Benefits of Overparameterized Convolutional Residual Networks: Function Approximation under Smoothness Constraint

Hao Liu [1] **Minshuo Chen** [2] **Siawpeng Er** [2] **Wenjing Liao** [3] **Tong Zhang** [4 5] **Tuo Zhao** [2]

Abstract

Overparameterized neural networks enjoy great representation power on complex data, and more importantly yield sufficiently smooth output, which is crucial to their generalization and robustness. Most existing function approximation theories suggest that with sufficiently many parameters, neural networks can well approximate certain classes of functions in terms of the function value. The neural network themselves, however, can be highly nonsmooth. To bridge this gap, we take convolutional residual networks (ConvResNets) as an example, and prove that large ConvResNets can not only approximate a target function in terms of function value, but also exhibit sufficient first-order smoothness. Moreover, we extend our theory to approximating functions supported on a low-dimensional manifold. Our theory partially justifies the benefits of using deep and wide networks in practice. Numerical experiments on adversarial robust image classification are provided to support our theory.

1. Introduction

Deep neural networks of enormous sizes have achieved remarkable success in various applications. Some well-known examples include ViT-Huge of 632 million parameters (Dosovitskiy et al., 2020), BERT-Large of 336 million parameters (Devlin et al., 2018), and the gigantic GPT-3 of 175 billion parameters (Brown et al., 2020). In addition to outstanding testing accuracy, there has been evidence that large neural networks favor smoothness and yield good robustness (Madry et al., 2017; Bubeck & Sellke, 2021).

Among vast literature on explaining the success of neural networks, universal approximation theories analyze how well neural networks can represent complex data models (see literature in related work section). These works focus on approximating a target function in terms of its function value (i.e., in function L_∞ norm). However, other important properties, espcifically the smoothness of the neural networks, are less investigated. A few early results provide asymptotic results on two-layer networks with smooth activation for approximating both function value and derivatives (Hornik et al., 1990; Cardaliaguet & Euvrard, 1992). Recently, Gühring et al. (2020); Hon & Yang (2021) established nonasymptotic approximation theory of feedforward networks in terms of Sobolev norms.

In real-world applications, on the other hand, practitioners empirically demonstrated a close tie between the smoothness of a trained neural network to its adversarial robustness (Gu & Rigazio, 2014; Hein & Andriushchenko, 2017; Weng et al., 2018; Miyato et al., 2018). The intuition behind is relatively clear. Consider, for instance, adding some adversarial perturbation to an input. A network of small (local) Lipschitz constant produces less deviation to the original output, and therefore, is often resilient to adversarial attackes. On the contrary, a network that is vulnerable to adversarial attacks usually has a large Lipschitz constant. Over the years, many computational methods are proposed and extensively tested in experiments for promoting network smoothness (Goodfellow et al., 2014; Madry et al., 2017; Miyato et al., 2018; Zhang et al., 2019). Apart from these explicit training methodologies, the size of a network is also recognized as a critical factor to its generalization and robustness (Zagoruyko & Komodakis, 2016; Madry et al., 2017; Wu et al., 2020). Yet, theoretical understanding is largely missing.

In this paper, we investigate universal approximation ability of neural networks with smoothness guarantees. We consider the convolutional residual networks (ConvResNet, see a description in Section 2.2) with ReLU activation as an

[1]Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Kowloon, Hong Kong. [2]School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA. [3]School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332 USA. [4]Department of Mathematics and Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong. [5]Google Research. Correspondence to: Tuo Zhao <tourzhao@gatech.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

example. We measure the approximation error of ConvResNet in terms of not only the function value, but also higher order smoothness. Specifically, suppose given a target function f belonging to a Sobolev space in a D-dimensional hypercube. We provide an approximation error estimate in terms of Sobolev norm as a function of the size of ConvResNet. We also extend our theory to functions supported on a d-dimensional Riemannian manifold ($d \ll D$). We summarize our main results in the following informal theorem.

Theorem 1.1 (informal)**.** *Consider a ConvResNet architecture with $\widetilde{M}$ residual blocks and each convolutional filter having at most $\widetilde{J}$ channels. Let $\alpha \geq 2$ and $1 \leq p \leq \infty$ be positive integers. Then*

• *(Euclidean) for any target function in a Sobolev space $W^{\alpha,p}((0,1)^D)$ with Sobolev norm $\|f\|_{W^{\alpha,p}((0,1)^D)} \leq 1$, there exists $\widetilde{f}$ yielded by the ConvResNet architecture, such that*

$$\|\widetilde{f} - f\|_{W^{s,p}} \leq \mathsf{const} \cdot (\widetilde{M}\widetilde{J})^{-\frac{\alpha-s}{D}} \quad \text{for} \quad s \in [0,1]$$

with the constant depending on D, α, p;

• *(Manifold) given $\mathcal{M} \subset \mathbb{R}^D$ a d-dimensional Riemannian manifold satisfying mild regularity conditions, for any target function in a Sobolev space $W^{\alpha,\infty}(\mathcal{M})$ with $\|f\|_{W^{\alpha,\infty}(\mathcal{M})} \leq 1$, there exists $\widetilde{f}$ yielded by the ConvResNet architecture, such that*

$$\|\widetilde{f} - f\|_{W^{k,\infty}} \leq \mathsf{const} \cdot (\widetilde{M}\widetilde{J})^{-\frac{\alpha-k}{d}} \quad \text{for} \quad k \in \{0,1\}$$

with the constant depending on $\alpha, p, \mathcal{M}$.

Our theory restricts to $s \leq 1$, since only first-order weak derivatives exist for ReLU networks. Moreover, setting $s = 0$ or $s = 1$ is of particular interest, as $s = 0$ recovers the function value approximation guarantee and $s = 1$ extends the guarantee to first-order derivatives. As can be seen, to achieve the same function value approximation error, $s = 1$ requires a larger network, but enjoys good smoothness. This can partially explain that larger networks are often more robust. We refer readers to Corollary 3.3 for more discussion.

Theorem 1.1 implies that as the number of residual blocks increases or each filter having more channels, ConvResNet gives better approximation of the target function. In order to achieve an ϵ-error, we may set $\widetilde{M}\widetilde{J} = O(\epsilon^{-\frac{D}{\alpha-s}})$ ($O(\epsilon^{-\frac{d}{\alpha-s}})$ for the manifold case), while there is no scaling restriction between $\widetilde{M}$ and $\widetilde{J}$. See an explicit configuration of ConvResNet architecture depending on $\widetilde{M}$ and $\widetilde{J}$ in Theorem 3.2 and Theorem 4.5. (Although the rate in the manifold case is independent of D, the network size inevitably weakly depends on D.)

Our result on Euclidean spaces is related to Gühring et al. (2020); Hon & Yang (2021), nonetheless, they focus on approximation guarantees of feedforward networks in terms of $W^{s,p}$ norm. It is also worth mentioning that our results are complementary to Bubeck & Sellke (2021), which provides a lower bound on network Lipschitz continuity. Bubeck & Sellke (2021) suggest that small network suffers from bad Lipschitz continuity, in fitting isoperimetric random data. However, whether large network enjoys good smoothness is questionable. Our result proves that large network indeed yields appealing Lipschitz continuity from a function approximation perspective.

The manifold case draws motivation from the fact that data in real applications are often governed by a small number of free parameters (Tenenbaum et al., 2000; Roweis & Saul, 2000; Coifman et al., 2005; Allard et al., 2012). As a concrete example, Pope et al. (2021) estimate the intrinsic dimension of many benchmark data sets, including MNIST, CIFAR-10/100, and ImageNet. A striking finding is that the intrinsic dimension of ImageNet is merely around 43, in a sharp contrast to its $224 \times 224 \times 3$ total pixels. Therefore, it is reasonable to model data as a low-dimensional Riemannian manifold, and we show ConvResNet can adapt to data geometric structures and does not suffer from the curse of ambient dimensionality.

Related work Approximation theories of feedforward neural network have been studied for a long time, most of which dedicate to function value approximation. The earliest literature dates back to late 1980s. For example, Irie & Miyake (1988); Funahashi (1989); Cybenko (1989); Hornik (1991); Chui & Li (1992); Leshno et al. (1993) investigated the approximation power of two-layer feedforward neural networks with sigmoidal activation for square integrable functions and established some asymptotic results, where the number of neurons goes to infinity. Barron (1993); Mhaskar (1996) established nonasymptotic results for the so-called "Barron" function space. For multi-layer feedforward neural networks with ReLU activation, Yarotsky (2017) analyzed the approximation of Sobolev $W^{\alpha,\infty}$ functions in a D-dimensional hypercube, and proved nonasymptotic results that given a pre-specified approximation error ϵ, the depth and width of neural networks need to be at most of the order $O(\epsilon^{-D/\alpha})$ and $O(\log(1/\epsilon))$, respectively. More recently, Suzuki (2019); Suzuki & Nitanda (2019); Liu et al. (2021) extended to more general function classes such as Besov spaces.

Approximation theories for convolutional networks are established by Zhou (2020b;a); Petersen & Voigtlaender (2020). In Zhou (2020b), the authors consider CNN with ReLU activation whose width increases linearly from the first layer to the last. They show that such a CNN can approximate functions in Sobolev $W^{\alpha,2}$ space with arbitrary accuracy for integer $\alpha \geq 2 + D/2$. To have a better control on the width of the network, the authors of Zhou (2020a) studied downsampled CNNs, and show that the downsam-

pled CNN can approximate Lipschitz ridge functions with an arbitrary accuracy. In Petersen & Voigtlaender (2020), the authors show that any approximation bounds of FNN can be achieved by CNNs. The results in Oono & Suzuki (2019); Liu et al. (2021) dedicate to convolutional residual networks. In Oono & Suzuki (2019), the authors show that ConvResNets is able to approximate Hölder functions with an arbitrary accuracy.

Theoretical results on approximating or learning functions on low-dimensional manifold can be found in Shaham et al. (2018); Chui & Mhaskar (2018); Schmidt-Hieber (2019); Chen et al. (2019a;b; 2020); Nakada & Imaizumi (2019); Cloninger & Klock (2020); Shen et al. (2019); Montanelli & Yang (2020); Liu et al. (2021; 2022). These works show that when the target function is defined on or around a low-dimensional manifold, to achieve an approximation error ϵ, the network size mainly depends on the intrinsic dimension and weakly depends on the ambient dimension.

Notations: We use lower case letters to denote scalars, bold lower case letters to denote vectors, upper case letters to denote matrices, and calligraphic letters to denote tensors and sets. For $\mathbf{x} = [x_1, ..., x_D]^\top, \mathbf{v} = [v_1, ..., v_D]^\top$, we denote $\mathbf{x}^\mathbf{v} = x_1^{v_1} \cdots x_D^{v_D}$ (if well-defined) and $|\mathbf{v}| = \sum_{i=1}^D |v_i|$. Let $\boldsymbol{\alpha} = [\alpha_1, ..., \alpha_D]^\top \in \mathbb{N}^D$ be a multi-index and f be a function, we denote $D^{\boldsymbol{\alpha}} f = \frac{\partial^{|\boldsymbol{\alpha}|} f}{\partial x_1^{\alpha_1} \cdots \partial x_D^{\alpha_D}}$. Let Ω be a subset in $\mathbb{R}^D$, we denote $\overline{\Omega}$ as its closure and $\mathrm{ch}(\Omega)$ as its convex hull. We use $B_r(\mathbf{c})$ to denote the closed Euclidean ball with radius r and centered at $\mathbf{c}$.

2. Preliminary

2.1. Sobolev Functions

We focus on studying neural networks for approximating Sobolev functions. We provide a formal definition of Sobolev functions in both Euclidean spaces and on manifolds. We begin with Sobolev functions in Euclidean spaces (Brezis & Brézis, 2011, Chapter 8).

Definition 2.1 (Sobolev spaces)**.** Let $\alpha \geq 0, 1 \leq p \leq \infty$ be integers, and domain $\Omega \subset \mathbb{R}^D$. We define Sobolev space $W^{\alpha,p}(\Omega)$ as

$$W^{\alpha,p}(\Omega) = \big\{ f \in L^p(\Omega) : D^{\boldsymbol{\alpha}} f \in L^p(\Omega) \text{ for all } |\boldsymbol{\alpha}| \leq \alpha \big\},$$

where $\boldsymbol{\alpha}$ is a multi-index.

For $f \in W^{\alpha,p}(\Omega)$, we define its Sobolev norm as

$$\|f\|_{W^{\alpha,p}(\Omega)} = \Big(\sum_{|\boldsymbol{\alpha}| \leq \alpha} \|D^{\boldsymbol{\alpha}} f\|_{L^p(\Omega)}^p \Big)^{1/p}.$$

In the special case of $p = \infty$, the Sobolev norm can be rewritten as $\|f\|_{W^{\alpha,\infty}(\Omega)} = \max_{|\boldsymbol{\alpha}| \leq \alpha} \|D^{\boldsymbol{\alpha}} f\|_{L^\infty(\Omega)}$. In this case, $\|f\|_{W^{0,\infty}} < \infty$ implies the function value is bounded, and $\|f\|_{W^{1,\infty}} < \infty$ implies both the function value and its gradient are bounded.

Our later approximation theories will provide error estimate in terms of Sobolev norms. To allow more flexibility, we define fractional Sobolev norms, which can be viewed as a generalization of Sobolev norms to non-integer α. The fractional Sobolev functions are defined as follows.

Definition 2.2 (Sobolev–Slobodeckij spaces (Slobodeckij, 1958))**.** For $0 < s < 1$ and $1 \leq p \leq \infty$, we define $W^{s,p}(\Omega)$ as

$$W^{s,p}(\Omega) = \big\{ f \in L^p(\Omega) : \|f\|_{W^{s,p}(\Omega)} < \infty \big\}$$

with

$$\|f\|_{W^{s,p}(\Omega)} = \Big(\|f\|_{L^p(\Omega)}^p + \int_\Omega \int_\Omega \Big(\frac{|f(\mathbf{x}) - f(\mathbf{y})|}{\|\mathbf{x} - \mathbf{y}\|_2^{s + D/p}} \Big)^p d\mathbf{x} d\mathbf{y} \Big)^{1/p}$$

for $1 \leq p < \infty$ and

$$\|f\|_{W^{s,\infty}(\Omega)} = \max \left\{ \|f\|_{L^\infty(\Omega)}, \operatorname{ess\,sup}_{\mathbf{x},\mathbf{y} \in \Omega} \frac{|f(\mathbf{x}) - f(\mathbf{y})|}{\|\mathbf{x} - \mathbf{y}\|_2^s} \right\}.$$

We restrict our attention to $s < 1$ for simplicity, as we focus on approximation guarantees up to first-order continuity.

Next, we extend Sobolev spaces to Riemannian manifolds. We provide a brief introduction to manifold; a more detailed description can be found in Appendix A. Roughly speaking, a Riemannian manifold $\mathcal{M}$ is a collection of local neighborhoods, each of which is diffeomorphic to a low-dimensional Euclidean space. These local neighborhoods are termed charts, and a collection of which is an atlas. We provide a formal definition.

Definition 2.3 (Atlas)**.** A smooth atlas for a d-dimensional manifold $\mathcal{M} \subset \mathbb{R}^D$ is a collection of charts $\{(U_\alpha, \phi_\alpha)\}_{\alpha \in \mathcal{A}}$, which verifies $\bigcup_{\alpha \in \mathcal{A}} U_\alpha = \mathcal{M}$ and $\phi_\alpha : U_\alpha \mapsto \mathbb{R}^d$ being diffeomorphic and pairwise compatible, i.e.,

$$\phi_\alpha \circ \phi_\beta^{-1} : \phi_\beta(U_\alpha \cap U_\beta) \to \phi_\alpha(U_\alpha \cap U_\beta) \quad \text{and}$$
$$\phi_\beta \circ \phi_\alpha^{-1} : \phi_\alpha(U_\alpha \cap U_\beta) \to \phi_\beta(U_\alpha \cap U_\beta)$$

are both smooth for any $\alpha, \beta \in \mathcal{A}$. An atlas is called finite if it contains finitely many charts.

To define Sobolev spaces on a manifold $\mathcal{M}$, we shall consider function regularity on each chart, as charts are geometrically "akin" to a Eulidean space through the chart mapping ϕ_α. One caveat, however, is that the chart mapping ϕ_α can be arbitrarily rescaled, which results in potential unboundedness. We therefore, fix an atlas on $\mathcal{M}$ to mitigate this issue. We are ready to define Sobolev spaces on a manifold (Driver, 2003, Definition 48.17).

Definition 2.4 (Sobolev spaces on manifold)**.** Let $\mathcal{M}$ be a compact Riemannian manifold of dimension d. Let $\{(U_i, \phi_i)\}_{i=1}^{C_\mathcal{M}}$ be a finite atlas on $\mathcal{M}$ and $\{\rho_i\}_{i=1}^{C_\mathcal{M}}$ be a partition of unity on $\mathcal{M}$ such that $\mathrm{supp}(\rho_i) \subset U_i$. For integers $k \geq 0$ and $1 \leq p \leq \infty$, a function $f : \mathcal{M} \to \mathbb{R}$ is in the Sobolev space $W^{k,p}(\mathcal{M})$ if

$$\|f\|_{W^{k,p}(\mathcal{M})} := \sum_{i=1}^{C_\mathcal{M}} \|(f\rho_i) \circ \phi_i^{-1}\|_{W^{k,p}(\phi_i(U_i))} < \infty.$$

Since $\mathcal{M}$ is compact, a finite altas exists on $\mathcal{M}$. Besides, we introduce the partition of unity ρ_i to follow the standard definition in Tu (2010, Definition 13.4). The existence of a smooth partition of unity is shown in Appendix A. From Definition 2.4, we observe that a Sobolev function on $\mathcal{M}$ is locally Sobolev on each chart.

2.2. Convolutional Residual Networks

We consider one-sided stride-one convolution in our network. Let $\mathcal{W} = \{\mathcal{W}_{j,k,l}\} \in \mathbb{R}^{C' \times K \times C}$ be a filter where C' is the output channel size, K is the filter size and C is the input channel size. For $Z \in \mathbb{R}^{D \times C}$, the convolution of $\mathcal{W}$ with Z gives $Y = \mathcal{W} * Z \in \mathbb{R}^{D \times C'}$ with

$$Y_{i,j} = \sum_{k=1}^{K} \sum_{l=1}^{C} \mathcal{W}_{j,k,l} Z_{i+k-1,l},$$

where we set $Z_{i+k-1,l} = 0$ for $i + k - 1 > D$. See a graphical demonstration in Figure 1(a).

In this paper, we study convolutional residual networks (ConvResNets) equipped with the rectified linear unit (ReLU) activation function ($\mathrm{ReLU}(z) = \max(z, 0)$). The ConvResNet we consider consists consecutively of a padding layer, several residual blocks, and finally a fully connected output layer.

Given an input vector $\mathbf{x} \in \mathbb{R}^D$, the network first applies a padding operator $P : \mathbb{R}^D \to \mathbb{R}^{D \times C}$ for some integer $C \geq 1$ such that

$$Z = P(\mathbf{x}) = \begin{bmatrix} \mathbf{x} & \mathbf{0} & \cdots & \mathbf{0} \end{bmatrix} \in \mathbb{R}^{D \times C}.$$

Then the matrix Z is passed through M residual blocks. To ease the notation, we denote the input matrix to the m-th block as Z_m and its output as Z_{m+1} (Consequently, $Z_1 = Z$).

In the m-th block, let $\mathcal{W}_m = \{\mathcal{W}_m^{(1)}, ..., \mathcal{W}_m^{(L_m)}\}$ and $\mathcal{B}_m = \{B_m^{(1)}, ..., B_m^{(L_m)}\}$ be a collection of filters and biases of proper sizes. The m-th residual block maps its input matrix Z_m from $\mathbb{R}^{D \times C}$ to $\mathbb{R}^{D \times C}$ by the operator

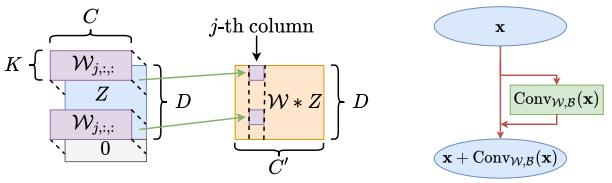
$$\mathrm{Conv}_{\mathcal{W}_m, \mathcal{B}_m} + \mathrm{id},$$

where id is the identity mapping (also known as the shortcut connection) and

$$\mathrm{Conv}_{\mathcal{W}_m, \mathcal{B}_m}(Z_m) = \mathrm{ReLU}\Big(\mathcal{W}_m^{(L_m)} * \cdots \\ \cdots * \mathrm{ReLU}\left(\mathcal{W}_m^{(1)} * Z_m + B_m^{(1)}\right) \cdots + B_m^{(L_m)}\Big), \quad (1)$$

with ReLU applied entrywise. We denote the mapping from input $\mathbf{x}$ to the output of the M-th residual block as

$$Q(\mathbf{x}) = (\mathrm{Conv}_{\mathcal{W}_M, \mathcal{B}_M} + \mathrm{id}) \circ \cdots \\ \cdots \circ (\mathrm{Conv}_{\mathcal{W}_1, \mathcal{B}_1} + \mathrm{id}) \circ P(\mathbf{x}). \quad (2)$$



(a) Convolution. (b) A residual block.

Figure 1: (a) Convolution of $\mathcal{W} * Z$, where the input is $Z \in \mathbb{R}^{D \times C}$, and the output is $\mathcal{W} * Z \in \mathbb{R}^{D \times C'}$. Here $\mathcal{W} = \{\mathcal{W}_{j,k,l}\} \in \mathbb{R}^{C' \times K \times C}$ is a filter where C' is the output channel size, K is the filter size and C is the input channel size. $\mathcal{W}_{j,:,:}$ is a $D \times C$ matrix for the j-th output channel. (b) A convolutional residual block.

Given (2), a ConvResNet applies an additional fully connected layer to Q and outputs

$$f(\mathbf{x}) = W \otimes Q(\mathbf{x}) + b,$$

where $W \in \mathbb{R}^{D \times C}$ and $b \in \mathbb{R}$ are a weight matrix and a bias, respectively, and $\otimes$ denotes sum of entrywise product, i.e., $W \otimes Q(\mathbf{x}) = \sum_{i,j} W_{i,j}[Q(\mathbf{x})]_{i,j}$. To this end, we define a class of ConvResNets of the same architecture as

$$\mathcal{C}(M, L, J, K, \kappa_1, \kappa_2) = \\ \big\{ f \mid f(\mathbf{x}) = W \otimes Q(\mathbf{x}) + b \text{ with } \|W\|_\infty \vee |b| \leq \kappa_2, \\ Q(\mathbf{x}) \text{ in the form of (2) with } M \text{ residual blocks.} \\ \text{The number of filters per block is bounded by } L; \\ \text{filter size is bounded by } K; \\ \text{the number of channels is bounded by } J; \\ \max_{m,l} \|\mathcal{W}_m^{(l)}\|_\infty \vee \|B_m^{(l)}\|_\infty \leq \kappa_1 \big\}. \quad (3)$$

Here $\|\cdot\|_\infty$ denotes the entrywise maximum norm, i.e., when the input argument is a vector, it returns the vector ℓ^∞ norm; when the input is a matrix or a tensor, it returns the maximum magnitude of its entries, e.g., for a 3-dimensional tensor $\mathcal{W}$, $\|\mathcal{W}\|_\infty = \max_{j,k,l} |\mathcal{W}_{j,k,l}|$.

3. Approximation in Euclidean Space

Consider a Sobolev function class defined on a unit hypercube $(0, 1)^D$. We aim to use convolutional residual networks for approximating functions in the target class in

terms of the $W^{s,p}$ norm. Here p is a positive integer and s can vary in $[0,1]$; in particular, $s=0$ corresponds to function value approximation, and $s=1$ resembles the result Section 1. We formally define our target function class as a Sobolev norm ball.

Assumption 3.1. Let $\alpha \geq 2, 1 \leq p \leq +\infty$ be integers. Assume the target function f satisfies

$$f \in W^{\alpha,p}\left((0,1)^D\right) \quad \text{and} \quad \|f\|_{W^{\alpha,p}((0,1)^D)} \leq 1.$$

We set the norm ball of radius 1 for the sake of simplicity, while the results in the sequel hold for any constant radius. We also let $\alpha \geq 2$ for techincal convenience. In the following theorem, we show that ConvResNets can approximate any functions in a Sobolev norm ball in terms of $W^{s,p}$ norm ($s \leq 1$). The approximation error is obtained as a function of the network configuration.

Theorem 3.2. *For any positive integers $K \in [2, D]$, $\widetilde{M}$, and $\widetilde{J} > 0$, we choose*

$$L = O(\log(\widetilde{M}\widetilde{J})),\ J = O(\widetilde{J}),\ \kappa_1 = O((\widetilde{M}\widetilde{J})^{1/D}),$$
$$\kappa_2 = O((\widetilde{M}\widetilde{J})^{1/D}),\ M = O(\widetilde{M}).$$

Then given $s \in [0,1]$, the ConvResNet architecture $\mathcal{C}(M, L, J, K, \kappa_1, \kappa_2)$ can approximate any function f satisfying Assumption 3.1, i.e., there exists $\widetilde{f} \in \mathcal{C}(M, L, J, K, \kappa_1, \kappa_2)$ with

$$\|\widetilde{f} - f\|_{W^{s,p}((0,1)^D)} \leq C_1(\widetilde{M}\widetilde{J})^{-\frac{\alpha-s}{D}}$$

for some constant C_1 depending on D, α, p.

Theorem 3.2 says that the approximation power of ConvResNet amplifies as its width and depth increase. To better interpret the result, we choose $s=1$ and $p=\infty$, which corresponds to simultaneously approximating function value and first-order derivatives.

Corollary 3.3. *In the setup of Theorem 3.2, taking $s=1$ and $p=\infty$, the ConvResNet architecture $\mathcal{C}(M, L, J, K, \kappa_1, \kappa_2)$ can approximate any f satisfying Assumption 3.1 up to first-order, i.e., there exists $\widetilde{f} \in \mathcal{C}(M, L, J, K, \kappa_1, \kappa_2)$ with*

$$\|\widetilde{f} - f\|_\infty \leq C_2(\widetilde{M}\widetilde{J})^{-\frac{\alpha-1}{D}} \quad \text{and}$$
$$\sup_i \left\| \frac{\partial \widetilde{f}}{\partial x_i} - \frac{\partial f}{\partial x_i} \right\|_\infty \leq C_2(\widetilde{M}\widetilde{J})^{-\frac{\alpha-1}{D}},$$

where the constant C_2 depends on D and α. In particular, we have Lipschitz continuity bound

$$\|\widetilde{f}\|_{\text{Lip}} \leq 1 + C_2\sqrt{D}(\widetilde{M}\widetilde{J})^{-\frac{\alpha-1}{D}}.$$

Theorem 3.2 and Corollary 3.3 have rich implications.

Large network for smooth approximation. Taking $s=0$ in Theorem 3.2 recovers function approximation in terms of L_∞ norm. The corresponding approximation error scales as $O((\widetilde{M}\widetilde{J})^{-\frac{\alpha}{D}})$. A quick comparison to Corollary 3.3 indicates that in order to additionally capture the first-order information of a target function, large network is needed to achieve the same function value error bound.

Arbitrary width and depth. Gühring et al. (2020); Hon & Yang (2021) provide approximation guarantees of feedforward networks in terms of $W^{s,p}$ norm. Despite different network architectures, we remark that our theory covers general networks with arbitrary width and depth. More specifically, for a given approximation error ϵ, Gühring et al. (2020) set the network depth and width as $O(\log 1/\epsilon)$ and $O(\epsilon^{-D/(\alpha-s)})$, respectively. Yet in our result, we only need to ensure $\widetilde{M}\widetilde{J} = O(\epsilon^{-D/(\alpha-s)})$, which does not require any scaling relation between $\widetilde{M}$ and $\widetilde{J}$.

Theorem 3.2 can be used as a tool to analyze the empirical residual error. Specifically, assume the response in the data set contains bounded zero–mean noise, we have the following probability bound on the upper bound of the empirical residual error (see a proof in Appendix D)

Theorem 3.4. *Let $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ be a given data set where $\mathbf{x}_i$'s are i.i.d. samples from some distribution defined on $[0,1]^D$ and*

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

with i.i.d. noise ξ_i's satisfying $\mathbb{E}[\xi_i] = 0$ and $|\xi_i| \leq \sigma$ for all $i = 1, ..., n$. Assume f satisfy Assumption 3.1 with $p = +\infty$. For $0 < \varepsilon < \min\{\sigma, 1\}$, let $\mathcal{C} = \mathcal{C}(M, L, J, K, \kappa_1, \kappa_2)$ be the network architecture in Theorem 3.2 with $\widetilde{M}\widetilde{J} = \left(\frac{\varepsilon}{C_1}\right)^{-D/\alpha} = O(\varepsilon^{-D/\alpha})$. We have

$$\mathbb{P}\Bigg(\exists \widetilde{f} \in \mathcal{C} : \left\|\widetilde{f}\right\|_{\text{Lip}} \leq 1 + \sqrt{D}\varepsilon^{\frac{\alpha-1}{\alpha}} \ \text{and} \ \frac{1}{n}\sum_{i=1}^n (\widetilde{f}(\mathbf{x}_i) - y_i)^2 \leq 2\varepsilon^2 + \sigma^2\Bigg)$$
$$\geq 1 - \exp\left(-\frac{3n\varepsilon^2}{104\sigma^4}\right). \tag{4}$$

Theorem 3.4 implies that with high probability, larger network architectures ensure the existence of a network that has small empirical residual error as well as certain smoothness, i.e., a bounded Lipschitz constant whcih is close to that of the underlying function. Our result is an upper bound counterpart of Bubeck & Sellke (2021, Theorem 3), in which a high probability lower bound of the Lipschitz constant is derived.

Connection to adversarial robustness. Consider, for example, the supervised learning scenario. Noisy or noiseless response is generated by a ground truth function satisfying Assumption 3.1. Corollary 3.3 then indicates the existence of a properly large ConvResNet capable of smoothly approximating the data model, and the network's Lipschitz

constant is approximately that of the ground truth function. Such Lipschitz continuity should be considered nearly optimal, in viewing of the smoothness of the ground truth function. The network's Lipschitz continuity closely relates to adversarial risk (Uesato et al., 2018; Zhao et al., 2021) defined as

Definition 3.5 (Adversarial risk)**.** Given a data distribution ρ, and a loss function $l(\cdot,\cdot)$, for a positive constant $\delta > 0$, we define the adversarial risk of a network $\widetilde{f}$ as

$$R(\widetilde{f},\delta) = \mathbb{E}_{(\mathbf{x},y)\in \mathrm{supp}(\rho)} \left[\sup_{\mathbf{x}'\in B_\delta(\mathbf{x})} \ell\left(\widetilde{f}(\mathbf{x}'),y\right)\right], \quad (5)$$

where $B_\delta(\mathbf{x})$ is the Euclidean ball with radius δ centered at $\mathbf{x}$.

In the case $\delta = 0$, the adversarial risk $R(\widetilde{f},0)$ reduces to the population risk $\mathbb{E}_{(\mathbf{x},y)\in\mathrm{supp}(\rho)}\left[\ell\left(\widetilde{f}(\mathbf{x}),y\right)\right]$. Based on Theorem 3.2 and Corollary 3.3, we have the following theorem on adversarial risk (see a proof in Appendix E):

Theorem 3.6. *Let ρ be a data distribution defined on $[0,1]^D \times [-R,R]$ for some constant R and $l(\cdot,\cdot)$ be a loss function with Lipschitz constant L_{Lip}. Denote the population risk minimizer by f:*

$$f = \operatorname*{argmin}_g \mathbb{E}_{(\mathbf{x},y)\in\mathrm{supp}(\rho)} l(g(\mathbf{x}),y). \quad (6)$$

Assume f satisfies Assumption 3.1 with $p = +\infty$. For $0 < \varepsilon < 1$, let $\mathcal{C}(M,L,J,K,\kappa_1,\kappa_2)$ be the network architecture in Theorem 3.2 with $\widetilde{M}\widetilde{J} = \left(\frac{\varepsilon}{C_1}\right)^{-D/\alpha} = O(\varepsilon^{-D/\alpha})$. Then there exists $\widetilde{f} \in \mathcal{C}(M,L,J,K,\kappa_1,\kappa_2)$ so that

$$\|\widetilde{f} - f\|_\infty \le \varepsilon, \quad \left\|\widetilde{f}\right\|_{\mathrm{Lip}} \le 1 + \sqrt{D}\varepsilon^{\frac{\alpha-1}{\alpha}} \quad (7)$$

and

$$R(\widetilde{f},\delta) \le R(\widetilde{f},0) + L_{\mathrm{Lip}}\left(1+\sqrt{D}\varepsilon^{\frac{\alpha-1}{\alpha}}\right)\delta. \quad (8)$$

In Theorem 3.6, the difference between the adversarial risk and population risk depends on the Lipschitz constant of the network $\widetilde{f}$, the Lipscthiz constant of the loss function and the adversarial parameter δ. It implies that large networks can give rise to smooth functions with a small adversarial risk, i.e., adversarially robust. This partially explains the empirical observation that large networks are often smooth with respect to input, and hence, tend to have better robustness. However, how to use practical training algorithms to find such networks remains curiously unclear.

4. Approximation on Manifold

Theorem 3.2 indicates a curse of data dimensionality: When data dimension D is large, such as image data, Theorem 3.2 converges extremely slowly and becomes less attractive. Motivated by applications, we model data as a low-dimensional Riemannian manifold $\mathcal{M}$ and extend our approximation theory to functions defined on $\mathcal{M}$. We will show that ConvResNet is adaptable to manifold structures. We first impose some mild regularity conditions.

Assumption 4.1. $\mathcal{M}$ is a d-dimensional compact Riemannian manifold isometrically embedded in $\mathbb{R}^D$. It's range is bounded by B, i.e., there exists a constant $B > 0$ such that for any $\mathbf{x} \in \mathcal{M}$, we have $\|\mathbf{x}\|_\infty \le B$.

Besides boundedness, we characterize the curvature of manifold by the following geometric notion.

Definition 4.2 (Reach (Federer, 1959; Niyogi et al., 2008))**.** Define the set

$$G = \{\mathbf{x}\in\mathbb{R}^D : \exists \text{ distinct } \mathbf{p},\mathbf{q}\in\mathcal{M} \text{ such that } d(\mathbf{x},\mathcal{M}) = \|\mathbf{x}-\mathbf{p}\|_2 = \|\mathbf{x}-\mathbf{q}\|_2\}.$$

Then the reach of $\mathcal{M}$ is defined as

$$\mathrm{reach}(\mathcal{M}) = \inf_{\mathbf{x}\in\mathcal{M}} \inf_{\mathbf{y}\in G} \|\mathbf{x}-\mathbf{y}\|_2.$$

To roughly put, a large reach implies that the manifold is flat. While a manifold with a small reach can be highly zigzagging. Therefore, the reach is highly relevant to the difficulty of capturing the local structures on a manifold. We assume a positive reach on $\mathcal{M}$.

Assumption 4.3. The reach of $\mathcal{M}$ is $\tau > 0$.

Similar to Section 3, we consider a Sobolev norm ball on $\mathcal{M}$ as target function class.

Assumption 4.4. Let $\alpha \ge 2$ be an integer. Assume the target function f satisfies

$$f \in W^{\alpha,\infty}(\mathcal{M}) \quad \text{and} \quad \|f\|_{W^{\alpha,\infty}(\mathcal{M})} \le 1.$$

We now present a counterpart of Theorem 3.2, showing an efficient approximation of functions in a Sobolev norm ball on $\mathcal{M}$.

Theorem 4.5. *For any positive integers $K \in [2,D]$, $\widetilde{M}$, and $\widetilde{J} > 0$, we choose*

$$L = O(\log(\widetilde{M}\widetilde{J})) + D,\ J = O(D\widetilde{J}),$$
$$\kappa_1 = O((\widetilde{M}\widetilde{J})^{1/d}),\ \kappa_2 = O((\widetilde{M}\widetilde{J})^{1/d}),\ M = O(\widetilde{M}).$$

Then given $k \in \{0,1\}$, the ConvResNet architecture $\mathcal{C}(M,L,J,K,\kappa_1,\kappa_2)$ can approximate any function f satisfying Assumption 4.4, i.e., there exists $\widetilde{f} \in \mathcal{C}(M,L,J,K,\kappa_1,\kappa_2)$ with

$$\|\widetilde{f} - f\|_{W^{k,\infty}(\mathcal{M})} \le C_3(\widetilde{M}\widetilde{J})^{-\frac{\alpha-k}{d}},$$

where constant C_3 depends on d,α,B,τ, and the surface area of $\mathcal{M}$.

As can be seen, the approximation error decays at a rate only depending on intrinsic data dimension d, which is a significant improvement over Theorem 3.2 given $d \ll D$. We also note that the size of ConvResNet has a weak dependence on D, yet it is inevitable due to the residual connection preserves input dimensionality.

Theorem 4.5 can be viewed as further results of recent advances on the adaptability of neural networks for approximating functions on low-dimensional structures. In particular, Chen et al. (2019a) and Schmidt-Hieber (2019) share a very similar setup as Theorem 4.5, and established function value approximation theories.

5. Numerical Experiments

We verify our theory by numerical experiments. Due to the complex structure of convolutional residual networks, directly estimating the Lipschitz constant is rather difficult. We instead testing the adversarial robustness as an indication of the network smoothness.

We consider the TRADES model which uses a data driven smoothness regularization and encourages model smoothness. By keeping the same clean testing accuracy, we can compare model smoothness through the robust testing accuracy. We follow the setup in TRADES (Zhang et al., 2019), and report the performance of WideResNet (Zagoruyko & Komodakis, 2016) with different widening factor (WF) and number of convolutional layers per residual block (we term as "depth" in the sequel). We use the CIFAR-10 data set. Hyperparameters in training are set as follows: perturbation diameter $\epsilon = 0.031$ under the ℓ_∞ norm, step size for generating perturbation 0.007, number of iterations 10, learning rate 0.1, batch size $b = 128$ and run 76 epochs on the training dataset. We run the White-box attacks by applying PGD attack with 20 iterations (PGD-20) and the step size is 0.003. We report the robust accuracy $\mathcal{A}_{\rm rob}$ and the natural accuracy $\mathcal{A}_{\rm nat}$ on the test data set.

The training objective is

$$\min_f \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}} \mathcal{L}(f(\mathbf{x}), y) + \max_{\|\widetilde{\mathbf{x}}-\mathbf{x}\|_\infty \le \epsilon} \mathcal{R}\left(f(\mathbf{x}), f\left(\widetilde{\mathbf{x}}\right)\right)/\lambda,$$

where $\mathcal{L}$ is the cross entropy loss, $\mathcal{R}$ is the KL-divergence, $\mathbf{x}$ is the clean input, $\widetilde{\mathbf{x}}$ is the adversarial input, y is the label, λ is the tuning parameter controlling the strength of the regularizer, and $\mathcal{D}$ denotes the training dataset $\{\mathbf{x}_i, y_i\}_{i=1}^n$.

For a fair comparison, we tune λ such that networks of different sizes achieve approximately the same natural accuracy. This can be understood as achieving approximately the same L_∞ approximation error to the data model. As can be seen in Table 1, $\mathcal{A}_{\rm nat}$ of different models about matches the performance in Zhang et al. (2019), indicating the network has been sufficiently trained. By comparing the robust accuracy $\mathcal{A}_{\rm rob}$, we observe that wider and deeper WideResNet attains better robustness. When fixing the depth, a wider network can achieve a higher robust accuracy. Similarly, when fixing the widening factor, a deeper network can achieve a higher robust accuracy.

Table 1: Performance of Wide Residual Networks with different widening factors and depths under PGD-20 attacks.

Depth	WF	$\mathcal{A}_{\rm nat}$	$\mathcal{A}_{\rm rob}$
16	1	$78.87 \pm 0.47\%$	$34.31 \pm 0.45\%$
	2	$79.34 \pm 0.28\ \%$	$46.14 \pm 0.21\%$
	4	$79.97 \pm 0.04\%$	$51.40 \pm 0.16\%$
22	1	$78.51 \pm 0.25\%$	$41.47 \pm 0.11\%$
	2	$79.49 \pm 0.48\%$	$49.63 \pm 0.07\%$
	4	$80.81 \pm 0.44\%$	$53.36 \pm 0.21\%$
28	1	$79.46 \pm 0.06\%$	$43.33 \pm 0.57\%$
	2	$79.01 \pm 0.11\%$	$50.85 \pm 0.07\%$
	4	$80.90 \pm 0.71\%$	$54.45 \pm 0.14\%$
34	1	$78.58 \pm 0.09\%$	$46.14 \pm 0.16\%$
	2	$79.29 \pm 0.35\%$	$51.63 \pm 0.28\%$
	4	$80.79 \pm 0.71\%$	$55.28 \pm 0.35\%$

6. Proof Sketch

We highlight key steps in establishing Theorem 3.2 and 4.5 in this section. Full proofs are deferred to Appendix C and F, respectively.

6.1. Proof Sketch of Theorem 3.2

The main idea consists of two stages: 1) Approximating target function f in terms of $W^{s,p}$ norm using a sum of averaged Taylor polynomials; 2) Implementing the sum of averaged Taylor polynomials by a given width and depth ConvResNet up to a certain error. In stage 1), we rely on tools from the finite element anaylsis to quantify approximation error. In stage 2), we first represent polynomials using convolutional networks, and then assemble them according to the specified width and depth as a ConvResNet. We dive into the following four steps.

Step 1: Decompose f using a partition of unity. Given the network size parameter $\widetilde{M}$ and $\widetilde{J}$, we define a partition of unity $\{\phi_j\}_{j=1}^{N^D}$ on $(0,1)^D$ for an integer $N = O((\widetilde{M}\widetilde{J})^{1/D})$, so that each ϕ_j is supported on a small hypercube of edge length $\frac{4}{3N}$. The function f is decomposed into $f = \sum_{j=1}^{N^D} f_j$ with $f_j = f\phi_j$. See Figure 2(a) for an illustration.

Step 2: Averaged Taylor polynomial approximation. Each f_j is a Sobolev function, which may not have classical derivatives but weak derivatives. Similar to approximating differentiable functions by Taylor polynomials, we approximate f_j by an averaged Taylor polynomial $\widehat{f}_j$, which is

defined in an integral form and indeed is a polynomial. The approximation error of averaged Taylor polynomial is similar to that of using Taylor polynomial, and can be found in Lemma C.6.

Step 3: Network implementation. As shown in Lemma G.6 and C.9, CNN can approximate multiplication and compositions of muliplications well. Since a polynomial is a sum of compositions of multiplication, each $\widehat{f}_i$ can be approximated by a sum of $O(1)$ CNNs, and therefore $\sum_{i=1}^{N^D} \widehat{f}_i$ is approximated by a sum of $O(N^D)$ CNNs, each of which has width of $O(1)$. We prove in Lemma C.11 that such a sum can be realized by a sum of $\widetilde{M}$ CNNs with width $\widetilde{J}$. The new sum can be realized by a ConvResNet with $\widetilde{M}$ residual blocks (Lemma C.12), where each summand corresponds to a residual block and the sum is realized using skip-layer connections.

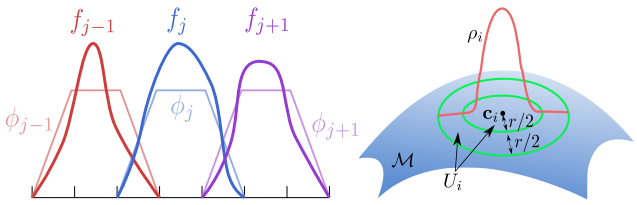
Step 4: Error estimation. To estimate the approximation error of $\widetilde{f}$, we decompose the error as

$$\|\widetilde{f} - f\|_{W^{s,p}(0,1)^D} \leq \sum_{j=1}^{N^D} \|\widetilde{f}_j - \widehat{f}_j\|_{W^{s,p}((0,1)^D)} + \sum_{j=1}^{N^D} \|\widehat{f}_j - f_j\|_{W^{s,p}((0,1)^D)}. \quad (9)$$

On the right-hand side of (9), the second term is the approximation error of averaged Taylor polynomial, whose upper bound is given by Lemma C.8.

The first term is the network implementation error. We derive an upper bound of it in Lemma C.10. In the proof of Lemma C.10, we first derive an upper bound with respect to the $W^{k,p}$ norm for $k = 0, 1$. The case $k = 0$ corresponds to the error of function value approximation, and the case $k = 1$ corresponds to the error of first order weak derivative approximation. Note that each $\widehat{f}_j$ is a polynomial, and each $\widetilde{f}_j$ consists of compositions of $\widetilde{\times}$, the network approximation of multiplication $\times$. The error indeed is the approximation error of compositions of $\widetilde{\times}$. We first derive the $W^{k,\infty}$ approximation error of $\widetilde{\times}$ and then show that compositions of $\widetilde{\times}$ have $W^{k,p}$ approximation errors of the same order. After the upper bounds of $W^{0,p}$ and $W^{1,p}$ errors are derived, these upper bounds are generalized to $W^{s,p}$ errors using an argument on interpolation spaces, which is discussed in Appendix G.2.

Combining the upper bounds of both terms in (9) gives rise to the total approximation error as a function of N. Utilizing the relation $\widetilde{M}\widetilde{J} = O(N^D)$, we can further express the approximation error in terms of number of blocks and width of the ConvResNet.



(a) Partition of unity on $(0,1)^D$. (b) Partition of unity on $\mathcal{M}$.

Figure 2: (a) Illustration of ϕ_j's and f_j's in Step 1 of the proof of Theorem 3.2. (b) Illustration of the construction of charts and paritition of unity in Step 1 of the proof of Theorem 4.5. The red curve represents a cross section of ρ_i.

6.2. Proof Sketch of Theorem 4.5

We exploit the geometric nature of manifold $\mathcal{M}$ and Sobolev functions on it to prove Theorem 4.5. By an explicit construction of a finite atlas on $\mathcal{M}$ based on the curvature condition in Assumption 4.3, we first restrict ourselves to a single chart on $\mathcal{M}$. Recall Definition 2.4 that a Sobolev function f on $\mathcal{M}$ is locally Sobolev on a chart. We are thus, able to locally approximate f on each chart by the results in Theorem 3.2. However, the main challenge stems from combining these local approximations to obtain a global guarantee. This requires to determine which charts a given input belongs to. We develop a chart determination subnetwork for approximating indicator functions of charts, nonetheless, its Lipschitz continuity is troublesome due to the sharp jump on the boundary of a chart. We resolve such an issue by carefully constructing a partition of unity vanishing at a neighborhood of the boundary of charts. We provide more details in the following four steps.

Step 1: Decompose f using an atlas and partition of unity of $\mathcal{M}$. We first construct an atlas and a partition of unity of $\mathcal{M}$ so that each function in the partition of unity is compactly supported in a chart (Lemma F.1). To construct an atlas of $\mathcal{M}$, we use a set of D-dimensional Euclidean balls $\{B_{r/2}(\mathbf{c}_i)\}_{i=1}^{C_\mathcal{M}}$ with centers $\{\mathbf{c}_i\}_{i=1}^{C_\mathcal{M}} \subset \mathcal{M}$ and radius $r/2$ satisfying $0 < r < \tau/4$ to cover $\mathcal{M}$. Since $\mathcal{M}$ is compact, $C_\mathcal{M}$ is finite. The collection of intersections between each ball and $\mathcal{M}$, denoted by $\{\widetilde{U}_i\}_{i=1}^{C_\mathcal{M}}$ with $\widetilde{U}_i = B_{r/2}(\mathbf{c}_i) \cap \mathcal{M}$, forms an open cover of $\mathcal{M}$. It is guaranteed that there exists a C^∞ partition of unity $\{\rho_i\}_{i=1}^{\mathcal{M}}$ so that ρ_i is supported in $\widetilde{U}_i$ (Lemma H.1). We then double the radius and denote $U_i = B_r(\mathbf{c}_i) \cap \mathcal{M}$. The collection $\{U_i\}_{i=1}^{C_\mathcal{M}}$ is also an open cover of $\mathcal{M}$. Since $\widetilde{U}_i \subset U_i$, ρ_i is compactly supported in U_i and the distance between the support of ρ_i and ∂U_i is at least $r/2$. For each U_i, an orthogonal projection φ_i with proper scaling and shifting, which projects any $\mathbf{x} \in U_i$ to a tangent plane, is constructed so that $\varphi_i(U_i) \subset (0,1)^d$. See the proof of Lemma F.1 for details. With this construction, we illustrate U_i and ρ_i in Figure 2(b). We then focus on

the atlas $\{U_i, \varphi_i\}_{i=1}^{C_{\mathcal{M}}}$ and partition of unity $\{\rho_i\}_{i=1}^{\mathcal{M}}$. We decompose f as $f = \sum_{i=1}^{C_{\mathcal{M}}} (f_i \circ \varphi_i^{-1}) \circ \varphi_i$ with $f_i = f\rho_i$.

Step 2: Averaged Taylor polynomial approximation. In the decomposition in Step 1, each $f_i \circ \varphi_i^{-1}$ is a Sobolev function compactly supported in $\varphi_i(U_i) \subset (0,1)^d$. Extend $f_i \circ \varphi_i^{-1}$ to $(0,1)^d$ by 0. The extended function has the same smoothness as $f_i \circ \varphi_i^{-1}$, and can be approximated by a sum of local averaged Taylor polynomials $\sum_{i=1}^{N^d} \widehat{f}_{i,j}$, as what has been done in the proof of Theorem 3.2.

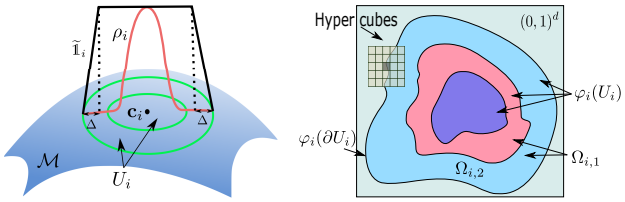
Step 3: Network implementation. Each polynomial $\widehat{f}_{i,j}$ can be approximated by a CNN $\widetilde{f}_{i,j}$. Since we are only interested in the value of $\widetilde{f}_{i,j} \circ \varphi_i(\mathbf{x})$ when $\mathbf{x} \in U_i$, we need to determine the chart it belongs to. We accomplish this by introducing a chart determination function $\mathbb{1}_i(\mathbf{x}) = \mathbb{1}_{[0,r^2]} \circ d_i^2(\mathbf{x})$, where $\mathbb{1}_{[0,r^2]}(a)$ is a step function which outputs 1 when $a \in [0, r^2]$ and outputs 0 otherwise, $d_i^2(\mathbf{x})$ computes the squared Euclidean distance between $\mathbf{x}$ and $\mathbf{c}_i$. The squared distance function d_i^2 can be approximated by a CNN with high accuracy. To approximate the step function $\mathbb{1}_{[0,r^2]}$, we construct a CNN which outputs 1 on $[0, r^2 - \Delta]$, 0 on $[r^2, \infty)$ and is linear on $[r^2 - \Delta, r^2]$ for some small Δ. The CNN approximation of $\mathbb{1}_i$, denoted by $\widetilde{\mathbb{1}}_i$, is illustrated in Figure 3(a). Our network approximation of f is constructed as

$$\widetilde{f}(\mathbf{x}) = \sum_{i=1}^{C_{\mathcal{M}}} \sum_{j=1}^{N^d} \widetilde{\times}(\widetilde{f}_{i,j} \circ \varphi_i(\mathbf{x}), \widetilde{\mathbb{1}}_i(\mathbf{x})),$$

where $\widetilde{\times}$ denotes the CNN approximation of multiplication. By Lemma C.11 and C.12, $\widetilde{f}$ can be realized by a ConvResNet with $\widetilde{M}$ blocks and width of $O(\widetilde{J})$ as long as $\widetilde{M}\widetilde{J} = O(N^d)$.

Step 4: Error estimation. We decompose the error into two parts: 1) the error between f and its averaged Taylor polynomial approximation, and 2) the error between the averaged Taylor polynomial and its network approximation, see (42) in Appendix F. The first part can be bounded using Lemma C.8. The second part is characterized by the approximation error of $\widetilde{\times}$ for multiplication, of $\widetilde{f_{i,j}}$ for averaged Taylor polynomials, and of $\widetilde{\mathbb{1}}_i$ for chart determination $\mathbb{1}_i$. The first two errors can be bounded using techniques similar to those in the proof of Theorem 3.2.

For the approximation error of $\widetilde{\mathbb{1}}_i$, bounding its $W^{1,\infty}$ norm is the most challenging task. To derive an upper bound, one needs to bound $|(\widetilde{f}_{i,j} \circ \varphi_i) \times (\partial(\widetilde{\mathbb{1}}_i \circ \varphi_i^{-1})/\partial z_l)|$ for $l = 1, ..., d$ and $\mathbf{z} \in \varphi_i(U_i)$. In our network construction, $\widetilde{\mathbb{1}}_i \circ \varphi_i^{-1}$ is linear on a narrow band, denoted by $\Omega_{i,2}$, with width of $O(\Delta)$. Its weak derivative on the narrow band is of $O(1/\Delta)$, which blows up as $\Delta \to 0$ and causes problems. To eliminate the effect of Δ, we show that the value of $\widetilde{f}_{i,j} \circ \varphi_i$ is small enough so that its product with $\partial(\widetilde{\mathbb{1}}_i \circ \varphi_i^{-1})/\partial z_l$ does not blow up as $\Delta \to 0$. Specifically, thanks to the fact that f_i is compactly supported on U_i, we have $f_i \circ \varphi^{-1}$ is compactly supported on $\varphi_i(U_i)$. Therefore there exists another band $\Omega_{i,1}$ adjacent to $\varphi_i(\partial U_i)$ so that $f_i \circ \varphi^{-1} = 0$ on $\Omega_{i,1}$. We choose Δ small enough so that $\Omega_{i,2} \subset \Omega_{i,1}$, and $\widehat{f}_{i,j}$ and all of its first order weak derivatives vanish on $\Omega_{i,2}$, see Figure 3(a) and (b) for illustrations. Note that $\widetilde{f}_{i,j}$ is an approximation of $\widehat{f}_{i,j}$. We can show that $\widetilde{f}_i = 0$ on $\varphi_i(\partial U_i)$, and all of its first order weak derivatives on $\Omega_{i,2}$ are in the same order of other error terms. Since the width of $\Omega_{i,2}$ is of $O(\Delta)$, by Taylor's theorem, $|\widetilde{f}_{i,j} \circ \varphi_i|$ is bounded by a linear function of Δ on $\Omega_{i,2}$. With such a construction and proper choice of Δ, the resulting upper bound is in the same order of those of other terms. See Lemma F.3 for details.



(a) Chart determination. (b) Projected region in $(0,1)^d$.

Figure 3: (a) Illustration of an element of a chart and partition of unity. The red curve represents a cross section of ρ_i. (b) Illustration of the chart determination network $\widetilde{\mathbb{1}}_i$. The black curve represents a cross section of $\widetilde{\mathbb{1}}_i$. (c) Illustration of the projected regions in $(0,1)^d$.

Combining all of the error bounds, we can express the error in terms of N. Substituting the relation $\widetilde{M}\widetilde{J} = O(N^d)$ proves Theorem 4.5.

7. Conclusion

We provide universal approximation theories of Convolutional Residual Networks in terms of Sobolev norms. Our theory applies to Sobolev function spaces defined on a high-dimensional hypercube or low-dimensional Riemannian manifold. We demonstrate that deep and wide ConvResNets can provide approximation with good first-order smoothness properties. This partially justifies why using large networks in practice often leads to better performance and robustness.

Acknowledgment

The work of Hao Liu is partially supported by HKBU 162784 and HKBU 179356. The work of Wenjing Liao is partially supported by DMS 2012652 and NSF CAREER 2145167. The work of Wenjing Liao and Tuo Zhao is partially supported by DMS 2012652.

References

Allard, W. K., Chen, G., and Maggioni, M. Multi-scale geometric methods for data sets II: Geometric multi-resolution analysis. *Appl. Comput. Harmon. Anal.*, 32(3): 435–462, 2012.

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

Brenner, S. C., Scott, L. R., and Scott, L. R. *The Mathematical Theory of Finite Element Methods*, volume 3. Springer, 2008.

Brezis, H. and Brézis, H. *Functional Analysis, Sobolev Spaces and Partial Differential Equations*, volume 2. Springer, 2011.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

Bubeck, S. and Sellke, M. A universal law of robustness via isoperimetry. *arXiv preprint arXiv:2105.12806*, 2021.

Cardaliaguet, P. and Euvrard, G. Approximation of a function and its derivative with a neural network. *Neural Networks*, 5(2):207–220, 1992.

Chen, M., Jiang, H., Liao, W., and Zhao, T. Efficient approximation of deep ReLU networks for functions on low dimensional manifolds. *Advances in Neural Information Processing Systems*, 32:8174–8184, 2019a.

Chen, M., Jiang, H., Liao, W., and Zhao, T. Nonparametric regression on low-dimensional manifolds using deep ReLU networks. *arXiv preprint arXiv:1908.01842*, 2019b.

Chen, M., Liu, H., Liao, W., and Zhao, T. Doubly robust off-policy learning on low-dimensional manifolds by deep neural networks. *arXiv preprint arXiv:2011.01797*, 2020.

Chui, C. K. and Li, X. Approximation by ridge functions and neural networks with one hidden layer. *J. Approx. Theory*, 70(2):131–141, 1992.

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

Cloninger, A. and Klock, T. ReLU nets adapt to intrinsic dimensionality beyond the target domain. *arXiv e-prints*, pp. arXiv–2008, 2020.

Coifman, R. R., Lafon, S., Lee, A. B., Maggioni, M., Nadler, B., Warner, F., and Zucker, S. W. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proc. Natl. Acad. Sci.*, 102(21): 7426–7431, 2005.

Conway, J. H. and Sloane, N. J. A. *Sphere Packings, Lattices and Groups*. Springer Science & Business Media, 1988.

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

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Driver, B. K. Analysis tools with applications. *Lecture notes*, 2003.

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

Funahashi, K.-I. On the approximate realization of continuous mappings by neural networks. *Neural Networks*, 2 (3):183–192, 1989.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Gu, S. and Rigazio, L. Towards deep neural network architectures robust to adversarial examples. *arXiv preprint arXiv:1412.5068*, 2014.

Gühring, I., Kutyniok, G., and Petersen, P. Error bounds for approximations with deep ReLU neural networks in $W^{s,p}$ norms. *Analysis and Applications*, 18(05):803–859, 2020.

Hein, M. and Andriushchenko, M. Formal guarantees on the robustness of a classifier against adversarial manipulation. *arXiv preprint arXiv:1705.08475*, 2017.

Hon, S. and Yang, H. Simultaneous neural network approximations in Sobolev spaces. *arXiv preprint arXiv:2109.00161*, 2021.

Hornik, K. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2):251–257, 1991.

Hornik, K., Stinchcombe, M., and White, H. Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks. *Neural Networks*, 3(5):551–560, 1990.

Irie, B. and Miyake, S. Capabilities of three-layered perceptrons. In *IEEE International Conference on Neural Networks*, volume 1, pp. 218, 1988.

Lee, J. M. *Riemannian Manifolds: An Introduction to Curvature*, volume 176. Springer Science & Business Media, 2006.

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

Liu, H., Chen, M., Zhao, T., and Liao, W. Besov function approximation and binary classification on low-dimensional manifolds using convolutional residual networks. In *International Conference on Machine Learning*, pp. 6770–6780. PMLR, 2021.

Liu, H., Yang, H., Chen, M., Zhao, T., and Liao, W. Deep nonparametric estimation of operators between infinite dimensional spaces. *arXiv preprint arXiv:2201.00217*, 2022.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

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

Miyato, T., Maeda, S.-i., Koyama, M., and Ishii, S. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41 (8):1979–1993, 2018.

Montanelli, H. and Yang, H. Error bounds for deep ReLU networks using the Kolmogorov–Arnold superposition theorem. *Neural Networks*, 129:1–6, 2020.

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

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

Oono, K. and Suzuki, T. Approximation and non-parametric estimation of ResNet-type convolutional neural networks. In *International Conference on Machine Learning*, pp. 4922–4931. PMLR, 2019.

Petersen, P. and Voigtlaender, F. Equivalence of approximation by convolutional neural networks and fully-connected networks. *Proceedings of the American Mathematical Society*, 148(4):1567–1581, 2020.

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

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

Schmidt-Hieber, J. Deep ReLU network approximation of functions on a manifold. *arXiv preprint arXiv:1908.00695*, 2019.

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

Shen, Z., Yang, H., and Zhang, S. Deep network approximation characterized by number of neurons. *arXiv preprint arXiv:1906.05497*, 2019.

Slobodeckij, L. Generalized Sobolev spaces and their applications to boundary value problems of partial differential equations. *Gos. Ped. Inst. Ucep. Zap*, 197:54–112, 1958.

Spivak, M. A comprehensive introduction to differential geometry. *Bull. Amer. Math. Soc*, 79:303–306, 1973.

Stein, E. M. *Singular Integrals and Differentiability Properties of Functions*, volume 2. Princeton University Press, 1970.

Suzuki, T. 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*, 2019.

Suzuki, T. and Nitanda, A. Deep learning is adaptive to intrinsic dimensionality of model smoothness in anisotropic Besov space. *arXiv preprint arXiv:1910.12799*, 2019.

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

Tu, L. *An Introduction to Manifolds*. Universitext. Springer New York, 2010. ISBN 9781441973993.

Uesato, J., O'donoghue, B., Kohli, P., and Oord, A. Adversarial risk and the dangers of evaluating against weak attacks. In *International Conference on Machine Learning*, pp. 5025–5034. PMLR, 2018.

Weng, T.-W., Zhang, H., Chen, P.-Y., Yi, J., Su, D., Gao, Y., Hsieh, C.-J., and Daniel, L. Evaluating the robustness of neural networks: An extreme value theory approach. *arXiv preprint arXiv:1801.10578*, 2018.

Wu, B., Chen, J., Cai, D., He, X., and Gu, Q. Do wider neural networks really help adversarial robustness? *arXiv preprint arXiv:2010.01279*, 2020.

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

Zagoruyko, S. and Komodakis, N. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

Zhang, H., Yu, Y., Jiao, J., Xing, E. P., Ghaoui, L. E., and Jordan, M. I. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning*, 2019.

Zhao, Z., Zuo, S., Zhao, T., and Zhao, Y. Adversarially regularized policy learning guided by trajectory optimization. *arXiv preprint arXiv:2109.07627*, 2021.

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

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

Appendix

A. A Brief Introduction to Manifold

We introduce some concepts and quantities that characterize a low-dimensional Riemannian manifold. (Some are restatements of the main text for completeness.) These concepts and quantities are used in our theorems and proofs. We refer readers to Lee (2006); Tu (2010) for more details.

Let $\mathcal{M}$ be a d-dimensional manifold embedded in $\mathbb{R}^D$ with $d \leq D$. The first concept related to manifolds is chart, which defines a local coordinate neighborhood of a manifold.

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

In a chart (U, ϕ), U is called a coordinate neighborhood and ϕ is a coordinate system on U. A collection of charts which covers $\mathcal{M}$ is called an atlas of $\mathcal{M}$.

Definition A.2 (C^k Atlas)**.** A C^k atlas for $\mathcal{M}$ is a collection of charts $\{(U_\alpha, \phi_\alpha)\}_{\alpha \in \mathcal{A}}$ which satisfies $\bigcup_{\alpha \in \mathcal{A}} U_\alpha = \mathcal{M}$, and are pairwise C^k compatible, i.e.,

$$\phi_\alpha \circ \phi_\beta^{-1} : \phi_\beta(U_\alpha \cap U_\beta) \to \phi_\alpha(U_\alpha \cap U_\beta) \quad \text{and}$$
$$\phi_\beta \circ \phi_\alpha^{-1} : \phi_\alpha(U_\alpha \cap U_\beta) \to \phi_\beta(U_\alpha \cap U_\beta)$$

are both C^k for any $\alpha, \beta \in \mathcal{A}$. An atlas is called finite if it contains finitely many charts.

With the concept of atlas, we then define smooth manifolds:

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

Simple examples of smooth manifold include the Euclidean space, the torus and the unit sphere. C^s functions on a smooth manifold $\mathcal{M}$ are defined as follows:

Definition A.4 (C^s functions on $\mathcal{M}$)**.** Let $\mathcal{M}$ be a smooth manifold and $f : \mathcal{M} \to \mathbb{R}$ be a function on $\mathcal{M}$. We say f is a C^s function defined on $\mathcal{M}$, if for every chart (U, ϕ) on $\mathcal{M}$, the function $f \circ \phi^{-1} : \phi(U) \to \mathbb{R}$ is a C^s function.

We next define the C^∞ partition of unity which is an important tool for the study of functions on manifolds.

Definition A.5 (Partition of Unity)**.** A C^∞ partition of unity on a manifold $\mathcal{M}$ is a collection of C^∞ functions $\{\rho_\alpha\}_{\alpha \in \mathcal{A}}$ with $\rho_\alpha : \mathcal{M} \to [0, 1]$ such that for any $\mathbf{x} \in \mathcal{M}$,

1. there is a neighbourhood of $\mathbf{x}$ where only a finite number of the functions in $\{\rho_\alpha\}_{\alpha \in \mathcal{A}}$ are nonzero, and
2. $\sum_{\alpha \in \mathcal{A}} \rho_\alpha(\mathbf{x}) = 1.$

An open cover of $\mathcal{M}$ is called locally finite if every $\mathbf{x} \in \mathcal{M}$ has a neighbourhood that intersects with a finite number of sets in the cover. For a locally finite cover of a smooth manifold $\mathcal{M}$, there always exists a C^∞ partition of unity subordinate to the cover (Spivak, 1973, Chapter 2, Theorem 15).

Proposition A.6 (Existence of a C^∞ partition of unity)**.** *Let $\{U_\alpha\}_{\alpha \in \mathcal{A}}$ be a locally finite cover of a smooth manifold $\mathcal{M}$. There is a C^∞ partition of unity $\{\rho_\alpha\}_{\alpha=1}^\infty$ such that* $\operatorname{supp}(\rho_\alpha) \subset U_\alpha$.

Let $\{(U_\alpha, \phi_\alpha)\}_{\alpha \in \mathcal{A}}$ be a C^∞ atlas of $\mathcal{M}$. Proposition A.6 guarantees the existence of a partition of unity $\{\rho_\alpha\}_{\alpha \in \mathcal{A}}$ such that ρ_α is supported on U_α.

B. Convolutional neural networks and multi-layer perceptions

Our proofs are based on approximation theories of convolutional neural networks (CNN) and their relations to multi-layer perceptions (MLP). In this section, we introduce related notations and definitions. For the convenience of notation, we use $\cdot$ to denote $\otimes$, the sum of entrywise product.

We consider CNNs in the form of

$$f(\mathbf{x}) = W \cdot \mathrm{Conv}_{\mathcal{W},\mathcal{B}}(\mathbf{x}), \tag{10}$$

where $\mathrm{Conv}_{\mathcal{W},\mathcal{B}}(Z)$ is defined in (1), W is the weight matrix of the fully connected layer, $\mathcal{W}, \mathcal{B}$ are sets of filters and biases, respectively. We define the class of CNNs as

$$\begin{aligned}\mathcal{F}^{\mathrm{CNN}}(L, J, K, \kappa_1, \kappa_2) = \big\{ f \,|\, & f(\mathbf{x}) \text{ in the form (10) with } L \text{ layers.} \\ & \text{Each convolutional layer has filter size bounded by } K. \\ & \text{The number of channels of each layer is bounded by } J. \\ & \max_l \|\mathcal{W}^{(l)}\|_\infty \vee \|B^{(l)}\|_\infty \le \kappa_1,\ \|W\|_\infty \le \kappa_2 \big\}.\end{aligned}$$

For MLP, we consider the following form

$$f(\mathbf{x}) = W_L \cdot \mathrm{ReLU}(W_{L-1} \cdots \mathrm{ReLU}(W_1 \mathbf{x} + \mathbf{b}_1) \cdots + \mathbf{b}_{L-1}) + \mathbf{b}_L, \tag{11}$$

where $W_1, \ldots, W_L$ and $\mathbf{b}_1, \ldots, \mathbf{b}_L$ are weight matrices and bias vectors of proper sizes, respectively. The class of MLP is defined as

$$\begin{aligned}\mathcal{F}^{\mathrm{MLP}}(L, J, \kappa) = \big\{ f \,|\, & f(\mathbf{x}) \text{ in the form (11) with } L\text{-layers and width bounded by } J. \\ & \|W_i\|_{\infty,\infty} \le \kappa, \|\mathbf{b}_i\|_\infty \le \kappa \text{ for } i = 1, \ldots, L \big\}.\end{aligned}$$

In some cases it is necessary to enforce the output of the MLP to be bounded. We define such a class as

$$\mathcal{F}^{\mathrm{MLP}}(L, J, \kappa, R) = \left\{ f \,|\, f(\mathbf{x}) \in \mathcal{F}^{\mathrm{MLP}}(L, J, \kappa) \text{ and } \|f\|_\infty \le R \right\}.$$

In some case we do not need the constraint on the output, we denote such MLP class as $\mathcal{F}^{\mathrm{MLP}}(L, J, \kappa)$.

C. Proof of Theorem 3.2

Before we prove Theorem 3.2, we define the Sobolev semi-norm:

Definition C.1. For any integers $0 \le k \le \alpha$, $1 \le p < \infty$ and function $f \in W^{\alpha,p}(\Omega)$, we define its Sobolev semi-norm as

$$\begin{aligned}|f|_{W^{k,p}(\Omega)} &= \Big(\sum_{|\boldsymbol{\alpha}|=k} \|D^{\boldsymbol{\alpha}} f\|^p_{L^p(\Omega)} \Big)^{1/p}, \\ |f|_{W^{k,\infty}(\Omega)} &= \max_{|\boldsymbol{\alpha}|=k} \|D^{\boldsymbol{\alpha}} f\|_{L^\infty(\Omega)},\end{aligned}$$

Now we prove Theorem 3.2.

Proof of Theorem 3.2. We prove Theorem 3.2 in four steps.

Step 1: Decompose $(0,1)^D$ using locally supported functions. We define

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

and

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

with $\mathbf{m} = (m_1, m_2, ..., m_D) \in \{0, ..., N\}^D$. We have $\sum_{\mathbf{m}} \phi_{\mathbf{m}} = 1$ on $(0,1)^D$ and $\phi_{\mathbf{m}}$ is supported on $B_{\frac{2}{3}N, \|\cdot\|_\infty}(\frac{\mathbf{m}}{N}) \subset B_{1/N, \|\cdot\|_\infty}(\frac{\mathbf{m}}{N})$. We denote $\mathcal{S}_N = \{0, 1, ..., N\}^D$. The following lemma shows that each $\psi\left(3N\left(x_k - \frac{m_k}{N}\right)\right)$ can be realized by a CNN (see a proof in Appendix G.3).

Lemma C.2. *There exists a CNN architecture $\mathcal{F}^{\mathrm{CNN}}(L, J, K, \kappa_1, \kappa_2)$ such that for any N, m, such an architecture yields a CNN $\widetilde{\psi}$ with*

$$\widetilde{\psi}_{m,N}(x) = \psi\left(3N\left(x_k - \frac{m}{N}\right)\right), \tag{12}$$

$$\|\widetilde{\psi}_{m,N}\|_{W^{k,\infty}(0,1)} \leq (3N)^k. \tag{13}$$

Such an architecture has

$$L = 2,\ J = 16,\ K = 2,\ \kappa_1 = \kappa_2 = O(N).$$

Further more, the weight matrix in the fully connected layer of $\mathcal{F}^{\mathrm{CNN}}$ has nonzero entries only in the first row.

We then decompose f as

$$f = \sum_{\mathbf{m}} \phi_{\mathbf{m}} f.$$

Step 2: Approximate each $\phi_{\mathbf{m}} f$ using averaged Taylor polynomials. On each $B_{1/N,\|\cdot\|_\infty}(\frac{\mathbf{m}}{N})$, we approximate $\phi_{\mathbf{m}} f$ by an averaged Taylor polynomial. The averaged Taylor polynomial is defined as follows:

Definition C.3 (Averaged Taylor polynomials)**.** Let $\alpha > 0, 1 \leq p \leq +\infty$ be integers and $f \in W^{\alpha-1,p}(\Omega)$. For $\mathbf{x}_0 \in \Omega, r > 0$ such that $\overline{B_{r,\|\cdot\|}(\mathbf{x}_0)}$ is compact in Ω, the corresponding Taylor polynomial of order α of f averaged over $B_{r,\|\cdot\|}(\mathbf{x}_0)$ is defined as

$$Q^\alpha_{\mathbf{x}_0} f(\mathbf{x}) = \int_{B_{r,\|\cdot\|}(\mathbf{x}_0)} T^\alpha f(\mathbf{x}, \mathbf{z}) \phi(\mathbf{z}) d\mathbf{z}$$

with

$$T^\alpha f(\mathbf{x}, \mathbf{y}) = \sum_{|\mathbf{v}| \leq \alpha - 1} \frac{1}{\mathbf{v}} \partial^{\mathbf{v}} f(\mathbf{z}) (\mathbf{x} - \mathbf{z})^{\mathbf{v}}$$

and ϕ being arbitrary cut-off function satisfying

$$\phi \in C_c^\infty(\mathbb{R}^D) \text{ with } \phi(\mathbf{x}) \geq 0 \text{ for all } \mathbf{x} \in \mathbb{R}^D,$$
$$\mathrm{supp}(\phi) = \overline{B_{r,\|\cdot\|}(\mathbf{x}_0)} \text{ and } \int_{\mathbb{R}^D} \phi(\mathbf{x}) d\mathbf{x} = 1,$$

where $C_c^\infty(\mathbb{R}^D)$ denotes the space of infinitely differentiable functions on $\mathbb{R}^D$ with compact support.

Under proper assumptions, the averaged Taylor polynomial can approximate f and its partial derivatives well. We first define the star-shaped sets and chunkiness parameter, which are used in the error estimation result.

Definition C.4 (Star-shaped sets, Definition 4.2.2 of (Brenner et al., 2008))**.** Let $\Omega, \widetilde{\Omega} \subset \mathbb{R}^D$. Then Ω is called star-shaped with respect to $\widetilde{\Omega}$ if for all $\mathbf{x} \in \Omega$, we have

$$\overline{\mathrm{ch}(\{\mathbf{x}\} \cup \widetilde{\Omega})} \subset \Omega.$$

Definition C.5 (Chunkiness parameter, Definition 4.2.16 of (Brenner et al., 2008))**.** Let $\Omega \subset \mathbb{R}^D$ be bounded. Define

$$\mathcal{R} = \left\{r > 0 : \text{ there exists } \mathbf{x} \in \Omega \text{ such that } \Omega \text{ is star-shaped with respect to } B_{r,|\cdot|}(\mathbf{x})\right\}.$$

For $\mathcal{R} \neq \emptyset$, we define

$$r^*_{\max} = \sup \mathcal{R} \quad \text{and} \quad \gamma = \frac{\mathrm{diam}(\Omega)}{r^*_{\max}},$$

where γ is called the chunkiness parameter of Ω.

The following lemma gives an error estimation of averaged Taylor polynomials:

Lemma C.6 (Bramble-Hilbert, Lemma 4.3.8 of (Brenner et al., 2008)). *Let $\Omega \subset \mathbb{R}^D$ be open and bounded, $\mathbf{x} \in \Omega$ and $r > 0$ such that Ω is star-shaped with respect to $B_{r,\|\cdot\|}(\mathbf{x}_0)$ and $r > \frac{1}{2} r^*_{\max}$, with $r^*_{\max}$ defined in Definition C.5. Let $n > 0, 1 \le p \le +\infty$ be integers and γ be the chunkiness parameter of Ω. Then we have*

$$|f - Q^\alpha_{\mathbf{x}_0} f|_{W^{\alpha,p}(\Omega)} \le C h^{\alpha-k} |f|_{W^{\alpha,p}(\Omega)}$$

for $k = 0, 1, ..., \alpha$, where $h = \operatorname{diam}(\Omega)$ and C is a constant depending on D, α, γ.

Lemma C.7 below shows that $Q^\alpha f$ can be written as a weighted sum of polynomials.

Lemma C.7 (Lemma B.9 of (Gühring et al., 2020)). *Let $\alpha > 0, 1 \le p \le +\infty$ be integers and $f \in W^{\alpha-1,p}(\Omega)$. Let $\mathbf{x}_0 \in \Omega, r > 0$ such that $\overline{B_{r,\|\cdot\|}(\mathbf{x}_0)}$ is compact in Ω, and there exists $\widetilde{r} > 0$ with $B_{r,\|\cdot\|}(\mathbf{x}_0) \subset B_{\widetilde{r},\|\cdot\|_\infty}(0)$. Then the averaged Taylor polynomial $Q^\alpha_{\mathbf{x}_0}(f)$ can be written as*

$$Q^\alpha_{\mathbf{x}_0} f(\mathbf{x}) = \sum_{|\mathbf{v}| \le \alpha - 1} c_{\mathbf{v}} \mathbf{x}^{\mathbf{v}} \tag{14}$$

for $\mathbf{x} \in \Omega$. There exists a constant C depending on $\alpha, D, \widetilde{r}$ such that

$$|c_{\mathbf{v}}| \le C r^{-D/p} \|f\|_{W^{\alpha-1,p}}(\Omega)$$

for all $|\mathbf{v}| \le \alpha - 1$.

Using averaged Taylor polynomials, we approximate $\phi_{\mathbf{m}} f$ by

$$\phi_{\mathbf{m}} f \approx (\phi_{\mathbf{m}} Q^\alpha_{\mathbf{m}/N} f)(\mathbf{x}) = \phi_{\mathbf{m}} \sum_{|\mathbf{v}| \le \alpha - 1} c_{\mathbf{m},\mathbf{v}} \mathbf{x}^{\mathbf{v}} = \sum_{|\mathbf{v}| \le \alpha - 1} c_{\mathbf{m},\mathbf{v}} \phi_{\mathbf{m}} \mathbf{x}^{\mathbf{v}}. \tag{15}$$

Define

$$\widehat{f} = \sum_{\mathbf{m} \in S_N} \sum_{|\mathbf{v}| \le \alpha - 1} c_{\mathbf{m},\mathbf{v}} \phi_{\mathbf{m}} \mathbf{x}^{\mathbf{v}}, \tag{16}$$

where $c_{\mathbf{m},\mathbf{v}}$'s are the coefficients in (14). Then $\widehat{f}$ is an approximation of f. The following lemma gives an upper bound on the approximation error

Lemma C.8 (Lemma C.4 of (Gühring et al., 2020)). *Let $\alpha \ge 2$ be an integer and $1 \le p \le \infty$. For any $s \in [0, 1]$ and $f \in W^{\alpha,p}((0,1)^D)$, one has*

$$\|\widehat{f} - f\|_{W^{s,p}((0,1)^D)} \le C \left(\frac{1}{N}\right)^{\alpha - s} \|f\|_{W^{\alpha,p}((0,1)^D)},$$

where C is a constant depending on α, p, D. Furthermore, the coefficients in $\widehat{f}$ satisfies

$$|c_{\mathbf{m},\mathbf{v}}| \le C_1 N^{D/p} \|f\|_{W^{\alpha,p}((0,1)^D)}$$

for some constant C_1 depending on D, α, p.

Step 3: Network approximation Note that $\widehat{f}$ is a sum of functions in the form of $\phi_{\mathbf{m}} \mathbf{x}^{\mathbf{v}}$ with weights $c_{\mathbf{m},\mathbf{v}}$'s. We next approximate each $\phi_{\mathbf{m}} \mathbf{x}^{\mathbf{v}}$ by a CNN.

Lemma C.9. *For any $0 < \varepsilon < 1, \mathbf{x} \in (0,1)^D, N > 0, \mathbf{m} \in \{0, 1, ..., N\}^D, |\mathbf{v}| < \alpha$, there exists a CNN architecture $\mathcal{F}^{\mathrm{CNN}}(L, J, K, \kappa, \kappa)$ that yields a CNN $\widetilde{g}$ with*

$$\|\widetilde{g}_{\mathbf{m},\mathbf{v}}(\mathbf{x}) - \phi_{\mathbf{m}} \mathbf{x}^{\mathbf{v}}\|_{W^{k,\infty}((0,1)^D} \le C_2 N^k \varepsilon, \tag{17}$$

$$\widetilde{g}_{\mathbf{m},\mathbf{v}}(\mathbf{x}) = 0 \text{ if } \phi_{\mathbf{m}} \mathbf{x}^{\mathbf{v}} = 0 \tag{18}$$

for $k = 0, 1$, where C_2 is a constant depending on α, k. Such an architecture has

$$L = O\left(D \log \frac{1}{\varepsilon}\right),\ J = O(D),\ \kappa = 3N.$$

The constants hidden in O depends on α, k. Further more, the weight matrix in the fully connected layer of $\mathcal{F}^{\mathrm{CNN}}$ has nonzero entries only in the first row.

Lemma C.9 is proved in Appendix G.4. By Lemma C.9, each $\phi_{\mathbf{m}}\mathbf{x}^{\mathbf{v}}$ can be approximated by a CNN. Denote the network approximation of $\phi_{\mathbf{m}}\mathbf{x}^{\mathbf{v}}$ by $\widetilde{g}_{\mathbf{m},\mathbf{v}}(\mathbf{x})$. We approximate $\widehat{f}$ by $\widetilde{f}$ defined as

$$\widetilde{f} = \sum_{\mathbf{m}} \sum_{|\mathbf{v}| \le \alpha - 1} c_{\mathbf{m},\mathbf{v}} \widetilde{g}_{\mathbf{m},\mathbf{v}}(\mathbf{x}). \tag{19}$$

The following lemma gives an upper bound of the approximation error of $\widetilde{f}$ (see a proof in Appendix G.6).

Lemma C.10. *Let $\alpha \ge 2$ and $1 \le p \le \infty$ be integers. For any $f \in W^{\alpha,p}((0,1)^D)$, let $\phi_{\mathbf{m}} Q^{\alpha}_{\mathbf{m}/N} f(\mathbf{x})$ be the averaged Taylor approximation of $\phi_{\mathbf{m}} f$ defined in (15). For any $0 < \eta < 1$, let $\widetilde{g}_{\mathbf{m},\mathbf{v}}$ be the CNN approximation of $\phi_{\mathbf{m}} Q^{\alpha}_{\mathbf{m}/N} f(\mathbf{x})$ constructed in Lemma C.9 with accuracy η. For $0 \le s \le 1$, we have*

$$\| \sum_{\mathbf{m} \in \mathcal{S}_N} \phi_{\mathbf{m}} Q^{\alpha}_{\mathbf{m}/N} f - \sum_{\mathbf{m} \in \mathcal{S}_N} \sum_{|\mathbf{v}| \le \alpha - 1} c_{\mathbf{m},\mathbf{v}} \widetilde{g}_{\mathbf{m},\mathbf{v}} \|_{W^{s,p}((0,1)^D)} \le C_3 \|f\|_{W^{\alpha,p}((0,1)^D)} N^s \eta, \tag{20}$$

where $c_{\mathbf{m},\mathbf{v}}$'s are coefficients defined in (15), C_3 is a constant depending on D, α, s, p.

Note the $\widetilde{f}$ is the sum of no more than $N^D(D+1)^{\alpha-1}$ CNNs of which the width is of $J = O(D)$. The following lemma shows that under appropriate conditions, the sum of n_0 CNNs with width in the same order can be realized by the sum of n_1 CNNs with a proper width (see a proof in Appendix G.8):

Lemma C.11. *Let $\{f_i\}_{i=1}^{n_0}$ be a set of CNNs with architecture $\mathcal{F}^{\mathrm{CNN}}(L_0, J_0, K_0, \kappa_0, \kappa_0)$. For any integers $1 \le n \le n_0$ and $\widetilde{J}$ satisfying $n\widetilde{J} = O(n_0 J_0)$ and $\widetilde{J} \ge J_0$, there exists a CNN architecture $\mathcal{F}^{\mathrm{CNN}}(L, J, K, \kappa, \kappa)$ that gives a set of CNNs $\{g_i\}_{i=1}^{n}$ such that*

$$\sum_{i=1}^{n} g_i(\mathbf{x}) = \sum_{i=1}^{n_0} f_i(\mathbf{x}).$$

Such an architecture has

$$L = O(L_0), J = O(\widetilde{J}), K = K_0, \kappa = \kappa_0.$$

Furthermore, the fully connected layer of f has nonzero elements only in the first row.

By Lemma C.11, for any $\widetilde{M}, \widetilde{J}$ satisfying $\widetilde{M}\widetilde{J} = O(N^D)$, there exists a CNN architecture $\mathcal{F}^{\mathrm{CNN}}(L, J, K, \kappa, \kappa)$ that gives rise to $\{g_i\}_{i=1}^{n}$ with

$$\widetilde{f} = \sum_{i=1}^{\widetilde{M}} g_i,$$

where

$$L = O\left(\log \frac{1}{\eta}\right),\ J = O(\widetilde{J}),\ \kappa = 3N.$$

The following lemma shows that the sum of CNNs can be realized by a ConvResNet:

Lemma C.12 (Lemma 18 in (Liu et al., 2021)). *Let $\mathcal{F}^{\mathrm{CNN}}(L,J,K,\kappa_1,\kappa_2)$ be any CNN architecture from $\mathbb{R}^D$ to $\mathbb{R}$. Assume the weight matrix in the fully connected layer of $\mathcal{F}^{\mathrm{CNN}}(L,J,K,\kappa_1,\kappa_2)$ has nonzero entries only in the first row. Let M be a positive integer. There exists a ConvResNet architecture $\mathcal{C}(M,L,J,\kappa_1,\kappa_2(1\vee\kappa_1^{-1}))$ such that for any $\{f_i(\mathbf{x})\}_{i=1}^M\subset\mathcal{F}^{\mathrm{CNN}}(L,J,K,\kappa_1,\kappa_2)$, there exists $\widetilde{f}\in\mathcal{C}(M,L,J,\kappa_1,\kappa_2(1\vee\kappa_1^{-1}))$ with*

$$\widetilde{f}(\mathbf{x})=\sum_{i=1}^M f_i(\mathbf{x}).$$

By Lemma C.12, there exits a ConvResNet architecture $\mathcal{C}(M,L,J,K,\kappa_1,\kappa_2)$ with

$$L=O(\log 1/\eta),\ J=O(\widetilde{J}),\ \kappa_1=O(3N),\ \kappa_2=O(3N),\ M=O(\widetilde{M}) \tag{21}$$

and $\widetilde{J},\widetilde{M}$ satisfying

$$\widetilde{M}\widetilde{J}=O(N^D), \tag{22}$$

that yields a ConvResNet realizing $\widetilde{f}$.

Step 4: Error estimation. We compute

$$\begin{aligned}
&\|f-\widetilde{f}\|_{W^{s,p}((0,1)^D)}\\
\leq&\left\|f-\left(\sum_{\mathbf{m}}\phi_{\mathbf{m}}Q^\alpha_{\mathbf{m}/N}f\right)\right\|_{W^{s,p}((0,1)^D)}+\left\|\left(\sum_{\mathbf{m}}\phi_{\mathbf{m}}Q^\alpha_{\mathbf{m}/N}f\right)-\widetilde{f}\right\|_{W^{s,p}((0,1)^D)}\\
\leq&C_4\left(\frac{1}{N}\right)^{\alpha-s}\|f\|_{W^{\alpha,p}((0,1)^D)}+C_5N^s\eta\|f\|_{W^{\alpha,p}((0,1)^D)}\\
\leq&(C_4+C_5)N^{-(\alpha-s)},
\end{aligned}\tag{23}$$

where C_4,C_5 are two constants depending on D,α,s,p,R. In the second inequality, we use Lemma C.8 and C.10 for the first and second term, respectively. In the third inequality, we set $\eta=N^{-\alpha}$ to balance the two terms. Using the relation (22), we have

$$N=(\widetilde{M}\widetilde{J})^{1/D},\quad \eta=(\widetilde{M}\widetilde{J})^{-\frac{\alpha}{D}}. \tag{24}$$

Substituting (24) into (23) gives rise to

$$\|f-\widetilde{f}\|_{W^{s,p}((0,1)^D)}\leq C_6(\widetilde{M}\widetilde{J})^{-\frac{\alpha-s}{D}} \tag{25}$$

for some constant C_6 depending on D,α,s,p,R. Substituting (24) into (21) and (22) gives rise to the network architecture

$$L=O(\log(\widetilde{M}\widetilde{J})),\ J=O(\widetilde{J}),\ \kappa_1=O((\widetilde{M}\widetilde{J})^{1/D}),\ \kappa_2=O((\widetilde{M}\widetilde{J})^{1/D}),\ M=O(\widetilde{M}).$$

□

D. Proof of Theorem 3.4

Proof of Theorem 3.4. By Theorem 3.2 and the choice of $\widetilde{M}\widetilde{J}$, there exits $\widetilde{f}\in\mathcal{C}$ so that $\|\widetilde{f}-f\|_\infty\leq\varepsilon$ and

$$\max_j\left\|\frac{\partial\widetilde{f}}{\partial x_j}-\frac{\partial f}{\partial x_j}\right\|\leq\varepsilon^{\frac{\alpha-1}{\alpha}}, \tag{26}$$

which implies

$$\left\|\widetilde{f}\right\|_{\mathrm{Lip}}\leq 1+\sqrt{D}\varepsilon^{\frac{\alpha-1}{\alpha}}. \tag{27}$$

We have

$$\begin{aligned}&\mathbb{E}\left[(\widetilde{f}(\mathbf{x}_1)-y_1)^2\right]\\ \leq&\mathbb{E}\left[(\widetilde{f}(\mathbf{x}_1)-f(\mathbf{x}_1))^2\right]+\mathbb{E}\left[(f(\mathbf{x}_1)-y_1)^2\right]\\ \leq&\varepsilon^2+\sigma^2.\end{aligned}\tag{28}$$

Denote $X_i=\frac{1}{n}(\widetilde{f}(\mathbf{x}_i)-y_i)^2-\mathbb{E}\left[(\widetilde{f}(\mathbf{x}_i)-y_i)^2\right]$. We have

$$|X_i|\leq\frac{2(\varepsilon^2+\sigma^2)}{n},\ \mathbb{E}[X_i]=0,\tag{29}$$

and

$$\mathbb{E}[X_i^2]\leq\frac{8(\varepsilon^4+\sigma^4)}{n^2}.\tag{30}$$

By Bernstein inequality, we deduce

$$\begin{aligned}\mathbb{P}\left(\sum_{i=1}^n X_i\geq t\right)\leq&\exp\left(-\frac{\frac{1}{2}t^2}{\frac{8(\varepsilon^4+\sigma^4)}{n}+\frac{2(\varepsilon^2+\sigma^2)}{3n}t}\right)\\ =&\exp\left(-\frac{3nt^2}{48(\varepsilon^4+\sigma^4)+4(\varepsilon^2+\sigma^2)t}\right).\end{aligned}\tag{31}$$

Therefore

$$\begin{aligned}&\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n(\widetilde{f}(\mathbf{x}_i)-y_i)^2\geq\varepsilon^2+\sigma^2+t\right)\\ \leq&\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n(\widetilde{f}(\mathbf{x}_i)-y_i)^2\geq\mathbb{E}\left[(\widetilde{f}(\mathbf{x}_1)-y_1)^2\right]+t\right)\\ \leq&\exp\left(-\frac{3nt^2}{48(\varepsilon^4+\sigma^4)+4(\varepsilon^2+\sigma^2)t}\right).\end{aligned}\tag{32}$$

Setting $t=\varepsilon^2$ gives rise to

$$\begin{aligned}&\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n(\widetilde{f}(\mathbf{x}_i)-y_i)^2\geq2\varepsilon^2+\sigma^2\right)\\ \leq&\exp\left(-\frac{3n\varepsilon^2}{104\sigma^4}\right).\end{aligned}\tag{33}$$

□

E. Proof of Theorem 3.6

Proof of Theorem 3.6. By Theorem 3.2 and the choice of $\widetilde{M}\widetilde{J}$, there exits $\widetilde{f}\in\mathcal{C}$ so that $\|\widetilde{f}-f\|_\infty\leq\varepsilon$ and

$$\max_j\left\|\frac{\partial\widetilde{f}}{\partial x_j}-\frac{\partial f}{\partial x_j}\right\|\leq\varepsilon^{\frac{\alpha-1}{\alpha}}.\tag{34}$$

Since $\|f\|_{W^{\alpha,\infty}}\leq1$, we have

$$\|\widetilde{f}\|_{\mathrm{Lip}}\leq1+\sqrt{D}\varepsilon^{\frac{\alpha-1}{\alpha}}.\tag{35}$$

We have

$$
\begin{aligned}
&R(\widetilde{f},\delta)-R(\widetilde{f},0)\\
=&\mathbb{E}_{(\mathbf{x},y)\in\mathrm{supp}(\rho)}\left[\sup_{\mathbf{x}'\in B_\delta(\mathbf{x})}\ell\left(\widetilde{f}(\mathbf{x}'),y\right)\right]-\mathbb{E}_{(\mathbf{x},y)\in\mathrm{supp}(\rho)}\left[\ell\left(\widetilde{f}(\mathbf{x}'),y\right)\right]\\
\leq&\mathbb{E}_{(\mathbf{x},y)\in\mathrm{supp}(\rho)}\left[\sup_{\mathbf{x}'\in B_\delta(\mathbf{x})}\left|\ell\left(\widetilde{f}(\mathbf{x}'),y\right)-\ell\left(\widetilde{f}(\mathbf{x}),y\right)\right|\right]\\
\leq&\mathbb{E}_{(\mathbf{x},y)\in\mathrm{supp}(\rho)}\sup_{\mathbf{x}'\in B_\delta(\mathbf{x})}L_{\mathrm{Lip}}|\widetilde{f}(\mathbf{x}')-\widetilde{f}(\mathbf{x})|\\
\leq&\mathbb{E}_{(\mathbf{x},y)\in\mathrm{supp}(\rho)}\sup_{\mathbf{x}'\in B_\delta(\mathbf{x})}L_{\mathrm{Lip}}\|\widetilde{f}\|_{\mathrm{Lip}}\|\mathbf{x}'-\mathbf{x}\|_2\\
\leq&L_{\mathrm{Lip}}(1+\sqrt{D}\varepsilon^{\frac{\alpha-1}{\alpha}})\delta
\end{aligned}
\tag{36}
$$

□

F. Proof of Theorem 4.5

Proof of Theorem 4.5. We prove Theorem 4.5 in three steps.

Step 1: Decomposition of f

• **Construct an atlas on $\mathcal{M}$.** According to Assumption 4.1, $\mathcal{M}$ is bounded. Therefore, for any given $0<r<\tau/2$, we can find a finite collection of points $\{\mathbf{c}_i\}_{i=1}^{C_\mathcal{M}}\subset\mathcal{M}$ such that

$$\mathcal{M}\subset\bigcup_{i=1}^{C_\mathcal{M}}B_r(\mathbf{c}_i).$$

Denote $U_i=B_r(\mathbf{c}_i)\cap\mathcal{M}$. Then $\{U_i\}_{i=1}^{C_\mathcal{M}}$ form an open cover of $\mathcal{M}$ and each U_i is diffeomorphic to an open subset of $\mathbb{R}^d$. The total number of partitions if bounded by $C_\mathcal{M}\leq\left\lceil\frac{\mathrm{SA}(\mathcal{M})}{r^d}T_d\right\rceil$, where $\mathrm{SA}(\mathcal{M})$ is the surface area of $\mathcal{M}$ and T_d is the average number of U_i's that contain a given point on $\mathcal{M}$.

On each U_i, we define a transformation ϕ_i that projects any $\mathbf{x}\in U_i$ to $T_{\mathbf{c}_i}(\mathcal{M})$, the tangent space of $\mathcal{M}$ at $\mathbf{c}_i$. Let $V_i\in\mathbb{R}^{D\times d}$ be an orthogonal matrix whose columns form an orthonomal basis of $T_{\mathbf{c}_i}(\mathcal{M})$. Define

$$\varphi_i(\mathbf{x})=a_iV_i^\top(\mathbf{x}-\mathbf{c}_i)+\boldsymbol{b}_i\text{ for }\mathbf{x}\in U_i,\tag{37}$$

where $a_i\in\mathbb{R}$ is a scaling factor and $\boldsymbol{b})i\in\mathbb{R}^d$ is a shifting vector that ensure $\varphi_i(U_i)\subseteq[0,1]^d$. Then $\{(U_i,\varphi_i)\}_{i=1}^{C_\mathcal{M}}$ form an atlas of $\mathcal{M}$.

•**Decomposition of f by a partition of unity.** The following lemma shows that under proper assumption, there exists a partition of unity $\{\rho_i\}_{i=1}^{C_\mathcal{M}}$ subordinate to $\{(U_i,\varphi_i)\}_{i=1}^{C_\mathcal{M}}$ (see Appendix H.1 for a proof).

Lemma F.1. *Let $\{(U_i,\varphi_i)\}_{i=1}^{C_\mathcal{M}}$ be the atlas of $\mathcal{M}$ defined above with $r<\tau/4$. There exist a finite number $C_\mathcal{M}$ and a C^∞ partition of unity $\{\rho_i\}_{i=1}^{C_\mathcal{M}}$ satisfying*

(i) $\mathrm{supp}(\rho_i)$ *is compact in* U_i.

(ii) $\sum_{i=1}^{C_\mathcal{M}}\rho_i(\mathbf{x})=1$ *for any* $\mathbf{x}\in\mathcal{M}$.

(iii) There exists a constant $c>0$ depending on r such that for any i, we have

$$\inf_{\mathbf{x}\in\mathrm{supp}(\rho_i),\ \widetilde{\mathbf{x}}\in\partial U_i}\|\mathbf{x}-\widetilde{\mathbf{x}}\|_2\geq c.$$

Here $C_\mathcal{M}$ depends on the surface area of $\mathcal{M}$ and the average number of U_i's that contain a given point on $\mathcal{M}$.

Let $\{\rho_i\}_{i=1}^{C_{\mathcal{M}}}$ be the partition of unity from Lemma F.1. Since for each i, φ_i is a bijection from U_i to a subset of $[0,1]^d$, φ^{-1} exists and is a linear operator. We decompose f as

$$f = \sum_{i=1}^{C_{\mathcal{M}}} f_i \quad \text{with} \quad f_i = (f\rho_i).$$

Here each f_i is compactly supported on U_i and each $f_i \circ \varphi_i^{-1}$ is compactly supportedin $\varphi_i(U_i) \subseteq [0,1]^d$. We extend $f_i \circ \varphi_i^{-1}$ by 0 on $[0,1]^d \backslash \varphi_i(U_i)$. The extended function is in $W^{\alpha,k}([0,1]^d)$. To simplify the notation, we still use $f_i \circ \varphi_i^{-1}$ to denote the extended function. For each i, we use averaged Taylor polynomials to approximate $f_i \circ \varphi^{-1}$ on $[0,1]^d$ as in (16):

$$f_i \circ \varphi_i^{-1} \approx \widehat{f_i} = \sum_{\mathbf{m},\mathbf{v}} c_{i,\mathbf{m},\mathbf{v}} \phi_{\mathbf{m}} \mathbf{x}^{\mathbf{v}}.$$

Step 2: Network approximation

•**Approximate $\widehat{f_i}$ by CNNs.** Since each $\widehat{f_i}$ is the averaged Taylor polynomial approximation of $f_i \circ \varphi_i^{-1}$, by Lemma C.9, it can be approximated by a sum of $(d+1)^{\alpha-1}N^d$ CNNs. Denote the approximation accuracy by η as in Lemma C.9, each CNN has depth $O(\log(1/\eta))$, width $O(1)$, all weight parameters are of $O(N)$.

•**Chart determination** For any input $\mathbf{x}$, to determine the chart it belongs to, we are going to construct an indicator function. With our construction of charts, we have $\mathbf{x} \in U_i$ if and only if $\|\mathbf{x} - \mathbf{c}_i\|_2^2 \leq r^2$. Define the indicator function

$$\mathbb{1}_{[0,r^2]}(a) = \begin{cases} 1 & \text{if } a \leq r^2, \\ 0 & \text{otherwise,} \end{cases}$$

and the squared distance function

$$d_i^2(\mathbf{x}) = \|\mathbf{x} - \mathbf{c}_i\|_2^2 = \sum_{j=1}^{D} (x_j - c_{i,j})^2, \tag{38}$$

where we used the expression $\mathbf{x} = [x_1, ..., x_D]^\top$ and $\mathbf{c}_i = [c_{i,1}, ..., c_{i,D}]^\top$. The composition $\mathbb{1}_i = \mathbb{1}_{[0,r^2]} \circ d_i^2$ outputs 1 if $\mathbf{x} \in U_i$ and outputs 0 otherwise. We are going to construct a CNN to approximate $\mathbb{1}_i$.

In (38), the function d_i^2 is a sum of D square functions. By Lemma G.6, For any $0 < \theta < 1/2$, $x \in [-B, B]$, and $K \geq 2$, there is a CNN architecture $\mathcal{F}^{\text{CNN}}(L, J, K, \kappa, \kappa)$ that yields a CNN, denoted by $\widetilde{d^2}$, such that

$$\|\widetilde{d^2}(x) - x^2\|_{W^{1,\infty}([-B,B])} < \theta, \ \widetilde{d}(0) = 0.$$

Such a network has

$$L = O\left(\log \frac{1}{\theta}\right), \ J = 24, \ \kappa = 1.$$

Furthermore, one has

$$\|\widetilde{d^2}\|_{W^{1,\infty}((-B,B))} \leq C_7 B \tag{39}$$

for some absolute constant C_7. We approximate d_i by

$$\widetilde{d_i^2}(\mathbf{x}) = \sum_{j=1}^{D} \widetilde{d^2}(x_j - c_{i,j}).$$

According to Lemma C.11, $\widetilde{d_i}$ can be realized by a CNN with $O\left(\log \frac{1}{\theta}\right)$ layers, $O(D)$ width and all weight parameters of $O(1)$. The approximation error is bounded as

$$\|\widetilde{d_i^2} - d_i^2\|_{L^\infty} \leq 4B^2 D\theta.$$

The following Lemma shows that $\mathbb{1}_{[0,r^2]}$ can be approximated by a CNN:

Lemma F.2 (Lemma 9 of Liu et al. (2021)). *For any $0 < \theta < 1$ and $\Delta \geq 8B^2D\theta$, there exists a CNN $\widetilde{\mathbb{1}}_\Delta$ approximating $\mathbb{1}_{[0,\omega^2]}$ with*

$$\widetilde{\mathbb{1}}_\Delta(\mathbf{x}) = \begin{cases} 1, & \text{if } a \leq (1-2^{-w})(r^2 - 4B^2D\theta), \\ 0, & \text{if } a \geq r^2 - 4B^2D\theta, \\ 2^w((r^2-4B^2D\theta)^{-1}a - 1), & \text{otherwise} \end{cases}$$

for $\mathbf{x} \in \mathcal{M}$, where $w = \lceil \log(r^2/\Delta) \rceil$ such that $(1-2^{-k})(\omega^2 - 4B^2D\theta) \geq \omega^2 - \Delta + 4B^2D\theta$. Such a CNN has $\lceil \log(r^2/\Delta) \rceil + D$ layers, 2 channels. All weight parameters are of $O(1)$.

Let $\mathbb{1}_\Delta$ be the CNN defined in Lemma F.2. We have

$$\frac{\partial \widetilde{\mathbb{1}}_\Delta(a)}{\partial a} = \begin{cases} 0, & \text{if } a \leq (1-2^{-w})(r^2 - 4B^2D\theta) \text{ or } a \geq r^2 - 4B^2D\theta, \\ C_8/\Delta, & \text{otherwise} \end{cases} \tag{40}$$

for some constant C_8 depending on r.

The function $\mathbb{1}_i$ is approximated by

$$\widetilde{\mathbb{1}}_i(\mathbf{x}) = \widetilde{\mathbb{1}}_\Delta \circ \widetilde{d}_i^2(\mathbf{x}).$$

Combining (40) and (39) gives rise to

$$\left|\frac{\partial \widetilde{\mathbb{1}}_i}{\partial x_j}\right| = \left|\frac{\partial \widetilde{\mathbb{1}}_\Delta}{a}\Big|_{\widetilde{d}_i^2(\mathbf{x})}\right| \left|\frac{\widetilde{d}_i}{\partial x_j}\right| \leq \begin{cases} 0, & \text{if } d_i(\mathbf{x})^2 \geq r^2 \text{ or } d_i^2(\mathbf{x}) \leq r^2 - \Delta, \\ CB/\Delta, & \text{otherwise.} \end{cases}$$

Step 3: Error analysis. Our network approximation of f is

$$\widetilde{f} = \sum_{i=1}^{C_\mathcal{M}} \widetilde{f}_i \quad \text{with} \quad \widetilde{f}_i(\mathbf{x}) = \sum_{\mathbf{m},\mathbf{v}} c_{i,\mathbf{m},\mathbf{v}} (\widetilde{g}_{\mathbf{m},\mathbf{v}} \circ \varphi_i(\mathbf{x})) \widetilde{\times} \widetilde{\mathbb{1}}_i(\mathbf{x}), \tag{41}$$

where $\widetilde{g}_{\mathbf{m},\mathbf{v}}$ is the CNN approximation of $\phi_\mathbf{m}\mathbf{z}^\mathbf{v}$ for $\mathbf{z} \in [0,1]^d$ as in (19). We decompose the error as

$$\begin{aligned}
\|\widetilde{f} - f\|_{W^{k,\infty}(\mathcal{M})} &\leq \sum_{i=1}^{C_\mathcal{M}} \|\widetilde{f}_i - f_i\|_{W^{k,\infty}(U_i)} \\
&= \sum_{i=1}^{C_\mathcal{M}} \|\widetilde{f}_i \circ \varphi_i^{-1} \circ \varphi_i - f_i \circ \varphi_i^{-1} \circ \varphi_i\|_{W^{k,\infty}(U_i)} \\
&\leq \sum_{i=1}^{C_\mathcal{M}} \|\widetilde{f}_i \circ \varphi_i^{-1}(\mathbf{z}) - f_i \circ \varphi_i^{-1}(\mathbf{z})\|_{W^{k,\infty}(\varphi_i(U_i))} \qquad \text{(set } \mathbf{z} = \varphi_i(\mathbf{x})\text{)} \\
&\leq \sum_{i=1}^{C_\mathcal{M}} \|\widetilde{f}_i \circ \varphi_i^{-1}(\mathbf{z}) - f_i \circ \varphi_i^{-1}(\mathbf{z})\|_{W^{k,\infty}(\varphi_i(U_i))} \\
&\leq \sum_{i=1}^{C_\mathcal{M}} \|\widetilde{f}_i \circ \varphi_i^{-1}(\mathbf{z}) - \widehat{f}_i(\mathbf{z})\|_{W^{k,\infty}(\varphi_i(U_i))} + \|\widehat{f}_i(\mathbf{z}) - f_i \circ \varphi_i^{-1}(\mathbf{z})\|_{W^{k,\infty}([0,1]^d)}.
\end{aligned} \tag{42}$$

The second term can be bounded using Lemma C.8. We next focus on the first term

$$
\begin{aligned}
&\|\widetilde{f}_i \circ \varphi_i^{-1}(\mathbf{z}) - \widehat{f}_i(\mathbf{z})\|_{W^{k,\infty}(\varphi_i(U_i))} \\
\leq &\left\| \sum_{\mathbf{m},\mathbf{v}} c_{i,\mathbf{m},\mathbf{v}} \left[(\widetilde{g}_{\mathbf{m},\mathbf{v}}(\mathbf{z})) \widetilde{\times} (\widetilde{\mathbb{1}}_i \circ \varphi_i^{-1}(\mathbf{z})) - \phi_{\mathbf{m}}(\mathbf{z}) \mathbf{z}^{\mathbf{v}} \right] \right\|_{W^{k,\infty}(\varphi_i(U_i))} \\
\leq &\left\| \sum_{\mathbf{m},\mathbf{v}} c_{i,\mathbf{m},\mathbf{v}} \left[(\widetilde{g}_{\mathbf{m},\mathbf{v}}(\mathbf{z})) \widetilde{\times} (\widetilde{\mathbb{1}}_i \circ \varphi_i^{-1}(\mathbf{z})) - (\widetilde{g}_{\mathbf{m},\mathbf{v}}(\mathbf{z})) \times (\widetilde{\mathbb{1}}_i \circ \varphi_i^{-1}(\mathbf{z})) \right] \right\|_{W^{k,\infty}(\varphi_i(U_i))} \\
&+ \left\| \sum_{\mathbf{m},\mathbf{v}} c_{i,\mathbf{m},\mathbf{v}} \left[(\widetilde{g}_{\mathbf{m},\mathbf{v}}(\mathbf{z})) \times (\widetilde{\mathbb{1}}_i \circ \varphi_i^{-1}(\mathbf{z})) - (\widetilde{g}_{\mathbf{m},\mathbf{v}}(\mathbf{z})) \times (\mathbb{1}_i \circ \varphi_i^{-1}(\mathbf{z})) \right] \right\|_{W^{k,\infty}(\varphi_i(U_i))} \\
&+ \left\| \sum_{\mathbf{m},\mathbf{v}} c_{i,\mathbf{m},\mathbf{v}} \left[(\widetilde{g}_{\mathbf{m},\mathbf{v}}(\mathbf{z})) \times (\mathbb{1}_i \circ \varphi_i^{-1}(\mathbf{z})) - \phi_{\mathbf{m}}(\mathbf{z}) \mathbf{z}^{\mathbf{v}} \right] \right\|_{W^{k,\infty}(\varphi_i(U_i))} \\
= &\|A_1\|_{W^{k,\infty}(\varphi_i(U_i))} + \|A_2\|_{W^{k,\infty}(\varphi_i(U_i))} + \|A_3\|_{W^{k,\infty}(\varphi_i(U_i))}
\end{aligned} \tag{43}
$$

with

$$
A_1 = \sum_{\mathbf{m},\mathbf{v}} c_{i,\mathbf{m},\mathbf{v}} \left[(\widetilde{g}_{\mathbf{m},\mathbf{v}}(\mathbf{z})) \widetilde{\times} (\widetilde{\mathbb{1}}_i \circ \varphi_i^{-1}(\mathbf{z})) - (\widetilde{g}_{\mathbf{m},\mathbf{v}}(\mathbf{z})) \times (\widetilde{\mathbb{1}}_i \circ \varphi_i^{-1}(\mathbf{z})) \right], \tag{44}
$$

$$
A_2 = \sum_{\mathbf{m},\mathbf{v}} c_{i,\mathbf{m},\mathbf{v}} \left[(\widetilde{g}_{\mathbf{m},\mathbf{v}}(\mathbf{z})) \times (\widetilde{\mathbb{1}}_i \circ \varphi_i^{-1}(\mathbf{z})) - (\widetilde{g}_{\mathbf{m},\mathbf{v}}(\mathbf{z})) \times (\mathbb{1}_i \circ \varphi_i^{-1}(\mathbf{z})) \right], \tag{45}
$$

$$
A_3 = \sum_{\mathbf{m},\mathbf{v}} c_{i,\mathbf{m},\mathbf{v}} \left[\widetilde{g}_{\mathbf{m},\mathbf{v}}(\mathbf{z}) - \phi_{\mathbf{m}}(\mathbf{z}) \mathbf{z}^{\mathbf{v}} \right]. \tag{46}
$$

Denote the $W^{1,\infty}$ error of $\widetilde{\times}$ by δ. We first derive an upper bound for A_1. We can show that $\|\widetilde{g}_{\mathbf{m},\mathbf{v}}\|_\infty \leq \alpha + d$ (see (69)) and $\|\widetilde{\mathbb{1}}_i \circ \varphi_i^{-1}\|_{L^\infty} = 1$. Therefore by Lemma G.7, we have for $k = 0$

$$
\begin{aligned}
|A_1|_{W^{0,\infty}([-\alpha-d,\alpha+d])} \leq &\sum_{\mathbf{m},\mathbf{v}} c_{i,\mathbf{m},\mathbf{v}} |\widetilde{\times}(a,b) - ab|_{W^{0,\infty}([-\alpha-D,\alpha+D])} \\
\leq &C_9 N^d \delta,
\end{aligned} \tag{47}
$$

and for $k = 1$

$$
\begin{aligned}
&|A_1|_{W^{1,\infty}([-\alpha-d,\alpha+d])} \\
\leq &\sum_{\mathbf{m},\mathbf{v}} c_{i,\mathbf{m},\mathbf{v}} C' |\widetilde{\times}(a,b) - ab|_{W^{1,\infty}([-\alpha-D,\alpha+D])} |\widetilde{g}_{\mathbf{m},\mathbf{v}}|_{W^{1,\infty}(\varphi_i(U_i))} \left| \widetilde{\mathbb{1}}_i \circ \varphi_i^{-1} \right|_{W^{1,\infty}(\varphi_i(U_i))} \\
\leq &C_{10} N^{d+1} \delta / \Delta
\end{aligned} \tag{48}
$$

for some constants C_9, C_{10}, C' depending on r, α, d, where we used Lemma C.8 and (65) in the last inequality. Combining (47) and (48) gives rise to

$$
\|A_1\|_{W^{k,\infty}([-\alpha-d,\alpha+d])} \leq C_{11} N^{d+k} \delta / \Delta \tag{49}
$$

for $k = 0, 1$ and a constant C_{11} depending on d, α, r.

Before we derive upper bounds for A_2 and A_3, we define some sets which will be used in our following proof.

Define the set

$$
\widetilde{\Omega}_{i,1} = \left\{ \mathbf{x} \in U_i : \min_{\widetilde{\mathbf{x}} \in \partial U_i} \|\mathbf{x} - \widetilde{\mathbf{x}}\|_2 \leq c \right\},
$$

where c is the constant from Lemma F.1. Denote $\Omega_{i,1} = \varphi_i(\widetilde{\Omega}_{i,1})$. According to Lemma F.1, we have $f_i|_{\widetilde{\Omega}_{i,1}} = f_i \circ \phi_i|_{\Omega_{i,1}} = 0$. Since φ_i is a bijection, both $\Omega_{i,1}$ and $\widetilde{\Omega}_{i,1}$ have two disjoint boundaries. Denote the two boundaries of $\Omega_{i,1}$ by $\lambda_{i,1,1}$ and $\lambda_{i,1,2}$. We define the thickness of $\Omega_{i,1}$ as

$$\chi_{i,1} = \min_{\mathbf{z}\in\lambda_{i,1,1},\, \widetilde{\mathbf{z}}\in\lambda_{i,1,2}} \|\mathbf{z} - \widetilde{\mathbf{z}}\|_2.$$

Since each φ_i is a bijection, there exists a constant c_1 depending on c and the atlas such that $\chi_{i,1} \geq c_1$ for all i's. Again since ϕ_i is a linear bijection, its inverse exists and is linear, and there exists a constant c_2 such that

$$\|\varphi_i^{-1}(\mathbf{z}) - \varphi_i^{-1}(\widetilde{\mathbf{z}})\|_2 \geq c_2\|\mathbf{z} - \widetilde{\mathbf{z}}\|_2. \tag{50}$$

We will choose θ and Δ small enough such that

$$\frac{8B^2D\theta}{c_2} \leq \frac{\Delta}{c_2 r} \leq \frac{c_1}{2}. \tag{51}$$

Define the region

$$\Omega_{i,2} = \left\{\mathbf{z} \in \varphi_i(U_i) : \min_{\widetilde{\mathbf{z}}\in\phi_i(\partial U_i)} \|\mathbf{z} - \widetilde{\mathbf{z}}\|_2 \leq \frac{\Delta}{c_2 r}\right\}. \tag{52}$$

According to (50), (51) and the definition of $\Omega_{i,1}$, we have $\Omega_{i,2} \subset \Omega_{i,1}$. For any $\mathbf{z} \in \varphi_i(U_i)\backslash\Omega_{i,2}$, denote $\mathbf{z}^* = \operatorname{argmin}_{\widetilde{\mathbf{z}}\in\varphi_i(\partial U_i)} \|\mathbf{z} - \widetilde{\mathbf{z}}\|_2$. We have

$$\min_{\widetilde{\mathbf{x}}\in\partial U_i} \|\varphi_i^{-1}(\mathbf{z}) - \widetilde{\mathbf{x}}\|_2 \geq c_2\|\mathbf{z} - \mathbf{z}^*\|_2 \geq \Delta/r.$$

Therefore

$$\|\varphi_i^{-1}(\mathbf{z}) - \mathbf{c}_i\|_2^2 \leq (r - \Delta/r)^2 = r^2 + \left(\frac{\Delta}{r}\right)^2 - 2\Delta \leq r^2 - \Delta$$

when $\Delta \leq r^2$ and

$$\widetilde{\mathbb{1}}_i \circ \varphi^{-1}(\mathbf{z}) = 1, \quad \left.\frac{\partial \widetilde{\mathbb{1}}_i \circ \varphi^{-1}}{z_j}\right|_{\varphi_i(U_i)\backslash\Omega_{i,2}} = 0$$

for $j = 1, ..., d$, where we used the notation $\mathbf{z} = [z_1, ..., z_d]^\top$.

Note that each $\widetilde{g}_{\mathbf{m},\mathbf{v}}$ and $\phi_{\mathbf{m}}\mathbf{z}^{\mathbf{v}}$ is supported on $B_{1/N,\|\cdot\|_\infty}(\mathbf{m}/N)$, a hyper cube with edge length $2/N$. We will choose N large enough such that

$$\frac{2}{N} \leq \frac{\Delta}{4c_2 r} \leq \frac{c_1}{8}.$$

Such a choice of N ensures that along any directions of z_j for $j = 1, ..., d$, there are at least 2 hypercubes that entirely locate inside $\Omega_{i,2}$. Since any $\mathbf{z} \in [0,1]^d$ is only covered by 2 hypercubes along each coordinate direction, we have

$$\{c_{i,\mathbf{m},\mathbf{v}} : \text{ there exits } \mathbf{z} \in \varphi_i(\partial U_i) \text{ such that } \mathbf{z} \in B_{1/N,\|\cdot\|_\infty}(\mathbf{m}/N)\} = 0 \tag{53}$$

and $\widetilde{f}_i \circ \varphi_i(\mathbf{z}) = 0$ for any $\mathbf{z} \in \varphi_i(\partial U_i)$. See Figure 4 for an illustration.

We have the following lemma on the bound of $\|A_2\|_{W^{k,\infty}(\varphi_i(U_i))}$ (see Appendix H.2 for a proof):

Lemma F.3. *Let A_2 be defined as in (45). Assume $\Delta \leq r^2$. We have*

$$\|A_2\|_{W^{1,\infty}(\varphi_i(U_i))} \leq C_{12} N\eta\Delta^{1-k} \tag{54}$$

for $k = 1, 2$.

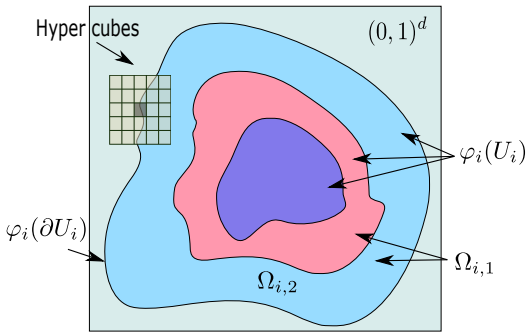


Figure 4: Illustration of the relations of $\Omega_{i,1},\ \Omega_{i,2}$ and $\varphi_i(U_i)$.

The term A_3 can be bounded using Lemma C.10:

$$\begin{aligned}\|A_3\|_{W^{k,\infty}(\varphi_i(U_i))} =& \left|\sum_{\mathbf{m},\mathbf{v}} c_{i,\mathbf{m},\mathbf{v}}\left[\widetilde{g}_{\mathbf{m},\mathbf{v}}(\mathbf{z}) - \phi_{\mathbf{m}}(\mathbf{z})\mathbf{z}^{\mathbf{v}}\right]\right|_{W^{k,\infty}(\varphi_i(U_i))}\\ \leq& \left|\sum_{\mathbf{m},\mathbf{v}} c_{i,\mathbf{m},\mathbf{v}}\left[\widetilde{g}_{\mathbf{m},\mathbf{v}}(\mathbf{z}) - \phi_{\mathbf{m}}(\mathbf{z})\mathbf{z}^{\mathbf{v}}\right]\right|_{W^{k,\infty}([0,1]^d)}\\ \leq& C_{13}N^k\eta \end{aligned} \tag{55}$$

for some constant C_{13} depending on d, α, R. Substituting (49), (54) and (55) into (43) gives rise to

$$\|\widetilde{f}_i \circ \varphi_i^{-1}(\mathbf{z}) - \widehat{f}_i(\mathbf{z})\|_{W^{k,\infty}(\varphi_i(U_i))} \leq C_{11}N^{d+k}\delta/\Delta + C_{12}N^k\eta + C_{13}N\eta\Delta^{1-k}. \tag{56}$$

The second term in (42) can be bounded by Lemma C.8 as

$$\|\widehat{f}_i(\mathbf{z}) - f_i \circ \varphi_i^{-1}(\mathbf{z})\|_{W^{k,\infty}([0,1]^d)} \leq C_{14}N^{-(\alpha-k)}. \tag{57}$$

Substituting (56) and (57) into (42) gives rise to

$$\|\widetilde{f} - f\|_{W^{k,\infty}(\mathcal{M})} \leq C_{\mathcal{M}}C_{11}N^{d+k}\delta/\Delta + C_{\mathcal{M}}C_{12}N^k\eta + C_{\mathcal{M}}C_{13}N\eta\Delta^{1-k} + C_{\mathcal{M}}C_{14}N^{-(\alpha-k)}.$$

Setting

$$\eta = N^{-\alpha},\ \Delta = 8c_2rN^{-1},\ \delta = N^{-(\alpha+d+1)},\ \theta = (8B^2D)^{-1}\Delta,$$

we have

$$\|\widetilde{f} - f\|_{W^{k,\infty}(\mathcal{M})} \leq C_{15}N^{-(\alpha-k)} \tag{58}$$

for $k = 0, 1$ and a constant C_{15} depending on d, α, τ and the surface area of $\mathcal{M}$.

•**Network size** We analyze the network size for each $\widetilde{f}_i$:

- $\widetilde{\mathbb{1}}_i$: The chart dermination network is the composition of $\widetilde{d}_i$ and $\widetilde{\mathbb{1}}_\Delta$, where $\widetilde{d}_i$ has $O(\log\frac{1}{\theta}) = O(\log N + \log D)$ layers and $O(D)$ width, $\widetilde{\mathbb{1}}_\Delta$ has $O(\log\frac{1}{\delta}) + D = O(\log N) + D$ layers and $O(1)$ width. In both subnetworks, all parameters are of $O(1)$. By Lemma G.2, the chart dermination network has $O(\log N + \log D) + D$ layers, $O(D)$ width and all weight parameters are of $O(1)$.
- $\widetilde{\times}$: The multiplication network has $O(\log\frac{1}{\delta}) = O(\log N)$ layers, $O(1)$ width. All weight parameters are bounded by $2(\alpha + d + 1)$.
- φ_i: the projection φ_i can be realized by a single layer with width d. All parameters are of $O(1)$.

- $\widetilde{g}_{i,\mathbf{m},\mathbf{v}}$: By Lemma C.9, each $\widetilde{g}_{i,\mathbf{m},\mathbf{v}}$ has $O(\log N)$ layers and $O(d)$ width. All parameters are of $O(N)$.
- $c_{i,\mathbf{m},\mathbf{v}}$: By Lemma C.8 with $p=\infty$, each $c_{i,\mathbf{m},\mathbf{v}}$ is of $O(1)$.

By Lemma G.2, each $c_{i,\mathbf{m},\mathbf{v}}(\widetilde{g}_{\mathbf{m},\mathbf{v}}\circ\varphi_i(\mathbf{x}))\widetilde{\times}\widetilde{\mathbb{1}}_i(\mathbf{x})$ is a CNN with $O(\log N+\log D)+D$ layers, $O(D)$ width and all parameters of $O(N)$. According to (41), $\widetilde{f}$ can be written as a sum of $C_{\mathcal{M}}N^d(d+1)^\alpha$ CNNs

$$\widetilde{f}=\sum_{i=1}^{C_{\mathcal{M}}}\sum_{\mathbf{m},\mathbf{v}}c_{i,\mathbf{m},\mathbf{v}}(\widetilde{g}_{\mathbf{m},\mathbf{v}}\circ\varphi_i(\mathbf{x}))\widetilde{\times}\widetilde{\mathbb{1}}_i(\mathbf{x}). \tag{59}$$

By Lemma C.11, for any $\widetilde{M},\widetilde{J}$ satisfying $\widetilde{M}\widetilde{J}=O(N^d)$, there exists a CNN architecture $\mathcal{F}^{\rm CNN}(L,J,K,\kappa,\kappa)$ that gives rise to $\{g_i\}_{i=1}^n$ with

$$\widetilde{f}=\sum_{i=1}^{\widetilde{M}}g_i$$

and

$$L=O\left(\log N+\log D\right)+D,\ J=O(D\widetilde{J}),\ \kappa=O(N).$$

By Lemma C.12, there exits a ConvResNet architecture $\mathcal{C}(M,L,J,K,\kappa_1,\kappa_2)$ with

$$L=O(\log N)+D, J=O(D\widetilde{J}), \kappa_1=\kappa_2=O(N), M=O(\widetilde{M}) \tag{60}$$

and $\widetilde{J},\widetilde{M}$ satisfying

$$\widetilde{M}\widetilde{J}=O(N^d), \tag{61}$$

that yields a ConvResNet realizing $\widetilde{f}$. Setting $N=O((\widetilde{M}\widetilde{J})^{1/d})$ in (58) and (60) gives rise to

$$\|\widetilde{f}-f\|_{W^{k,\infty}(\mathcal{M})}\le C_{15}(\widetilde{M}\widetilde{J})^{-\frac{\alpha-k}{d}} \tag{62}$$

and the network size

$$L=O\left(\log(\widetilde{M}\widetilde{J})+\log D\right)+D,\ J=O(D\widetilde{J}),\ \kappa=O((\widetilde{M}\widetilde{J})^{1/d}).$$

□

G. Definitions, Lemmas and their proofs used in Section C

G.1. Existing lemmas on CNNs

Lemma G.1 shows that any MLP can be realized by a CNN.

Lemma G.1 (Theorem 1 in Oono & Suzuki (2019))**.** *Let D be the dimension of the input. Let L,J be positive integers and $\kappa>0$. For any $2\le K'\le D$, any MLP architectures $\mathcal{F}^{\rm MLP}(L,J,\kappa)$ can be realized by a CNN architecture $\mathcal{F}^{\rm CNN}(L',J',K',\kappa_1',\kappa_2')$ with*

$$L'=L+D, J'=4J, \kappa_1'=\kappa_2'=\kappa.$$

Specifically, any $\bar{f}^{\rm MLP}\in\mathcal{F}^{\rm MLP}(L,J,\kappa)$ can be realized by a CNN $\bar{f}^{\rm CNN}\in\mathcal{F}^{\rm CNN}(L',J',K',\kappa_1',\kappa_2')$. Furthermore, the weight matrix in the fully connected layer of $\bar{f}^{\rm CNN}$ has nonzero entries only in the first row.

Lemma G.2 shows that the composition of two CNNs can be realized by a CNN.

Lemma G.2 (Lemma 13 in Liu et al. (2021))**.** *Let $\mathcal{F}_1^{\rm CNN}(L_1,J_1,K_1,\kappa_1,\kappa_1)$ be a CNN architecture from $\mathbb{R}^D\to\mathbb{R}$ and $\mathcal{F}_2^{\rm CNN}(L_2,J_2,K_2,\kappa_2,\kappa_2)$ be a CNN architecture from $\mathbb{R}\to\mathbb{R}$. Assume the weight matrix in the fully connected layer of $\mathcal{F}_1^{\rm CNN}(L_1,J_1,K_1,\kappa_1,\kappa_1)$ and $\mathcal{F}_2^{\rm CNN}(L_2,J_2,K_2,\kappa_2,\kappa_2)$ has nonzero entries only in the first row. Then there exists a CNN architecture $\mathcal{F}^{\rm CNN}(L,J,K,\kappa,\kappa)$ from $\mathbb{R}^D\to\mathbb{R}$ with*

$$L=L_1+L_2,\ J=\max(J_1,J_2),\ K=\max(K_1,K_2),\kappa=\max(\kappa_1,\kappa_2)$$

such that for any $f_1\in\mathcal{F}^{\rm CNN}(L_1,J_1,K_1,\kappa_1,\kappa_1)$ and $f_2\in\mathcal{F}^{\rm CNN}(L_2,J_2,K_2,\kappa_2,\kappa_2)$, there exists $f\in\mathcal{F}^{\rm CNN}(L,J,K,\kappa,\kappa)$ such that $f(\mathbf{x})=f_2\circ f_1(\mathbf{x})$. Furthermore, the weight matrix in the fully connected layer of $\mathcal{F}^{\rm CNN}(L,J,K,\kappa,\kappa)$ has nonzero entries only in the first row.

G.2. Interpolation spaces

Definition G.3 (Interpolation spaces)**.** Let (B_0, B_1) be an interpolation couple. For any $u \in B_1$, define

$$K(t, u, B_0, B_1) = \inf_{v \in B_1} \left(\|u - v\|_{B_0} + t\|v\|_{B_1} \right)$$

and the norm

$$\|u\|_{(B_0,B_1)_{\theta,p}} = \begin{cases} \left(\int_0^\infty t^{-\theta p} K(t, u, B_0, B_1)^p \frac{dt}{t} \right)^{1/p}, & \text{for } 1 \le p < \infty, \\ \sup_{0<t<\infty} t^{-\theta} K(t, u, B_0, B_1), & \text{for } p = \infty. \end{cases}$$

Then the interpolation space $(B_0, B_1)_{\theta,p}$ is defined by

$$(B_0, B_1)_{\theta,p} = \left\{ u \in B_0 : \|u\|_{(B_0,B_1)_{\theta,p}} < \infty \right\}.$$

The following lemma shows that the fractional Sobolev space is an interpolation space:

Lemma G.4 (Theorem 14.2.3 of Brenner et al. (2008))**.** *Let $\Omega \in \mathbb{R}^D$ be an Lipschitz domain. Then for any $0 < s < 1$ and $1 \le p \le \infty$, we have*

$$W^{s,p}(\Omega) = (L^p(\Omega), W^{1,p}(\Omega))_{s,p}.$$

The following lemma shows that the norm of the interpolation space of $(B_0, B_1)_{\theta,p}$ can be bounded using $\|\cdot\|_{B_0}$ and $\|\cdot\|_{B_1}$:

Lemma G.5. *Let (B_0, B_1) be an interpolation couple. Moreover, let $0 < \theta < 1$ and $1 \le p \le \infty$. Then there exists a constant C depending on θ and p such that for all $u \in B_1$, we have*

$$\|u\|_{B_{\theta,p}} \le C \|u\|_{B_0}^{1-\theta} \|u\|_{B_1}^{\theta}.$$

In particular, when $p = \infty$, we have $C = 1$.

G.3. Proof of Lemma C.2

Proof of Lemma C.2. Note that $\psi(x)$ can be realized by a two-layer MLP

$$\psi(x) = \mathrm{ReLU}(A_2 \cdot \mathrm{ReLU}\left(A_1 x + \boldsymbol{b}_1\right))$$

with

$$A_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \ \boldsymbol{b}_1 = \begin{bmatrix} 2 \\ 1 \\ -1 \\ -2 \end{bmatrix}, \ A_2 = \begin{bmatrix} 1 & -1 & -1 & 1 \end{bmatrix}.$$

According to Lemma G.1, for any $2 \le K$, such an MLP can be realized by a CNN in $\mathcal{F}^{\mathrm{CNN}}(2, 16, 2, 2, 2)$. According to the expression of the right-hand-side of (12), we have $\widetilde{\psi}_{m,N}(x) \in \mathcal{F}^{\mathrm{CNN}}(2, 16, 2, 3N, 3N)$.

To prove (13), the case $k = 0$ follows by the definition of ψ. For $k = 1$, we have

$$\frac{d\widetilde{\psi}_{m,N}(x)}{dx} = \frac{d\psi\left(3N\left(x_k - \frac{m}{N}\right)\right)}{dx} = 3N.$$

□

G.4. Proof of Lemma C.9

Proof of Lemma C.9. For any given $\mathbf{m}$ and $\mathbf{v}$, $\phi_{\mathbf{m}} \mathbf{x}^{\mathbf{v}}$ is a product of at most $\alpha + D$ quantities each of which can be realized by a CNN. The following lemma shows that the multiplication operator $\times$ can be well approximated by a CNN (see a proof in Appendix G.5):

Lemma G.6. *For any $0 < \eta < 1/2$, $x, y \in [-B, B]$, and $K \geq 2$, there is a CNN architecture $\mathcal{F}^{\mathrm{CNN}}(L, J, K, \kappa, \kappa)$ that yields a CNN, denoted by $\widetilde{\times}(\cdot, \cdot)$, such that*

$$\|\widetilde{\times}(x, y) - xy\|_{W^{1,\infty}[-B,B]^2} < \eta, \ \widetilde{\times}(x, 0) = \widetilde{\times}(y, 0) = 0.$$

Such a network has

$$L = O\left(\log \frac{1}{\eta}\right), \ J = 24, \ \kappa = 1.$$

Furthermore, one has

$$\|\widetilde{\times}(x, y)\|_{W^{1,\infty}((-B,B)^2)} \leq CB$$

for some absolute constant C.

For simplicity, we denote $\psi_{m_k}(x) = \psi\left(3N\left(x_k - \frac{m_k}{N}\right)\right)$ for $k = 1, ..., D$. Then we construct $\widetilde{g}_{\mathbf{m},\mathbf{v}}(\mathbf{x})$ as

$$\widetilde{g}_{\mathbf{m},\mathbf{v}}(\mathbf{x}) = \widetilde{\times}(\widetilde{\times}(...\widetilde{\times}(\widetilde{\times}(\widetilde{p}_{\mathbf{v}}(\mathbf{x}), \psi_{m_1}(x_1)), \psi_{m_2}(x_2)), ...,), \psi_{m_D}(x_D)),$$

where $\widetilde{p}_{\mathbf{v}}(\mathbf{x})$ is the network approximation of $\mathbf{x}^{\mathbf{v}}$ defined by

$$\widetilde{p}_{\mathbf{v}}(\mathbf{x}) = \widetilde{\times}(...\widetilde{\times}(x_1, x_1), ..., x_D).$$

The structure of $\widetilde{g}_{\mathbf{m},\mathbf{v}}$ is visualized in Figure 5. Here $\widetilde{g}_{\mathbf{m},\mathbf{v}}$ consists of no more than $\alpha + D - 1$ compositions of $\widetilde{\times}$ and $2D$ additional channels. These additional channels are used to pass the information $\mathbf{x}_+$ and $\mathbf{x}_-$.

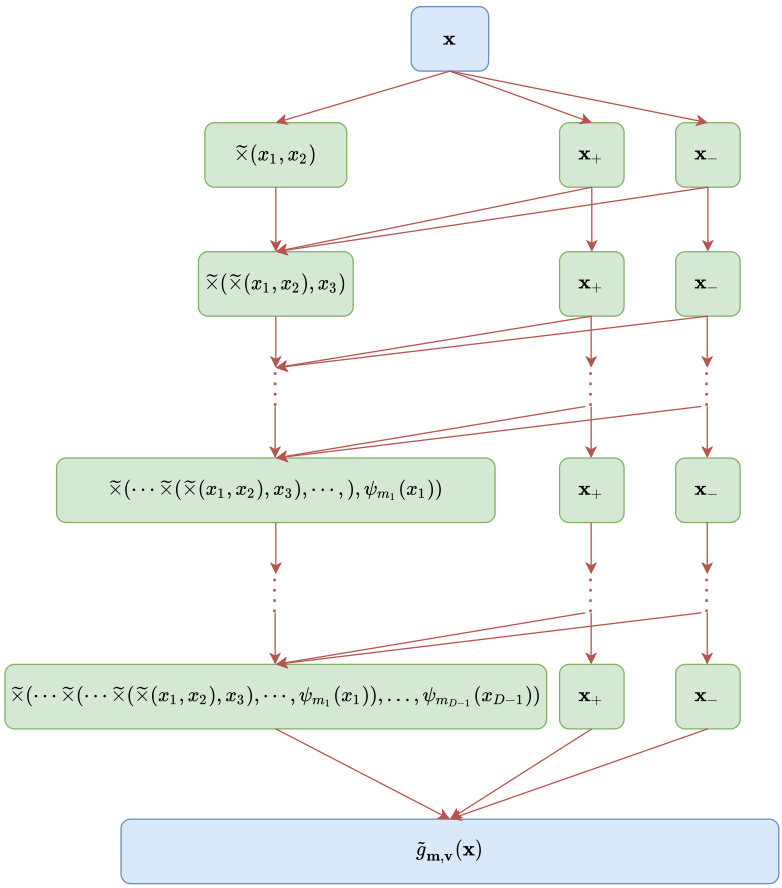


Figure 5: Illustration of $\widetilde{g}_{\mathbf{m},\mathbf{v}}$.

By applying Lemma G.2 $\alpha + D - 2$ times, we have $\widetilde{g}_{\mathbf{m},\mathbf{v}} \in \mathcal{F}^{\mathrm{CNN}}(L, J, K, \kappa, \kappa)$ with

$$L = O\left(D \log \frac{1}{\varepsilon}\right),\ J = O(D),\ \kappa = 3N.$$

We next prove (17) and (18). First note that we can express

$$\phi_{\mathbf{m}\mathbf{x}^{\mathbf{v}}} = g_n \equiv \prod_{i=1}^{n} h_i(\mathbf{x}),$$
$$\widetilde{g}_{\mathbf{m},\mathbf{v}}(\mathbf{x}) = \widetilde{g}_n \equiv \widetilde{\times}(\widetilde{\times}(\cdots \widetilde{\times}(h_1(\mathbf{x}), h_2(\mathbf{x})), ...), h_n(\mathbf{x}))$$

for some $n \leq \alpha + D$, where each h_i can be realized by one layer and satisfies

$$\|h_i(\mathbf{x})\|_{W^{k,\infty}(0,1)} \leq (3N)^k.$$

To prove (17) and (18), it is enough to show

$$\|\widetilde{g}_n(\mathbf{x}) - g_n(\mathbf{x})\|_{W^{k,\infty}((0,1)^D} \leq n^{1-k} c_n^k N^k \varepsilon \tag{63}$$
$$\widetilde{g}_n(\mathbf{x}) = 0 \text{ if } g_n(\mathbf{x}) = 0, \tag{64}$$
$$|\widetilde{g}_n(\mathbf{x})|_{W^{1,\infty}((0,1)^D)} \leq C_{16} N^k \tag{65}$$

for any $1 \leq n \leq \alpha + D - 1$, where $\{c_n\}_{n=1}^{\alpha+D-1}$ and C_{16} are constants depending on D and α.

For $n = 1$, we have

$$|\widetilde{g}_n - g_n|_{W^{k,\infty}((0,1)^D)} = |\widetilde{\times}(h_1, 1) - h_1|_{W^{k,\infty}((0,1)^D)}.$$

By Lemma G.6 with $B = \alpha + D + 1$, we have for $k = 0$,

$$|\widetilde{\times}(h_1, 1) - h_1|_{W^{0,\infty}((0,1)^D)} \leq \varepsilon.$$

For $k = 1$, by Lemma G.6, we deduce

$$|\widetilde{\times}(h_1, 1) - h_1|_{W^{1,\infty}((0,1)^D)} \leq C' |\widetilde{\times}(x, y) - x \cdot y|_{W^{1,\infty}([0,1]^2)} |h_1|_{W^{1,\infty}([0,1]^2)} \leq 3C' N\varepsilon,$$

where C' is a constant depending on D. We set $c_1 = 3C'$. Furthermore,

$$|\widetilde{g}_1(\mathbf{x})|_{W^{1,\infty}((0,1)^D)} = |\widetilde{\times}(h_1, 1)|_{W^{1,\infty}((0,1)^D)} \leq C_4 |\widetilde{\times}(x, y) - x \cdot y|_{W^{1,\infty}((0,1)^2)} |h_1|_{W^{1,\infty}((0,1)^2)} \leq C_5 N,$$

where C_{17}, C_{18} are constants depending on D, α.

Therefore, the inequalities (63) and (65) hold for $n = 1$.

For (64), if $g_1(\mathbf{x}) = 0$, then $h_1(\mathbf{x}) = 0$. By Lemma G.6, $\widetilde{g}_1(\mathbf{x}) = 0$.

Assume (63)–(65) hold for any $1 \leq n \leq t$ for some integer t satisfying $1 \leq t \leq \alpha + D - 2$, i.e., for any $1 \leq n \leq t$, we have

$$|\widetilde{g}_n - g_n|_{W^{k,\infty}((0,1)^D)} \leq n^{1-k} c_n^k N^k \varepsilon, \tag{66}$$
$$\widetilde{g}_n = 0 \text{ if } g_n = 0, \tag{67}$$
$$|\widetilde{g}_n|_{W^{1,\infty}((0,1)^D)} \leq C_{19} N. \tag{68}$$

We also deduce that

$$|\widetilde{g}_t|_{W^{0,\infty}((0,1)^D)} = |\widetilde{g}_t - g_t|_{W^{0,\infty}((0,1)^D)} + |g_t|_{W^{0,\infty}((0,1)^D)} \leq t\varepsilon + 1 \leq t + 1. \tag{69}$$

For $n = t + 1$, we have

$$\begin{aligned} |\widetilde{g}_{t+1} - g_{t+1}|_{W^{k,\infty}((0,1)^D)} &= |\widetilde{\times}(\widetilde{g}_t, h_{t+1}) - g_t \cdot h_{t+1}|_{W^{k,\infty}((0,1)^D)} \\ &\leq |\widetilde{\times}(\widetilde{g}_t, h_{t+1}) - \widetilde{g}_t \cdot h_{t+1}|_{W^{k,\infty}((0,1)^D)} + |\widetilde{g}_t \cdot h_{t+1} - g_t \cdot h_{t+1}|_{W^{k,\infty}((0,1)^D)}. \end{aligned} \tag{70}$$

Consider the first term in (70). For $k=0$, we have

$$|\widetilde{\times}(\widetilde{g}_t, h_{t+1}) - \widetilde{g}_t \cdot h_{t+1}|_{W^{0,\infty}((0,1)^D)} \leq |\widetilde{\times}(x,y) - x\cdot y|_{W^{0,\infty}([-t-1,t+1]^2)} \leq \varepsilon. \tag{71}$$

For $k=1$, we have

$$\begin{aligned}&|\widetilde{\times}(\widetilde{g}_t, h_{t+1}) - \widetilde{g}_t \cdot h_{t+1}|_{W^{1,\infty}((0,1)^D)}\\ \leq& C'|\widetilde{\times}(x,y) - x\cdot y|_{W^{1,\infty}([-t-1,t+1]^2)}|\widetilde{g}_t|_{W^{1,\infty}([-t-1,t+1]^2)} \leq 3C' c_t N\varepsilon,\end{aligned} \tag{72}$$

where (66) with $k=1$ is used in the last inequality, C' is a constant depending on D.

For the second term in (70), we first consider $k=0$:

$$|\widetilde{g}_t \cdot h_{t+1} - g_t \cdot h_{t+1}|_{W^{0,\infty}((0,1)^D)} \leq |h_{t+1}|_\infty |\widetilde{g}_t - g_t|_\infty \leq t\varepsilon, \tag{73}$$

where (66) with $k=0$ is used.

For $k=1$, we have

$$\begin{aligned}&|\widetilde{g}_t \cdot h_{t+1} - g_t \cdot h_{t+1}|_{W^{1,\infty}((0,1)^D)}\\ =&|h_{t+1}(\widetilde{g}_t - g_t)|_{W^{1,\infty}((0,1)^D)}\\ \leq& C_{20}|h_{t+1}|_{W^{1,\infty}((0,1)^D)}\|\widetilde{g}_t - g_t\|_\infty + C_{20}\|h_{t+1}\|_\infty |\widetilde{g}_t - g_t|_{W^{1,\infty}((0,1)^D)}\\ \leq& 3C_{20}Nt\varepsilon + C_{20}c_t N\varepsilon \leq C_{21}N\varepsilon,\end{aligned} \tag{74}$$

where C_{20}, C_{21} are constants depending on D and α. In (74), (66) with $k=0$ and $k=1$ are used in the second inequality.

Combining (71)–(74) and setting $c_{t+1} = 3C_{20}c_t + C_{21}$ gives rise to

$$|\widetilde{g}_{t+1} - g_{t+1}|_{W^{k,\infty}((0,1)^D)} \leq (t+1)^{1-k} c_{t+1}^k N^k \varepsilon.$$

Therefore, (63) holds for $n = t+1$.

To prove (64), note that if $g_{t+1}=0$, then either $h_{t+1}=0$, or $g_t = 0$. By our induction assumption, when $g_t = 0$, we have $\widetilde{g}_t = 0$. Since $\widetilde{g}_{t+1} = \widetilde{\times}(\widetilde{g}_t, h_{t+1})$, by Lemma C.9, we have $\widetilde{g}_{t+1} = 0$ and (64) holds for $n = t+1$.

For (65), we deduce

$$\begin{aligned}&|\widetilde{g}_{t+1}(\mathbf{x})|_{W^{1,\infty}((0,1)^D)}\\ =&|\widetilde{\times}(\widetilde{g}_t, h_{t+1})|_{W^{1,\infty}((0,1)^D)}\\ \leq& C'|\widetilde{\times}(x,y) - x\cdot y|_{W^{1,\infty}((-t-1,t+1)^2)} \max\left\{|\widetilde{g}_t|_{W^{1,\infty}((0,1)^2)}, |h_{t+1}|_{W^{1,\infty}((0,1)^2)}\right\}\\ \leq& C_{22}N,\end{aligned}$$

where C_{22} is a constant depending on D and α.

Therefore, (63)–(65) hold for $n = t+1$. By mathematical induction, (63)–(65) hold for any $1 \leq n \leq D+\alpha+1$, and (17) and (18) are proved. □

G.5. Proof of Lemma G.6

Proof of Lemma G.6. The proof of Lemma G.6 is based on the following lemma.

Lemma G.7 (Proposition C.2 in Gühring et al. (2020))**.** *For any $0<\eta<1/2$, $x, y \in [-B, B]$. There is an MLP, denoted by $\widetilde{\times}(\cdot,\cdot)$, such that*

$$\|\widetilde{\times}(x,y) - xy\|_{W^{1,\infty}[-B,B]^2} < \eta,\ \widetilde{\times}(x,0) = \widetilde{\times}(y,0) = 0.$$

Such a network has $O\left(\log \frac{1}{\eta}\right)$ layers and parameters. The width of each layer is bounded by 6 and all parameters are bounded by 2. Furthermore, we have

$$\|\widetilde{\times}(x,y)\|_{W^{1,\infty}((-B,B)^2)} \leq CM,$$

for some absolute constant C.

Combing Lemma G.7 and G.1, for any $\varepsilon > 0$, $K \geq 2$, there exits a CNN $\widetilde{\times} \in \mathcal{F}^{\mathrm{CNN}}(L, J, K, \kappa, \kappa)$ such that for any $|x| \leq B, |y| \leq B$, we have

$$|\widetilde{\times}(x,y) - xy| < \varepsilon,\ \widetilde{\times}(x,0) = \widetilde{\times}(y,0) = 0,$$
$$\|\widetilde{\times}(x,y)\|_{W^{1,\infty}((-B,B)^2)} \leq C_{23}B,$$

where C_{23} is an absolute constant. Such an architecture has

$$L = O\left(\log \frac{1}{\varepsilon}\right),\ J = 24,\ \kappa = 1.$$

□

G.6. Proof of Lemma C.10

Proof of Lemma C.10. Denote $\Omega_{\mathbf{m},N} = B_{\frac{1}{N},\|\cdot\|_\infty}\left(\frac{\mathbf{m}}{N}\right)$. We have

$$\begin{aligned}
&\| \sum_{\mathbf{m}\in\mathcal{S}_N} \phi_{\mathbf{m}} Q^\alpha_{\mathbf{m}/N} f - \sum_{\mathbf{m}\in\mathcal{S}_N} \sum_{|\mathbf{v}|\leq\alpha-1} c_{\mathbf{m},\mathbf{v}} \widetilde{g}_{\mathbf{m},\mathbf{v}} \|_{W^{k,p}((0,1)^D)} \\
=&\left\| \sum_{\mathbf{m}\in\mathcal{S}_N} \sum_{|\mathbf{v}|\leq\alpha-1} c_{\mathbf{m},\mathbf{v}} \phi_{\mathbf{m}} \mathbf{x}^{\mathbf{v}} - \sum_{\mathbf{m}\in\mathcal{S}_N} \sum_{|\mathbf{v}|\leq\alpha-1} c_{\mathbf{m},\mathbf{v}} \widetilde{g}_{\mathbf{m},\mathbf{v}} \right\|^p_{W^{k,p}((0,1)^D)} \\
=&\left\| \sum_{\mathbf{m}\in\mathcal{S}_N} \sum_{|\mathbf{v}|\leq\alpha-1} c_{\mathbf{m},\mathbf{v}} \left(\phi_{\mathbf{m}} \mathbf{x}^{\mathbf{v}} - \widetilde{g}_{\mathbf{m},\mathbf{v}}\right) \right\|^p_{W^{k,p}((0,1)^D)} \\
\leq& \sum_{\widetilde{\mathbf{m}}\in\mathcal{S}_N} \left\| \sum_{\mathbf{m}\in\mathcal{S}_N} \sum_{|\mathbf{v}|\leq\alpha-1} c_{\mathbf{m},\mathbf{v}} \left(\phi_{\mathbf{m}} \mathbf{x}^{\mathbf{v}} - \widetilde{g}_{\mathbf{m},\mathbf{v}}\right) \right\|^p_{W^{k,p}(\Omega_{\widetilde{\mathbf{m}},N}\cap(0,1)^D)},
\end{aligned} \tag{75}$$

where the first equality follows from (15), the last inequality holds since $(0,1)^D \subset \cup_{\widetilde{\mathbf{m}}\in\mathcal{S}_N} \Omega_{\widetilde{\mathbf{m}},N}$.

For each $\widetilde{\mathbf{m}}$, we have

$$\begin{aligned}
&\left\| \sum_{\mathbf{m}\in\mathcal{S}_N} \sum_{|\mathbf{v}|\leq\alpha-1} c_{\mathbf{m},\mathbf{v}} \left(\phi_{\mathbf{m}} \mathbf{x}^{\mathbf{v}} - \widetilde{g}_{\mathbf{m},\mathbf{v}}\right) \right\|_{W^{k,p}(\Omega_{\widetilde{\mathbf{m}},N}\cap(0,1)^D)} \\
\leq& \sum_{\mathbf{m}\in\mathcal{S}_N} \sum_{|\mathbf{v}|\leq\alpha-1} |c_{\mathbf{m},\mathbf{v}}| \, \|\phi_{\mathbf{m}} \mathbf{x}^{\mathbf{v}} - \widetilde{g}_{\mathbf{m},\mathbf{v}}\|_{W^{k,p}(\Omega_{\widetilde{\mathbf{m}},N}\cap(0,1)^D)} \\
\leq& C_{24} N^{d/p} \sum_{\mathbf{m}\in\mathcal{S}_N} \sum_{|\mathbf{v}|\leq\alpha-1} \|\bar{f}\|_{W^{\alpha-1,p}(\Omega_{\mathbf{m},N})} \, \|\phi_{\mathbf{m}} \mathbf{x}^{\mathbf{v}} - \widetilde{g}_{\mathbf{m},\mathbf{v}}\|_{W^{k,p}(\Omega_{\widetilde{\mathbf{m}},N}\cap(0,1)^D)},
\end{aligned} \tag{76}$$

where C_{21} is the constant in Lemma C.7, $\bar{f}$ is the extension of f to $\mathbb{R}^D$ from Stein (1970, Theorem VI.3.1.5), which satisfies

$$\|\bar{f}\|_{W^{\alpha,p}(\mathbb{R}^D)} \leq C_{25} \|f\|_{W^{\alpha,p}((0,1)^D)} \tag{77}$$

for some constant C_{25} depending on D, p, α.

We next derive an upper bound of the summand of (76). We first deduce that

$$\begin{aligned}
&\|\phi_{\mathbf{m}} \mathbf{x}^{\mathbf{v}} - \widetilde{g}_{\mathbf{m},\mathbf{v}}\|_{W^{k,p}(\Omega_{\widetilde{\mathbf{m}},N}\cap(0,1)^D)} \\
\leq& \left|\Omega_{\widetilde{\mathbf{m}},N} \cap (0,1)^D\right|^{1/p} (D+1)^{1/p} \|\phi_{\mathbf{m}} \mathbf{x}^{\mathbf{v}} - \widetilde{g}_{\mathbf{m},\mathbf{v}}\|_{W^{k,\infty}(\Omega_{\widetilde{\mathbf{m}},N}\cap(0,1)^D)} \\
\leq& C_{26} \left(\frac{1}{N}\right)^{d/p} \|\phi_{\mathbf{m}} \mathbf{x}^{\mathbf{v}} - \widetilde{g}_{\mathbf{m},\mathbf{v}}\|_{W^{k,\infty}(\Omega_{\widetilde{\mathbf{m}},N}\cap(0,1)^D)} \\
\leq& C_{27} \left(\frac{1}{N}\right)^{d/p} N^k \eta,
\end{aligned} \tag{78}$$

where $\left|\Omega_{\widetilde{\mathbf{m}},N}\cap(0,1)^D\right|$ denotes the volume of $\Omega_{\widetilde{\mathbf{m}},N}\cap(0,1)^D$, C_{26}, C_{27} are constants depending on D,α and p. We used Lemma C.9 in the last inequality. Substituting (78) into (76) gives rise to

$$\begin{aligned}
&\left\|\sum_{\mathbf{m}\in\mathcal{S}_N}\sum_{|\mathbf{v}|\le\alpha-1} c_{\mathbf{m},\mathbf{v}}\left(\phi_{\mathbf{m}}\mathbf{x}^{\mathbf{v}}-\widetilde{g}_{\mathbf{m},\mathbf{v}}\right)\right\|_{W^{k,p}(\Omega_{\widetilde{\mathbf{m}},N}\cap(0,1)^D)}\\
=&C_{24}\sum_{\substack{\mathbf{m}\in\mathcal{S}_N\\ \|\mathbf{m}-\widetilde{\mathbf{m}}\|_\infty\le 1}}\sum_{|\mathbf{v}|\le\alpha-1}\|\bar{f}\|_{W^{\alpha-1,p}(\Omega_{\mathbf{m},N})}\left\|\phi_{\mathbf{m}}\mathbf{x}^{\mathbf{v}}-\widetilde{g}_{\mathbf{m},\mathbf{v}}\right\|_{W^{k,p}(\Omega_{\widetilde{\mathbf{m}},N}\cap(0,1)^D)}\\
\le&C_{24}C_{27}N^k\eta\sum_{\substack{\mathbf{m}\in\mathcal{S}_N\\ \|\mathbf{m}-\widetilde{\mathbf{m}}\|_\infty\le 1}}\sum_{|\mathbf{v}|\le\alpha-1}\|\bar{f}\|_{W^{\alpha-1,p}(\Omega_{\mathbf{m},N})}\\
\le&C_{28}N^k\eta\sum_{\substack{\mathbf{m}\in\mathcal{S}_N\\ \|\mathbf{m}-\widetilde{\mathbf{m}}\|_\infty\le 1}}\|\bar{f}\|_{W^{\alpha-1,p}(\Omega_{\mathbf{m},N})},
\end{aligned}\tag{79}$$

where $C_{28}=C_{24}C_{27}(D+1)^{\alpha-1}$. By Hölder's inequality, we have

$$\begin{aligned}
&\sum_{\substack{\mathbf{m}\in\mathcal{S}_N\\ \|\mathbf{m}-\widetilde{\mathbf{m}}\|_\infty\le 1}}\|\bar{f}\|_{W^{\alpha-1,p}(\Omega_{\mathbf{m},N})}\\
=&\sum_{\substack{\mathbf{m}\in\mathcal{S}_N\\ \|\mathbf{m}-\widetilde{\mathbf{m}}\|_\infty\le 1}}\|\bar{f}\|_{W^{\alpha-1,p}(\Omega_{\mathbf{m},N})}\cdot 1\\
\le&\left(\sum_{\substack{\mathbf{m}\in\mathcal{S}_N\\ \|\mathbf{m}-\widetilde{\mathbf{m}}\|_\infty\le 1}}\|\bar{f}\|^p_{W^{\alpha-1,p}(\Omega_{\mathbf{m},N})}\right)^{\frac{1}{p}}\left(\sum_{\substack{\mathbf{m}\in\mathcal{S}_N\\ \|\mathbf{m}-\widetilde{\mathbf{m}}\|_\infty\le 1}}1^q\right)^{\frac{1}{q}}\\
\le&3^{\frac{D}{q}}\left(\sum_{\substack{\mathbf{m}\in\mathcal{S}_N\\ \|\mathbf{m}-\widetilde{\mathbf{m}}\|_\infty\le 1}}\|\bar{f}\|^p_{W^{\alpha-1,p}(\Omega_{\mathbf{m},N})}\right)^{\frac{1}{p}},
\end{aligned}\tag{80}$$

where $q=1/(1-1/p)$. Substituting (79), (80) into (75) gives rise to

$$\begin{aligned}
&\|\sum_{\mathbf{m}\in\mathcal{S}_N}\phi_{\mathbf{m}}Q^\alpha_{\mathbf{m}/N}f-\sum_{\mathbf{m}\in\mathcal{S}_N}\sum_{|\mathbf{v}|\le\alpha-1}c_{\mathbf{m},\mathbf{v}}\widetilde{g}_{\mathbf{m},\mathbf{v}}\|^p_{W^{k,p}((0,1)^D)}\\
\le&\left(C_{28}3^{\frac{D}{q}}N^k\eta\right)^p\left(\sum_{\widetilde{\mathbf{m}}\in\mathcal{S}_N}\sum_{\substack{\mathbf{m}\in\mathcal{S}_N\\ \|\mathbf{m}-\widetilde{\mathbf{m}}\|_\infty\le 1}}\|\bar{f}\|^p_{W^{\alpha-1,p}(\Omega_{\mathbf{m},N})}\right)\\
\le&\left(C_{28}3^{\frac{D}{q}}N^k\eta\right)^p 3^D\left(\sum_{\widetilde{\mathbf{m}}\in\mathcal{S}_N}\|\bar{f}\|^p_{W^{\alpha-1,p}(\Omega_{\widetilde{\mathbf{m}},N})}\right)\\
\le&\left(C_{28}3^{\frac{D}{q}}N^k\eta\right)^p 3^D2^D\|\bar{f}\|^p_{W^{\alpha-1,p}(\cup_{\widetilde{\mathbf{m}}\in\mathcal{S}_N}\Omega_{\widetilde{\mathbf{m}},N})}\\
\le&C_{29}N^{kp}\eta^p\|f\|_{W^{\alpha-1,p}((0,1)^D)},
\end{aligned}$$

where C_{29} is a constant depending on D,α,p. In the above, we used (77) in the last inequality. Lemma C.10 is proved for $s=0$ and $s=1$. For any $0<s<1$ and $1\le p\le\infty$, by Lemma G.5, we have

$$\begin{aligned}
\left\|\sum_{\mathbf{m}\in\mathcal{S}_N}\phi_{\mathbf{m}}Q^\alpha_{\mathbf{m}/N}f-\sum_{\mathbf{m}\in\mathcal{S}_N}\sum_{|\mathbf{v}|\le\alpha-1}c_{\mathbf{m},\mathbf{v}}\widetilde{g}_{\mathbf{m},\mathbf{v}}\right\|^p_{W^{k,p}((0,1)^D)}&\le C_{30}N^{kp}\eta^p\|f\|_{W^{\alpha-1,p}((0,1)^D)}\\
&\le C_{30}N^{sp}\eta^p\|f\|_{W^{\alpha,p}((0,1)^D)}
\end{aligned}$$

for some constant C_{30} depending on D, α, s, p. The proof is finished. □

G.7. Lemma G.8 and its proof

Lemma G.8. *Let $\{f_i\}_{i=1}^n$ be a set of CNNs with architecture $\mathcal{F}^{\mathrm{CNN}}(L_0, J_0, K_0, \kappa_0, \kappa_0)$. Then there for any integer $1 \le w \le n$, there exists a CNN architecture $\mathcal{F}^{\mathrm{CNN}}(L_w, J_w, K_w, \kappa_w, \kappa_w)$ that gives rise to a CNN g_w such that*

$$g_w(\mathbf{x}) = \sum_{i=1}^{w} f_i(\mathbf{x}).$$

Such an architecture has

$$L = O(L_0),\ J = wJ_0,\ K = K_0,\ \kappa = \kappa_0.$$

Furthermore, the fully connected layer of f has nonzero elements only in the first row.

Proof of Lemma G.8. The idea of the proof is similar to Liu et al. (2021, proof of Lemma 14). Following the proof of Liu et al. (2021, Lemma 14), we can show that there exist a set of filters $\mathcal{W}$ and biases $\mathcal{B}$ such that

$$\mathrm{Conv}_{\mathcal{W},\mathcal{B}}(\mathbf{x}) = \begin{bmatrix} (f_1(\mathbf{x}))_+ & (f_1(\mathbf{x}))_- & (f_2(\mathbf{x}))_+ & (f_2(\mathbf{x}))_- & \cdots & (f_w(\mathbf{x}))_+ & (f_w(\mathbf{x}))_- \\ \star & \star & \star & \star & \cdots & \star & \star \end{bmatrix},$$

where $\mathrm{Conv}_{\mathcal{W},\mathcal{B}}$ has depth bounded by L_0, number of channels bounded by wj_0 and all weight parameters bounded by κ_0. We write g_w as

$$g_w = W_1 \cdot \mathrm{Conv}_{\mathcal{W},\mathcal{B}},$$

where W_1 is given as

$$W_1 = \begin{bmatrix} 1 & -1 & 1 & -1 & \cdots & 1 & -1 \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \end{bmatrix}.$$

The proof is finished. □

G.8. Proof of Lemma C.11

Proof of Lemma C.11. For any given $\widetilde{J}$, let c be the smallest integer such that $\widetilde{J} \le cJ_0$. Then we set $J = cJ_0$ and $n = \lceil n_0/c \rceil$. By Lemma G.8, there exists a CNN architecture $\mathcal{F}^{\mathrm{CNN}}(L, J, K, \kappa, \kappa)$ with

$$L = O(L_0),\ J = cJ_0,\ K = K_0,\ \kappa = \kappa_0.$$

Such an architecture gives rise to CNNs $\{g_j\}_{j=1}^{\lceil n_0/c \rceil}$ such that

$$g_j = \sum_{i=c(j-1)+1}^{\min\{cj,n\}} f_i.$$

The lemma is proved. □

H. Proof of lemmas in Appendix F

H.1. Proof of Lemma F.1

Proof of Lemma F.1. Following the construction in **Step 1** of the proof of Theorem 4.5, for $\widetilde{r} = r/2 < \tau/8$, there exists a collection of points atlas of $\mathcal{M}$ denoted by $\{\widetilde{U}_i, \widetilde{\varphi}_i\}_{i=1}^{\widetilde{C}_{\mathcal{M}}}$, where $\widetilde{U}_i = B_{\widetilde{r}}(\widetilde{\mathbf{c}}_i)$ for some $\widetilde{\mathbf{c}}_i \in \mathcal{M}$, and $\widetilde{\varphi}_i$ is defined according to (37). By Conway & Sloane (1988, Chapter 2 Equation (1)), the number of charts is bounded by

$$\widetilde{C}_{\mathcal{M}} \le \left\lceil \frac{\mathrm{SA}(\mathcal{M})}{\widetilde{r}^d} T_d \right\rceil = \left\lceil \frac{\mathrm{SA}(\mathcal{M})}{r/2^d} T_d \right\rceil.$$

The following lemma shows that for any locally finite cover of a smooth manifold, a C^∞ partition of unity always exists:

Lemma H.1 (Chapter 2 Theorem 15 of (Spivak, 1973)). *Let $\{U_\alpha\}_{\alpha\in\mathcal{A}}$ be a locally finite cover of a smooth manifold $\mathcal{M}$. There is a C^∞ partition of unity $\{\rho_\alpha\}_{\alpha=1}^\infty$ such that* $\mathrm{supp}(\rho_\alpha)\subset U_\alpha$.

Let $\{\rho_i\}_{i=1}^{\widetilde{C}_\mathcal{M}}$ be the partition of unity in Lemma H.1 with respect to $\{\widetilde{U}_i\}_{i=1}^{C_\mathcal{M}}$.

We set $C_\mathcal{M}=\widetilde{C}_\mathcal{M}$ and define $U_i=B_r(\widetilde{\mathbf{c}}_i)$ and φ_i according to (37). Since $\widetilde{r}<r$, $\widetilde{U}_i\subset U_i$, we have $\widetilde{U}_i\subset U_i$ and

$$\mathcal{M}\subseteq\bigcup_{i=1}^{\widetilde{C}_\mathcal{M}}\widetilde{U}_i\subseteq\bigcup_{i=1}^{C_\mathcal{M}}U_i.$$

Therefore $\{U_i\}_{i=1}^{C_\mathcal{M}}$ is an open cover of $\mathcal{M}$ and $\{U_i,\varphi_i\}_{i=1}^{C_\mathcal{M}}$ is an atlas of $\mathcal{M}$. Since $\mathrm{supp}(\rho_i)\subseteq\widetilde{U_i}$, we have $\mathrm{supp}(\rho_i)\subset U_i$ and

$$\inf_{\mathbf{x}\in\mathrm{supp}(\rho_i),\,\widetilde{\mathbf{x}}\in\partial U_i}\|\mathbf{x}-\widetilde{\mathbf{x}}\|_2\geq\inf_{\mathbf{x}\in\widetilde{U}_i,\,\widetilde{\mathbf{x}}\in\partial U_i}\|\mathbf{x}-\widetilde{\mathbf{x}}\|_2=r/2.$$

The lemma is proved. □

H.2. Proof of Lemma F.3

Proof of Lemma F.3. We deduce

$$\begin{aligned}
|A_2|_{W^{k,\infty}(\varphi_i(U_i))}&=\left|\left(\sum_{\mathbf{m},\mathbf{v}}c_{i,\mathbf{m},\mathbf{v}}\widetilde{g}_{\mathbf{m},\mathbf{v}}(\mathbf{z})\right)\times\left(\widetilde{\mathbb{1}}_i\circ\varphi_i^{-1}(\mathbf{z})-\mathbb{1}_i\circ\varphi_i^{-1}(\mathbf{z})\right)\right|_{W^{k,\infty}(\varphi_i(U_i))}\\
&=\left|\widetilde{f}_i\circ\varphi_i^{-1}(\mathbf{z})\times\left(\widetilde{\mathbb{1}}_i\circ\varphi_i^{-1}(\mathbf{z})-\mathbb{1}_i\circ\varphi_i^{-1}(\mathbf{z})\right)\right|_{W^{k,\infty}(\varphi_i(U_i))}\\
&\leq\left|\widetilde{f}_i\circ\varphi_i^{-1}(\mathbf{z})\times\left(\widetilde{\mathbb{1}}_i\circ\varphi_i^{-1}(\mathbf{z})-\mathbb{1}_i\circ\varphi_i^{-1}(\mathbf{z})\right)\right|_{W^{k,\infty}(\Omega_{i,2})}\\
&\quad+\left|\widetilde{f}_i\circ\varphi_i^{-1}(\mathbf{z})\times\left(\widetilde{\mathbb{1}}_i\circ\varphi_i^{-1}(\mathbf{z})-\mathbb{1}_i\circ\varphi_i^{-1}(\mathbf{z})\right)\right|_{W^{k,\infty}(\varphi_i(U_i)\setminus\Omega_{i,2})}\\
&=\left|\widetilde{f}_i\circ\varphi_i^{-1}(\mathbf{z})\times\left(\widetilde{\mathbb{1}}_i\circ\varphi_i^{-1}(\mathbf{z})-\mathbb{1}_i\circ\varphi_i^{-1}(\mathbf{z})\right)\right|_{W^{k,\infty}(\Omega_{i,2})}
\end{aligned}$$

for $k=0,1$, where the last equality holds since

$$\widetilde{\mathbb{1}}_i\circ\varphi_i^{-1}(\mathbf{z})=\mathbb{1}_i\circ\varphi_i^{-1}(\mathbf{z})=1$$

on $\varphi_i(U_i)\setminus\Omega_{i,2}$.

According to (53), $\widetilde{f}_i\circ\varphi_i^{-1}(\mathbf{z})=\widehat{f}_i\circ\varphi_i^{-1}(\mathbf{z})=0$ for $\mathbf{z}\in\varphi_i(\partial U_i)$. For any $\mathbf{z}\in\Omega_{i,2}$, let

$$\mathbf{z}^*=\operatorname*{argmin}_{\widetilde{\mathbf{z}}\in\varphi_i(\partial U_i)}\|\mathbf{z}-\widetilde{\mathbf{z}}\|_2.$$

According to (52), we have $\|\mathbf{z}-\mathbf{z}^*\|_2\leq\Delta/(c_2r)$.

By Lemma C.10 with some small $\eta>0$ and for $s=k=0,1$, we have

$$\|\widetilde{f}_i\circ\varphi_i^{-1}-\widehat{f}_i\|_{W^{k,\infty}([0,1]^d)}\leq C_{31}N^k\eta,\tag{81}$$

where C_{31} is a constant depending on d,α,R. Since $\|\widehat{f}_i\|_{W^{1,\infty}(\Omega_{i,2})}=0$, we have $\max_j\left|\frac{\partial\widetilde{f}_i}{\partial z_j}\right|\leq C_{31}N\eta$ for any $\mathbf{z}\in\Omega_{i,2}$. Therefore

$$|\widetilde{f}_i\circ\varphi_i^{-1}(\mathbf{z})|\leq\widetilde{f}_i\circ\varphi_i^{-1}(\mathbf{z}^*)+C_{31}N\eta\|\mathbf{z}-\mathbf{z}^*\|_2\leq\frac{C_{31}}{c_2r}N\eta\Delta\tag{82}$$

for any $\mathbf{z}\in\Omega_{i,2}$.

Using $|\widetilde{\mathbb{1}}_i \circ \varphi_i^{-1}(\mathbf{z}) - \mathbb{1}_i \circ \varphi_i^{-1}(\mathbf{z})| \leq 1$, we bound A_2 as

$$\begin{aligned}|A_2|_{W^{0,\infty}(\varphi_i(U_i))} =& \left|\widetilde{f}_i \circ \varphi_i^{-1}(\mathbf{z}) \times \left(\widetilde{\mathbb{1}}_i \circ \varphi_i^{-1}(\mathbf{z}) - \mathbb{1}_i \circ \varphi_i^{-1}(\mathbf{z})\right)\right|_{W^{0,\infty}(\Omega_{i,2})} \\ \leq& \left|\widetilde{f}_i \circ \varphi_i^{-1}(\mathbf{z})\right|_{W^{0,\infty}(\Omega_{i,2})} \times \left|\widetilde{\mathbb{1}}_i \circ \varphi_i^{-1}(\mathbf{z}) - \mathbb{1}_i \circ \varphi_i^{-1}(\mathbf{z})\right|_{W^{0,\infty}(\Omega_{i,2})} \\ \leq& \frac{C_{11}}{c_2 r} N\eta\Delta \end{aligned} \tag{83}$$

for $k = 0$ and

$$\begin{aligned}|A_2|_{W^{1,\infty}(\varphi_i(U_i))} =& \left|\widetilde{f}_i \circ \varphi_i^{-1}(\mathbf{z}) \times \left(\widetilde{\mathbb{1}}_i \circ \varphi_i^{-1}(\mathbf{z}) - \mathbb{1}_i \circ \varphi_i^{-1}(\mathbf{z})\right)\right|_{W^{1,\infty}(\Omega_{i,2})} \\ \leq& \left|\widetilde{f}_i \circ \varphi_i^{-1}(\mathbf{z})\right|_{W^{0,\infty}(\Omega_{i,2})} \times \left|\widetilde{\mathbb{1}}_i \circ \varphi_i^{-1}(\mathbf{z}) - \mathbb{1}_i \circ \varphi_i^{-1}(\mathbf{z})\right|_{W^{1,\infty}(\Omega_{i,2})} \\ &+ \left|\widetilde{f}_i \circ \varphi_i^{-1}(\mathbf{z})\right|_{W^{1,\infty}(\Omega_{i,2})} \times \left|\widetilde{\mathbb{1}}_i \circ \varphi_i^{-1}(\mathbf{z}) - \mathbb{1}_i \circ \varphi_i^{-1}(\mathbf{z})\right|_{W^{0,\infty}(\Omega_{i,2})} \\ \leq& \frac{C_{31}C_8}{c_2 r} N\eta\Delta/\Delta + C_{11}N\eta \\ =& C_{32}N\eta \end{aligned} \tag{84}$$

for $k = 1$, where C_{12} is a constant depending on α, R, τ. In the first inequality of (84), we used (82), the inequality

$$\left|\widetilde{\mathbb{1}}_i \circ \varphi_i^{-1}(\mathbf{z}) - \mathbb{1}_i \circ \varphi_i^{-1}(\mathbf{z})\right|_{W^{1,\infty}(\Omega_{i,2})} = \left|\widetilde{\mathbb{1}}_i \circ \varphi_i^{-1}(\mathbf{z})\right|_{W^{1,\infty}(\Omega_{i,2})} \leq C_8/\Delta$$

by (40) and the fact $\mathbb{1}_i \circ \varphi_i^{-1}(\mathbf{z}) = 1$ for $\mathbf{z} \in \Omega_{i,2}$, and the inequality

$$\left|\widetilde{f}_i \circ \varphi_i^{-1}(\mathbf{z})\right|_{W^{1,\infty}(\Omega_{i,2})} = \left|\widetilde{f}_i \circ \varphi_i^{-1}(\mathbf{z}) - 0\right|_{W^{1,\infty}(\Omega_{i,2})} = \|\widetilde{f}_i \circ \varphi_i^{-1} - f_i \circ \varphi_i^{-1}\|_{W^{1,\infty}(\Omega_{i,2})} \leq C_{31}N\eta$$

by (81).

Combining (83) and (84) gives rise to

$$\|A_2\|_{W^{1,\infty}(\varphi_i(U_i))} \leq C_{32}N\eta\Delta^{1-k} \tag{85}$$

for $k = 0, 1$. □