpha}$ and by a similar argument, we set $\beta_{i,\alpha} = \sqrt{d} \lambda_i (2r)^{1-\alpha}$. We complete the proof by taking $L_i = 2^{n+1} \sqrt{d} \lambda_i \mu_i (2r)^{1-\alpha}$. ☐

Lemma 2 is crucial for the error estimation in the local approximation of $f_i \circ \phi_i^{-1}$ by Taylor polynomials. This error estimate is given in the following theorem, where some of the proof techniques are from Theorem 1 in Yarotsky (2017).

Theorem 3. Let $f_i = f\rho_i$ as in **Step 4**. For any $\delta \in (0, 1)$, there exists a ReLU network structure that, if the weight parameters are properly chosen, the network yields an approximation of $f_i \circ \phi_i^{-1}$ uniformly with error δ. Such a network has no more than $c\left(\log \frac{1}{\delta} + 1\right)$ layers, and at most $c' \delta^{-\frac{d}{n+\alpha}} \left(\log \frac{1}{\delta} + 1\right)$ neurons and weight parameters with c, c' depending on $n, d, f_i \circ \phi_i^{-1}$.

Proof Sketch. The detailed proof is provided in Appendix B.2. The proof consists of two steps: **1)**. Approximate $f_i \circ \phi_i^{-1}$ using a weighted sum of Taylor polynomials; **2)**. Implement the weighted sum of Taylor polynomials using ReLU networks. Specifically, we set up a uniform grid and divide $[0, 1]^d$

into small cubes, and then approximate $f_i \circ \phi_i^{-1}$ by its n-th order Taylor polynomial in each cube. To implement such polynomials by ReLU networks, we recursively apply the multiplication $\widehat{\times}$ operator in Corollary 1, since these polynomials are sums of the products of different variables. $\square$

Step 5. Estimating the total error. We have collected all the ingredients to implement the entire ReLU network to approximate f on $\mathcal{M}$. Recall that the network structure consists of 3 main sub-networks as demonstrated in Figure 2. Let $\widehat{\times}$ be an approximation to the multiplication operator in the pairing sub-network with error η. Accordingly, the function given by the whole network is

$$\widehat{f} = \sum_{i=1}^{C_{\mathcal{M}}} \widehat{\times}(\widehat{f}_i, \widehat{\mathbb{1}}_\Delta \circ \widehat{d}_i^2) \quad \text{with } \widehat{f}_i = \widetilde{f}_i \circ \phi_i,$$

where $\widetilde{f}_i$ is the approximation of $f_i \circ \phi_i^{-1}$ using Taylor polynomials in Theorem 3. The total error can be decomposed to three components according to the following theorem.

Theorem 4. For any $i = 1, \ldots, C_{\mathcal{M}}$, we have $\|\widehat{f} - f\|_\infty \le \sum_{i=1}^{C_{\mathcal{M}}} (A_{i,1} + A_{i,2} + A_{i,3})$, where

$$\begin{aligned} A_{i,1} &= \left\|\widehat{\times}(\widehat{f}_i, \widehat{\mathbb{1}}_\Delta \circ \widehat{d}_i^2) - \widehat{f}_i \times (\widehat{\mathbb{1}}_\Delta \circ \widehat{d}_i^2)\right\|_\infty \le \eta, \\ A_{i,2} &= \left\|\widehat{f}_i \times (\widehat{\mathbb{1}}_\Delta \circ \widehat{d}_i^2) - f_i \times (\widehat{\mathbb{1}}_\Delta \circ \widehat{d}_i^2)\right\|_\infty \le \delta, \\ A_{i,3} &= \left\|f_i \times (\widehat{\mathbb{1}}_\Delta \circ \widehat{d}_i^2) - f_i \times \mathbb{1}(\mathbf{x} \in U_i)\right\|_\infty \le \frac{c(\pi+1)}{r(1-r/\tau)}\Delta \quad \text{for some constant } c. \end{aligned}$$

Here $\mathbb{1}(\mathbf{x} \in U_i)$ is the indicator function on U_i. Theorem 4 is proved in Appendix B.3. In order to achieve an ϵ total approximation error, i.e., $\|f - \widehat{f}\|_\infty \le \epsilon$, we need to control the errors in the three sub-networks. In other words, we need to decide ν for $\widehat{d}_i^2$, Δ for $\widehat{\mathbb{1}}_\Delta$, δ for $\widetilde{f}_i$, and η for $\widehat{\times}$. Note that $A_{i,1}$ is the error from the pairing sub-network, $A_{i,2}$ is the approximation error in the Taylor approximation sub-network, and $A_{i,3}$ is the error from the chart determination sub-network. The error bounds on $A_{i,1}, A_{i,2}$ are straightforward from the constructions of $\widehat{\times}$ and $\widehat{f}_i$. The estimate of $A_{i,3}$ involves some technical analysis since $\|\widehat{\mathbb{1}}_\Delta \circ \widehat{d}_i^2 - \mathbb{1}(\mathbf{x} \in U_i)\|_\infty = 1$. Note that $\widehat{\mathbb{1}}_\Delta \circ \widehat{d}_i^2(\mathbf{x}) - \mathbb{1}(\mathbf{x} \in U_i) = 0$ whenever $\|\mathbf{x} - \mathbf{c}_i\|_2^2 < r^2 - \Delta$ or $\|\mathbf{x} - \mathbf{c}_i\|_2^2 > r^2$, so we only need to prove that $|f_i(\mathbf{x})|$ is sufficiently small in the region $\mathcal{K}_i$ defined below.

Lemma 3. For any $i = 1, \ldots, C_{\mathcal{M}}$, denote $\mathcal{K}_i = \{\mathbf{x} \in \mathcal{M} : r^2 - \Delta \le \|\mathbf{x} - \mathbf{c}_i\|_2^2 \le r^2\}$. Then there exists a constant c depending on f_i's and ϕ_i's such that

$$\max_{\mathbf{x} \in \mathcal{K}_i} |f_i(\mathbf{x})| \le \frac{c(\pi+1)}{r(1-r/\tau)}\Delta.$$

Proof Sketch. The detailed proof is in Appendix B.4. The function $f_i \circ \phi_i^{-1}$ is defined on $\phi_i(U_i) \subset [0,1]^d$. We extend $f_i \circ \phi_i^{-1}$ to $[0,1]^d$ by letting $f_i \circ \phi_i^{-1}(\mathbf{x}) = 0$ for $\mathbf{x} \in [0,1]^d \setminus \phi_i(U_i)$. It is easy to verify that such an extension preserves the regularity of $f_i \circ \phi_i^{-1}$, since $\mathrm{supp}(f_i)$ is a compact subset of U_i. By the mean value theorem, for any $\mathbf{x}, \mathbf{y} \in \mathcal{K}_i$, there exists $\mathbf{z} = \beta\phi_i(\mathbf{x}) + (1-\beta)\phi_i(\mathbf{y})$ for some $\beta \in (0,1)$ such that

$$|f_i(\mathbf{x}) - f_i(\mathbf{y})| \le \|\nabla f_i \circ \phi_i^{-1}(\mathbf{z})\|_2 \|\phi_i(\mathbf{x}) - \phi_i(\mathbf{y})\|_2 \le \|\nabla f_i \circ \phi_i^{-1}(\mathbf{z})\|_2 b_i \|V_i\|_2 \|\mathbf{x} - \mathbf{y}\|_2.$$

We pick $\mathbf{y} \in \partial U_i$ (the boundary of U_i) so that $f_i(\mathbf{y}) = 0$. Since $f_i \in H^{n,\alpha}$ and $\mathcal{M}$ is compact, $\left\|\nabla f_i \circ \phi_i^{-1}(\mathbf{z})\right\|_2 b_i \|V_i\|_2 \le c$ for some $c > 0$. To bound $|f_i(\mathbf{x})|$, the key is to estimate $\|\mathbf{x} - \mathbf{y}\|_2$. We next prove that, for any $\mathbf{x} \in \mathcal{K}_i$, there exists $\mathbf{y} \in \partial U_i$ satisfying $\|\mathbf{x} - \mathbf{y}\|_2 \le \frac{\pi+1}{r(1-r/\tau)}\Delta$.

The idea is to consider a geodesic[6] $\gamma(t)$ parameterized by the arc length from $\mathbf{x}$ to ∂U_i in Figure 5. Denote $\mathbf{y} = \partial U_i \bigcap \gamma$. Without loss of generality, we shift the center $\mathbf{c}_i$ to $\mathbf{0}$ in the following analysis. To utilize polar coordinates, we define two auxiliary quantities: $\theta(t) = \gamma(t)^\top \dot{\gamma}(t) / \|\gamma(t)\|_2$ and $\ell(t) = \|\gamma(t)\|_2$, where $\dot{\gamma}$ denotes the derivative of γ.

---

[6] A geodesic is the shortest path between two points on the manifold. We refer readers to Chapter 6 in Lee (2006) for a formal introduction.

We show that there exists a geodesic $\gamma(t)$ satisfying $\inf_t \dot{\ell}(t) \geq \frac{1-r/\tau}{\pi+1} > 0$. This implies that the geodesic continuously moves away from the center. Denote T such that $\gamma(T) = \mathbf{y}$. By the definition of geodesic, T is the arc length of $\gamma(t)$ between $\mathbf{x}$ and $\mathbf{y}$. We have $T \inf_t \dot{\ell}(t) \leq \ell(T) - \ell(0) \leq r - \sqrt{r^2 - \Delta} \leq \frac{\Delta}{r}$. Therefore, $\|\mathbf{x} - \mathbf{y}\|_2 \leq T \leq \frac{\Delta}{r \inf_t \dot{\ell}(t)} \leq \frac{\pi+1}{r(1-r/\tau)}\Delta$. □

Given Theorem 4, we choose

$$\eta = \delta = \frac{\epsilon}{3C_{\mathcal{M}}} \text{ and } \Delta = \frac{r(1-r/\tau)\epsilon}{3c(\pi+1)C_{\mathcal{M}}} \tag{3}$$

so that the approximation error is bounded by ϵ. Moreover, we choose $\nu = \frac{\Delta}{16B^2D}$ to guarantee $\Delta > 8B^2D\nu$ so that the definition of $\widehat{\mathbb{1}}_\Delta$ is valid.

Finally we quantify the size of the ReLU network. Recall that the chart determination sub-network has $c_1 \log \frac{1}{\nu}$ layers, the Taylor approximation sub-network has $c_2 \log \frac{1}{\delta}$ layers, and the pairing sub-network has $c_3 \log \frac{1}{\eta}$ layers. Here c_2 depends on d, n, f, and c_1, c_3 are absolute constants. Combining these with (3) yields the depth in Theorem 1. By a similar argument, we can obtain the number of neurons and weight parameters. A detailed analysis is given in Appendix B.5.

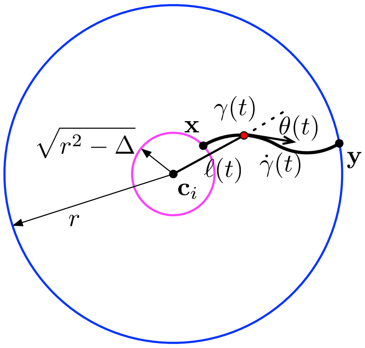

Figure 5: A geometric illustration of θ and ℓ.

5 Discussions

ReLU activations. We consider neural networks with ReLU activations for a practical concern — ReLU activations are widely used in deep networks. Moreover, ReLU networks are easier to train compared with sigmoid or hyperbolic tangent activations, which are known for their notorious vanishing gradient problem (Goodfellow et al., 2016; Glorot et al., 2011).

Low Dimensional Manifolds. The low dimensional manifold model plays a vital role to reduce the network size. As shown in Theorem 2, to approximate functions in $F^{n,D}$ with accuracy ϵ, the minimal number of weight parameters is $O(\epsilon^{-\frac{D}{n}})$. This lower bound is huge, and can not be improved without low dimensional structures of data.

Existence vs. Learnability and Generalization. Our Theorem 1 shows the existence of a ReLU network structure that gives efficient approximations of functions on low dimensional manifolds, if the weight parameters are properly chosen. In practice, it is observed that larger neural networks are easier to train and yield better generalization performances (Li et al., 2018; Zhang et al., 2016; Arora et al., 2018). This is referred to as overparameterization. Establishing the connection between learnability and generalization is an important future direction.

Convolutional Filters. Convolutional neural networks (CNNs, Krizhevsky et al. (2012)) are widely used in computer vision, language modeling, etc. Empirical results reveal that different convolutional filters can capture various patterns in images, e.g., edge detection filters. An interesting question is whether convolutional filters serve as charts in our framework.

Equivalent Networks. The ReLU network identified in Theorem 1 is sparsely connected. Several other network structures can yield the same function as our ReLU network. It is interesting to investigate whether these network structures also possess the universal approximation property.

6 Acknowledgements

This work is supported by NSF grants DMS 1818751 and III 1717916. The authors would like to thank Ryan Tibshirani for his helpful discussions and insightful comments.

References

Aamari, E., Kim, J., Chazal, F., Michel, B., Rinaldo, A., Wasserman, L. et al. (2019). Estimating the reach of a manifold. *Electronic Journal of Statistics*, **13** 1359–1399.

AMODEI, D., ANANTHANARAYANAN, S., ANUBHAI, R., BAI, J., BATTENBERG, E., CASE, C., CASPER, J., CATANZARO, B., CHENG, Q., CHEN, G. ET AL. (2016). Deep speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning*.

ARORA, S., COHEN, N. and HAZAN, E. (2018). On the optimization of deep networks: Implicit acceleration by overparameterization. *arXiv preprint arXiv:1802.06509*.

BAHDANAU, D., CHO, K. and BENGIO, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

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

BICKEL, P. J., LI, B. ET AL. (2007). Local polynomial regression on unknown manifolds. In *Complex datasets and inverse problems*. Institute of Mathematical Statistics, 177–186.

CHUI, C. K. and LI, X. (1992). Approximation by ridge functions and neural networks with one hidden layer. *Journal of Approximation Theory*, **70** 131–141.

CHUI, C. K. and MHASKAR, H. N. (2016). Deep nets for local manifold learning. *arXiv preprint arXiv:1607.07110*.

COIFMAN, R. R. and MAGGIONI, M. (2006). Diffusion wavelets. *Applied and Computational Harmonic Analysis*, **21** 53–94.

CONWAY, J. H., SLOANE, N. J. A. and BANNAI, E. (1987). *Sphere-packings, Lattices, and Groups*. Springer-Verlag, Berlin, Heidelberg.

CYBENKO, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, **2** 303–314.

DEVORE, R. A., HOWARD, R. and MICCHELLI, C. (1989). Optimal nonlinear approximation. *Manuscripta mathematica*, **63** 469–478.

FUNAHASHI, K.-I. (1989). On the approximate realization of continuous mappings by neural networks. *Neural networks*, **2** 183–192.

GLOROT, X., BORDES, A. and BENGIO, Y. (2011). Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*.

GOODFELLOW, I., BENGIO, Y. and COURVILLE, A. (2016). *Deep Learning*. MIT Press. `http://www.deeplearningbook.org`.

GOODFELLOW, I., POUGET-ABADIE, J., MIRZA, M., XU, B., WARDE-FARLEY, D., OZAIR, S., COURVILLE, A. and BENGIO, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems*.

GRAVES, A., MOHAMED, A.-R. and HINTON, G. (2013). Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*. IEEE.

GU, S., HOLLY, E., LILLICRAP, T. and LEVINE, S. (2017). Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates. In *2017 IEEE international conference on robotics and automation (ICRA)*. IEEE.

GYÖRFI, L., KOHLER, M., KRZYZAK, A. and WALK, H. (2006). *A distribution-free theory of nonparametric regression*. Springer Science & Business Media.

HANIN, B. (2017). Universal function approximation by deep neural nets with bounded width and relu activations. *arXiv preprint arXiv:1708.02691*.

HORNIK, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural networks*, **4** 251–257.

HU, J., SHEN, L. and SUN, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

IRIE, B. and MIYAKE, S. (1988). Capabilities of three-layered perceptrons. In *IEEE International Conference on Neural Networks*, vol. 1.

JIANG, F., JIANG, Y., ZHI, H., DONG, Y., LI, H., MA, S., WANG, Y., DONG, Q., SHEN, H. and WANG, Y. (2017). Artificial intelligence in healthcare: past, present and future. *Stroke and vascular neurology*, **2** 230–243.

KRIZHEVSKY, A., SUTSKEVER, I. and HINTON, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*.

LEE, J. M. (2006). *Riemannian manifolds: an introduction to curvature*, vol. 176. Springer Science & Business Media.

LESHNO, M., LIN, V. Y., PINKUS, A. and SCHOCKEN, S. (1993). Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural networks*, **6** 861–867.

LI, H., XU, Z., TAYLOR, G., STUDER, C. and GOLDSTEIN, T. (2018). Visualizing the loss landscape of neural nets. In *Advances in Neural Information Processing Systems*.

LONG, J., SHELHAMER, E. and DARRELL, T. (2015). Fully convolutional networks for semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

LU, Z., PU, H., WANG, F., HU, Z. and WANG, L. (2017). The expressive power of neural networks: A view from the width. In *Advances in Neural Information Processing Systems*.

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

MIOTTO, R., WANG, F., WANG, S., JIANG, X. and DUDLEY, J. T. (2017). Deep learning for healthcare: review, opportunities and challenges. *Briefings in bioinformatics*, **19** 1236–1246.

NIYOGI, P., SMALE, S. and WEINBERGER, S. (2008). Finding the homology of submanifolds with high confidence from random samples. *Discrete & Computational Geometry*, **39** 419–441.

PANAYOTOV, V., CHEN, G., POVEY, D. and KHUDANPUR, S. (2015). Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

ROWEIS, S. T. and SAUL, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *science*, **290** 2323–2326.

SHAHAM, U., CLONINGER, A. and COIFMAN, R. R. (2018). Provable approximation properties for deep neural networks. *Applied and Computational Harmonic Analysis*, **44** 537–557.

SIMONYAN, K. and ZISSERMAN, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

TENENBAUM, J. B., DE SILVA, V. and LANGFORD, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, **290** 2319–2323.

TU, L. (2010). *An Introduction to Manifolds*. Universitext, Springer New York.
`https://books.google.com/books?id=br1KngEACAAJ`

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

YOUNG, T., HAZARIKA, D., PORIA, S. and CAMBRIA, E. (2018). Recent trends in deep learning based natural language processing. *ieee Computational intelligenCe magazine*, **13** 55–75.

ZHANG, C., BENGIO, S., HARDT, M., RECHT, B. and VINYALS, O. (2016). Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*.

Supplementary Material for Efficient Approximation of Deep Neural Networks

A Proofs of Preliminary Results in Section 4

A.1 Proof of Lemma 1

Proof. We partition the interval $[0, 1]$ uniformly into 2^N subintervals $I_k = [\frac{k}{2^N}, \frac{k+1}{2^N}]$ for $k = 0, \ldots, 2^N - 1$. We approximate $f(x) = x^2$ on these subintervals by a linear interpolation

$$\widehat{f}_k = \frac{2k+1}{2^N}\left(x - \frac{k}{2^N}\right) + \frac{k^2}{2^{2N}}, \quad \text{for } x \in I_k.$$

It is straightforward to check that $\widehat{f}_k$ meets f at the endpoints $\frac{k}{2^N}, \frac{k+1}{2^N}$ of I_k.

We evaluate the approximation error of $\widehat{f}_k$ on the interval I_k:

$$\begin{aligned}\max_{x \in I_k}\left|f(x) - \widehat{f}_k(x)\right| &= \max_{x \in I_k}\left|x^2 - \frac{2k+1}{2^N}x + \frac{k^2+k}{2^{2N}}\right| \\ &= \max_{x \in I_k}\left|\left(x - \frac{2k+1}{2^{2N}}\right)^2 - \frac{1}{2^{4N}}\right| \\ &= \frac{1}{2^{4N}}.\end{aligned}$$

Note that this approximation error does not depend on k. Thus, in order to achieve an ϵ approximation error, we only need

$$\frac{1}{2^{4N}} \leq \epsilon \Longrightarrow N \geq \frac{\log \frac{1}{\epsilon}}{4}.$$

Let $N = \lceil \frac{\log \frac{1}{\epsilon}}{4} \rceil$ and denote $f_N = \sum_{k=0}^{2^N-1} \widehat{f}_k \mathbb{1}\{x \in I_k\}$. We compute the increment from f_{N-1} to f_N for $x \in \left[\frac{k}{2^{N-1}}, \frac{k+1}{2^{N-1}}\right]$ as follows,

$$\begin{aligned}f_{N-1} - f_N &= \begin{cases} \frac{k^2}{2^{2(N-1)}} + \frac{2k+1}{2^{N-1}}\left(x - \frac{k}{2^{N-1}}\right) - \frac{k^2}{2^{2(N-1)}} - \frac{4k+1}{2^N}\left(x - \frac{k}{2^{N-1}}\right), & x \in \left[\frac{k}{2^{N-1}}, \frac{2k+1}{2^N}\right) \\ \frac{k^2}{2^{2(N-1)}} + \frac{2k+1}{2^{N-1}}\left(x - \frac{k}{2^{N-1}}\right) - \frac{(2k+1)^2}{2^{2N}} - \frac{4k+3}{2^N}\left(x - \frac{2k+1}{2^N}\right), & x \in \left[\frac{2k+1}{2^N}, \frac{k+1}{2^{N-1}}\right) \end{cases} \\ &= \begin{cases} \frac{1}{2^N}x - \frac{k}{2^{2N-1}}, & x \in \left[\frac{k}{2^{N-1}}, \frac{2k+1}{2^N}\right) \\ -\frac{1}{2^N}x + \frac{k+1}{2^{2N-1}}, & x \in \left[\frac{2k+1}{2^N}, \frac{k+1}{2^{N-1}}\right) \end{cases}.\end{aligned}$$

We observe that $f_{N-1} - f_N$ is a triangular function on $\left[\frac{k}{2^{N-1}}, \frac{k+1}{2^{N-1}}\right]$. The maximum is $\frac{1}{2^{2N}}$ independent of k attained at $x = \frac{2k+1}{2^N}$. The minimum is 0 attained at the endpoints $\frac{k}{2^{N-1}}, \frac{k+1}{2^{N-1}}$. To implement f_N, we consider a triangular function representable by a one-layer ReLU network:

$$g(x) = 2\sigma(x) - 4\sigma(x - 0.5) + 2\sigma(x - 1).$$

Denote by $g_m = g \circ g \circ \cdots \circ g$ the composition of totally m functions g. Observe that g_m is a sawtooth function with 2^{m-1} peaks at $\frac{2k+1}{2^m}$ for $k = 0, \ldots, 2^{m-1} - 1$, and we have $g_m\left(\frac{2k+1}{2^m}\right) = 1$ for $k = 0, \ldots, 2^{m-1} - 1$. Then we have $f_{N-1} - f_N = \frac{1}{2^{2N}} g_N$. By induction, we have

$$\begin{aligned}f_N &= f_{N-1} - \frac{1}{2^{2N}} g_N \\ &= f_{N-2} - \frac{1}{2^{2N}} g_N - \frac{1}{2^{2N-2}} g_{N-1} \\ &= \cdots \\ &= x - \sum_{k=1}^{N} \frac{1}{2^{2k}} g_k.\end{aligned}$$

Therefore, f_N can be implemented by a ReLU network of depth $\lceil \frac{\log \frac{1}{\epsilon}}{4} \rceil \leq c \log \frac{1}{\epsilon}$ for an absolute constant c. Each layer consists of at most 3 neurons, hence, the total number of neurons and weight parameters is no more than $c' \log \frac{1}{\epsilon}$. $\square$

A.2 Proof of Corollary 1

Proof. Let $\widehat{f}_\delta$ be an approximation of the quadratic function on $[0, 1]$ with error $\delta \in (0, 1)$. We set

$$\widehat{\times}(x, y) = C^2 \left(\widehat{f}_\delta \left(\frac{|x + y|}{2C} \right) - \widehat{f}_\delta \left(\frac{|x - y|}{2C} \right) \right).$$

Now we determine δ. We bound the error of $\widehat{\times}$

$$\begin{aligned} \left| \widehat{\times}(x, y) - xy \right| &= C^2 \left| \widehat{f}_\delta \left(\frac{|x+y|}{2C} \right) - \frac{|x+y|^2}{4C^2} - \widehat{f}_\delta \left(\frac{|x-y|}{2C} \right) + \frac{|x-y|^2}{4C^2} \right| \\ &\leq C^2 \left| \widehat{f}_\delta \left(\frac{|x+y|}{2C} \right) - \frac{|x+y|^2}{4C^2} \right| + \left| \widehat{f}_\delta \left(\frac{|x-y|}{2C} \right) - \frac{|x-y|^2}{4C^2} \right| \\ &\leq 2C^2 \delta. \end{aligned}$$

Thus, we pick $\delta = \frac{\epsilon}{2C^2}$ to ensure $\left| \widehat{\times}(x, y) - xy \right| \leq \epsilon$ for any inputs x and y. As shown in Lemma 1, we can implement $\widehat{f}_\delta$ using a ReLU network of depth at most $c' \log \frac{1}{\delta} = c \log \frac{C^2}{\epsilon}$ with absolute constants c', c. The proof is complete. $\square$

B Proofs of Construction of Neural Networks in Section 4

B.1 Proof of Lemma 2

Proof. We rewrite $f_i \circ \phi_i^{-1}$ as

$$\underbrace{(f \circ \phi_i^{-1})}_{g_1} \times \underbrace{(\rho_i \circ \phi_i^{-1})}_{g_2}. \tag{4}$$

By the definition of the partition of unity, we know g_2 is C^∞. This implies that g_2 is $(n + 1)$ continuously differentiable. Since $\text{supp}(\rho_i)$ is compact, the k-th derivative of g_2 is uniformly bounded by $\lambda_{i,k}$ for any $k \leq n + 1$. Let $\lambda_i = \max_{k \leq n+1} \lambda_{i,k}$. We have for any $|\mathbf{n}| \leq n$ and $\mathbf{x}_1, \mathbf{x}_2 \in U_i$,

$$\begin{aligned} |D^{\mathbf{n}} g_2(\phi_i(\mathbf{x}_1)) - D^{\mathbf{n}} g_2(\phi_i(\mathbf{x}_2))| &\leq \sqrt{d} \lambda_i \left\| \phi_i(\mathbf{x}_1) - \phi_i(\mathbf{x}_2) \right\|_2 \\ &\leq \sqrt{d} \lambda_i b_i^{1-\alpha} \left\| \mathbf{x}_1 - \mathbf{x}_2 \right\|_2^{1-\alpha} \left\| \phi_i(\mathbf{x}_1) - \phi_i(\mathbf{x}_2) \right\|_2^{\alpha}. \end{aligned}$$

The last inequality follows from $\phi_i(\mathbf{x}) = b_i(V_i^\top (\mathbf{x} - \mathbf{c}_i) + \mathbf{s}_i)$ and $\|V_i\|_2 = 1$. Observe that U_i is bounded, hence, we have $\|\mathbf{x}_1 - \mathbf{x}_2\|_2^{1-\alpha} \leq (2r)^{1-\alpha}$. Absorbing $\|\mathbf{x}_1 - \mathbf{x}_2\|_2^{1-\alpha}$ into $\sqrt{d} \lambda_i b_i^{1-\alpha}$, we have the derivative of g_2 is Hölder continuous. We denote $\beta_{i,\alpha} = \sqrt{d} \lambda_i b_i^{1-\alpha} (2r)^{1-\alpha} \leq \sqrt{d} \lambda_i (2r)^{1-\alpha}$. Similarly, g_1 is C^{n-1} by Assumption 3. Then there exists a constant μ_i such that the k-th derivative of g_1 is uniformly bounded by μ_i for any $k \leq n - 1$. These derivatives are also Hölder continuous with coefficient $\theta_{i,\alpha} \leq \sqrt{d} \mu_i (2r)^{1-\alpha}$.

By the Leibniz rule, for any $|\mathbf{n}| = n$, we expand the n-th derivative of $f_i \circ \phi_i^{-1}$ as

$$D^{\mathbf{n}}(g_1 \times g_2) = \sum_{|\mathbf{p}| + |\mathbf{q}| = n} \binom{n}{|\mathbf{p}|} D^{\mathbf{p}} g_1 D^{\mathbf{q}} g_2.$$

Consider each summand in the above right-hand side. For any $\mathbf{x}_1, \mathbf{x}_2 \in U_i$, we derive

$$
\begin{aligned}
&\left|D^{\mathbf{p}} g_1(\phi_i(\mathbf{x}_1)) D^{\mathbf{q}} g_2(\phi_i(\mathbf{x}_1)) - D^{\mathbf{p}} g_1(\phi_i(\mathbf{x}_2)) D^{\mathbf{q}} g_2(\phi_i(\mathbf{x}_2))\right| \\
=&\left|D^{\mathbf{p}} g_1(\phi_i(\mathbf{x}_1)) D^{\mathbf{q}} g_2(\phi_i(\mathbf{x}_1)) - D^{\mathbf{p}} g_1(\phi_i(\mathbf{x}_1)) D^{\mathbf{q}} g_2(\phi_i(\mathbf{x}_2))\right. \\
&\quad \left. + D^{\mathbf{p}} g_1(\phi_i(\mathbf{x}_1)) D^{\mathbf{q}} g_2(\phi_i(\mathbf{x}_2)) - D^{\mathbf{p}} g_1(\phi_i(\mathbf{x}_2)) D^{\mathbf{q}} g_2(\phi_i(\mathbf{x}_2))\right| \\
\leq& |D^{\mathbf{p}} g_1(\phi_i(\mathbf{x}_1))| |D^{\mathbf{q}} g_2(\phi_i(\mathbf{x}_1)) - D^{\mathbf{q}} g_2(\phi_i(\mathbf{x}_2))| \\
&\quad + |D^{\mathbf{q}} g_2(\phi_i(\mathbf{x}_2))| |D^{\mathbf{p}} g_1(\phi_i(\mathbf{x}_1)) - D^{\mathbf{p}} g_1(\phi_i(\mathbf{x}_2))| \\
\leq& \mu_i \theta_{i,\alpha} \left\|\phi_i(\mathbf{x}_1) - \phi_i(\mathbf{x}_2)\right\|_2^{\alpha} + \lambda_i \beta_{i,\alpha} \left\|\phi_i(\mathbf{x}_1) - \phi_i(\mathbf{x}_2)\right\|_2^{\alpha} \\
\leq& 2\sqrt{d} \mu_i \lambda_i (2r)^{1-\alpha} \left\|\phi_i(\mathbf{x}_1) - \phi_i(\mathbf{x}_2)\right\|_2^{\alpha}.
\end{aligned}
$$

Observe that there are totally 2^n summands in the right hand side of (4). Therefore, for any $\mathbf{x}_1, \mathbf{x}_2 \in U_i$ and $|\mathbf{n}| = n$, we have

$$
\left| D^{\mathbf{n}}(f_i \circ \phi_i^{-1})\big|_{\phi_i(\mathbf{x}_1)} - D^{\mathbf{n}}(f_i \circ \phi_i^{-1})\big|_{\phi_i(\mathbf{x}_2)} \right| \leq 2^{n+1} \sqrt{d} \mu_i \lambda_i (2r)^{1-\alpha} \left\|\phi_i(\mathbf{x}_1) - \phi_i(\mathbf{x}_2)\right\|_2^{\alpha}.
$$

□

B.2 Proof of Theorem 3

Proof. The proof consists of two steps. We first approximate $f_i \circ \phi_i^{-1}$ by a Taylor polynomial, and then implement the Taylor polynomial using a ReLU network. To ease the analysis, we extend $f_i \circ \phi_i^{-1}$ to the whole cube $[0, 1]^d$ by assigning $f_i \circ \phi_i^{-1}(\mathbf{x}) = 0$ for $\phi_i(\mathbf{x}) \in [0, 1]^d \setminus \phi_i(U_i)$. It is straightforward to check that this extension preserves the regularity of $f_i \circ \phi_i^{-1}$, since f_i vanishes on the complement of the compact set $\mathrm{supp}(\rho_i) \subset U_i$. For notational simplicity, we denote $f_i^{\phi} = f_i \circ \phi_i^{-1}$ with the extension.

Step 1. We define a trapezoid function

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

Note that we have $\|\psi\|_{\infty} = 1$. Let N be a positive integer, we form a uniform grid on $[0, 1]^d$ by dividing each coordinate into N subintervals. We then consider a partition of unity on these grid defined by

$$
\zeta_{\mathbf{m}}(\mathbf{x}) = \prod_{k=1}^{d} \psi\left(3N\left(x_k - \frac{m_k}{N}\right)\right).
$$

We can check that $\sum_{\mathbf{m}} \zeta_{\mathbf{m}}(\mathbf{x}) = 1$ as in Figure 6.

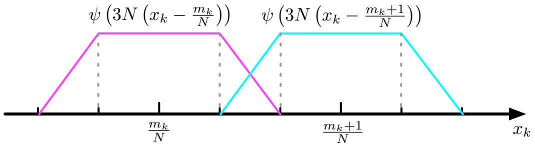
$\psi\left(3N\left(x_k - \frac{m_k}{N}\right)\right)$
$\psi\left(3N\left(x_k - \frac{m_k+1}{N}\right)\right)$
$\frac{m_k}{N}$
$\frac{m_k+1}{N}$
x_k

Figure 6: Illustration of the construction of $\zeta_{\mathbf{m}}$ on the k-th coordinate.

We also observe that $\mathrm{supp}(\zeta_{\mathbf{m}}) = \left\{\mathbf{x} : \left|x_k - \frac{m_k}{N}\right| \leq \frac{1}{N}, k = 1, \ldots, d\right\}$. Now we construct a Taylor polynomial of degree n for approximating f_i^{ϕ} at $\frac{\mathbf{m}}{N}$:

$$
P_{\mathbf{m}}(\mathbf{x}) = \sum_{|\mathbf{n}| \leq n} \frac{D^{\mathbf{n}} f_i^{\phi}}{\mathbf{n}!}\bigg|_{\mathbf{x} = \frac{\mathbf{m}}{N}} \left(\mathbf{x} - \frac{\mathbf{m}}{N}\right)^{\mathbf{n}}.
$$

Define $\bar{f}_i = \sum_{\mathbf{m} \in \{0,\dots,N\}^d} \zeta_{\mathbf{m}} P_{\mathbf{m}}$. We bound the approximation error $\left\| \bar{f}_i - f_i^{\phi} \right\|_\infty$:

$$
\begin{aligned}
\max_{\mathbf{x} \in [0,1]^d} \left| \bar{f}_i(\mathbf{x}) - f_i^{\phi}(\mathbf{x}) \right| &= \max_{\mathbf{x}} \left| \sum_{\mathbf{m}} \phi_{\mathbf{m}}(\mathbf{x}) (P_{\mathbf{m}}(\mathbf{x}) - f_i^{\phi}(\mathbf{x})) \right| \\
&\leq \max_{\mathbf{x}} \sum_{\mathbf{m}: \left| x_k - \frac{m_k}{N} \right| \leq \frac{1}{N}} \left| P_{\mathbf{m}}(\mathbf{x}) - f_i^{\phi}(\mathbf{x}) \right| \\
&\leq \max_{\mathbf{x}} 2^d \max_{\mathbf{m}: \left| x_k - \frac{m_k}{N} \right| \leq \frac{1}{N}} \left| P_{\mathbf{m}}(\mathbf{x}) - f_i^{\phi}(\mathbf{x}) \right| \\
&\leq \max_{\mathbf{x}} \frac{2^d d^n}{n!} \left(\frac{1}{N} \right)^n \max_{|\mathbf{n}| = n} \left| D^{\mathbf{n}} f_i^{\phi} \big|_{\frac{\mathbf{m}}{N}} - D^{\mathbf{n}} f_i^{\phi} \big|_{\mathbf{y}} \right| \\
&\leq \max_{\mathbf{x}} \frac{2^d d^n}{n!} \left(\frac{1}{N} \right)^n 2^{n+1} \sqrt{d} \mu_i \lambda_i (2r)^{1-\alpha} \left\| \frac{\mathbf{m}}{N} - \mathbf{x} \right\|_2^\alpha \\
&\leq \sqrt{d} \mu_i \lambda_i (2r)^{1-\alpha} \frac{2^{d+n+1} d^{n+\alpha/2}}{n!} \left(\frac{1}{N} \right)^{n+\alpha}.
\end{aligned}
$$

Here $\mathbf{y}$ is the linear interpolation of $\frac{\mathbf{m}}{N}$ and $\mathbf{x}$, determined by the Taylor remainder. The second last inequality is obtained by the Hölder continuity in Lemma 2. By setting $\sqrt{d} \mu_i \lambda_i (2r)^{1-\alpha} \frac{2^{d+n+1} d^{n+\alpha/2}}{n!} \left(\frac{1}{N} \right)^{n+\alpha} \leq \frac{\delta}{2}$, we get $N \geq \left(\frac{\sqrt{d} \mu_i \lambda_i (2r)^{1-\alpha} 2^{d+n+2} d^{n+\alpha/2}}{\delta n!} \right)^{\frac{1}{n+\alpha}}$. Accordingly, the approximation error is bounded by $\| \bar{f}_i - f_i^{\phi} \|_\infty \leq \frac{\delta}{2}$.

Step 2. We next implement $\widetilde{f}_i$ by a ReLU network that approximates $\bar{f}_i$ up to an error $\frac{\delta}{2}$. We denote

$$
P_{\mathbf{m}}(\mathbf{x}) = \sum_{|\mathbf{n}| \leq n} a_{\mathbf{m},\mathbf{n}} \left(\mathbf{x} - \frac{\mathbf{m}}{N} \right)^{\mathbf{n}},
$$

where $a_{\mathbf{m},\mathbf{n}} = \left. \frac{D^{\mathbf{n}} f_i^{\phi}}{\mathbf{n}!} \right|_{\mathbf{x} = \frac{\mathbf{m}}{N}}$. Then we rewrite $\bar{f}_i$ as

$$
\bar{f}_i(\mathbf{x}) = \sum_{\mathbf{m} \in \{0,\dots,N\}^d} \sum_{|\mathbf{n}| \leq n} a_{\mathbf{m},\mathbf{n}} \zeta_{\mathbf{m}}(\mathbf{x}) \left(\mathbf{x} - \frac{\mathbf{m}}{N} \right)^{\mathbf{n}}. \tag{5}
$$

Note that (5) is a linear combination of products $\zeta_{\mathbf{m}} \left(\mathbf{x} - \frac{\mathbf{m}}{N} \right)^{\mathbf{n}}$. Each product involves at most $d + n$ univariate terms: d terms for $\zeta_{\mathbf{m}}$ and n terms for $\left(\mathbf{x} - \frac{\mathbf{m}}{N} \right)^{\mathbf{n}}$. We recursively apply Corollary 1 to implement the product. Specifically, let $\widehat{\times}_\epsilon$ be the approximation of the product operator in Corollary 1 with error ϵ, which will be chosen later. Consider the following chain application of $\widehat{\times}_\epsilon$:

$$
\widetilde{f}_{\mathbf{m},\mathbf{n}}(\mathbf{x}) = \widehat{\times}_\epsilon \left(\psi(3N x_1 - 3m_1), \widehat{\times}_\epsilon \left(\dots, \widehat{\times}_\epsilon (\psi(3N_d x_d - m_d), \widehat{\times}_\epsilon (x_1 - \frac{m_1}{N}, \dots)) \right) \right).
$$

Now we estimate the error of the above approximation. Note that we have $|\psi(3N x_k - 3m_k)| \leq 1$ and $\left| x_k - \frac{m_k}{N} \right| \leq 1$ for all $k \in \{1, \dots, d\}$ and $\mathbf{x} \in [0,1]^d$. We then have

$$
\begin{aligned}
\left| \widetilde{f}_{\mathbf{m},\mathbf{n}}(\mathbf{x}) - \zeta_{\mathbf{m}} \left(\mathbf{x} - \frac{\mathbf{m}}{N} \right)^{\mathbf{n}} \right| &= \left| \widehat{\times}_\epsilon \Big(\psi(3N x_1 - 3m_1), \widehat{\times}_\epsilon (\dots, \widehat{\times}_\epsilon (x_1 - \frac{m_1}{N}, \dots)) \Big) - \zeta_{\mathbf{m}} \left(\mathbf{x} - \frac{\mathbf{m}}{N} \right)^{\mathbf{n}} \right| \\
&\leq \Big| \widehat{\times}_\epsilon \left(\psi(3N x_1 - 3m_1), \widehat{\times}_\epsilon (\psi(3N x_2 - 3m_2), \dots) \right) \\
&\qquad - \psi(3N_1 - 3m_1) \widehat{\times}_\epsilon (\psi(3N x_2 - 3m_2), \dots) \Big| \\
&\quad + |\psi(3N x_1 - m_1)| \left| \widehat{\times}_\epsilon (\psi(3N x_2 - 3m_2), \dots) - \psi(3N x_2 - 3m_2) \widehat{\times}_\epsilon (\dots) \right| \\
&\quad + \dots \\
&\leq (d + n) \delta.
\end{aligned}
$$

Moreover, we have $\widetilde{f}_{\mathbf{m},\mathbf{n}}(\mathbf{x}) = \zeta_{\mathbf{m}} \left(\mathbf{x} - \frac{\mathbf{m}}{N} \right)^{\mathbf{n}} = 0$, if $\mathbf{x} \notin \operatorname{supp}(\zeta_{\mathbf{m}})$. Now we define

$$
\widetilde{f}_i = \sum_{\mathbf{m} \in \{0,\dots,N\}^d} \sum_{|\mathbf{n}| \leq n} a_{\mathbf{m},\mathbf{n}} \widetilde{f}_{\mathbf{m},\mathbf{n}}.
$$

The approximation error is bounded by

$$\begin{aligned}\max_{\mathbf{x}} \left|\widetilde{f}_i(\mathbf{x}) - \bar{f}_i(\mathbf{x})\right| &= \left|\sum_{\mathbf{m}\in\{0,\dots,N\}^d} \sum_{|\mathbf{n}|\leq n} a_{\mathbf{m},\mathbf{n}} \left(\widetilde{f}_{\mathbf{m},\mathbf{n}}(\mathbf{x}) - \zeta_{\mathbf{m}}\left(\mathbf{x} - \frac{\mathbf{m}}{N}\right)^{\mathbf{n}}\right)\right| \\ &\leq \max_{\mathbf{x}} \lambda_i \mu_i 2^{d+n+1} \max_{\mathbf{m}:\mathbf{x}\in \operatorname{supp}(\zeta_{\mathbf{m}})} \sum_{|\mathbf{n}|\leq n} \left|\widetilde{f}_{\mathbf{m},\mathbf{n}}(\mathbf{x}) - \zeta_{\mathbf{m}}\left(\mathbf{x} - \frac{\mathbf{m}}{N}\right)^{\mathbf{n}}\right| \\ &\leq \lambda_i \mu_i 2^{d+n+1} d^n (d+n)\epsilon.\end{aligned}$$

We choose $\epsilon = \frac{\delta}{\lambda_i \mu_i 2^{d+n+2} d^n (d+n)}$, so that $\|\bar{f}_i - \widetilde{f}_i\|_\infty \leq \frac{\delta}{2}$. Thus, we eventually have $\|\widetilde{f}_i - f_i^{\phi}\|_\infty \leq \delta$. Now we compute the depth and computational units for implement $\widetilde{f}_i$. $\widetilde{f}_i$ can be implemented by a collection of parallel sub-networks that compute each $\widetilde{f}_{\mathbf{m},\mathbf{n}}$. The total number of parallel sub-networks is bounded by $d^n(N+1)^d$. For each sub-network, we observe that ψ can be exactly implemented by a single layer ReLU network, i.e., $\psi(x) = \text{ReLU}(x+2) - \text{ReLU}(x+1) - \text{ReLU}(x-1) + \text{ReLU}(x-2)$. Corollary 1 shows that $\widehat{\times}_\epsilon$ can be implemented by a depth $c_1 \log \frac{1}{\epsilon}$ ReLU network. Therefore, the whole network for implementing $\widetilde{f}_i$ has no more than $c_1' \left(\log \frac{1}{\epsilon} + 1\right)$ layers and $c_1' d^n (N+1)^d \left(\log \frac{1}{\epsilon} + 1\right)$ neurons and weight parameters. With $\epsilon = \frac{\delta}{\lambda_i \mu_i 2^{d+n+2} d^n (d+n)}$ and $N = \left\lceil \left(\frac{\mu_i \lambda_i (2r)^{1-\alpha} 2^{d+n+2} d^{n+\alpha/2}}{\delta n!}\right)^{\frac{1}{n+\alpha}} \right\rceil$, we obtain that the whole network has no more than $c_1 \log \frac{1}{\delta}$ layers, and at most $c_2 \delta^{-\frac{d}{n+\alpha}} \left(\log \frac{1}{\delta} + 1\right)$ neurons and weight parameters, for constants c_1, c_2 depending on d, n, and $f_i \circ \phi_i^{-1}$. $\square$

B.3 Proof of Theorem 4

Proof. We expand the estimation error as

$$\begin{aligned}\left\|\widehat{f} - f\right\|_\infty &= \left\|\sum_{i=1}^{C_{\mathcal{M}}} \widehat{\times}(\widehat{f}_i, \widehat{\mathbb{1}}_\Delta \circ \widehat{d}_i^2) - f\right\|_\infty \\ &= \left\|\sum_{i=1}^{C_{\mathcal{M}}} \widehat{\times}(\widehat{f}_i, \widehat{\mathbb{1}}_\Delta \circ \widehat{d}_i^2) - f\rho_i \mathbb{1}(\mathbf{x} \in U_i)\right\|_\infty \\ &\leq \sum_{i=1}^{C_{\mathcal{M}}} \left\|\widehat{\times}(\widehat{f}_i, \widehat{\mathbb{1}}_\Delta \circ \widehat{d}_i^2) - f_i \mathbb{1}(\mathbf{x} \in U_i)\right\|_\infty \\ &\leq \sum_{i=1}^{C_{\mathcal{M}}} \left\|\widehat{\times}(\widehat{f}_i, \widehat{\mathbb{1}}_\Delta \circ \widehat{d}_i^2) - \widehat{f}_i \times (\widehat{\mathbb{1}}_\Delta \circ \widehat{d}_i^2) + \widehat{f}_i \times (\widehat{\mathbb{1}}_\Delta \circ \widehat{d}_i^2) - f_i \times (\widehat{\mathbb{1}}_\Delta \circ \widehat{d}_i^2)\right. \\ &\qquad \left. + f_i \times (\widehat{\mathbb{1}}_\Delta \circ \widehat{d}_i^2) - f_i \times \mathbb{1}(\mathbf{x} \in U_i)\right\|_\infty \\ &\leq \sum_{i=1}^{C_{\mathcal{M}}} \underbrace{\left\|\widehat{\times}(\widehat{f}_i, \widehat{\mathbb{1}}_\Delta \circ \widehat{d}_i^2) - \widehat{f}_i \times (\widehat{\mathbb{1}}_\Delta \circ \widehat{d}_i^2)\right\|_\infty}_{A_{i,1}} + \underbrace{\left\|\widehat{f}_i \times (\widehat{\mathbb{1}}_\Delta \circ \widehat{d}_i^2) - f_i \times (\widehat{\mathbb{1}}_\Delta \circ \widehat{d}_i^2)\right\|_\infty}_{A_{i,2}} \\ &\qquad + \underbrace{\left\|f_i \times (\widehat{\mathbb{1}}_\Delta \circ \widehat{d}_i^2) - f_i \times \mathbb{1}(\mathbf{x} \in U_i)\right\|_\infty}_{A_{i,3}}.\end{aligned}$$

The first two terms $A_{i,1}, A_{i,2}$ are straightforward to handle, since by the construction we have

$$\begin{aligned}A_{i,1} &= \left\|\widehat{\times}(\widehat{f}_i, \widehat{\mathbb{1}}_\Delta \circ \widehat{d}_i^2) - \widehat{f}_i \times (\widehat{\mathbb{1}}_\Delta \circ \widehat{d}_i^2)\right\|_\infty \leq \eta, \quad \text{and} \\ A_{i,2} &= \left\|\widehat{f}_i \times (\widehat{\mathbb{1}}_\Delta \circ \widehat{d}_i^2) - f_i \times (\widehat{\mathbb{1}}_\Delta \circ \widehat{d}_i^2)\right\|_\infty \leq \left\|\widehat{f}_i - f_i\right\|_\infty \left\|\widehat{\mathbb{1}}_\Delta \circ \widehat{d}_i^2\right\|_\infty \leq \delta.\end{aligned}$$

By Lemma 3, we have $\max_{\mathbf{x}\in\mathcal{K}_i} |f_i(\mathbf{x})| \leq \frac{c(\pi+1)}{r(1-r/\tau)}\Delta$ for a constant c depending on f_i. Then we bound $A_{i,3}$ as

$$A_{i,3} = \left\| f_i \times (\widehat{\mathbb{1}}_\Delta \circ \widehat{d}_i^2) - f_i \times \mathbb{1}(\mathbf{x} \in U_i) \right\|_\infty \leq \max_{\mathbf{x}\in\mathcal{K}_i} |f_i(\mathbf{x})| \leq \frac{c(\pi+1)}{r(1-r/\tau)}\Delta.$$

$\square$

B.4 Proof of Lemma 3

Proof. We extend $f_i \circ \phi_i^{-1}$ to the whole cube $[0,1]^d$ as in the proof of Theorem 3. We also have $f_i(\mathbf{x}) = 0$ for $\|\mathbf{x} - \mathbf{c}_i\|_2 = r$. By the first order Taylor expansion, we have for any $\mathbf{x}, \mathbf{y} \in U_i$

$$\begin{aligned} |f_i(\mathbf{x}) - f_i(\mathbf{y})| &= \left| f_i \circ \phi_i^{-1}(\phi_i(\mathbf{x})) - f_i \circ \phi_i^{-1}(\phi_i(\mathbf{y})) \right| \\ &\leq \left\| \nabla (f_i \circ \phi_i^{-1})(\mathbf{z}) \right\|_2 \|\phi_i(\mathbf{x}) - \phi_i(\mathbf{y})\|_2 \\ &\leq \left\| \nabla (f_i \circ \phi_i^{-1})(\mathbf{z}) \right\|_2 b_i \|V_i\|_2 \|\mathbf{x} - \mathbf{y}\|_2 , \end{aligned}$$

where $\mathbf{z}$ is a linear interpolation of $\phi_i(\mathbf{x})$ and $\phi_i(\mathbf{y})$ satisfying the mean value theorem. Since $f_i \circ \phi_i^{-1}$ is C^n in $[0,1]^d$, the first derivative is uniformly bounded, i.e., $\left\| \nabla f_i \circ \phi_i^{-1}(\mathbf{z}) \right\|_2 \leq \alpha_i$ for any $\mathbf{z} \in [0,1]^d$. Let $\mathbf{y} \in U_i$ satisfying $f_i(\mathbf{y}) = 0$. In order to bound the function value for any $\mathbf{x} \in \mathcal{K}_i$, we only need to bound the Euclidean distance between $\mathbf{x}$ and $\mathbf{y}$. More specifically, for any $\mathbf{x} \in \mathcal{K}_i$, we need to show that there exists $\mathbf{y} \in U_i$ satisfying $f_i(\mathbf{y}) = 0$, such that $\|\mathbf{x} - \mathbf{y}\|_2$ is sufficiently small.

Before continuing with the proof, we introduce some notations. Let $\gamma(t)$ be a geodesic on $\mathcal{M}$ parameterized by the curve length. In the following context, we use $\dot{\gamma}$ and $\ddot{\gamma}$ to denote the first and second derivatives of γ with respect to t. By the definition of geodesic, we have $\|\dot{\gamma}(t)\|_2 = 1$ (unit speed) and $\ddot{\gamma}(t) \perp \dot{\gamma}(t)$.

Without loss of generality, we shift $\mathbf{c}_i$ to $\mathbf{0}$. We consider a geodesic starting from $\mathbf{x}$ with initial "velocity" $\dot{\gamma}(0) = \mathbf{v}$ in the tangent space of $\mathcal{M}$ at $\mathbf{x}$. To utilize polar coordinate, we define two auxiliary quantities: $\ell(t) = \|\gamma(t)\|_2$ and $\theta(t) = \arccos \frac{\gamma(t)^\top \dot{\gamma}(t)}{\|\gamma(t)\|_2} \in [0, \pi]$. As can be seen in Figure 5, ℓ and θ have clear geometrical interpretations: ℓ is the radial distance from the center $\mathbf{c}_i$, and θ is the angle between the velocity and $\gamma(t)$.

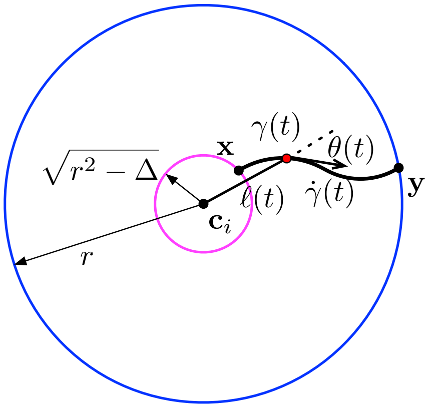

Figure 7: Illustration of ℓ and θ along a parametric curve γ.

Suppose $\mathbf{y} = \gamma(T)$, we need to upper bound T. Note that $\ell(T) - \ell(0) \leq r - \sqrt{r^2 - \Delta} \leq \Delta / r$. Moreover, observe that the derivative of ℓ is $\dot{\ell}(t) = \cos\theta(t)$, since γ has unit speed. It suffices to find a lower bound on $\dot{\ell}(t) = \cos\theta(t)$ so that $T \leq \frac{\Delta}{r \inf_t \dot{\ell}(t)}$.

We immediately have the second derivative of ℓ as $\ddot{\ell}(t) = -\sin\theta(t)\dot{\theta}(t)$. Meanwhile, using the equation $\ell(t) = \sqrt{\gamma(t)^\top\gamma(t)}$, we also have

$$\ddot{\ell}(t) = \frac{\left(\ddot{\gamma}(t)^\top\gamma(t) + \dot{\gamma}(t)^\top\dot{\gamma}(t)\right)\sqrt{\gamma(t)^\top\gamma(t)} - \left(\gamma(t)^\top\dot{\gamma}(t)\right)^2/\sqrt{\gamma(t)^\top\gamma(t)}}{\gamma(t)^\top\gamma(t)}. \tag{6}$$

Note that by definition, we have $\dot{\gamma}(t)^\top\dot{\gamma}(t) = 1$ and $\gamma(t)^\top\dot{\gamma}(t) = \cos\theta(t)\sqrt{\gamma(t)^\top\gamma(t)}$. Plugging into (6), we can derive

$$\ddot{\ell}(t) = \frac{1 + \ddot{\gamma}(t)^\top\gamma(t) - \cos^2\theta(t)}{\ell(t)} = \frac{\sin^2\theta(t) + \ddot{\gamma}(t)^\top\gamma(t)}{\ell(t)}. \tag{7}$$

Now we find a lower bound on $\ddot{\gamma}(t)^\top\gamma(t)$. Specifically, by Cauchy-Schwarz inequality, we have

$$\begin{aligned}\ddot{\gamma}(t)^\top\gamma(t) &\geq -\left\|\ddot{\gamma}(t)\right\|_2\left\|\gamma(t)\right\|_2\left|\cos\angle\left(\ddot{\gamma}(t),\gamma(t)\right)\right| \\ &\geq -\frac{r}{\tau}\left|\cos\angle\left(\ddot{\gamma}(t),\gamma(t)\right)\right|.\end{aligned}$$

The last inequality follows from $\left\|\ddot{\gamma}(t)\right\|_2 \leq \frac{1}{\tau}$ (Niyogi et al., 2008) and $\left\|\gamma(t)\right\|_2 \leq r$. We now need to bound $\angle(\ddot{\gamma}(t),\gamma(t))$, given $\angle\left(\gamma(t),\dot{\gamma}(t)\right) = \theta(t)$ and $\ddot{\gamma}(t) \perp \dot{\gamma}(t)$. Consider the following optimization problem,

$$\begin{aligned}\min \quad & a^\top x, \\ \text{subject to} \quad & x^\top x = 1, \\ & b^\top x = 0.\end{aligned} \tag{8}$$

By assigning $a = \frac{\gamma(t)}{\|\gamma(t)\|_2}$ and $b = \frac{\dot{\gamma}(t)}{\|\dot{\gamma}(t)\|_2}$, the optimal objective value is exactly the minimum of $\cos\angle\left(\ddot{\gamma}(t),\gamma\right)$. Additionally, we can find the maximum of $\cos\angle\left(\ddot{\gamma}(t),\gamma\right)$ by replacing the minimization in (8) by maximization. We solve (8) by the Lagrangian method. More precisely, let

$$\mathcal{L}(x,\lambda,\mu) = -a^\top x + \lambda(x^\top x - 1) + \mu(b^\top x).$$

We have the optimal solution x^* satisfying $\nabla_x\mathcal{L} = 0$, which implies $x^* = \frac{1}{2\lambda^*}(a - \mu^* b)$ with μ^* and λ^* being the optimal dual variable. By the primal feasibility, we have $\mu^* = a^\top b$ and $\lambda^* = -\frac{1}{2}\sqrt{1-(a^\top b)^2}$. Therefore, the optimal objective value is $-\sqrt{1-(a^\top b)^2}$. Similarly, the maximum is $\sqrt{1-(a^\top b)^2}$. Note that $a^\top b = \cos\theta(t)$, we then get

$$\ddot{\gamma}(t)^\top\gamma(t) \geq -\frac{r}{\tau}\sin\theta(t).$$

Substituting into (7), we have the following lower bound

$$\ddot{\ell}(t) = \frac{\sin\theta^2(t) + \ddot{\gamma}(t)^\top\gamma(t)}{\ell(t)} \geq \frac{1}{\ell(t)}\left(\sin^2\theta(t) - \frac{r}{\tau}\sin\theta(t)\right).$$

Now combining with $\ddot{\ell}(t) = -\sin\theta(t)\dot{\theta}(t)$, we can derive

$$\dot{\theta}(t) \leq -\frac{1}{\ell(t)}\left(\sin\theta(t) - \frac{r}{\tau}\right). \tag{9}$$

Inequality (9) has an important implication: When $\sin\theta(t) > \frac{r}{\tau}$, as t increasing, $\theta(t)$ is monotone decreasing until $\sin\theta(t') = \frac{r}{\tau}$ for some $t' = t$. Thus, we distinguish two cases depending on the value of $\theta(0)$. Indeed, we only need to consider $\theta(0) \in [0,\pi/2]$. The reason behind is that if $\theta(0) \in (\pi/2,\pi]$, we only need to set the initial velocity in the opposite direction.

Case 1: $\theta(0) \in \left[0, \arcsin\frac{r}{\tau}\right]$. We claim that $\theta(t) \in \left[0, \arcsin\frac{r}{\tau}\right]$ for all $t \leq T$. In fact, suppose there exists some $t_1 \leq T$ such that $\theta(t_1) > \arcsin\frac{r}{\tau}$. By the continuity of θ, there exists $t_0 < t_1$, such that $\theta(t_0) = \arcsin\frac{r}{\tau}$ and $\theta(t) \geq \arcsin\frac{r}{\tau}$ for $t \in [t_0, t_1]$. This already gives us a contradiction:

$$\theta(t_0) < \theta(t_1) = \theta(t_0) + \underbrace{\int_{t_0}^{t_1}\dot{\theta}(t)dt}_{\leq 0} \leq \theta(t_0).$$

Therefore, we have $\dot{\ell}(t) \geq \cos \arcsin \frac{r}{\tau} = \sqrt{1 - \frac{r^2}{\tau^2}}$, and thus $T \leq \frac{\Delta}{r\sqrt{1-\frac{r^2}{\tau^2}}}$.

Case 2: $\theta(0) \in \left(\arcsin \frac{r}{\tau}, \pi/2\right]$. It is enough to show that $\theta(0)$ can be bounded sufficiently away from $\pi/2$. Let $\gamma_{\mathbf{c},\mathbf{x}} \subset \mathcal{M}$ be a geodesic from $\mathbf{c}_i$ to $\mathbf{x}$. We analogously define $\theta_{\mathbf{c},\mathbf{x}}$ and $\ell_{\mathbf{c},\mathbf{x}}$ as for the geodesic from $\mathbf{x}$ to $\mathbf{y}$. Let $T_{r/2} = \sup\{t : \ell_{\mathbf{c},\mathbf{x}}(t) \leq r/2 - \Delta/r\}$, and denote $\mathbf{z} = \gamma_{\mathbf{c},\mathbf{x}}(T_{r/2})$. We must have $\theta_{\mathbf{c},\mathbf{x}}(T_{r/2}) \in [0, \pi/2]$ and $\ell_{\mathbf{c},\mathbf{x}}(T_{r/2}) = r/2 - \Delta/r$, otherwise there exists $T'_{r/2} > T_{r/2}$ satisfying $\ell_{\mathbf{c},\mathbf{x}}(T'_{r/2}) \leq r/2$. Denote $T_{\mathbf{x}}$ satisfying $\mathbf{x} = \gamma_{\mathbf{c},\mathbf{x}}(T_{\mathbf{x}})$. We bound $\theta_{\mathbf{c},\mathbf{x}}(T_{\mathbf{x}})$ as follows,

$$\begin{aligned}\theta_{\mathbf{c},\mathbf{x}}(T_{\mathbf{x}}) &= \theta_{\mathbf{c},\mathbf{x}}(T_{r/2}) + \int_{T_{r/2}}^{T_{\mathbf{x}}} \dot{\theta}_{\mathbf{c},\mathbf{x}}(t)dt \\ &\leq \frac{\pi}{2} - \int_{T_{r/2}}^{T_{\mathbf{x}}} \frac{1}{\ell_{\mathbf{c},\mathbf{x}}(t)} \left(\sin\theta_{\mathbf{c},\mathbf{x}}(t) - \frac{r}{\tau}\right) dt.\end{aligned}$$

If there exists some $t \in (T_{r/2}, T_{\mathbf{x}}]$ such that $\sin\theta_{\mathbf{c},\mathbf{x}}(t) \leq \frac{r}{\tau}$, by the previous reasoning, we have $\sin\theta_{\mathbf{c},\mathbf{x}}(T_{\mathbf{x}}) \leq \frac{r}{\tau}$. Thus, we only need to handle the case when $\sin\theta_{\mathbf{c},\mathbf{x}}(t) > \frac{r}{\tau}$ for all $t \in (T_{r/2}, T_{\mathbf{x}}]$. In this case, $\theta_{\mathbf{c},\mathbf{x}}(t)$ is monotone decreasing, hence we further have

$$\begin{aligned}\theta_{\mathbf{c},\mathbf{x}}(T_{\mathbf{x}}) &\leq \frac{\pi}{2} - \int_{T_{r/2}}^{T_{\mathbf{x}}} \frac{1}{\ell_{\mathbf{c},\mathbf{x}}(t)} \left(\sin\theta_{\mathbf{c},\mathbf{x}}(T_{\mathbf{x}}) - \frac{r}{\tau}\right) dt \\ &\leq \frac{\pi}{2} - (T_{\mathbf{x}} - T_{r/2})\frac{1}{r}\left(\sin\theta_{\mathbf{c},\mathbf{x}}(T_{\mathbf{x}}) - \frac{r}{\tau}\right) \\ &\leq \frac{\pi}{2} - \frac{1}{2}\left(\sin\theta_{\mathbf{c},\mathbf{x}}(T_{\mathbf{x}}) - \frac{r}{\tau}\right).\end{aligned}$$

The last inequality follows from $T_{\mathbf{x}} - T_{r/2} \geq r/2$. Using the fact, $\sin x \geq \frac{2}{\pi}x$, we can derive

$$\begin{aligned}&\theta_{\mathbf{c},\mathbf{x}}(T_{\mathbf{x}}) \leq \frac{\pi}{2} - \frac{1}{2}\left(\frac{2}{\pi}\theta_{\mathbf{c},\mathbf{x}}(T_{\mathbf{x}}) - \frac{r}{\tau}\right) \\ \Longrightarrow \quad &\theta_{\mathbf{c},\mathbf{x}}(T_{\mathbf{x}}) \leq \frac{\pi}{2}\left(\frac{\pi + r/\tau}{\pi + 1}\right).\end{aligned}$$

We can then set $\theta(0) = \theta_{\mathbf{c},\mathbf{x}}(T_{\mathbf{x}})$, and thus

$$\begin{aligned}\cos\theta(0) \geq \cos\left(\frac{\pi}{2}\frac{\pi + r/\tau}{\pi + 1}\right) &= \cos\left(\frac{\pi}{2}\left(1 - \frac{1 - r/\tau}{\pi + 1}\right)\right) \\ &= \sin\left(\frac{\pi}{2}\frac{1 - r/\tau}{\pi + 1}\right) \\ &\geq \frac{1 - r/\tau}{\pi + 1}.\end{aligned}$$

Therefore, we have $T \leq \frac{\Delta}{r\cos\theta(0)} \leq \frac{\pi+1}{r(1-r/\tau)}\Delta$. By the choice of $r < \tau/2$, we immediately have $\frac{\tau}{\sqrt{\tau^2-r^2}} < \frac{\pi+1}{1-r/\tau}$. Hence, combining case 1 and case 2, we conclude

$$T \leq \frac{\pi + 1}{r(1 - r/\tau)}\Delta.$$

Therefore, the function value $f(\mathbf{x})$ on $\mathcal{K}_i$ is bounded by $\alpha_i \frac{\pi+1}{r(1-r/\tau)}\Delta$. It suffices to let $c = \max_i \alpha_i b_i \|V_i\|_2$, and we complete the proof. ☐

B.5 Characterization of the Size of the ReLU Network

Proof. We evenly split the error ϵ into 3 parts for $A_{i,1}, A_{i,2}$, and $A_{i,3}$, respectively. We pick $\eta = \frac{\epsilon}{3C_{\mathcal{M}}}$ so that $\sum_{i=1}^{C_{\mathcal{M}}} A_{i,1} \leq \frac{\epsilon}{3}$. The same argument yields $\delta = \frac{\epsilon}{3C_{\mathcal{M}}}$. Analogously, we can choose $\Delta = \frac{r(1-r/\tau)\epsilon}{3c(\pi+1)C_{\mathcal{M}}}$. Finally, we pick $\nu = \frac{\Delta}{16B^2D}$ so that $8B^2D\nu < \Delta$.

Now we compute the number of layers, and the number of neurons and weight parameters in the ReLU network identified by Theorem 1.

1. For the chart determination sub-network, $\widehat{\mathbb{1}}_\Delta$ can be implemented by a single layer ReLU network. The approximation of the distance function $\widehat{d_i^2}$ can be implemented by a network of depth $O\left(\log \frac{1}{\nu}\right)$ and the number of neurons and weight parameters is at most $O\left(\log \frac{1}{\nu}\right)$. Plugging in our choice of ν, we have the depth is no greater than $c_1\left(\log \frac{1}{\epsilon} + \log D\right)$ with c_1 depending on d, f, τ, and the surface area of $\mathcal{M}$. The number of neurons and weight parameters is also $c_1'\left(\log \frac{1}{\epsilon} + \log D\right)$ except for a different constant. Note that there are D parallel networks computing $\widehat{d_i^2}$ for $i = 1, \dots, C_{\mathcal{M}}$. Hence, the total number of neurons and weight parameters is $c_1' C_{\mathcal{M}} D\left(\log \frac{1}{\epsilon} + \log D\right)$ with c_1' depending on d, f, τ, and the surface area of $\mathcal{M}$.
2. For the Taylor polynomial sub-network, ϕ_i can be implemented by a linear network with at most Dd weight parameters. To implement each $\widehat{f}_i$, we need a ReLU network of depth $c_4 \log \frac{1}{\delta}$. The number of neurons and weight parameters is $c_4' \delta^{-\frac{d}{n+\alpha}} \log \frac{1}{\delta}$. Here c_4, c_4' depend on $n, d, f_i \circ \phi_i^{-1}$. Substituting $\delta = \frac{\epsilon}{3C_{\mathcal{M}}}$, we get the depth is $c_2 \log \frac{1}{\epsilon}$ and the number of neurons and weight parameters is $c_2' \epsilon^{-\frac{d}{n+\alpha}} \log \frac{1}{\epsilon}$. There are totally $C_{\mathcal{M}}$ parallel $\widehat{f}_i$'s, hence the total number of neurons and weight parameters is $c_2' C_{\mathcal{M}} \epsilon^{-\frac{d}{n+\alpha}} \log \frac{1}{\epsilon}$ with c_2' depending on $d, n, f_i \circ \phi_i^{-1}, \tau$, and the surface area of $\mathcal{M}$.
3. For the product sub-network, the analysis is similar to the chart determination sub-network. The depth is $O\left(\log \frac{1}{\eta}\right)$, and the number of neurons and weight parameters is $O\left(\log \frac{1}{\eta}\right)$. The choice of η yields the depth is $c_3 \log \frac{1}{\epsilon}$, and the number of neurons and weight parameters is $c_3' \log \frac{1}{\epsilon}$. There are $C_{\mathcal{M}}$ parallel pairs of outputs from the chart determination and the Taylor polynomial sub-networks. Hence, the total number of weight parameters is $c_3' C_{\mathcal{M}} \log \frac{1}{\epsilon}$ with c_3' depending on d, τ, and the surface area of $\mathcal{M}$.

Combining these 3 sub-networks, we see the depth of the full network is $c\left(\log \frac{1}{\epsilon} + \log D\right)$ for some constant c depending on d, n, f, τ, and the surface area of $\mathcal{M}$. The total number of neurons and weight parameters is $c'\left(\epsilon^{-\frac{d}{n+\alpha}} \log \frac{1}{\epsilon} + D \log \frac{1}{\epsilon} + D \log D\right)$ for some constant c' depending on d, n, f, τ, and the surface area of $\mathcal{M}$. □