Nonparametric Regression on Low-Dimensional Manifolds using Deep ReLU Networks

MINSHUO CHEN
ISYE, Georgia Institute of Technology
Atlanta, GA 30332

HAOMING JIANG
ISYE, Georgia Institute of Technology

WENJING LIAO*
Mathematics, Georgia Institute of Technology
*Corresponding author: wliao60@gatech.edu

AND

TUO ZHAO
ISYE, Georgia Institute of Technology

Real world data often exhibit low-dimensional geometric structures, and can be viewed as samples near a low-dimensional manifold. This paper studies nonparametric regression of Hölder functions on low-dimensional manifolds using deep ReLU networks. Suppose n training data are sampled from a Hölder function in $\mathscr{H}^{s,\alpha}$ supported on a d-dimensional Riemannian manifold isometrically embedded in $\mathbb{R}^D$, with sub-gaussian noise. A deep ReLU network architecture is designed to estimate the underlying function from the training data. The mean squared error of the empirical estimator is proved to converge in the order of $n^{-\frac{2(s+\alpha)}{2(s+\alpha)+d}} \log^3 n$. This result shows that deep ReLU networks give rise to a fast convergence rate depending on the data intrinsic dimension d, which is usually much smaller than the ambient dimension D. It therefore demonstrates the adaptivity of deep ReLU networks to low-dimensional geometric structures of data, and partially explains the power of deep ReLU networks in tackling high-dimensional data with low-dimensional geometric structures.

Keywords: Nonparametric regression, Low-dimensional manifolds, Deep ReLU networks, Sample complexity, Uniform approximation theory

1. Introduction

Deep learning has made astonishing breakthroughs in various 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. For example, in 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. In 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.

The empirical success of deep learning has brought new challenges to the conventional wisdom of

machine learning. Data sets in these applications are high-dimensional and highly complex. In existing literature, a minimax lower bound has been established for the optimal algorithm of learning C^s functions in $\mathbb{R}^D$ (Györfi et al., 2006; Tsybakov, 2008). Suppose the underlying function is f_0. The minimax lower bound suggests a pessimistic sample complexity: To obtain an estimator $\widehat{f}$ for C^s functions with an ε-error uniformly (i.e., $\sup_{f_0 \in C^s} \|\widehat{f} - f_0\|_{L_2} \leqslant \varepsilon$ with $\|\cdot\|_{L_2}$ denoting the function L_2 norm), the optimal algorithm requires the sample size $n \gtrsim \varepsilon^{-\frac{2s+D}{s}}$ in the worst scenario (i.e., when f_0 is the most difficult for the algorithm to estimate). We instantiate such a sample complexity bound for the ImageNet data set, which consists of RGB images with a resolution of 224×224. The theory above suggests that, to achieve an ε-error, the number of samples has to scale as $\varepsilon^{-224\times 224\times 3/s}$, where the modulus of smoothness s is significantly smaller compared to $224 \times 224 \times 3$. Setting $\varepsilon = 0.1$ already gives rise to a huge number of samples far beyond practical applications, which well exceeds 1.2 million labeled images in ImageNet.

To bridge the aforementioned gap between theory and practice, we take the low-dimensional geometric structures in data sets into consideration. This is motivated by the fact that real-world data sets often exhibit low-dimensional structures. Many images consist of projections of a three-dimensional object followed by some transformations, such as rotation, translation, and skeleton. This generating mechanism induces a small number of intrinsic parameters (Hinton and Salakhutdinov, 2006; Osher et al., 2017). Speech data are composed of words and sentences following the grammar, and therefore have small degrees of freedom (Djuric et al., 2015). More broadly, visual, acoustic, textual, and many other types of data often have low-dimensional geometric structures due to rich local regularities, global symmetries, repetitive patterns, or redundant sampling (Tenenbaum et al., 2000; Roweis and Saul, 2000). It is therefore reasonable to assume that data lie on a manifold $\mathscr{M}$ of dimension $d \ll D$.

In this paper, we study nonparametric regression problems (Wasserman, 2006; Györfi et al., 2006; Tsybakov, 2008) using neural networks in exploitation of low-dimensional geometric structures of data. Specifically, we model data as samples from a probability measure supported on a d-dimensional manifold $\mathscr{M}$ isometrically embedded in $\mathbb{R}^D$ where $d \ll D$. The goal is to recover the regression function f_0 supported on $\mathscr{M}$ using the samples $S_n = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ with $\mathbf{x} \in \mathscr{M}$ and $y \in \mathbb{R}$. The $\mathbf{x}_i$'s are i.i.d. sampled from a distribution $\mathscr{D}_x$ on $\mathscr{M}$, and the response y_i satisfies

$$y_i = f_0(\mathbf{x}_i) + \xi_i,$$

where ξ_i's are i.i.d. sub-Gaussian noise independent of the $\mathbf{x}_i$'s.

We use multi-layer ReLU (Rectified Linear Unit) neural networks to recover f_0. ReLU networks are widely used in computer vision, speech recognition, natural language processing, etc. (Nair and Hinton, 2010; Glorot et al., 2011; Maas et al., 2013). These networks can ease the notorious vanishing gradient issue during training, which commonly arises with sigmoid or hyperbolic tangent activations (Glorot et al., 2011; Goodfellow et al., 2016). Given an input $\mathbf{x}$, an L-layer ReLU neural network computes an output as

$$f(\mathbf{x}) = W_L \cdot \mathrm{ReLU}(W_{L-1} \cdots \mathrm{ReLU}(W_1 \mathbf{x} + \mathbf{b}_1) \cdots + \mathbf{b}_{L-1}) + \mathbf{b}_L, \tag{1.1}$$

where $W_1, \ldots, W_L$ and $\mathbf{b}_1, \ldots, \mathbf{b}_L$ are weight matrices and vectors of proper sizes, respectively, and $\mathrm{ReLU}(\cdot)$ denotes the entrywise rectified linear unit activation (i.e., $\mathrm{ReLU}(a) = \max\{0, a\}$). We fur-

ther denote $\mathscr{F}$ as a class of neural networks with bounded weight parameters and bounded output:

$$\mathscr{F}(R,\kappa,L,p,K) = \Big\{ f \mid f(\mathbf{x}) \text{ in the form (1.1) with } L\text{-layers and width bounded by } p, \\ \|f\|_\infty \leqslant R, \|W_i\|_{\infty,\infty} \leqslant \kappa, \|\mathbf{b}_i\|_\infty \leqslant \kappa \text{ for } i=1,\dots,L, \sum_{i=1}^{L} \|W_i\|_0 + \|\mathbf{b}_i\|_0 \leqslant K \Big\},$$

where $\|\cdot\|_0$ denotes the number of nonzero entries in a vector or a matrix, $\|\cdot\|_\infty$ denotes ℓ_∞ norm of a function or entrywise ℓ_∞ norm of a vector, and for a matrix M, $\|M\|_{\infty,\infty} = \max_{i,j} |M_{ij}|$.

To obtain an estimator $\widehat{f} \in \mathscr{F}(R,\kappa,L,p,K)$ of f_0, we minimize the empirical quadratic risk

$$\widehat{f}_n = \underset{f\in\mathscr{F}(R,\kappa,L,p,K)}{\operatorname{argmin}} \widehat{\mathscr{R}}_n(f) = \underset{f\in\mathscr{F}(R,\kappa,L,p,K)}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} (f(\mathbf{x}_i) - y_i)^2 . \tag{1.2}$$

The subscript n emphasizes that the estimator is obtained using n pairs of samples. Our theory shows that $\widehat{f}_n$ enjoys a fast rate of convergence to f_0, depending on the intrinsic dimension d. Let $\mathscr{M}$ be a d-dimensional compact Riemannian manifold isometrically embedded in $\mathbb{R}^D$ with $d \ll D$. Assume $\mathscr{M}$ satisfies some mild regularity conditions. For simplicity, we suppose f_0 is a C^s function on $\mathscr{M}$. Extensions to Hölder and Sobolev functions are given in Theorem 3.1. For the network class $\mathscr{F}(R,\kappa,L,p,K)$, we choose

$$L = \widetilde{O}\left(\frac{2s+d}{2s}\log n\right), \quad p = \widetilde{O}\left(n^{\frac{d}{2s+d}}\right), \quad K = \widetilde{O}\left(\frac{2s+d}{2s} n^{\frac{d}{2s+d}} \log n\right), \quad R = \|f_0\|_\infty ,$$

and set κ as a constant depending on s and $\mathscr{M}$. We use $\widetilde{O}$ to hide logarithmic factors (e.g., $\log D$) and polynomial factors in s and d. Let $\widehat{f}_n$ be the empirical minimizer of (1.2). Then we have

$$\mathbb{E}\left[\int_{\mathscr{M}} \left(\widehat{f}_n(\mathbf{x}) - f_0(\mathbf{x})\right)^2 d\mathscr{D}_x(\mathbf{x})\right] \leqslant c(R^2+\sigma^2) n^{-\frac{2s}{2s+d}} \log^3 n,$$

where the expectation is taken over the training samples S_n, σ^2 is the variance of the ξ_i's, and c is a constant depending on $\log D$, s, κ, and $\mathscr{M}$.

Our theory implies that, in order to estimate a C^s regression function up to an ε-error, the sample complexity is $n \gtrsim \varepsilon^{-\frac{2s+d}{s}}$ up to a log factor. This sample complexity depends on the intrinsic dimension d, and thus largely improves on existing theories of nonparametric regression using neural networks, where the sample complexity scales as $\widetilde{O}(\varepsilon^{-\frac{2s+D}{s}})$ (Hamers and Kohler, 2006; Kohler and Krzyżak, 2005, 2016; Kohler and Mehnert, 2011; Schmidt-Hieber, 2017). Our result partially explains the success of deep ReLU neural networks in tackling complext high-dimensional data with low-dimensional geometric structures, e.g., images and speech data.

A key ingredient in our analysis is an efficient universal approximation theory of deep ReLU networks for C^s functions on $\mathscr{M}$ (Theorem 3.2), which appeared in Chen et al. (2019). Specifically, we show that, in order to uniformly approximate C^s functions on a d-dimensional manifold with an ε-error, the network consists of at most $\widetilde{O}(\log 1/\varepsilon + \log D)$ layers and $\widetilde{O}(\varepsilon^{-d/s}\log 1/\varepsilon + D\log 1/\varepsilon + D\log D)$ neurons and weight parameters. Similar results also hold for functions in Hölder and Sobolev spaces (see Theorem 3.2). The network size in our approximation theory only weakly depends on data dimension D, which significantly improves existing universal approximation theories of neural networks (Barron, 1993; Mhaskar, 1996; Lu et al., 2017; Hanin, 2017; Yarotsky, 2017), where the network size scales as

$\widetilde{O}(\varepsilon^{-D/s})$. Figure 1 illustrates a huge gap between the network sizes used in practice (Tan and Le, 2019) and the required size predicted by existing theories (Yarotsky, 2017) for the ImageNet data set. Our network size also matches the lower bound up to logarithmic factors for a given manifold $\mathscr{M}$ (see Theorem 3.3). Our approximation theory partially justifies why networks of moderate size have achieved a great success in various applications.

1.1 *Related Work*

Nonparametric regression has been widely studied in statistics. A variety of methods has been proposed to estimate the regression function, including kernel method, wavelets, splines, and local polynomials (Wahba, 1990; Altman, 1992; Fan and Gijbels, 1996; Tsybakov, 2008; Györfi et al., 2006). Nonetheless, there is limited study on regression using deep ReLU networks until recently. The earliest works focused on neural networks with a single hidden layer and smooth activations (e.g., sigmoidal and sinusoidal functions, Barron (1991); McCaffrey and Gallant (1994)). Later results achieved the minimax lower bound for the mean squared error in the order of $O(n^{-\frac{2s}{2s+D}})$ up to a logarithmic factor for C^s functions in $\mathbb{R}^D$ (Hamers and Kohler, 2006; Kohler and Krzyżak, 2005, 2016; Kohler and Mehnert, 2011). Theories for deep ReLU networks were developed in Schmidt-Hieber (2017), where the estimate matches the minimax lower bound up to a logarithmic factor for Hölder functions. Extensions to more general function spaces, such as Besov spaces, can be found in (Suzuki, 2019) and results for classification problems can be found in Kim et al. (2018); Ohn and Kim (2019).

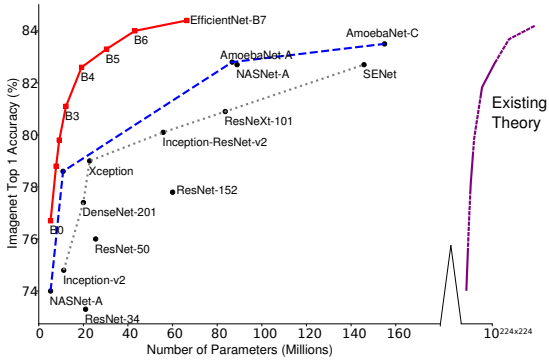

FIG. 1. Practical network sizes for the ImageNet data set (Tan and Le, 2019) versus the required size predicted by existing theories (Yarotsky, 2017).

The rate of convergence in the results above are insufficient to understand the success of deep learning due to the curse of data dimension with a large D. Fortunately, many real-world data sets exhibit low-dimensional geometric structures. It has been demonstrated that, some classical methods can automatically adapt to the low-dimensional structures of data, and perform as well as if the low-dimensional structures were known. Results in this direction include local linear regression (Bickel and Li, 2007), k-nearest neighbor (Kpotufe, 2011), kernel regression (Kpotufe and Garg, 2013), and Bayesian Gaussian process regression (Yang et al., 2015), where optimal rates depending on the intrinsic dimension were proved for functions having the second order of continuity (Bickel and Li, 2007), globally Lipschitz functions (Kpotufe, 2011), and Hölder functions with Hölder index no more than 1 (Kpotufe and Garg, 2013). However, it was not clear if deep ReLU networks are adaptable to low-dimensional geometric structures of data.

A crucial ingredient in the statistical analysis of neural networks is the universal approximation ability of neural networks. Early results justified the existence of two-layer networks with continuous sigmoidal activations (a function $\sigma(x)$ is sigmoidal, if $\sigma(x) \to 0$ as $x \to -\infty$, and $\sigma(x) \to 1$ as $x \to \infty$) for a universal approximation of continuous functions in a unit hypercube (Irie and Miyake, 1988; Funahashi, 1989; Cybenko, 1989; Hornik, 1991; Chui and Li, 1992; Leshno et al., 1993). In these works, the number of neurons was not explicitly given. Later, Barron (1993); Mhaskar (1996) proved that the

number of neurons can grow as $\varepsilon^{-D/2}$ where ε is the uniform approximation error. Recently, Lu et al. (2017), Hanin (2017) and Daubechies et al. (2019) extended the universal approximation theorem to networks of bounded width with ReLU activations. The depth of such networks has to grow exponentially with respect to the dimension of data. Yarotsky (2017) showed that ReLU neural networks can uniformly approximate functions in Sobolev spaces, where the network size scales exponentially with respect to the data dimension and matches the lower bound. Zhou (2019) also developed a universal approximation theory for deep convolutional neural networks (Krizhevsky et al., 2012), where the depth of the network scales exponentially with respect to the data dimension.

The aforementioned results focus on functions on a compact subset (e.g., $[0,1]^D$) in $\mathbb{R}^D$. Function approximation on manifolds has been well studied using classical methods, such as local polynomials (Bickel et al., 2007) and wavelets (Coifman and Maggioni, 2006). However, studies using neural networks are 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 ($\max^2\{0,x\}$). These smooth activations are rarely used in 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. This theory does not cover arbitrarily C^s functions. We are also aware of a concurrent work of ours, Shen et al. (2019), which established an approximation theory of ReLU networks for Hölder functions in terms of a modulus of continuity. When the target function belongs to $\mathscr{H}^{s,\alpha}$ and is supported in a neighborhood of a d-dimensional manifold embedded in $\mathbb{R}^D$, Shen et al. (2019) constructed a ReLU network which yields approximation error in the order of $N^{-2\min(s+\alpha,1)/d_\delta}L^{-2\min(s+\alpha,1)/d_\delta}$ where N and L are the width and depth of the network, and $d < d_\delta < D$. Their proof utilizes a different approach compared to ours: They first construct a piecewise constant function to approximate the target function, and then implement the piecewise constant function using a ReLU network. Unfortunately, the higher order smoothness (while $s+\alpha > 1$) is not exploited due to the use of piecewise constant approximations.

1.2 *Roadmap and Notations*

The rest of the paper is organized as follows: Section 2 presents a brief introduction to manifolds and functions on manifolds. Section 3 presents the main theory of efficient statistical recovery using deep ReLU neural networks on low-dimensional manifolds, and a new universal approximation theory of ReLU networks; Section 4 sketches the proofs of theories in Section 3, and the detailed proofs are deferred to the appendix; Section 5 provides a conclusion of the paper and discusses open questions and future directions.

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 $\mathbf{s} = [s_1,\ldots,s_d]^\top \in \mathbb{N}^d$, we define $\mathbf{s}! = \prod_{i=1}^d s_i!$ and $|\mathbf{s}| = \sum_{i=1}^d s_i$. We define $\mathbf{x}^\mathbf{s} = \prod_{i=1}^d x_i^{s_i}$. Given a function $f : \mathbb{R}^d \mapsto \mathbb{R}$, we denote its derivative as $D^\mathbf{s} f = \frac{\partial^{|\mathbf{s}|} f}{\partial x_1^{s_1}\ldots\partial x_d^{s_d}}$, 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 $\mathscr{M}$ be a d-dimensional Riemannian manifold isometrically embedded in $\mathbb{R}^D$.

DEFINITION 2.1 (Chart) A chart for $\mathscr{M}$ is a pair (U,ϕ) such that $U \subset \mathscr{M}$ 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 $\mathscr{M}$. We say two charts (U,ϕ) and (V,ψ) on $\mathscr{M}$ are C^k compatible if and only if the transition functions,

$$\phi \circ \psi^{-1} : \psi(U \cap V) \mapsto \phi(U \cap V) \quad \text{and} \quad \psi \circ \phi^{-1} : \phi(U \cap V) \mapsto \psi(U \cap V)$$

are both C^k. Then we give the definition of an atlas.

DEFINITION 2.2 (C^k Atlas) An atlas for $\mathscr{M}$ is a collection $\{(U_\alpha,\phi_\alpha)\}_{\alpha\in\mathscr{A}}$ of pairwise C^k compatible charts such that $\bigcup_{\alpha\in\mathscr{A}} U_\alpha = \mathscr{M}$.

DEFINITION 2.3 (Smooth Manifold) A smooth manifold is a manifold $\mathscr{M}$ 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 $\mathscr{M}$ allows us to define differentiable functions.

DEFINITION 2.4 (C^s Functions on $\mathscr{M}$) Let $\mathscr{M}$ be a smooth manifold in $\mathbb{R}^D$. A function $f : \mathscr{M} \mapsto \mathbb{R}$ is C^s if for any chart (U,ϕ), the composition $f \circ \phi^{-1} : \phi(U) \mapsto \mathbb{R}$ is continuously differentiable up to order s.

REMARK 2.1 The definition of C^s 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 $\mathscr{M}$ is a smooth manifold, (U,ϕ) and (V,ψ) are C^∞ compatible. Thus, $f \circ \phi^{-1}$ is C^s and $\phi \circ \psi^{-1}$ is C^∞, and their composition is C^s.

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

DEFINITION 2.5 (Partition of Unity) A C^∞ partition of unity on a manifold $\mathscr{M}$ is a collection of non-negative C^∞ functions $\rho_\alpha : \mathscr{M} \mapsto \mathbb{R}_+$ for $\alpha \in \mathscr{A}$ such that

1. the collection of supports, $\{\text{supp}(\rho_\alpha)\}_{\alpha\in\mathscr{A}}$ is locally finite, i.e., every point on $\mathscr{M}$ has a neighborhood that meets only finitely many of A_α's;
2. $\sum \rho_\alpha = 1$.

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

PROPOSITION 2.6 (Existence of a C^∞ partition of unity) Let $\{U_\alpha\}_{\alpha\in\mathscr{A}}$ be an open cover of a smooth manifold $\mathscr{M}$. Then there is a C^∞ partition of unity $\{\rho_i\}_{i=1}^\infty$ with every ρ_i having a compact support such that $\text{supp}(\rho_i) \subset U_\alpha$ for some $\alpha \in \mathscr{A}$.

Proposition 2.6 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 to control the bias of estimating f boils down to the approximation of 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 2.7 (Reach, Definition 2.1 in Aamari et al. (2019)) Denote

$$\mathscr{C}(\mathscr{M}) = \left\{ \mathbf{x} \in \mathbb{R}^D : \exists \mathbf{p} \neq \mathbf{q} \in \mathscr{M}, \|\mathbf{p}-\mathbf{x}\|_2 = \|\mathbf{q}-\mathbf{x}\|_2 = \inf_{\mathbf{y}\in\mathscr{M}} \|\mathbf{y}-\mathbf{x}\|_2 \right\}$$

as the set of points that have at least two nearest neighbors on $\mathscr{M}$. The reach $\tau > 0$ is defined as

$$\tau = \inf_{\mathbf{x}\in\mathscr{M}, \mathbf{y}\in\mathscr{C}(\mathscr{M})} \|\mathbf{x}-\mathbf{y}\|_2 .$$

Reach has a straightforward geometrical interpretation: At each point $\mathbf{x} \in \mathscr{M}$, the radius of the oscu-

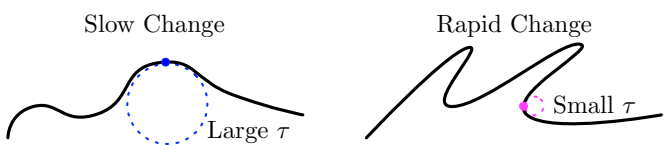

FIG. 2. Manifolds with large and small reaches.

lating circle is greater or equal to τ. A large reach for $\mathscr{M}$ essentially requires the manifold $\mathscr{M}$ not to change "rapidly" as shown in Figure 2.

Reach determines a proper choice of an atlas for $\mathscr{M}$. 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, reach of a smooth manifold reflects the complexity of the neural network for function approximation on $\mathscr{M}$.

3. Main Results – Statistical Recovery Theory

This section contains our main recovery theory of learning regression models on low dimensional manifolds using deep neural networks. We begin with some assumptions on the regression model and the manifold $\mathscr{M}$.

Assumption 1. $\mathscr{M}$ is a d-dimensional compact Riemannian manifold isometrically embedded in $\mathbb{R}^D$. There exists a constant $B > 0$ such that, for any point $\mathbf{x} \in \mathscr{M}$, we have $|x_i| \leqslant B$ for all $i = 1,\dots,D$.

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

Assumption 3. The ground truth function $f_0 : \mathscr{M} \mapsto \mathbb{R}$ belongs to the Hölder space $\mathscr{H}^{s,\alpha}(\mathscr{M})$ with a positive integer s and $\alpha \in (0,1]$, in the sense that for any chart (U,ϕ), we have

1. $f_0 \circ \phi^{-1} \in C^{s-1}$ with $|D^{\mathbf{s}} f_0 \circ \phi^{-1}| \leqslant 1$ for any $|\mathbf{s}| < s$, $\mathbf{x} \in U$;
2. for any $|\mathbf{s}| = s$ and $\mathbf{x}_1, \mathbf{x}_2 \in U$,

$$\left| D^{\mathbf{s}}(f_0 \circ \phi^{-1})\big|_{\phi(\mathbf{x}_1)} - D^{\mathbf{s}}(f_0 \circ \phi^{-1})\big|_{\phi(\mathbf{x}_2)} \right| \leqslant \|\phi(\mathbf{x}_1) - \phi(\mathbf{x}_2)\|_2^{\alpha} . \tag{3.1}$$

Assumption 3 requires that all s-th order derivatives of $f_0 \circ \phi^{-1}$ are Hölder continuous. 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.

The following theorem characterizes the convergence rate for the estimation of f_0 using ReLU neural networks. For simplicity, we state the result for Hölder functions. Extensions to Sobolev spaces are straightforward.

THEOREM 3.1 Suppose Assumptions 1 - 3 hold. Let $\widehat{f}_n$ be the minimizer of empirical risk (1.2) with the network class $\mathscr{F}(R,\kappa,L,p,K)$ properly designed such that

$$L = \widetilde{O}\left(\frac{2(s+\alpha)+d}{2(s+\alpha)}\log n\right), \quad p = \widetilde{O}\left(n^{\frac{d}{2(s+\alpha)+d}}\right), \quad K = \widetilde{O}\left(\frac{2(s+\alpha)+d}{2(s+\alpha)} n^{\frac{d}{2(s+\alpha)+d}} \log n\right),$$
$$R = \|f_0\|_\infty, \quad \text{and } \kappa = \max\{1, B, \sqrt{d}, \tau^2\}.$$

Then we have

$$\mathbb{E}\left[\int_{\mathscr{M}} \left(\widehat{f}_n(\mathbf{x}) - f_0(\mathbf{x})\right)^2 d\mathscr{D}_x(\mathbf{x})\right] \leqslant c(R^2+\sigma^2) n^{-\frac{2(s+\alpha)}{2(s+\alpha)+d}} \log^3 n,$$

where the expectation is taken over the training samples S_n, and c is a constant depending on $\log D$, s, κ, and $\mathscr{M}$.

Theorem 3.1 is established by a bias-variance trade-off (see its proof in Section 4). Here are some remarks:

1. The network class in Theorem 3.1 is sparsely connected, i.e. $K = O(Lp)$, while densely connected networks satisfy $K = O(Lp^2)$.

2. The network class $\mathscr{F}(R,\kappa,L,p,K)$ has outputs uniformly bounded by R. Such a requirement can be achieved by appending an additional clipping layer to the end of the network structure, i.e.,

$$g(a) = \max\{-R, \min\{a, R\}\} = \mathrm{ReLU}(a-R) - \mathrm{ReLU}(a+R) - R.$$

3. Each weight parameter in our network class is bounded by a constant κ only depending on the curvature τ, the range B of the manifold $\mathscr{M}$ and the smoothness s of f_0. Such boundedness condition is crucial to our theory and can be computationally realized by normalization after each step of the stochastic gradient descent.

Theorem 3.1 quantifies the network size for learning f_0. A natural question is: How is the network structure properly designed? The answer is given by the following universal approximation theory of ReLU networks for Hölder functions supported on the manifold $\mathscr{M}$.

THEOREM 3.2 Suppose Assumptions 1 and 2 hold. Given any $\varepsilon \in (0,1)$, there exists a ReLU network structure such that, for any $f : \mathscr{M} \to \mathbb{R}$ satisfying Assumption 3, if the weight parameters are properly chosen, the network yields a function $\widetilde{f}$ satisfying $\|\widetilde{f} - f\|_\infty \leqslant \varepsilon$. Such a network has

1. no more than $c_1(\log \frac{1}{\varepsilon} + \log D)$ layers,

2. at most $c_2(\varepsilon^{-\frac{d}{s+\alpha}} \log \frac{1}{\varepsilon} + D \log \frac{1}{\varepsilon} + D \log D)$ neurons and weight parameters,

where c_1, c_2 depend on d, s, f, τ, and the surface area of $\mathscr{M}$.

The network structure identified by Theorem 3.2 consists of three sub-networks as shown in Figure 3 (The detailed construction of each sub-network is postponed to Section 4):

• *Chart determination sub-network*, which assigns each input to its corresponding neighborhoods;

• *Taylor approximation sub-network*, which approximates f by polynomials in each neighborhood;

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

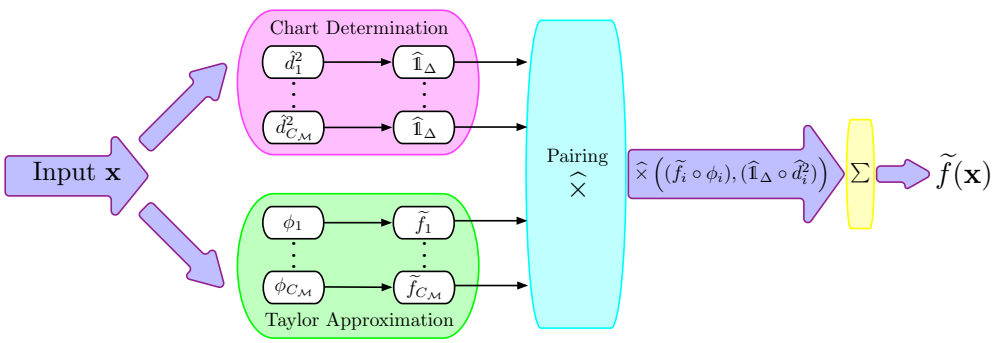

FIG. 3. The ReLU network identified by Theorem 3.2.

Our result significantly improves existing approximation theories Yarotsky (2017) where the network size grows exponentially with respect to the ambient dimension D, i.e. $\varepsilon^{-D/(s+\alpha)}$. Theorem 3.2 also improves Shaham et al. (2018) for C^s functions in the case that $s > 2$. When $s > 2$, our network size scales like $\varepsilon^{-d/s}$, which is significantly smaller than the one in Shaham et al. (2018) in the order of $\varepsilon^{-d/2}$.

Moreover, the size of our ReLU network in Theorem 3.2 matches its lower bound in DeVore et al. (1989) up to a logarithmic factor for the approximation of functions in the Hölder space $\mathscr{H}^{s-1,1}([0,1]^d)$ defined on $[0,1]^d$.

THEOREM 3.3 Fix d and s. Let W be a positive integer and $\mathscr{T} : \mathbb{R}^W \mapsto C([0,1]^d)$ be any mapping. Suppose there is a continuous map $\Theta : \mathscr{H}^{s-1,1}([0,1]^d) \mapsto \mathbb{R}^W$ such that $\|f - \mathscr{T}(\Theta(f))\|_\infty \leqslant \varepsilon$ for any $f \in \mathscr{H}^{s-1,1}([0,1]^d)$. Then $W \geqslant c\varepsilon^{-\frac{d}{s}}$ with c depending on s only.

We take $\mathbb{R}^W$ as the parameter space of a ReLU network, and $\mathscr{T}$ as the transformation given by the ReLU network. Theorem 3.3 implies that, to approximate any $f \in \mathscr{H}^{s-1,1}([0,1]^d)$, the ReLU network needs at least $c\varepsilon^{-\frac{d}{s}}$ weight parameters. Although Theorem 3.3 holds for functions defined 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.

On the other hand, the lower bound also reveals that the low-dimensional manifold model plays a vital role to reduce the network size. To approximate functions in $\mathscr{H}^{s-1,1}([0,1]^D)$ with accuracy ε, the minimal number of weight parameters is $O(\varepsilon^{-\frac{D}{s}})$. This lower bound cannot be improved without low-dimensional structures of data.

4. Proof of the Statistical Recovery Theory

This section contains the proof sketch of Theorem 3.1 and Theorem 3.2.

4.1 *Proof of Theorem 3.1*

We dilate the L_2 risk of $\widehat{f}_n$ using its empirical counterpart as

$$\mathbb{E}\left[\int_{\mathscr{M}}\left(\widehat{f}_n(\mathbf{x})-f_0(\mathbf{x})\right)^2 d\mathscr{D}_x(\mathbf{x})\right] = \underbrace{2\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(\widehat{f}_n(\mathbf{x}_i)-f_0(\mathbf{x}_i))^2\right]}_{T_1}$$
$$+\underbrace{\mathbb{E}\left[\int_{\mathscr{M}}\left(\widehat{f}_n(\mathbf{x})-f_0(\mathbf{x})\right)^2 d\mathscr{D}_x(\mathbf{x})\right]-2\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(\widehat{f}_n(\mathbf{x}_i)-f_0(\mathbf{x}_i))^2\right]}_{T_2}.$$

The proof of Theorem 3.1 contains two parts for the estimates of T_1 and T_2, respectively.

Bounding T_1. The first term T_1 reflects the bias of the estimation, and can be bounded through the following Lemma.

LEMMA 4.1 Fix the neural network class $\mathscr{F}(R,\kappa,L,p,K)$. For any constant $\delta\in(0,2R)$, we have

$$T_1 \leqslant 4\inf_{f\in\mathscr{F}(R,\kappa,L,p,K)}\int_{\mathscr{M}}(f(\mathbf{x})-f_0(\mathbf{x}))^2 d\mathscr{D}_x(\mathbf{x})+\frac{64\sigma^2\log\mathscr{N}(\delta,\mathscr{F}(R,\kappa,L,p,K),\|\cdot\|_\infty)}{n}+32\sigma\delta,$$

where $\mathscr{N}(\delta,\mathscr{F}(R,\kappa,L,p,K),\|\cdot\|_\infty)$ denotes the δ-covering number of $\mathscr{F}(R,\kappa,L,p,K)$ with respect to the ℓ_∞ norm.

Proof Sketch. The detailed proof is provided in Appendix C.1. We decompose T_1 into two parts:

$$\begin{aligned}
T_1 &= 2\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(\widehat{f}_n(\mathbf{x}_i)-f_0(\mathbf{x}_i)-\xi_i+\xi_i)^2\right]\\
&= 2\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(\widehat{f}_n(\mathbf{x}_i)-y_i+\xi_i)^2\right]\\
&\leqslant 2\mathbb{E}\left[\inf_{f\in\mathscr{F}(R,\kappa,L,p,K)}\frac{1}{n}\sum_{i=1}^{n}(f(\mathbf{x}_i)-y_i)^2+\frac{2}{n}\xi_i\widehat{f}_n(\mathbf{x}_i)\right]\\
&= \underbrace{2\inf_{f\in\mathscr{F}(R,\kappa,L,p,K)}\int_{\mathscr{M}}(f(\mathbf{x})-f_0(\mathbf{x}))^2 d\mathscr{D}_x(\mathbf{x})}_{(A)}+\underbrace{4\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\xi_i\widehat{f}_n(\mathbf{x}_i)\right]}_{(B)},
\end{aligned}$$

where the inequality is derived using the independence between the noise ξ_i and response y_i, and $\widehat{f}_n$ being the empirical risk minimizer.

As can be seen, term (A) is the smallest L_2 risk achieved by the network class $\mathscr{F}(R,\kappa,L,p,K)$, which can be quantified using our approximation theory (Theorem 3.2). Term (B) is a complexity measure of the network class $\mathscr{F}(R,\kappa,L,p,K)$. To upper bound (B), we discretize $\mathscr{F}(R,\kappa,L,p,K)$ as

$\{f_i^*\}_{i=1}^{\mathcal{N}(\delta,\mathcal{F}(R,\kappa,L,p,K),\|\cdot\|_\infty)}$. By the definition of covering, there exists f^* such that $\|\widehat{f}_n - f^*\|_\infty \leqslant \delta$. Denote $\|f - f_0\|_n = \frac{1}{n}\sum_{i=1}^n (f(\mathbf{x}_i) - f_0(\mathbf{x}_i))^2$, we cast (B) into

$$\begin{aligned}
(B) &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \xi_i(\widehat{f}_n(\mathbf{x}_i) - f^*(\mathbf{x}_i) + f^*(\mathbf{x}_i) - f_0(\mathbf{x}_i))\right] \\
&\overset{(i)}{\leqslant} \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \xi_i(f^*(\mathbf{x}_i) - f_0(\mathbf{x}_i))\right] + \delta\sigma \\
&= \mathbb{E}\left[\frac{\|f^* - f_0\|_n}{\sqrt{n}} \frac{\sum_{i=1}^n \xi_i(f^*(\mathbf{x}_i) - f_0(\mathbf{x}_i))}{\sqrt{n}\|f^* - f_0\|_n}\right] + \delta\sigma \\
&\overset{(ii)}{\leqslant} \sqrt{2}\mathbb{E}\left[\frac{\|\widehat{f}_n - f_0\|_n + \delta}{\sqrt{n}} \frac{\sum_{i=1}^n \xi_i(f^*(\mathbf{x}_i) - f_0(\mathbf{x}_i))}{\sqrt{n}\|f^* - f_0\|_n}\right] + \delta\sigma,
\end{aligned}$$

where (i) follows from Hölder's inequality and (ii) is obtained by some algebraic manipulation. To break the dependence between f^* and the samples, we replace f^* by any f_j^* in the δ-covering and observe that $\frac{\sum_{i=1}^n \xi_i(f^*(\mathbf{x}_i)-f_0(\mathbf{x}_i))}{\sqrt{n}\|f^*-f_0\|_n} \leqslant \max_j \frac{\sum_{i=1}^n \xi_i(f_j^*(\mathbf{x}_i)-f_0(\mathbf{x}_i))}{\sqrt{n}\|f_j^*-f_0\|_n}$. Note that $\frac{\sum_{i=1}^n \xi_i(f_j^*(\mathbf{x}_i)-f_0(\mathbf{x}_i))}{\sqrt{n}\|f_j^*-f_0\|_n}$ is a sub-Gaussian random variable with parameter σ^2 for given $\mathbf{x}_i$'s. It is well established in the literature on empirical processes (van der Vaart et al., 1996) that maximum of a collection of sub-Gaussian random variables satisfies

$$\mathbb{E}\left[\max_j \frac{\sum_{i=1}^n \xi_i(f_j^*(\mathbf{x}_i) - f_0(\mathbf{x}_i))}{\sqrt{n}\|f_j^* - f_0\|_n}\right] \leqslant 2\sigma\sqrt{\log \mathcal{N}(\delta, \mathcal{F}(R,\kappa,L,p,K), \|\cdot\|_\infty)}.$$

Substituting the above inequality into (B) and combining (A) and (B), we have

$$\begin{aligned}
T_1 = 2\mathbb{E}\left[\|\widehat{f}_n - f_0\|_n^2\right] &\leqslant 2 \inf_{f\in\mathcal{F}(R,\kappa,L,p,K)} \int_{\mathcal{M}} (f(\mathbf{x}) - f_0(\mathbf{x}))^2 d\mathcal{D}_x(\mathbf{x}) \\
&\quad + 8\sqrt{2}\sigma\left(\mathbb{E}\left[\|\widehat{f}_n - f_0\|_n\right] + \delta\right)\sqrt{\frac{\log \mathcal{N}(\delta, \mathcal{F}(R,\kappa,L,p,K), \|\cdot\|_\infty)}{n}} + 4\delta\sigma
\end{aligned}$$

Some algebra further gives rise to the desired result

$$T_1 \leqslant 4 \inf_{f\in\mathcal{F}(R,\kappa,L,p,K)} \int_{\mathcal{M}} (f(\mathbf{x}) - f_0(\mathbf{x}))^2 d\mathcal{D}_x(\mathbf{x}) + \frac{64\sigma^2 \log \mathcal{N}(\delta, \mathcal{F}(R,\kappa,L,p,K), \|\cdot\|_\infty)}{n} + 32\sigma\delta.$$

□

Bounding T_2. We observe that T_2 is the difference between the population L_2 risk of $\widehat{f}_n$ and its empirical counterpart. However, bounding such a difference is distinct from conventional concentration results due to the scaling factor 2 before the empirical risk.

LEMMA 4.2 For any constant $\delta \in (0, 2R)$, T_2 satisfies

$$T_2 \leqslant \frac{52R^2}{3n} \log \mathcal{N}(\delta/4R, \mathcal{F}(R,\kappa,L,p,K), \|\cdot\|_\infty) + \left(4 + \frac{1}{2R}\right)\delta.$$

Proof Sketch. The detailed proof is deferred to Appendix C.2. For notational simplicity, we denote $g(\mathbf{x}) = (\widehat{f}_n(\mathbf{x}) - f_0(\mathbf{x}))^2$ and $\|g\|_\infty \leqslant 4R^2$. Applying the inequality $\int_{\mathcal{M}} g^2 d\mathcal{D}_x(\mathbf{x}) \leqslant 4R^2 \int_{\mathcal{M}} g d\mathcal{D}_x(\mathbf{x})$

(Barron, 1991), we rewrite T_2 as

$$\begin{aligned}
T_2 &= \mathbb{E}\left[\int_{\mathscr{M}} g(\mathbf{x})d\mathscr{D}_x(\mathbf{x}) - \frac{2}{n}\sum_{i=1}^{n} g(\mathbf{x}_i)\right] \\
&= 2\mathbb{E}\left[\int_{\mathscr{M}} g(\mathbf{x})d\mathscr{D}_x(\mathbf{x}) - \frac{1}{n}\sum_{i=1}^{n} g(\mathbf{x}_i) - \frac{1}{2}\int_{\mathscr{M}} g(\mathbf{x})d\mathscr{D}_x(\mathbf{x})\right] \\
&\leqslant 2\mathbb{E}\left[\int_{\mathscr{M}} g(\mathbf{x})d\mathscr{D}_x(\mathbf{x}) - \frac{1}{n}\sum_{i=1}^{n} g(\mathbf{x}_i) - \frac{1}{8R^2}\int_{\mathscr{M}} g^2(\mathbf{x})d\mathscr{D}_x(\mathbf{x})\right].
\end{aligned}$$

We now utilize the symmetrization technique in existing literature on nonparametric statistics (van der Vaart et al., 1996; Györfi et al., 2006). Specifically, let $\bar{\mathbf{x}}_i$'s be independent replications of $\mathbf{x}_i$'s and U_i be i.i.d. Rademacher random variables, i.e., $\mathbb{P}(U_i = 1) = \mathbb{P}(U_i = -1) = 1/2$. We bound T_2 as

$$\begin{aligned}
T_2 &\leqslant 2\mathbb{E}\left[\sup_{g\in\mathscr{G}}\int_{\mathscr{M}} g(\mathbf{x})d\mathscr{D}_x(\mathbf{x}) - \frac{1}{n}\sum_{i=1}^{n} g(\mathbf{x}_i) - \frac{1}{8R^2}\int_{\mathscr{M}} g^2(\mathbf{x})d\mathscr{D}_x(\mathbf{x})\right] \\
&\leqslant 2\mathbb{E}_{\bar{\mathbf{x}},\mathbf{x}}\left[\sup_{g\in\mathscr{G}}\frac{1}{n}\sum_{i=1}^{n}(g(\bar{\mathbf{x}}_i) - g(\mathbf{x}_i)) - \frac{1}{8R^2}\int_{\mathscr{M}} g^2(\mathbf{x})d\mathscr{D}_x(\mathbf{x})\right] \\
&\leqslant 2\mathbb{E}_{\bar{\mathbf{x}},\mathbf{x},U}\left[\sup_{g\in\mathscr{G}}\frac{1}{n}\sum_{i=1}^{n} U_i(g(\bar{\mathbf{x}}_i) - g(\mathbf{x}_i)) - \frac{1}{16R^2}\mathbb{E}_{\mathbf{x},\bar{\mathbf{x}}}\left[g^2(\mathbf{x}) + g^2(\bar{\mathbf{x}})\right]\right],
\end{aligned}$$

where $\mathscr{G} = \{g = (f - f_0)^2 \mid f \in \mathscr{F}(R,\kappa,L,p,K)\}$ and $\mathbb{E}_{\mathbf{x}}$ denotes the expectation with respect to $\mathbf{x}$. Note here $g^2(\mathbf{x}) + g^2(\bar{\mathbf{x}})$ contributes as the variance term of $U_i(g(\bar{\mathbf{x}}_i) - g(\mathbf{x}_i))$, which yields a fast convergence of T_2 as n grows.

Similar to bounding T_1, we discretize the function space $\mathscr{G}$ using a δ-covering denoted by $\mathscr{G}^*$. This allows us to replace the supremum by the maximum over a finite set:

$$T_2 \leqslant 2\mathbb{E}_{\bar{\mathbf{x}},\mathbf{x},U}\left[\sup_{g^*\in\mathscr{G}^*}\frac{1}{n}\sum_{i=1}^{n} U_i(g^*(\bar{\mathbf{x}}_i) - g^*(\mathbf{x}_i)) - \frac{1}{16R^2}\mathbb{E}_{\mathbf{x},\bar{\mathbf{x}}}\left[(g^*)^2(\mathbf{x}) + (g^*)^2(\bar{\mathbf{x}})\right]\right] + \left(4 + \frac{1}{2R}\right)\delta.$$

We can bound the above maximum by the Bernstein's inequality, which yields

$$T_2 \leqslant \frac{52R^2}{3n}\log\mathscr{N}(\delta,\mathscr{G},\|\cdot\|_\infty) + \left(4 + \frac{1}{2R}\right)\delta$$

The last step is to relate the covering number of $\mathscr{G}$ to that of $\mathscr{F}(R,\kappa,L,p,K)$. Specifically, consider any $g_1, g_2 \in \mathscr{G}$ with $g_1 = (f_1 - f_0)^2$ and $g_2 = (f_2 - f_0)^2$, respectively. We can derive

$$\|g_1 - g_2\|_\infty = \sup_{\mathbf{x}\in\mathscr{M}} |f_1(\mathbf{x}) - f_2(\mathbf{x})|\,|f_1(\mathbf{x}) + f_2(\mathbf{x}) - 2f_0(\mathbf{x})| \leqslant 4R\|f_1 - f_2\|_\infty.$$

Therefore, the inequality $\mathscr{N}(\delta,\mathscr{G},\|\cdot\|_\infty) \leqslant \mathscr{N}(\delta/4R,\mathscr{F}(R,\kappa,L,p,K),\|\cdot\|_\infty)$ holds, which implies

$$T_2 \leqslant \frac{52R^2}{3n}\log\mathscr{N}(\delta/4R,\mathscr{F}(R,\kappa,L,p,K),\|\cdot\|_\infty) + \left(4 + \frac{1}{2R}\right)\delta.$$

□

Bounding the Covering Number $\mathscr{N}(\delta, \mathscr{F}(R,\kappa,L,p,K), \|\cdot\|_\infty)$. Since each weight parameter in the network is bounded by a constant κ, we construct a covering by partition the range of each weight parameter into a uniform grid. Consider $f, f' \in \mathscr{F}(R,\kappa,L,p,K)$ with each weight parameter differing at most h. By an induction on the number of layers in the network, we show that the ℓ_∞ norm of the difference $f - f'$ scales as

$$\left\|f - f'\right\|_\infty \leqslant hL(pB+2)(\kappa p)^{L-1}.$$

As a result, to achieve a δ-covering, it suffices to choose h such that $hL(pB+2)(\kappa p)^{L-1} = \delta$. Therefore, the covering number is bounded by

$$\mathscr{N}(\delta, \mathscr{F}(R,\kappa,L,p,K), \|\cdot\|_\infty) \leqslant \left(\frac{\kappa}{h}\right)^K \leqslant \left(\frac{\kappa L(pB+2)(\kappa p)^{L-1}}{\delta}\right)^K.$$

Choosing δ and Bounding the L_2 Risk. Combining T_1 and T_2 together and substituting the covering number, we derive

$$\begin{aligned}
\mathbb{E}\left[\int_{\mathscr{M}} \left(\widehat{f}_n(\mathbf{x}) - f_0(\mathbf{x})\right)^2 d\mathscr{D}_x(\mathbf{x})\right] &\leqslant 4 \inf_{f\in\mathscr{F}(R,\kappa,L,p,K)} \int_{\mathscr{M}} (f(\mathbf{x}) - f_0(\mathbf{x}))^2 d\mathscr{D}_x(\mathbf{x}) \\
&\quad + \frac{52R^2 + 192\sigma^2}{3n} \log \mathscr{N}(\delta/4R, \mathscr{F}(R,\kappa,L,p,K), \|\cdot\|_\infty) \\
&\quad + \left(4 + \frac{1}{2R} + 32\sigma\right)\delta \\
&\leqslant 4 \inf_{f\in\mathscr{F}(R,\kappa,L,p,K)} \int_{\mathscr{M}} (f(\mathbf{x}) - f_0(\mathbf{x}))^2 d\mathscr{D}_x(\mathbf{x}) \\
&\quad + \frac{52R^2 + 192\sigma^2}{3n} K \log \frac{4R\kappa L(pB+2)(\kappa p)^{L-1}}{\delta} \\
&\quad + \left(4 + \frac{1}{2R} + 32\sigma\right)\delta.
\end{aligned}$$

Choosing $\delta = 1/n$ gives rise to

$$\begin{aligned}
\mathbb{E}\left[\int_{\mathscr{M}} \left(\widehat{f}_n(\mathbf{x}) - f_0(\mathbf{x})\right)^2 d\mathscr{D}_x(\mathbf{x})\right] &\leqslant 4 \inf_{f\in\mathscr{F}(R,\kappa,L,p,K)} \int_{\mathscr{M}} (f(\mathbf{x}) - f_0(\mathbf{x}))^2 d\mathscr{D}_x(\mathbf{x}) \\
&\quad + \widetilde{O}\left(\frac{R^2+\sigma^2}{n} KL \log(R\kappa Lpn) + \frac{1}{n}\right).
\end{aligned}$$

We further set $\inf_{f\in\mathscr{F}(R,\kappa,L,p,K)} \|f(\mathbf{x}) - f_0(\mathbf{x})\|_\infty \leqslant \varepsilon$. Theorem 3.2 suggests that we choose $L = \widetilde{O}(\log \frac{1}{\varepsilon})$ and $K = \widetilde{O}(\varepsilon^{-\frac{d}{s+\alpha}} \log \frac{1}{\varepsilon})$. Plugging in L and K, we have

$$\mathbb{E}\left[\int_{\mathscr{M}} \left(\widehat{f}_n(\mathbf{x}) - f_0(\mathbf{x})\right)^2 d\mathscr{D}_x(\mathbf{x})\right] = \widetilde{O}\left(\varepsilon^2 + \frac{R^2+\sigma^2}{n} \varepsilon^{-\frac{d}{s+\alpha}} \log^2 \frac{1}{\varepsilon} \log(R\kappa Lpn) + \frac{1}{n}\right).$$

To balance the error terms, we pick ε satisfying $\varepsilon^2 = \frac{1}{n}\varepsilon^{-\frac{d}{s+\alpha}}$, which gives $\varepsilon = n^{-\frac{s+\alpha}{d+2(s+\alpha)}}$. The proof of Theorem 3.1 is complete by substituting $\varepsilon = n^{-\frac{s+\alpha}{d+2(s+\alpha)}}$ and rearranging terms.

4.2 *Proof of Theorem 3.2*

We next sketch the proof of Theorem 3.2. A preliminary version has appeared in Chen et al. (2019). 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 4.3 (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 $\varepsilon > 0$. The network has depth and the number of neurons and weight parameters no more than $c\log(1/\varepsilon)$ 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 4.3.

COROLLARY 4.1 (Proposition 3 in Yarotsky (2017)) Given a constant $C > 0$ and $\varepsilon \in (0, C^2)$, 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 $|x| \leqslant C$ and $|y| \leqslant C$, we have $|\widehat{\times}(x,y) - xy| \leqslant \varepsilon$; **2)**. The depth and the weight parameters of the network is no more than $c\log\frac{C^2}{\varepsilon}$ with an absolute constant c.

The ReLU network in Theorem 3.2 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 $\mathscr{B}(\mathbf{c}, r)$. For any r, the collection $\{\mathscr{B}(\mathbf{x}, r)\}_{\mathbf{x}\in\mathscr{M}}$ is an open cover of $\mathscr{M}$. Since $\mathscr{M}$ is compact, there exists a finite collection of points $\mathbf{c}_i$ for $i = 1, \ldots, C_{\mathscr{M}}$ such that $\mathscr{M} \subset \bigcup_i \mathscr{B}(\mathbf{c}_i, r)$.

Now we pick the radius $r < \tau/2$ so that

$$U_i = \mathscr{M} \cap \mathscr{B}(\mathbf{c}_i, r)$$

is diffeomorphic to a ball in $\mathbb{R}^d$ (Niyogi et al., 2008) as illustrated in Figure 4. Let $\{(U_i, \phi_i)\}_{i=1}^{C_{\mathscr{M}}}$ be an atlas on $\mathscr{M}$, where ϕ_i is to be defined in **Step 2**. The number of charts $C_{\mathscr{M}}$ is upper bounded by

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

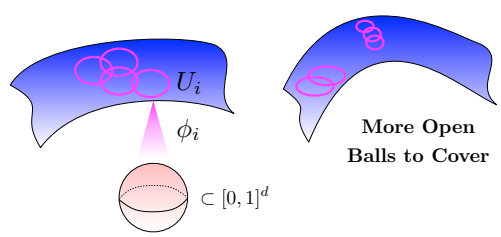

FIG. 4. Curvature decides the number of charts: smaller reach requires more chart.

where $SA(M)$ is the surface area of $\mathscr{M}$, and T_d is the thickness of the U_i's, which is defined as the average number of U_i's that contain a point on $\mathscr{M}$ (Conway et al., 1987).

REMARK 4.1 The thickness T_d scales approximately linear in d. As shown in Conway et al. (1987), there exists covering with $\frac{d}{e\sqrt{e}} \lesssim T_d \leqslant d\log d + d\log\log d + 5d$.

Step 2. Projection with rescaling and translation. We denote the tangent space at $\mathbf{c}_i$ as

$$T_{\mathbf{c}_i}(\mathscr{M}) = \mathrm{span}(\mathbf{v}_{i1}, \ldots, \mathbf{v}_{id}),$$

where $\{\mathbf{v}_{i1},\ldots,\mathbf{v}_{id}\}$ form an orthonormal basis. We obtain the matrix $V_i = [\mathbf{v}_{i1},\ldots,\mathbf{v}_{id}] \in \mathbb{R}^{D\times d}$ by concatenating $\mathbf{v}_{ij}$'s as column vectors.

Define

$$\phi_i(\mathbf{x}) = b_i(V_i^\top(\mathbf{x}-\mathbf{c}_i)+\mathbf{u}_i) \in [0,1]^d$$

for any $\mathbf{x}\in U_i$, where $b_i \in (0,1]$ is a scaling factor and $\mathbf{u}_i$ is a translation vector. Since U_i is bounded, we can choose proper b_i and $\mathbf{u}_i$ 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.

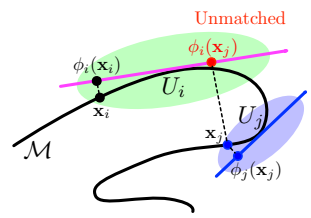

FIG. 5. Projecting $\mathbf{x}_j$ using a matched chart (blue) (U_j,ϕ_j), and an unmatched chart (green) (U_i,ϕ_i).

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 5.

An input $\mathbf{x}$ can belong to multiple charts, and the chart determination sub-network determines all these charts. This can be realized 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 $i=1,\ldots,C_{\mathscr{M}}$. The squared distance $d_i^2(\mathbf{x})$ is a sum of univariate quadratic functions, thus, we can apply Lemma 4.3 to approximate $d_i^2(\mathbf{x})$ by ReLU networks. Denote $\widehat{h}_{\text{sq}}$ as an approximation of the quadratic function x^2 on $[0,1]$ with an approximation error ν. Then we define

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

as an approximation of $d_i^2(\mathbf{x})$. The approximation error is $\|\widehat{d}_i^2-d_i^2\|_\infty \leqslant 4B^2D\nu$, by the triangle inequality. We consider an approximation of the indicator function of an interval as in Figure 6:

$$\widehat{\mathbb{1}}_\Delta(a) = \begin{cases} 1 & a \leqslant r^2-\Delta+4B^2m\nu \\ -\frac{1}{\Delta-8B^2m\nu}a+\frac{r^2-4B^2m\nu}{\Delta-8B^2m\nu} & a\in[r^2-\Delta+4B^2m\nu, r^2-4B^2m\nu] \\ 0 & a > r^2-4B^2m\nu \end{cases}, \tag{4.1}$$

where Δ ($\Delta \geqslant 8B^2m\nu$) will be chosen later according to the accuracy ε.

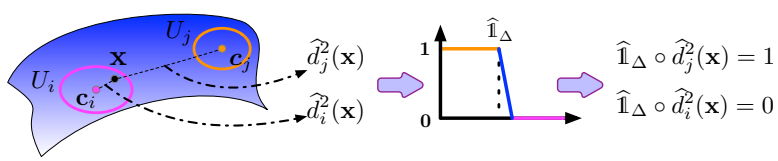

FIG. 6. Chart determination utilizes the composition of approximated distance function $\widehat{d}_i^2$ and the indicator function $\widehat{\mathbb{1}}_\Delta$.

To implement $\widehat{\mathbb{1}}_\Delta(a)$, we consider a basic step function $g = 2\text{ReLU}(x-0.5(r^2-4B^2m\nu))-2\text{ReLU}(x-$

$r^2+4B^2m\nu$). It is straightforward to check

$$g_k(a) = \underbrace{g\circ\cdots\circ g}_{k}(a) = \begin{cases} 0 & a < (1-2^{-k})(r^2-4B^2m\nu) \\ 2^k(a-r^2+4B^2m\nu)+r^2-4B^2m\nu & a \in \left[(1-\frac{1}{2^k})(r^2-4B^2m\nu), r^2-4B^2m\nu\right] \\ r^2-4B^2m\nu & a > r^2-4B^2m\nu \end{cases}.$$

Let $\widehat{\mathbb{1}}_\Delta = 1-\frac{1}{r^2-4B^2m\nu}g_k$. It suffices to choose k satisfying $(1-\frac{1}{2^k})(r^2-4B^2m\nu) \geqslant r^2-\Delta+4B^2m\nu$, which yields $k = \left\lceil \log\frac{r^2}{\Delta}\right\rceil$. 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., $d_i^2(\mathbf{x}) \geqslant r^2$, we have $\widehat{\mathbb{1}}_\Delta \circ \widehat{d_i^2}(\mathbf{x}) = 0$;
- if $\mathbf{x} \in U_i$ and $d_i^2(\mathbf{x}) \leqslant r^2-\Delta$, we have $\widehat{\mathbb{1}}_\Delta \circ \widehat{d_i^2}(\mathbf{x}) = 1$.

Step 4. Taylor approximation. In each chart (U_i,ϕ_i), we locally approximate f using Taylor polynomials of order n as shown in Figure 7. Specifically, we decompose f as

$$f = \sum_{i=1}^{C_{\mathscr{M}}} f_i \quad \text{with} \quad f_i = f\rho_i,$$

where ρ_i is an element in a C^∞ partition of unity on $\mathscr{M}$ which is supported inside U_i. The existence of such a partition of unity is guaranteed by Proposition 2.6. Since $\mathscr{M}$ is compact and ρ_i is C^∞, f_i preserves the regularity (smoothness) of f such that $f_i \in \mathscr{H}^{s,\alpha}(\mathscr{M})$ for $i = 1,\ldots,C_{\mathscr{M}}$.

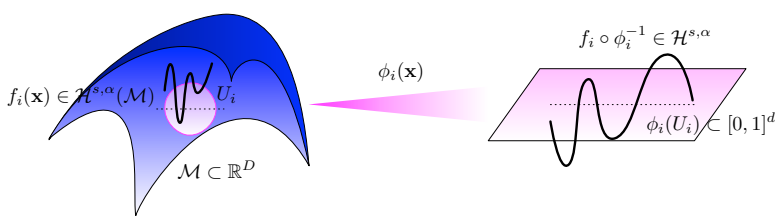

FIG. 7. Locally approximate f in each chart (U_i,ϕ_i) using Taylor polynomials.

LEMMA 4.4 Suppose Assumption 3 holds. For $i = 1,\ldots,C_{\mathscr{M}}$, the function f_i belongs to $H^{s,\alpha}_{\mathscr{M}}$: there exists a Hölder coefficient L_i depending on d, f_i, and ϕ_i such that for any $|\mathbf{s}| = s$, we have

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

Proof Sketch. We provide a sketch here. Details can be found in Appendix B.1. Denote $g_1 = f\circ\phi_i^{-1}$ and $g_2 = \rho_i\circ\phi_i^{-1}$. By the Leibniz rule, we have

$$D^{\mathbf{s}}(f_i\circ\phi_i^{-1}) = D^{\mathbf{s}}(g_1\times g_2) = \sum_{|\mathbf{p}|+|\mathbf{q}|=s} \binom{s}{|\mathbf{p}|} D^{\mathbf{p}}g_1 D^{\mathbf{q}}g_2.$$

Consider each term in the sum: for any $\mathbf{x}_1, \mathbf{x}_2 \in U_i$,

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

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

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

where the last inequality is due to the fact that $\|\phi_i(\mathbf{x}_1) - \phi_i(\mathbf{x}_2)\|_2 \leqslant b_i \|V_i\| \|\mathbf{x}_1 - \mathbf{x}_2\|_2 \leqslant 2r$. 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 $L_i = 2^{s+1}\sqrt{d}\lambda_i\mu_i(2r)^{1-\alpha}$. □

Lemma 4.4 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 4.1 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

1. no more than $c\left(\log \frac{1}{\delta} + 1\right)$ layers,
2. at most $c'\delta^{-\frac{d}{s+\alpha}}\left(\log \frac{1}{\delta} + 1\right)$ neurons and weight parameters,

where c, c' depend on $s, d, f_i \circ \phi_i^{-1}$.

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 s-th order Taylor polynomial in each cube. To implement such polynomials by ReLU networks, we recursively apply the multiplication $\widehat{\times}$ operator in Corollary 4.1, since these polynomials are sums of the products of different variables. □

Step 5. Estimating the total error. We have collected all the ingredients to implement the entire ReLU network to approximate f on $\mathscr{M}$. Recall that the network structure consists of 3 main sub-networks as demonstrated in Figure 3. 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

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

where $\widetilde{f}_i$ is the approximation of $f_i \circ \phi_i^{-1}$ using Taylor polynomials in Theorem 4.1. The total error can be decomposed to three components according to the following theorem.

THEOREM 4.2 For any $i = 1, \ldots, C_{\mathscr{M}}$, we have $\|\widetilde{f} - f\|_\infty \leqslant \sum_{i=1}^{C_{\mathscr{M}}} (A_{i,1} + A_{i,2} + A_{i,3})$, where

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

Here $\mathbb{1}(\mathbf{x} \in U_i)$ is the indicator function on U_i. Theorem 4.2 is proved in Appendix B.3. In order to achieve an ε total approximation error, i.e., $\|f - \widetilde{f}\|_\infty \leqslant \varepsilon$, 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 $\mathscr{K}_i$ defined below.

LEMMA 4.5 For any $i = 1, \ldots, C_{\mathscr{M}}$, denote

$$\mathscr{K}_i = \{\mathbf{x} \in \mathscr{M} : r^2 - \Delta \leqslant \|\mathbf{x} - \mathbf{c}_i\|_2^2 \leqslant r^2\}.$$

Then there exists a constant c depending on f_i's and ϕ_i's such that

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

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 $\operatorname{supp}(f_i)$ is a compact subset of U_i. By the mean value theorem, for any $\mathbf{x}, \mathbf{y} \in \mathscr{K}_i$, there exists $\mathbf{z} = \beta\phi_i(\mathbf{x}) + (1-\beta)\phi_i(\mathbf{y})$ for some $\beta \in (0,1)$ such that

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

We pick $\mathbf{y} \in \partial U_i$ (the boundary of U_i) so that $f_i(\mathbf{y}) = 0$. Since $f_i \in \mathscr{H}^{s,\alpha}(\mathscr{M})$ and $\mathscr{M}$ is compact, $\left\|\nabla f_i \circ \phi_i^{-1}(\mathbf{z})\right\|_2 b_i \|V_i\|_2 \leqslant c$ 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 $\mathbf{x} \in \mathscr{K}_i$, there exists $\mathbf{y} \in \partial U_i$ satisfying

$$\|\mathbf{x} - \mathbf{y}\|_2 \leqslant \frac{\pi+1}{r(1-r/\tau)}\Delta.$$

The idea is to consider a geodesic $\gamma(t)$ parameterized by the arc length from $\mathbf{x}$ to ∂U_i in Figure 8. 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. 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:

$$\theta(t) = \gamma(t)^\top \dot{\gamma}(t) / \|\gamma(t)\|_2 \quad \text{and} \quad \ell(t) = \|\gamma(t)\|_2,$$

where $\dot{\gamma}$ denotes the derivative of γ.

We show that there exists a geodesic $\gamma(t)$ satisfying

$$\inf_t \dot{\ell}(t) \geqslant \frac{1 - r/\tau}{\pi + 1} > 0.$$

This implies that the geodesic continuously moves away from the center $\mathbf{c}_i$. 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

$$T \inf_t \dot{\ell}(t) \leqslant \ell(T) - \ell(0) \leqslant r - \sqrt{r^2 - \Delta} \leqslant \frac{\Delta}{r}.$$

Therefore, we derive

$$\|\mathbf{x} - \mathbf{y}\|_2 \leqslant T \leqslant \frac{\Delta}{r \inf_t \dot{\ell}(t)} \leqslant \frac{\pi + 1}{r(1 - r/\tau)} \Delta.$$

□

Given Theorem 4.2, we choose

$$\eta = \delta = \frac{\varepsilon}{3C_{\mathscr{M}}} \quad \text{and} \quad \Delta = \frac{r(1 - r/\tau)\varepsilon}{3c(\pi + 1)C_{\mathscr{M}}} \tag{4.2}$$

so that the approximation error is bounded by ε. Moreover, we choose

$$\nu = \frac{\Delta}{16B^2 D} \tag{4.3}$$

to guarantee $\Delta > 8B^2 D\nu$ so that the definition of $\widehat{\mathbb{1}}_\Delta$ is valid.

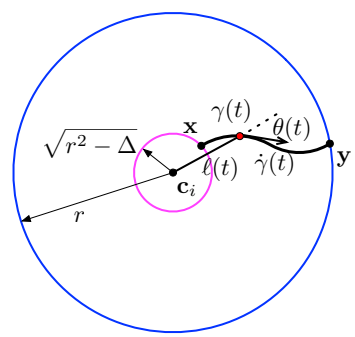

FIG. 8. A geometric illustration of θ and ℓ.

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, s, f, and c_1, c_3 are absolute constants. Combining these with (4.2) and (4.3) yields the depth in Theorem 3.2. By a similar argument, we can obtain the number of neurons and weight parameters. A detailed analysis is given in Appendix B.5.

5. Conclusion

We study nonparametric regression using deep ReLU neural networks, when data lie on a d-dimensional manifold $\mathscr{M}$ isometrically embedded in $\mathbb{R}^D$. Our result establishes an efficient recovery theory for general regression functions including C^s, Hölder, and Sobolev functions supported on manifolds. We show that the L_2 loss for the estimation of $f_0 \in \mathscr{H}^{s,\alpha}(\mathscr{M})$ converges in the order of $n^{-\frac{s+\alpha}{2(s+\alpha)+d}}$. This implies that, to obtain an ε-error estimation of f_0, the sample complexity scales in the order of $\varepsilon^{-\frac{2(s+\alpha)+d}{s+\alpha}}$. This fast rate depending on d reveals that deep neural networks are adaptive to low-dimensional geometric structures of data. Such results provide important insights in understanding why deep learning succeed in various real-world applications where data exhibit low-dimensional structures.

Acknowledgment

This work is supported by NSF grants DMS 1818751 and III 1717916.

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.

ALTMAN, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*.

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*.

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. (1991). Complexity regularization with application to artificial neural networks. In *Nonparametric functional estimation and related topics*. Springer, 561–576.

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

BICKEL, P. J. and LI, B. (2007). Local polynomial regression on unknown manifolds. *Lecture Notes-Monograph Series*, **54** 177–186.

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.

CHEN, M., JIANG, H., LIAO, W. and ZHAO, T. (2019). Efficient approximation of deep relu networks for functions on low dimensional manifolds. In *Advances in Neural Information Processing Systems*.

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.

DAUBECHIES, I., DEVORE, R., FOUCART, S., HANIN, B. and PETROVA, G. (2019). Nonlinear approximation and (deep) relu networks. *arXiv preprint arXiv:1905.02199*.

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

DJURIC, N., ZHOU, J., MORRIS, R., GRBOVIC, M., RADOSAVLJEVIC, V. and BHAMIDIPATI, N. (2015). Hate speech detection with comment embeddings. In *Proceedings of the 24th international conference on world wide web*. ACM.

FAN, J. and GIJBELS, I. (1996). *Local Polynomial Modelling and Its Applications*. Monographs on statistics and applied probability series, Chapman & Hall.

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.

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.

HAMERS, M. and KOHLER, M. (2006). Nonasymptotic bounds on the l 2 error of neural network regression estimates. *Annals of the Institute of Statistical Mathematics*, **58** 131–151.

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

HINTON, G. E. and SALAKHUTDINOV, R. R. (2006). Reducing the dimensionality of data with neural networks. *science*, **313** 504–507.

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.

KIM, Y., OHN, I. and KIM, D. (2018). Fast convergence rates of deep neural networks for classification. *arXiv preprint arXiv:1812.03599*.

KOHLER, M. and KRZYŻAK, A. (2005). Adaptive regression estimation with multilayer feedforward neural networks. *Nonparametric Statistics*, **17** 891–913.

KOHLER, M. and KRZYŻAK, A. (2016). Nonparametric regression based on hierarchical interaction models. *IEEE Transactions on Information Theory*, **63** 1620–1630.

KOHLER, M. and MEHNERT, J. (2011). Analysis of the rate of convergence of least squares neural network regression estimates in case of measurement errors. *Neural Networks*, **24** 273–279.

KPOTUFE, S. (2011). k-NN regression adapts to local intrinsic dimension. In *Advances in Neural Information Processing Systems 24*. 729–737.

KPOTUFE, S. and GARG, V. K. (2013). Adaptivity to local smoothness and dimension in kernel regression. In *Advances in Neural Information Processing Systems 26*. 3075–3083.

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.

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*.

MAAS, A. L., HANNUN, A. Y. and NG, A. Y. (2013). Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, vol. 30.

MCCAFFREY, D. F. and GALLANT, A. R. (1994). Convergence rates for single hidden layer feedforward networks.

Neural Networks, **7** 147–158.

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.

NAIR, V. and HINTON, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*.

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.

OHN, I. and KIM, Y. (2019). Smooth function approximation by deep neural networks with general activation functions. *Entropy*, **21** 627.

OSHER, S., SHI, Z. and ZHU, W. (2017). Low dimensional manifold model for image processing. *SIAM Journal on Imaging Sciences*, **10** 1669–1690.

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.

SCHMIDT-HIEBER, J. (2017). Nonparametric regression using deep neural networks with relu activation function. *arXiv preprint arXiv:1708.06633*.

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

SHEN, Z., YANG, H. and ZHANG, S. (2019). Deep network approximation characterized by number of neurons. *arXiv preprint arXiv:1906.05497*.

SUZUKI, T. (2019). Adaptivity of deep reLU network for learning in besov and mixed smooth besov spaces: optimal rate and curse of dimensionality. In *International Conference on Learning Representations*.

TAN, M. and LE, Q. V. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv preprint arXiv:1905.11946*.

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

TSYBAKOV, A. B. (2008). *Introduction to Nonparametric Estimation*. 1st ed. Springer Publishing Company, Incorporated.

TU, L. (2010). *An Introduction to Manifolds*. Universitext, Springer New York.

VAN DER VAART, A., VAN DER VAART, A., VAN DER VAART, A. and WELLNER, J. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics, Springer.

WAHBA, G. (1990). *Spline models for observational data*, vol. 59. Siam.

WASSERMAN, L. (2006). *All of nonparametric statistics*. Springer Science & Business Media.

YANG, Y., TOKDAR, S. T. ET AL. (2015). Minimax-optimal nonparametric regression in high dimensions. *The Annals of Statistics*, **43** 652–674.

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.

ZHOU, D.-X. (2019). Universality of deep convolutional neural networks. *Applied and Computational Harmonic Analysis*.

Supplementary Materials for Nonparametric Regression on Low-Dimensional Manifolds using Deep ReLU Networks

A. Proofs of Preliminary Results in Section 4

A.1 *Proof of Lemma 4.3*

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

$$\frac{1}{2^{4N}} \leqslant \varepsilon \Longrightarrow N \geqslant \frac{\log\frac{1}{\varepsilon}}{4}.$$

Let $N = \lceil \frac{\log\frac{1}{\varepsilon}}{4} \rceil$ 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 $\left\lceil \frac{\log \frac{1}{\varepsilon}}{4} \right\rceil \leqslant c \log \frac{1}{\varepsilon}$ 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 $c' \log \frac{1}{\varepsilon}$. □

A.2 *Proof of Corollary 4.1*

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 $\widehat{\times}$

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

Thus, we pick $\delta = \frac{\varepsilon}{2C^2}$ to ensure $\left| \widehat{\times}(x,y) - xy \right| \leqslant \varepsilon$ for any inputs x and y. As shown in Lemma 4.3, we can implement $\widehat{f}_\delta$ using a ReLU network of depth at most $c' \log \frac{1}{\delta} = c \log \frac{C^2}{\varepsilon}$ with absolute constants c', c. The proof is complete. □

B. Proof of Approximation Theory of ReLU Network (Theorem 3.2)

B.1 *Proof of Lemma 4.4*

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

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

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

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

The last inequality follows from $\phi_i(\mathbf{x}) = b_i(V_i^\top(\mathbf{x}-\mathbf{c}_i) + \mathbf{u}_i)$ and $\|V_i\|_2 = 1$. Observe that U_i is bounded, hence, we have $\|\mathbf{x}_1 - \mathbf{x}_2\|_2^{1-\alpha} \leqslant (2r)^{1-\alpha}$. 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 $\beta_{i,\alpha} = \sqrt{d}\lambda_i b_i^{1-\alpha}(2r)^{1-\alpha} \leqslant \sqrt{d}\lambda_i(2r)^{1-\alpha}$. Similarly, g_1 is C^{s-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 $k \leqslant n-1$. These derivatives are also Hölder continuous with coefficient $\theta_{i,\alpha} \leqslant \sqrt{d}\mu_i(2r)^{1-\alpha}$.

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

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

Observe that there are totally 2^n summands in the right hand side of (A.1). Therefore, for any $\mathbf{x}_1, \mathbf{x}_2 \in U_i$ and $|\mathbf{n}| = n$, we have

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

□

B.2 *Proof of Theorem 4.1*

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 $\operatorname{supp}(\rho_i) \subset U_i$. For notational simplicity, we denote $f_i^{\phi} = f_i \circ \phi_i^{-1}$ with the extension.

Step 1. We define a trapezoid function

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

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 A.9.

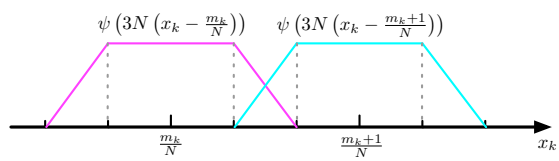

FIG. A.9. Illustration of the construction of $\zeta_{\mathbf{m}}$ on the k-th coordinate.

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

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

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

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

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 4.4. By setting

$$\sqrt{d}\mu_i\lambda_i(2r)^{1-\alpha}\frac{2^{d+s+1}d^{s+\alpha/2}}{s!}\left(\frac{1}{N}\right)^{s+\alpha}\leqslant\frac{\delta}{2},$$

we get $N\geqslant\left(\frac{\sqrt{d}\mu_i\lambda_i(2r)^{1-\alpha}2^{d+s+2}d^{s+\alpha/2}}{\delta s!}\right)^{\frac{1}{s+\alpha}}$. Accordingly, the approximation error is bounded by $\|\bar{f}_i-f_i^{\phi}\|_\infty\leqslant\frac{\delta}{2}$.

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

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

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

$$\bar{f}_i(\mathbf{x})=\sum_{\mathbf{m}\in\{0,\dots,N\}^d}\sum_{|\mathbf{s}|\leqslant s}a_{\mathbf{m},\mathbf{s}}\zeta_{\mathbf{m}}(\mathbf{x})\left(\mathbf{x}-\frac{\mathbf{m}}{N}\right)^{\mathbf{s}}. \tag{A.2}$$

Note that (A.2) is a linear combination of products $\zeta_{\mathbf{m}}\left(\mathbf{x}-\frac{\mathbf{m}}{N}\right)^{\mathbf{s}}$. Each product involves at most $d+n$ univariate terms: d terms for $\zeta_{\mathbf{m}}$ and n terms for $\left(\mathbf{x}-\frac{\mathbf{m}}{N}\right)^{\mathbf{s}}$. We recursively apply Corollary 4.1 to implement the product. Specifically, let $\widehat{\times}_\varepsilon$ be the approximation of the product operator in Corollary 4.1 with error ε, which will be chosen later. Consider the following chain application of $\widehat{\times}_\varepsilon$:

$$\widetilde{f}_{\mathbf{m},\mathbf{s}}(\mathbf{x})=\widehat{\times}_\varepsilon\left(\psi(3Nx_1-3m_1),\widehat{\times}_\varepsilon\left(\dots,\widehat{\times}_\varepsilon\left(\psi(3N_dx_d-m_d),\widehat{\times}_\varepsilon\left(x_1-\frac{m_1}{N},\dots\right)\right)\right)\right).$$

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

$$\begin{aligned}\left|\widetilde{f}_{\mathbf{m},\mathbf{s}}(\mathbf{x})-\zeta_{\mathbf{m}}\left(\mathbf{x}-\frac{\mathbf{m}}{N}\right)^{\mathbf{s}}\right|&=\left|\widehat{\times}_\varepsilon\left(\psi(3Nx_1-3m_1),\widehat{\times}_\varepsilon\left(\dots,\widehat{\times}_\varepsilon(x_1-\frac{m_1}{N},\dots)\right)\right)-\zeta_{\mathbf{m}}\left(\mathbf{x}-\frac{\mathbf{m}}{N}\right)^{\mathbf{s}}\right|\\&\leqslant\left|\widehat{\times}_\varepsilon\left(\psi(3Nx_1-3m_1),\widehat{\times}_\varepsilon(\psi(3Nx_2-3m_2),\dots)\right)\right.\\&\qquad\left.-\psi(3N_1-3m_1)\widehat{\times}_\varepsilon(\psi(3Nx_2-3m_2),\dots)\right|\\&\quad+|\psi(3Nx_1-m_1)|\left|\widehat{\times}_\varepsilon(\psi(3Nx_2-3m_2),\dots)-\psi(3Nx_2-3m_2)\widehat{\times}_\varepsilon(\dots)\right|\\&\quad+\dots\\&\leqslant(s+n)\delta.\end{aligned}$$

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

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

The approximation error is bounded by

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

We choose $\varepsilon=\frac{\delta}{\lambda_i\mu_i 2^{d+s+2}d^s(d+s)}$, so that $\|\bar{f}_i-\widetilde{f}_i\|_\infty\leqslant\frac{\delta}{2}$. Thus, we eventually have $\|\widetilde{f}_i-f_i^{\phi}\|_\infty\leqslant\delta$. 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 $\widetilde{f}_{\mathbf{m},\mathbf{s}}$. The total number of parallel sub-networks is bounded by $d^s(N+1)^d$. For each sub-network, we observe that ψ can be exactly implemented by a single layer ReLU network, i.e., $\psi(x)=\mathrm{ReLU}(x+2)-\mathrm{ReLU}(x+1)-\mathrm{ReLU}(x-1)+\mathrm{ReLU}(x-2)$. Corollary 4.1 shows that $\widehat{\times}_\varepsilon$ can be implemented by a depth $c_1\log\frac{1}{\varepsilon}$ ReLU network. Therefore, the whole network for implementing $\widetilde{f}_i$ has no more than $c_1'\left(\log\frac{1}{\varepsilon}+1\right)$ layers and $c_1'd^s(N+1)^d\left(\log\frac{1}{\varepsilon}+1\right)$ neurons and weight parameters. With $\varepsilon=\frac{\delta}{\lambda_i\mu_i 2^{d+s+2}d^s(d+s)}$ and $N=\left\lceil\left(\frac{\mu_i\lambda_i(2r)^{1-\alpha}2^{d+s+2}d^{s+\alpha/2}}{\delta s!}\right)^{\frac{1}{s+\alpha}}\right\rceil$, we obtain that the whole network has no more than $c_1\log\frac{1}{\delta}$ layers, and at most $c_2\delta^{-\frac{d}{s+\alpha}}\left(\log\frac{1}{\delta}+1\right)$ neurons and weight parameters, for constants c_1,c_2 depending on d,s, and $f_i\circ\phi_i^{-1}$. □

B.3 *Proof of Theorem 4.2*

Proof. We expand the estimation error as

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

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

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

By Lemma 4.5, we have $\max_{\mathbf{x} \in \mathscr{K}_i} |f_i(\mathbf{x})| \leqslant \frac{c(\pi+1)}{r(1-r/\tau)} \Delta$ for a constant c depending on f_i. Then we bound $A_{i,3}$ as

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

□

B.4 *Proof of Lemma 4.5*

Proof. We extend $f_i \circ \phi_i^{-1}$ to the whole cube $[0,1]^d$ as in the proof of Theorem 4.1. 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 $\mathbf{x}, \mathbf{y} \in U_i$

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

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^s in $[0,1]^d$, the first derivative is uniformly bounded, i.e., $\left\| \nabla f_i \circ \phi_i^{-1}(\mathbf{z}) \right\|_2 \leqslant \alpha_i$ 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 $\mathbf{x} \in \mathscr{K}_i$, we only need to bound the Euclidean distance between $\mathbf{x}$ and $\mathbf{y}$. More specifically, for any $\mathbf{x} \in \mathscr{K}_i$, 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 $\mathscr{M}$ parameterized by the arc 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 $\mathscr{M}$ 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 8, 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 $\gamma(t)$.

Suppose $\mathbf{y} = \gamma(T)$, we need to upper bound T. Note that $\ell(T) - \ell(0) \leqslant r - \sqrt{r^2 - \Delta} \leqslant \Delta / r$. 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 $T \leqslant \frac{\Delta}{r \inf_t \dot{\ell}(t)}$.

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

$$\ddot{\ell}(t) = \frac{\left(\ddot{\gamma}(t)^\top \gamma(t) + \dot{\gamma}(t)^\top \dot{\gamma}(t)\right) \sqrt{\gamma(t)^\top \gamma(t)} - \left(\gamma(t)^\top \dot{\gamma}(t)\right)^2 / \sqrt{\gamma(t)^\top \gamma(t)}}{\gamma(t)^\top \gamma(t)}. \tag{A.3}$$

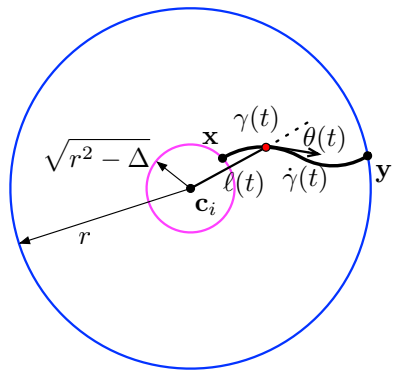

FIG. A.10. Illustration of ℓ and θ along a parametric curve γ.

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 (A.3), we can derive

$$\ddot{\ell}(t) = \frac{1+\ddot{\gamma}(t)^\top \gamma(t) - \cos^2\theta(t)}{\ell(t)} = \frac{\sin^2\theta(t) + \ddot{\gamma}(t)^\top \gamma(t)}{\ell(t)}. \tag{A.4}$$

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

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

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

$$\begin{aligned} \min \quad & a^\top x, \\ \text{subject to} \quad & x^\top x = 1, \\ & b^\top x = 0. \end{aligned} \tag{A.5}$$

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 $\cos\angle(\ddot{\gamma}(t), \gamma)$. Additionally, we can find the maximum of $\cos\angle(\ddot{\gamma}(t), \gamma)$ by replacing the minimization in (A.5) by maximization. We solve (A.5) by the Lagrangian method. More precisely, let

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

We have the optimal solution x^* satisfying $\nabla_x \mathscr{L} = 0$, 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

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

Substituting into (A.4), we have the following lower bound

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

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

$$\dot{\theta}(t) \leqslant -\frac{1}{\ell(t)}\left(\sin\theta(t) - \frac{r}{\tau}\right). \tag{A.6}$$

Inequality (A.6) 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 $t \leqslant T$. In fact, suppose there exists some $t_1 \leqslant T$ 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 $\theta(t) \geqslant \arcsin\frac{r}{\tau}$ for $t \in [t_0, t_1]$. This already gives us a contradiction:

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

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

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 $\gamma_{\mathbf{c},\mathbf{x}} \subset \mathscr{M}$ 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 $T_{r/2} = \sup\{t : \ell_{\mathbf{c},\mathbf{x}}(t) \leqslant r/2 - \Delta/r\}$, 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 $\ell_{\mathbf{c},\mathbf{x}}(T'_{r/2}) \leqslant r/2$. 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,

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

If there exists some $t \in (T_{r/2}, T_{\mathbf{x}}]$ such that $\sin\theta_{\mathbf{c},\mathbf{x}}(t) \leqslant \frac{r}{\tau}$, by the previous reasoning, we have $\sin\theta_{\mathbf{c},\mathbf{x}}(T_{\mathbf{x}}) \leqslant \frac{r}{\tau}$. 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

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

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

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

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

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

Therefore, we have $T \leqslant \frac{\Delta}{r\cos\theta(0)} \leqslant \frac{\pi+1}{r(1-r/\tau)}\Delta$. 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

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

Therefore, the function value $f(\mathbf{x})$ on $\mathscr{K}_i$ 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 $\eta = \frac{\varepsilon}{3C_{\mathscr{M}}}$ so that $\sum_{i=1}^{C_{\mathscr{M}}} A_{i,1} \leqslant \frac{\varepsilon}{3}$. The same argument yields $\delta = \frac{\varepsilon}{3C_{\mathscr{M}}}$. Analogously, we can choose $\Delta = \frac{r(1-r/\tau)\varepsilon}{3c(\pi+1)C_{\mathscr{M}}}$. 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 3.2.

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 $c_1\left(\log\frac{1}{\varepsilon} + \log D\right)$ with c_1 depending on d, f, τ, and the surface area of $\mathscr{M}$. The number of neurons and weight parameters is also $c_1'\left(\log\frac{1}{\varepsilon} + \log D\right)$ except for a different constant. Note that there are D parallel networks computing $\widehat{d_i^2}$ for $i = 1, \ldots, C_{\mathscr{M}}$. Hence, the total number of neurons and weight parameters is $c_1' C_{\mathscr{M}} D\left(\log\frac{1}{\varepsilon} + \log D\right)$ with c_1' depending on d, f, τ, and the surface area of $\mathscr{M}$.

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 $c_4'\delta^{-\frac{d}{s+\alpha}}\log\frac{1}{\delta}$. Here c_4, c_4' depend on $s, d, f_i \circ \phi_i^{-1}$.

Substituting $\delta = \frac{\varepsilon}{3C_{\mathscr{M}}}$, we get the depth is $c_2 \log \frac{1}{\varepsilon}$ and the number of neurons and weight parameters is $c_2' \varepsilon^{-\frac{d}{s+\alpha}} \log \frac{1}{\varepsilon}$. There are totally $C_{\mathscr{M}}$ parallel $\widehat{f}_i$'s, hence the total number of neurons and weight parameters is $c_2' C_{\mathscr{M}} \varepsilon^{-\frac{d}{s+\alpha}} \log \frac{1}{\varepsilon}$ with c_2' depending on $d, s, f_i \circ \phi_i^{-1}, \tau$, and the surface area of $\mathscr{M}$.

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 $c_3 \log \frac{1}{\varepsilon}$, and the number of neurons and weight parameters is $c_3' \log \frac{1}{\varepsilon}$. There are $C_{\mathscr{M}}$ parallel pairs of outputs from the chart determination and the Taylor polynomial sub-networks. Hence, the total number of weight parameters is $c_3' C_{\mathscr{M}} \log \frac{1}{\varepsilon}$ with c_3' depending on d, τ, and the surface area of $\mathscr{M}$.

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

C. Proof of Statistical Recovery of ReLU Network (Theorem 3.1)

C.1 *Proof of Lemma 4.1*

Proof. T_1 essentially reflects the bias of estimating f_0:

$$
\begin{aligned}
T_1 &= \mathbb{E}\left[\frac{2}{n}\sum_{i=1}^{n}(\widehat{f}_n(\mathbf{x}_i) - f_0(\mathbf{x}_i) - \xi_i + \xi_i)^2\right] \\
&= \frac{2}{n}\mathbb{E}\left[\sum_{i=1}^{n}(\widehat{f}_n(\mathbf{x}_i) - f_0(\mathbf{x}_i) - \xi_i)^2 + 2\xi_i(\widehat{f}_n(\mathbf{x}_i) - f_0(\mathbf{x}_i) - \xi_i) + \xi_i^2\right] \\
&\overset{(i)}{\leqslant} \frac{2}{n}\mathbb{E}\left[\sum_{i=1}^{n}(\widehat{f}_n(\mathbf{x}_i) - f_0(\mathbf{x}_i) - \xi_i)^2 + 2\xi_i\widehat{f}_n(\mathbf{x}_i)\right] \\
&= \frac{2}{n}\mathbb{E}\left[\sum_{i=1}^{n}(\widehat{f}_n(\mathbf{x}_i) - y_i)^2 + 2\xi_i\widehat{f}_n(\mathbf{x}_i)\right] \\
&= \frac{2}{n}\mathbb{E}\left[\inf_{f\in\mathscr{F}(R,\kappa,L,p,K)}\sum_{i=1}^{n}(f(\mathbf{x}_i) - y_i)^2 + 2\xi_i\widehat{f}_n(\mathbf{x}_i)\right] \\
&\overset{(ii)}{\leqslant} 2\inf_{f\in\mathscr{F}(R,\kappa,L,p,K)}\mathbb{E}\left[(f(\mathbf{x}) - f_0(\mathbf{x}))^2\right] + \mathbb{E}\left[\frac{4}{n}\sum_{i=1}^{n}\xi_i\widehat{f}_n(\mathbf{x}_i)\right],
\end{aligned}
\tag{A.1}
$$

where (i) holds since $\mathbb{E}[\xi_i f_0(\mathbf{x}_i)] = 0$, and (ii) holds due to Jensen's inequality and f being independent of $\mathbf{x}_i$'s. Now we need to bound $\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\xi_i\widehat{f}_n(\mathbf{x}_i)\right]$. We discretize the class $\mathscr{F}(R,\kappa,L,p,K)$ into $\mathscr{F}^*(R,\kappa,L,p,K) = \{f_i^*\}_{i=1}^{\mathscr{N}(\delta,\mathscr{F}(R,\kappa,L,p,K),\|\cdot\|_\infty)}$, where $\mathscr{N}(\delta,\mathscr{F}(R,\kappa,L,p,K),\|\cdot\|_\infty)$ denotes the δ-covering number with respect to the ℓ_∞ norm. Accordingly, there exists f^* such that $\|f^* - \widehat{f}_n\|_\infty \leqslant \delta$.

Denote $\|\widehat{f}_n - f_0\|_n^2 = \frac{1}{n}\sum_{i=1}^n (\widehat{f}_n(\mathbf{x}_i) - f_0(\mathbf{x}_i))^2$. Then we have

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \xi_i \widehat{f}_n(\mathbf{x}_i)\right] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \xi_i(\widehat{f}_n(\mathbf{x}_i) - f^*(\mathbf{x}_i) + f^*(\mathbf{x}_i) - f_0(\mathbf{x}_i))\right] \\
&\overset{(i)}{\leqslant} \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \xi_i(f^*(\mathbf{x}_i) - f_0(\mathbf{x}_i))\right] + \delta\sigma \\
&= \mathbb{E}\left[\frac{\|f^* - f_0\|_n}{\sqrt{n}} \frac{\sum_{i=1}^n \xi_i(f^*(\mathbf{x}_i) - f_0(\mathbf{x}_i))}{\sqrt{n}\|f^* - f_0\|_n}\right] + \delta\sigma \\
&\overset{(ii)}{\leqslant} \sqrt{2}\mathbb{E}\left[\frac{\|\widehat{f}_n - f_0\|_n + \delta}{\sqrt{n}} \frac{\sum_{i=1}^n \xi_i(f^*(\mathbf{x}_i) - f_0(\mathbf{x}_i))}{\sqrt{n}\|f^* - f_0\|_n}\right] + \delta\sigma.
\end{aligned}$$

Here (i) is obtained by applying Hölder's inequality to $\xi_i(\widehat{f}_n(\mathbf{x}_i) - f^*(\mathbf{x}_i))$ and invoking the Jensen's inequality:

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \xi_i(\widehat{f}_n(\mathbf{x}_i) - f^*(\mathbf{x}_i))\right] &\leqslant \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n |\xi_i| \left\|f^* - \widehat{f}_n\right\|_\infty\right] \\
&\leqslant \frac{1}{n}\sum_{i=1}^n \mathbb{E}[|\xi_i|]\delta \\
&\leqslant \frac{1}{n}\sum_{i=1}^n \sqrt{\mathbb{E}[|\xi_i|^2]}\delta \\
&\leqslant \delta\sigma.
\end{aligned}$$

Step (ii) holds, since by invoking the inequality $2ab \leqslant a^2 + b^2$, we have

$$\begin{aligned}
\|f^* - f_0\|_n &= \sqrt{\frac{1}{n}\sum_{i=1}^n (f^*(\mathbf{x}_i) - \widehat{f}_n(\mathbf{x}_i) + \widehat{f}_n(\mathbf{x}_i) - f_0(\mathbf{x}_i))^2} \\
&\leqslant \sqrt{\frac{2}{n}\sum_{i=1}^n (f^*(\mathbf{x}_i) - \widehat{f}_n(\mathbf{x}_i))^2 + (\widehat{f}_n(\mathbf{x}_i) - f_0(\mathbf{x}_i))^2} \\
&\leqslant \sqrt{\frac{2}{n}\sum_{i=1}^n \left[\delta^2 + \widehat{f}_n(\mathbf{x}_i) - f_0(\mathbf{x}_i))^2\right]} \\
&\leqslant \sqrt{2}\left\|\widehat{f}_n - f_0\right\|_n + \sqrt{2}\delta.
\end{aligned}$$

Now observe that $\frac{\sum_{i=1}^n \xi_i(f^*(\mathbf{x}_i) - f_0(\mathbf{x}_i))}{\sqrt{n}\|f^* - f_0\|_n} \leqslant \max_j \frac{\sum_{i=1}^n \xi_i(f_j^*(\mathbf{x}_i) - f_0(\mathbf{x}_i))}{\sqrt{n}\|f_j^* - f_0\|_n}$, and each $\frac{\sum_{i=1}^n \xi_i(f_j^*(\mathbf{x}_i) - f_0(\mathbf{x}_i))}{\sqrt{n}\|f_j^* - f_0\|_n}$ is sub-guassian with parameter σ^2. Given samples S_n, the quantity $\frac{\|\widehat{f}_n - f_0\|_n + \delta}{\sqrt{n}}$ is fixed. We then bound

$\mathbb{E}\left[\max_j \frac{\sum_{i=1}^n \xi_i(f_j^*(\mathbf{x}_i)-f_0(\mathbf{x}_i))}{\sqrt{n}\|f_j^*-f_0\|_n}\Big|S_n\right]$ by utilizing the moment generating function: For any t, we have

$$\begin{aligned}
\mathbb{E}\left[\max_j \frac{\sum_{i=1}^n \xi_i(f_j^*(\mathbf{x}_i)-f_0(\mathbf{x}_i))}{\sqrt{n}\|f_j^*-f_0\|_n}\Big|S_n\right] &= \frac{1}{t}\log\exp\left(t\mathbb{E}\left[\max_j \frac{\sum_{i=1}^n \xi_i(f_j^*(\mathbf{x}_i)-f_0(\mathbf{x}_i))}{\sqrt{n}\|f_j^*-f_0\|_n}\Big|S_n\right]\right)\\
&\leqslant \frac{1}{t}\log\mathbb{E}\left[\exp\left(t\max_j \frac{\sum_{i=1}^n \xi_i(f_j^*(\mathbf{x}_i)-f_0(\mathbf{x}_i))}{\sqrt{n}\|f_j^*-f_0\|_n}\right)\Big|S_n\right]\\
&\leqslant \frac{1}{t}\log\mathbb{E}\left[\sum_j \exp\left(t\frac{\sum_{i=1}^n \xi_i(f_j^*(\mathbf{x}_i)-f_0(\mathbf{x}_i))}{\sqrt{n}\|f_j^*-f_0\|_n}\right)\Big|S_n\right]\\
&\leqslant \frac{1}{t}\log\left(\mathscr{N}(\delta,\mathscr{F}(R,\kappa,L,p,K),\|\cdot\|_\infty)\exp(t^2\sigma^2)\right).
\end{aligned}$$

Taking $t=\sqrt{\sigma^{-1}\log\mathscr{N}(\delta,\mathscr{F}(R,\kappa,L,p,K),\|\cdot\|_\infty)}$, we have

$$\mathbb{E}\left[\max_j \frac{\sum_{i=1}^n \xi_i(f_j^*(\mathbf{x}_i)-f_0(\mathbf{x}_i))}{\sqrt{n}\|f_j^*-f_0\|_n}\right] \leqslant 2\sigma\sqrt{\log\mathscr{N}(\delta,\mathscr{F}(R,\kappa,L,p,K),\|\cdot\|_\infty)}.$$

This in turn yields

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \xi_i\widehat{f}_n(\mathbf{x}_i)\right] \leqslant 2\sqrt{2}\sigma\mathbb{E}\left[\|\widehat{f}_n-f_0\|_n+\delta\right]\sqrt{\frac{\log\mathscr{N}(\delta,\mathscr{F}(R,\kappa,L,p,K),\|\cdot\|_\infty)}{n}}+\delta\sigma.$$

Substituting back into (A.1), we have

$$\begin{aligned}
\mathbb{E}\left[\|\widehat{f}_n-f_0\|_n^2\right] &\leqslant \inf_{f\in\mathscr{F}(R,\kappa,L,p,K)}\mathbb{E}\left[(f(\mathbf{x})-f_0(\mathbf{x}))^2\right]+2\delta\sigma\\
&\quad+4\sqrt{2}\sigma\mathbb{E}\left[\|\widehat{f}_n-f_0\|_n+\delta\right]\sqrt{\frac{\log\mathscr{N}(\delta,\mathscr{F}(R,\kappa,L,p,K),\|\cdot\|_\infty)}{n}}\\
&\leqslant \inf_{f\in\mathscr{F}(R,\kappa,L,p,K)}\mathbb{E}\left[(f(\mathbf{x})-f_0(\mathbf{x}))^2\right]\\
&\quad+4\sqrt{2}\sigma\sqrt{\mathbb{E}\left[\|\widehat{f}_n-f_0\|_n^2\right]}\sqrt{\frac{\log\mathscr{N}(\delta,\mathscr{F}(R,\kappa,L,p,K),\|\cdot\|_\infty)}{n}}\\
&\quad+4\sqrt{2}\sigma\delta\sqrt{\frac{\log\mathscr{N}(\delta,\mathscr{F}(R,\kappa,L,p,K),\|\cdot\|_\infty)}{n}}+2\delta\sigma.
\end{aligned}$$

We only need to consider $\log\mathscr{N}(\delta,\mathscr{F}(R,\kappa,L,p,K),\|\cdot\|_\infty)<n$. Otherwise $\mathbb{E}\left[\|\widetilde{f}_n-f_0\|_n^2\right]$ is naturally bounded by $4R^2$, and thus the upper bound is trivial for large n. Invoking the fact that $x^2\leqslant 2ax+b$ implies $x^2\leqslant 4a^2+2b$. Letting $x^2=\mathbb{E}\left[\|\widehat{f}_n-f_0\|_n^2\right]$, $a=2\sqrt{2}\sigma\sqrt{\frac{\log\mathscr{N}(\delta,\mathscr{F}(R,\kappa,L,p,K),\|\cdot\|_\infty)}{n}}$ and $b=\inf_{f\in\mathscr{F}(R,\kappa,L,p,K)}\mathbb{E}\left[(f(\mathbf{x})-f_0(\mathbf{x}))^2\right]+(4\sqrt{2}+2)\sigma\delta$, we have

$$\begin{aligned}
T_1 \leqslant 2\mathbb{E}\left[\left\|\widehat{f}_n-f_0\right\|_n^2\right] \leqslant 4\inf_{f\in\mathscr{F}(R,\kappa,L,p,K)}\mathbb{E}\left[(f(\mathbf{x})-f_0(\mathbf{x}))^2\right]+64\sigma^2\frac{\log\mathscr{N}(\delta,\mathscr{F}(R,\kappa,L,p,K),\|\cdot\|_\infty)}{n}\\
+(16\sqrt{2}+8)\sigma\delta.
\end{aligned}$$

□

C.2 *Proof of Lemma 4.2*

Proof. In the following proofs, we use subscript to denote taking expectation with respect to a certain random variable. For example, $\mathbb{E}_{\mathbf{x}}$, $\mathbb{E}_{\xi}$ and $\mathbb{E}_{(\mathbf{x},y)}$ denote expectations with respect to $\mathbf{x}$, the noise, and the joint distribution of $(\mathbf{x},y)$, respectively. We rewrite T_2 as

$$\begin{aligned}
T_2 &= \mathbb{E}\left[\mathbb{E}_{\mathbf{x}}[g(\mathbf{x})|S_n] - \frac{2}{n}\sum_{i=1}^{n} g(\mathbf{x}_i)\right] \\
&= 2\mathbb{E}\left[\frac{1}{2}\mathbb{E}_{\mathbf{x}}[g(\mathbf{x})|S_n] - \frac{1}{n}\sum_{i=1}^{n} g(\mathbf{x}_i)\right] \\
&= 2\mathbb{E}\left[\mathbb{E}_{\mathbf{x}}[g(\mathbf{x})|S_n] - \frac{1}{n}\sum_{i=1}^{n} g(\mathbf{x}_i) - \frac{1}{2}\mathbb{E}_{\mathbf{x}}[g(\mathbf{x})|S_n]\right].
\end{aligned}$$

We lower bound $\mathbb{E}_{\mathbf{x}}[g(\mathbf{x})|S_n]$ by its second moment:

$$\begin{aligned}
\mathbb{E}_{\mathbf{x}}[g^2(\mathbf{x})|S_n] &= \mathbb{E}_{\mathbf{x}}\left[\left(\widehat{f}_n(\mathbf{x}) - f_0(\mathbf{x})\right)^4 \big| S_n\right] \\
&= \mathbb{E}_{\mathbf{x}}\left[\left(\widehat{f}_n(\mathbf{x}) - f_0(\mathbf{x})\right)^2 g(\mathbf{x}) \big| S_n\right] \\
&\leqslant \mathbb{E}_{\mathbf{x}}\left[4R^2 g(\mathbf{x}) \big| S_n\right].
\end{aligned}$$

The last inequality follows from $\left|\widehat{f}_n(\mathbf{x}) - f_0(\mathbf{x})\right| \leqslant 2R$. Now we cast T_2 into

$$T_2 \leqslant 2\mathbb{E}\left[\mathbb{E}_{\mathbf{x}}[g(\mathbf{x})|S_n] - \frac{1}{n}\sum_{i=1}^{n} g(\mathbf{x}_i) - \frac{1}{8R^2}\mathbb{E}_{\mathbf{x}}[g^2(\mathbf{x})|S_n]\right]. \tag{A.2}$$

Introducing the second moment allows us to establish a fast convergence of T_2. Specifically, we denote $\bar{\mathbf{x}}_i$'s as independent copies of $\mathbf{x}_i$'s following the same distribution. We also denote

$$\mathscr{G} = \left\{g(\mathbf{x}) = (f(\mathbf{x}) - f_0(\mathbf{x}))^2 \;\middle|\; f \in \mathscr{F}(R,\kappa,L,p,K)\right\}$$

as the function class induced by $\mathscr{F}(R,\kappa,L,p,K)$. Then we rewrite (A.2) as

$$\begin{aligned}
T_2 &\leqslant 2\mathbb{E}\left[\sup_{g\in\mathscr{G}} \mathbb{E}_{\bar{\mathbf{x}}}[g(\bar{\mathbf{x}})] - \frac{1}{n}\sum_{i=1}^{n} g(\mathbf{x}_i) - \frac{1}{8R^2}\mathbb{E}_{\mathbf{x}}[g^2(\mathbf{x})]\right] \\
&\leqslant 2\mathbb{E}\left[\sup_{g\in\mathscr{G}} \mathbb{E}_{\bar{\mathbf{x}}}\left[\frac{1}{n}\sum_{i=1}^{n} g(\bar{\mathbf{x}}_i) - g(\mathbf{x}_i)\right] - \frac{1}{16R^2}\mathbb{E}_{\bar{\mathbf{x}},\mathbf{x}}[g^2(\bar{\mathbf{x}}) + g^2(\mathbf{x})]\right] \\
&\overset{(i)}{\leqslant} 2\mathbb{E}_{\mathbf{x},\bar{\mathbf{x}},U}\left[\sup_{g\in\mathscr{G}} \frac{1}{n}\sum_{i=1}^{n} U_i\left(g(\bar{\mathbf{x}}_i) - g(\mathbf{x}_i)\right) - \frac{1}{16R^2}\mathbb{E}_{\bar{\mathbf{x}},\mathbf{x}}[g^2(\bar{\mathbf{x}}) + g^2(\mathbf{x})]\right].
\end{aligned}$$

Here U_i's are i.i.d. Rademacher random variables, i.e., $\mathbb{P}(U_i = 1) = \mathbb{P}(U_i = -1) = \frac{1}{2}$, independent of samples $\mathbf{x}_i$'s and $\bar{\mathbf{x}}_i$'s, and the inequality (i) follows from symmetrization.

We discretize $\mathscr{G}$ with respect to the ℓ_∞ norm. The δ-covering number is denoted as $\mathscr{N}(\delta, \mathscr{G}, \|\cdot\|_\infty)$ and the elements in the covering is denoted as $\mathscr{G}^* = \{g_i^*\}_{i=1}^{\mathscr{N}(\delta,\mathscr{G},\|\cdot\|_\infty)}$, that is, for any $g \in \mathscr{G}$, there exists a g^* satisfying $\|g - g^*\|_\infty \leqslant \delta$.

We replace $g \in \mathscr{G}$ by $g^* \in \mathscr{G}^*$ in bounding T_2, which then boils down to deriving concentration results on a finite concept class. Specifically, for g^* satisfying $\|g - g^*\|_\infty \leqslant \delta$, we have

$$\begin{aligned}
U_i\left(g(\bar{\mathbf{x}}_i) - g(\mathbf{x}_i)\right) &= U_i\left(g(\bar{\mathbf{x}}_i) - g^*(\bar{\mathbf{x}}_i) + g^*(\bar{\mathbf{x}}_i) - g^*(\mathbf{x}_i) + g^*(\mathbf{x}_i) - g(\mathbf{x}_i)\right) \\
&= U_i\left(g(\bar{\mathbf{x}}_i) - g^*(\bar{\mathbf{x}}_i)\right) + U_i\left(g^*(\bar{\mathbf{x}}_i) - g^*(\mathbf{x}_i)\right) + U_i\left(g^*(\mathbf{x}_i) - g(\mathbf{x}_i)\right) \\
&\leqslant U_i\left(g^*(\bar{\mathbf{x}}_i) - g^*(\mathbf{x}_i)\right) + 2\delta.
\end{aligned}$$

We also have

$$\begin{aligned}
g^2(\bar{\mathbf{x}}) + g^2(\mathbf{x}) &= \left[g^2(\bar{\mathbf{x}}) - (g^*)^2(\bar{\mathbf{x}})\right] + \left[(g^*)^2(\bar{\mathbf{x}}) + (g^*)^2(\mathbf{x})\right] - \left[(g^*)^2(\mathbf{x}) - g^2(\mathbf{x})\right] \\
&= (g^*)^2(\bar{\mathbf{x}}) + (g^*)^2(\mathbf{x}) + (g(\bar{\mathbf{x}}) - g^*(\bar{\mathbf{x}}))(g(\bar{\mathbf{x}}) + g^*(\bar{\mathbf{x}})) + (g^*(\mathbf{x}) - g(\mathbf{x}))(g^*(\mathbf{x}) + g(\mathbf{x})) \\
&\geqslant (g^*)^2(\bar{\mathbf{x}}) + (g^*)^2(\mathbf{x}) - |g(\bar{\mathbf{x}}) - g^*(\bar{\mathbf{x}})|\,|g(\bar{\mathbf{x}}) + g^*(\bar{\mathbf{x}})| - |g^*(\mathbf{x}) - g(\mathbf{x})|\,|g^*(\mathbf{x}) + g(\mathbf{x})| \\
&\geqslant (g^*)^2(\bar{\mathbf{x}}) + (g^*)^2(\mathbf{x}) - 2R\delta - 2R\delta.
\end{aligned}$$

Plugging the above two items into T_2, we upper bound T_2 as

$$\begin{aligned}
T_2 &\leqslant 2\mathbb{E}_{\mathbf{x},\bar{\mathbf{x}},U}\left[\sup_{g^*\in\mathscr{G}^*} \frac{1}{n}\sum_{i=1}^n U_i\left(g^*(\bar{\mathbf{x}}_i) - g^*(\mathbf{x}_i)\right) - \frac{1}{16R^2}\mathbb{E}_{\bar{\mathbf{x}},\mathbf{x}}[(g^*)^2(\bar{\mathbf{x}}) + (g^*)^2(\mathbf{x})]\right] + \left(4 + \frac{1}{2R}\right)\delta \\
&= 2\mathbb{E}_{\mathbf{x},\bar{\mathbf{x}},U}\left[\max_j \frac{1}{n}\sum_{i=1}^n U_i\left(g_j^*(\bar{\mathbf{x}}_i) - g_j^*(\mathbf{x}_i)\right) - \frac{1}{16R^2}\mathbb{E}_{\bar{\mathbf{x}},\mathbf{x}}[(g_j^*)^2(\bar{\mathbf{x}}) + (g_j^*)^2(\mathbf{x})]\right] + \left(4 + \frac{1}{2R}\right)\delta.
\end{aligned}$$

Denote $h_j(i) = U_i(g_j^*(\bar{\mathbf{x}}_i) - g_j^*(\mathbf{x}_i))$. By symmetry, it is straightforward to see $\mathbb{E}[h_j(i)] = 0$. The variance of $h_j(i)$ is computed as

$$\mathrm{Var}[h_j(i)] = \mathbb{E}\left[h_j^2(i)\right] = \mathbb{E}\left[U_i^2\left(g_j^*(\bar{\mathbf{x}}_i) - g_j^*(\mathbf{x}_i)\right)^2\right] \overset{(i)}{\leqslant} 2\mathbb{E}\left[(g_j^*)^2(\bar{\mathbf{x}}_i) + (g_j^*)^2(\mathbf{x}_i)\right].$$

The last inequality (i) utilizes the identity $(a-b)^2 \leqslant 2(a^2+b^2)$. Therefore, we derive the following upper bound for T_2,

$$T_2 \leqslant 2\mathbb{E}\left[\max_j \frac{1}{n}\sum_{i=1}^n h_j(i) - \frac{1}{32R^2}\frac{1}{n}\sum_{i=1}^n \mathrm{Var}[h_j(i)]\right] + \left(4 + \frac{1}{2R}\right)\delta.$$

We invoke the moment generating function to bound T_2. Note that we have $\|h_j\|_\infty \leqslant (2R)^2$. Then by

Taylor expansion, for $0 < t < \frac{3}{4R^2}$, we have

$$\begin{aligned}
\mathbb{E}[\exp(th_j(i))] &= \mathbb{E}\left[1 + th_j(i) + \sum_{k=2}^{\infty} \frac{t^k h_j^k(i)}{k!}\right] \\
&\leqslant \mathbb{E}\left[1 + th_j(i) + \sum_{k=2}^{\infty} \frac{t^k h_j^2(i)(4R^2)^{k-2}}{2 \times 3^{k-2}}\right] \\
&= \mathbb{E}\left[1 + th_j(i) + \frac{t^2 h_j^2(i)}{2} \sum_{k=2}^{\infty} \frac{t^{k-2}(4R^2)^{k-2}}{3^{k-2}}\right] \\
&= \mathbb{E}\left[1 + th_j(i) + \frac{t^2 h_j^2(i)}{2} \frac{1}{1 - 4tR^2/3}\right] \\
&= 1 + t^2 \operatorname{Var}[h_j(i)] \frac{1}{2 - 8tR^2/3} \\
&\overset{(i)}{\leqslant} \exp\left(\operatorname{Var}[h(i)] \frac{3t^2}{6 - 8tR^2}\right). \qquad \text{(A.3)}
\end{aligned}$$

Step (i) follows from the fact $1 + x \leqslant \exp(x)$ for $x \geqslant 0$. Given (A.3), we proceed to bound T_2. To ease the presentation, we temporarily neglect $\left(4 + \frac{1}{2R}\right)\delta$ term and denote $T_2' = T_2 - \left(4 + \frac{1}{2R}\right)\delta$. Then for $0 < t/n < \frac{3}{4R^2}$, we have

$$\begin{aligned}
\exp\left(t\frac{T_2'}{2}\right) &= \exp\left(t\mathbb{E}\left[\max_j \frac{1}{n}\sum_{i=1}^n h_j(i) - \frac{1}{32R^2}\frac{1}{n}\sum_{i=1}^n \operatorname{Var}[h_j(i)]\right]\right) \\
&\overset{(i)}{\leqslant} \mathbb{E}\left[\exp\left(t\sup_h \frac{1}{n}\sum_{i=1}^n h(i) - \frac{1}{32R^2}\frac{1}{n}\sum_{i=1}^n \operatorname{Var}[h(i)]\right)\right] \\
&\leqslant \mathbb{E}\left[\sum_h \exp\left(\frac{t}{n}\sum_{i=1}^n h(i) - \frac{1}{32R^2}\frac{t}{n}\sum_{i=1}^n \operatorname{Var}[h(i)]\right)\right] \\
&\overset{(ii)}{\leqslant} \mathbb{E}\left[\sum_h \exp\left(\sum_{i=1}^n \operatorname{Var}[h(i)]\frac{3(t/n)^2}{6 - 8tR^2/n} - \frac{1}{32R^2}\frac{t}{n}\operatorname{Var}[h(i)]\right)\right] \\
&= \mathbb{E}\left[\sum_h \exp\left(\sum_{i=1}^n \frac{t}{n}\operatorname{Var}[h(i)]\left(\frac{3t/n}{6 - 8tR^2/n} - \frac{1}{32R^2}\right)\right)\right].
\end{aligned}$$

Step (i) follows from Jensen's inequality, and step (ii) invokes (A.3). We now choose t so that $\frac{3t/n}{6-8tR^2/n} - \frac{1}{32R^2} = 0$, which yields $t = \frac{3n}{52R^2} < \frac{3n}{4R^2}$. Substituting our choice of t into $\exp(tT_2'/2)$, we have

$$t\frac{T_2'}{2} \leqslant \log\sum_h \exp(0) \Longrightarrow T_2' \leqslant \frac{2}{t}\log\mathscr{N}(\delta, \mathscr{G}, \|\cdot\|_\infty) = \frac{52R^2}{3n}\log\mathscr{N}(\delta, \mathscr{G}, \|\cdot\|_\infty).$$

To complete the proof, we relate the covering number of $\mathscr{G}$ to that of $\mathscr{F}(R, \kappa, L, p, K)$. Consider any $g_1, g_2 \in \mathscr{G}$ with $g_1 = (f_1 - f_0)^2$ and $g_2 = (f_2 - f_0)^2$, respectively, for $f_1, f_2 \in \mathscr{F}(R, \kappa, L, p, K)$. We can

derive

$$\begin{aligned}\|g_1 - g_2\|_\infty &= \sup_{\mathbf{x}} \left|(f_1(\mathbf{x}) - f_0(\mathbf{x}))^2 - (f_2(\mathbf{x}) - f_0(\mathbf{x}))^2\right| \\ &= \sup_{\mathbf{x}} |f_1(\mathbf{x}) - f_2(\mathbf{x})|\,|f_1(\mathbf{x}) + f_2(\mathbf{x}) - 2f_0(\mathbf{x})| \\ &\leqslant 4R\,\|f_1 - f_2\|_\infty .\end{aligned}$$

The above characterization immediately implies $\mathscr{N}(\delta, \mathscr{G}, \|\cdot\|_\infty) \leqslant \mathscr{N}(\delta/4R, \mathscr{F}(R,\kappa,L,p,K), \|\cdot\|_\infty)$. Therefore, we derive the desired upper bound on T_2:

$$T_2 \leqslant \frac{52R^2}{3n} \log \mathscr{N}(\delta/4R, \mathscr{F}(R,\kappa,L,p,K), \|\cdot\|_\infty) + \left(4 + \frac{1}{2R}\right)\delta.$$

□

C.3 *Proof of Theorem 3.1*

Proof. We dilate $\mathbb{E}\left[\mathbb{E}_{\mathbf{x}}\left(\widehat{f}_n(\mathbf{x}) - f_0(\mathbf{x})\right)^2 | S_n\right]$ using its empirical counterpart:

$$\begin{aligned}\mathbb{E}\left[\mathbb{E}_{\mathbf{x}}\left[\left(\widehat{f}_n(\mathbf{x}) - f_0(\mathbf{x})\right)^2 | S_n\right]\right] &= \underbrace{\mathbb{E}\left[\frac{2}{n}\sum_{i=1}^n (\widehat{f}_n(\mathbf{x}_i) - f_0(\mathbf{x}_i))^2\right]}_{T_1} \\ &\quad + \underbrace{\mathbb{E}\left[\mathbb{E}_{\mathbf{x}}\left[\left(\widehat{f}_n(\mathbf{x}) - f_0(\mathbf{x})\right)^2 | S_n\right]\right] - \mathbb{E}\left[\frac{2}{n}\sum_{i=1}^n (\widehat{f}_n(\mathbf{x}_i) - f_0(\mathbf{x}_i))^2\right]}_{T_2}.\end{aligned}$$

Combining the upper bounds on T_1 and T_2, we can derive

$$\begin{aligned}\mathbb{E}\left[\mathbb{E}_{\mathbf{x}}\left(\widetilde{f}_n(\mathbf{x}) - f_0(\mathbf{x})\right)^2 | S_n\right] &\leqslant 4 \inf_{f \in \mathscr{F}(R,\kappa,L,p,K)} \mathbb{E}\left[(f(\mathbf{x}) - f_0(\mathbf{x}))^2\right] \\ &\quad + \frac{52R^2 + 192\sigma^2}{3n} \log \mathscr{N}(\delta/4R, \mathscr{F}(R,\kappa,L,p,K), \|\cdot\|_\infty) \\ &\quad + \left(4 + \frac{1}{2R} + (16\sqrt{2} + 8)\sigma\right)\delta.\end{aligned}$$

The remaining step is to characterize the covering number $\mathscr{N}(\delta/4R, \mathscr{F}(R,\kappa,L,p,K), \|\cdot\|_\infty)$. To construct a covering for $\mathscr{F}(R,\kappa,L,p,K)$, we discretize each weight parameter by a uniform grid with grid size h. Recall we write $f \in \mathscr{F}(R,\kappa,L,p,K)$ as $f = W_L \cdot \mathrm{ReLU}(W_{L-1} \cdots \mathrm{ReLU}(W_1\mathbf{x} + \mathbf{b}_1) \cdots + \mathbf{b}_{L-1}) + \mathbf{b}_L$. Let $f, f' \in \mathscr{F}$ with all the weight parameters at most h from each other. Denoting the weight matrices in f, f' as $W_L, \ldots, W_1, \mathbf{b}_L, \ldots, \mathbf{b}_1$ and $W'_L, \ldots, W'_1, \mathbf{b}'_L, \ldots, \mathbf{b}'_1$, respectively, we bound the ℓ_∞ difference

$\|f-f'\|_\infty$ as

$$\begin{aligned}
\left\|f-f'\right\|_\infty &= \left\|W_L\cdot \mathrm{ReLU}(W_{L-1}\cdots \mathrm{ReLU}(W_1\mathbf{x}+\mathbf{b}_1)\cdots+\mathbf{b}_{L-1})+\mathbf{b}_L\right.\\
&\quad \left.-(W'_L\cdot \mathrm{ReLU}(W'_{L-1}\cdots \mathrm{ReLU}(W'_1\mathbf{x}+\mathbf{b}'_1)\cdots+\mathbf{b}'_{L-1})-\mathbf{b}'_L)\right\|_\infty\\
&\leqslant \left\|\mathbf{b}_L-\mathbf{b}'_L\right\|_\infty+\left\|W_L-W'_L\right\|_1\|W_{L-1}\cdots \mathrm{ReLU}(W_1\mathbf{x}+\mathbf{b}_1)\cdots+\mathbf{b}_{L-1}\|_\infty\\
&\quad +\|W_L\|_1\left\|W_{L-1}\cdots \mathrm{ReLU}(W_1\mathbf{x}+\mathbf{b}_1)\cdots+\mathbf{b}_{L-1}-(W'_{L-1}\cdots \mathrm{ReLU}(W'_1\mathbf{x}+\mathbf{b}'_1)\cdots+\mathbf{b}'_{L-1})\right\|_\infty\\
&\leqslant h+hp\|W_{L-1}\cdots \mathrm{ReLU}(W_1\mathbf{x}+\mathbf{b}_1)\cdots+\mathbf{b}_{L-1}\|_\infty\\
&\quad +\kappa p\left\|W_{L-1}\cdots \mathrm{ReLU}(W_1\mathbf{x}+\mathbf{b}_1)\cdots+\mathbf{b}_{L-1}-(W'_{L-1}\cdots \mathrm{ReLU}(W'_1\mathbf{x}+\mathbf{b}'_1)\cdots+\mathbf{b}'_{L-1})\right\|_\infty.
\end{aligned}$$

We derive the following bound on $\|W_{L-1}\cdots \mathrm{ReLU}(W_1\mathbf{x}+\mathbf{b}_1)\cdots+\mathbf{b}_{L-1}\|_\infty$:

$$\begin{aligned}
\|W_{L-1}\cdots \mathrm{ReLU}(W_1\mathbf{x}+\mathbf{b}_1)\cdots+\mathbf{b}_{L-1}\|_\infty &\leqslant \|W_{L-1}(\cdots \mathrm{ReLU}(W_1\mathbf{x}+\mathbf{b}_1)\cdots)\|_\infty+\|\mathbf{b}_{L-1}\|_\infty\\
&\leqslant \|W_{L-1}\|_1\|W_{L-2}(\cdots \mathrm{ReLU}(W_1\mathbf{x}+\mathbf{b}_1)\cdots)+\mathbf{b}_{L-2}\|_\infty+\kappa\\
&\leqslant \kappa p\|W_{L-2}(\cdots \mathrm{ReLU}(W_1\mathbf{x}+\mathbf{b}_1)\cdots)+\mathbf{b}_{L-2}\|_\infty+\kappa\\
&\overset{(i)}{\leqslant} (\kappa p)^{L-1}B+\kappa\sum_{i=0}^{L-3}(\kappa p)^i\\
&\leqslant (\kappa p)^{L-1}B+\kappa(\kappa p)^{L-2},
\end{aligned}$$

where (i) is obtained by induction and $\|\mathbf{x}\|_\infty\leqslant B$. The last inequality holds, since $\kappa p>1$. Substituting back into the bound for $\|f-f'\|_\infty$, we have

$$\begin{aligned}
\left\|f-f'\right\|_\infty &\leqslant \kappa p\left\|W_{L-1}\cdots \mathrm{ReLU}(W_1\mathbf{x}+\mathbf{b}_1)\cdots+\mathbf{b}_{L-1}-(W'_{L-1}\cdots \mathrm{ReLU}(W'_1\mathbf{x}+\mathbf{b}'_1)\cdots+\mathbf{b}'_{L-1})\right\|_\infty\\
&\quad +h+hp\left[(\kappa p)^{L-1}B+\kappa(\kappa p)^{L-2}\right]\\
&\leqslant \kappa p\left\|W_{L-1}\cdots \mathrm{ReLU}(W_1\mathbf{x}+\mathbf{b}_1)\cdots+\mathbf{b}_{L-1}-(W'_{L-1}\cdots \mathrm{ReLU}(W'_1\mathbf{x}+\mathbf{b}'_1)\cdots+\mathbf{b}'_{L-1})\right\|_\infty\\
&\quad +h(pB+2)(\kappa p)^{L-1}\\
&\overset{(i)}{\leqslant} (\kappa p)^{L-1}\left\|W_1\mathbf{x}+\mathbf{b}_1-W'_1\mathbf{x}-\mathbf{b}'_1\right\|_\infty+h(L-1)(pB+2)(\kappa p)^{L-1}\\
&\leqslant hL(pB+2)(\kappa p)^{L-1},
\end{aligned}$$

where (i) is obtained by induction. We choose h satisfying $hL(pB+2)(\kappa p)^{L-1}=\frac{\delta}{4R}$. Then discretizing each parameter uniformly into κ/h grids yields a $\frac{\delta}{4R}$-covering on $\mathscr{F}$. Therefore, the covering number is upper bounded by

$$\mathscr{N}(\delta/4R,\mathscr{F}(R,\kappa,L,p,K),\|\cdot\|_\infty)\leqslant\left(\frac{\kappa}{h}\right)^{\#\text{ of parameters}}. \tag{A.4}$$

By our choice of $\mathscr{F}(R,\kappa,L,p,K)$, there exists a network which yields f satisfying $\|f-f_0\|_\infty\leqslant\varepsilon$. Such a network network consists of $\widetilde{O}\left(\log\frac{1}{\varepsilon}\right)$ layers and $\widetilde{O}\left(\varepsilon^{-\frac{d}{s+\alpha}}\log\frac{1}{\varepsilon}\right)$ weight parameters. Then we have

$$\mathbb{E}\left[\mathbb{E}_{\mathbf{x}}\left(\widehat{f}_n(\mathbf{x})-f_0(\mathbf{x})\right)^2\Big|S_n\right]\leqslant\widetilde{O}\left(4\varepsilon^2+\frac{R^2+\sigma^2}{n}\varepsilon^{-\frac{d}{s+\alpha}}\log\frac{1}{\varepsilon}\log\frac{\kappa}{h}+\delta\right).$$

Now we choose ε to satisfy $\varepsilon^2 = \frac{1}{n}\varepsilon^{-\frac{d}{s+\alpha}}$, which gives $\varepsilon = n^{-\frac{s+\alpha}{d+2(s+\alpha)}}$. It suffices to pick $\delta = \frac{1}{n}$, then we have the desired estimation error bound

$$\begin{aligned}\mathbb{E}\left[\mathbb{E}_{\mathbf{x}}\left(\widehat{f}_n(\mathbf{x}) - f_0(\mathbf{x})\right)^2 \middle| S_n\right] &\leqslant 4n^{-\frac{2s+2\alpha}{d+2s+2\alpha}} + (R^2+\sigma^2)n^{-\frac{2s+2\alpha}{d+2s+2\alpha}} \log n \log \frac{\kappa}{h} + \frac{1}{n} \\ &\leqslant c(R^2+\sigma^2)n^{-\frac{2s+2\alpha}{d+2s+2\alpha}} \log^3 n,\end{aligned}$$

where c depends on $d, s, f_0, \tau, \log D$. □

List of Figures