Journal of Machine Learning Research 26 (2025) 1-56 Submitted 7/24; Revised 8/25; Published 9/25

Deep Neural Networks are Adaptive to Function Regularity and Data Distribution in Approximation and Estimation

Hao Liu[1] HAOLIU@HKBU.EDU.HK
Jiahui Cheng[2] JCHENG328@GATECH.EDU
Wenjing Liao[2*] WLIAO60@GATECH.EDU
[*] *Corresponding author*
[1] *Department of Mathematics, Hong Kong Baptist University, Hong Kong, China*
[2] *School of Mathematics, Georgia Institute of Technology, USA*

Editor: Lorenzo Rosasco

Abstract

Deep learning has exhibited remarkable results across diverse areas. To understand its success, substantial research has been directed towards its theoretical foundations. Nevertheless, the majority of these studies examine how well deep neural networks can model functions with uniform regularities. In this paper, we explore a different angle: how deep neural networks can adapt to varying degrees of smoothness in functions and nonuniform data distributions across different locations and scales. More precisely, we focus on a broad class of functions defined by nonlinear tree-based approximation methods. This class encompasses a range of function types, such as functions with uniform regularities and discontinuous functions. We develop nonparametric approximation and estimation theories for this class using deep ReLU networks. Our results show that deep neural networks are adaptive to the nonuniform smoothness of functions and nonuniform data distributions at different locations and scales. We apply our results to several function classes, and derive the corresponding approximation and generalization errors. The validity of our results is demonstrated through numerical experiments.

Keywords: deep neural networks, adaptive approximation, nonparametric regression

1. Introduction

Deep learning has achieved significant success in practical applications with high-dimensional data, such as computer vision (Krizhevsky et al., 2012), natural language processing (Graves et al., 2013; Young et al., 2018), health care (Miotto et al., 2018; Jiang et al., 2017) and bioinformatics (Alipanahi et al., 2015; Zhou and Troyanskaya, 2015). The success of deep learning demonstrates the power of neural networks in representing and learning complex operations on high-dimensional data.

In the past decades, the representation power of neural networks has been extensively studied. Early works in literature focused on shallow (two-layer) networks with continuous sigmoidal activations (a function $\sigma(x)$ is sigmoidal, if $\sigma(x) \to 0$ as $x \to -\infty$, and $\sigma(x) \to 1$ as $x \to \infty$) for a universal approximation of continuous functions in a unit hypercube (Irie and Miyake, 1988; Funahashi, 1989; Cybenko, 1989; Hornik, 1991; Chui and Li, 1992; Leshno et al., 1993; Barron, 1993; Mhaskar, 1996). The universal approximation theory of feedforward neural networks with a ReLU activation $\sigma(x) = \max(0, x)$ was studied in

©2025 Hao Liu, Jiahui Cheng, Wenjing Liao.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v26/24-1148.html.

Lu et al. (2017); Hanin (2017); Daubechies et al. (2022); Yarotsky (2017); Schmidt-Hieber (2017); Suzuki (2018). In particular, the approximation theories of ReLU networks have been established for the Sobolev $W^{k,\infty}$ (Yarotsky, 2017), Hölder (Schmidt-Hieber, 2017) and Besov (Suzuki, 2018) functions. Recently, it has been shown that shallow ReLU and ReLU^k networks can approximate Hölder functions well (Mao and Zhou, 2023; Yang and Zhou, 2024b; Siegel and Xu, 2024), and can achieve a minimax learning rate (Yang and Zhou, 2024b,a). These works guarantee that, the Sobolev, Hölder, or Besov function class can be well approximated by a ReLU network function class with a properly chosen network architecture. The approximation error in these works was given in terms of certain function norm. Furthermore, the works in Gühring et al. (2020); Hon and Yang (2021); Liu et al. (2022a) proved the approximation error in terms of the Sobelev norm, which guaranteed the approximation error for the function and its derivatives simultaneously. In terms of the network architecture, feedforward neural networks were considered in the vast majority of approximation theories. Convolutional neural networks were considered in Zhou (2020); Petersen and Voigtlaender (2020), and convolutional residual networks were considered in Oono and Suzuki (2019); Liu et al. (2021).

It has been widely believed that deep neural networks are adaptive to complex data structures. Recent progresses have been made towards theoretical justifications that deep neural networks are adaptive to low-dimensional structures in data. Specifically, function approximation theories have been established for Hölder and Sobolev functions supported on low-dimensional manifolds (Chen et al., 2019a; Schmidt-Hieber, 2019; Cloninger and Klock, 2020; Nakada and Imaizumi, 2020; Liu et al., 2021). The network size in these works crucially depends on the intrinsic dimension of data, instead of the ambient dimension. In the task of regression and classification, the sample complexity of neural networks (Chen et al., 2019b; Nakada and Imaizumi, 2020; Liu et al., 2021) depends on the intrinsic dimension of data, while the ambient dimension does not affect the rate of convergence.

This paper answers another interesting question about the adaptivity of deep neural networks: *How does deep neural networks adapt to the function regularity and data distribution at different locations and scales?* The answer of this question is beyond the scope of existing function approximation and estimation theory of neural networks. The Sobolev $W^{k,\infty}$ and Hölder functions are uniformly regular within the whole domain. The analytical technique to build the approximation theory of these functions relies on accurate local approximations everywhere within the domain. In real applications, functions of interests often exhibit different regularity at different locations and scales. Empirical experiments have demonstrated that deep neural networks are capable of extracting interesting information at various locations and scales (Chung et al., 2016; Haber et al., 2018). However, there are limited works on theoretical justifications of the adaptivity of neural networks.

In this paper, we re-visit the nonlinear approximation theory (DeVore, 1998) in the classical multi-resolution analysis (Mallat, 1999; Daubechies, 1992). Nonlinear approximations allow one to approximate functions beyond linear spaces. The smoothness of the function can be defined according to the rate of approximation error versus the complexity of the elements. In many settings, such characterization of smoothness is significantly weaker than the uniform regularity condition in the Sobolev or Hölder class (DeVore, 1998).

We focus on the tree-based nonlinear approximations with piecewise polynomials (Binev et al., 2007, 2005; Cohen et al., 2001). Specifically, the domain of functions is partitioned to

multiscale dyadic cubes associated with a master tree. If we build piecewise polynomials on these multiscale dyadic cubes, we naturally obtain multiscale piecewise polynomial approximations. A refinement quantity is defined at every node to quantify how much the error decreases when the node is refined to its children. A thresholding of the master tree based on this refinement quantity gives rise to a truncated tree, as well as an adaptive partition of the domain. Thanks to the thresholding technique, we can define a function class whose regularity is characterized by how fast the size of the truncated tree grows with respect to the level of the threshold. This is a large function class containing the Hölder and piecewise Hölder functions as special cases.

Our main contributions can be summarized as:

1. We establish the approximation theory of deep ReLU networks for a large class of functions whose regularity is defined according to the nonlinear tree-based approximation theory. This function class allows the regularity of the function to vary at different locations and scales.

2. We provide several examples of functions in this class which exhibit different information at different locations and scales. These examples are beyond the characterization of function classes with uniform regularity, such as the Hölder class. In particular, our results apply to piecewise smooth functions with a weaker assumption on the discontinuity hypersurface compared to existing works, and also apply to functions which are irregular on a set of measure 0.

3. A nonparametric estimation theory for this large function class is established with deep ReLU networks, which is validated by numerical experiments.

4. Our results demonstrate that, when deep neural networks are representing functions, it does not require a uniform regularity everywhere on the domain. Deep neural networks are automatically adaptive to the regularity of functions at different locations and scales.

In literature, adaptive function approximation and estimation has been studied for classical methods (DeVore, 1998), including free-knot spline (Jupp, 1978), adaptive smoothing splines (Wahba, 1995; Pintore et al., 2006; Liu and Guo, 2010; Wang et al., 2013), nonlinear wavelet (Cohen et al., 2001; Donoho and Johnstone, 1998; Donoho et al., 1995), and adaptive piecewise polynomial approximation (Binev et al., 2007, 2005). Based on traditional methods for estimating functions with uniform regularity, these methods allow the smoothing parameter, the kernel band width or knots placement to vary spatially to adapt to the varying regularity. Kernel methods with variable bandwidth were studied in Muller and Stadtmuller (1987) and local polynomial estimators were studied in Fan and Gijbels (1996). Based on traditional smoothing splines with a global smoothing parameter, Wahba (1995) suggested to replace the smoothing parameter by a roughness penalty function. This idea was then studied in Pintore et al. (2006); Liu and Guo (2010) by using piecewise constant roughness penalty, and in Wang et al. (2013) with a more general roughness penalty. A locally penalized spline estimator was proposed and studied in Ruppert and Carroll (2000), in which a penalty function was applied to spline coefficients and was knot-dependent. Adaptive wavelet shrinkage was studied in Donoho and Johnstone (1994, 1995, 1998), in

which the authors used selective wavelet reconstruction, adaptive thresholding and nonlinear wavelet shrinkage to achieve adaptation to spatially varying regularity, and proved the minimax optimality. A Bayesian mixture of splines method was proposed in Wood et al. (2002), in which each component spline had a locally defined smoothing parameter. Other methods include regression splines (Friedman, 1991; Smith and Kohn, 1996; Denison et al., 1998), hybrid smoothing splines and regression splines (Luo and Wahba, 1997), and the trend filtering method (Tibshirani, 2014). The minimax theory for adaptive nonparametric estimator was established in Cai (2012). Most of the works mentioned above focused on one-dimensional problems. For high dimensional problems, an additive model was considered in Ruppert and Carroll (2000). Recently, the Bayesian additive regression trees were studied in Jeong and Rockova (2023) for estimating a class of sparse piecewise heterogeneous anisotropic Hölder continuous functions in high dimension.

Classical methods mentioned above adapt to varying regularity of the target functions through a careful selection of some adaptive parameter, such as the location of knots, kernel bandwidth, roughness penalty and adaptive tree structure. These methods require the knowledge about or an estimation of how the regularity of the target function changes. Compared to classical methods, deep learning solves the regression problem by minimizing the empirical risk in (20), so the same optimization problem can be applied to various functions without explicitly figuring out where the regularity of the underlying function changes. Such kind of automatic adaptivity is crucial for real-world applications.

The connection between neural networks and adaptive spline approximation has been studied in Daubechies et al. (2022); DeVore et al. (2021); Liu et al. (2022b); Petersen and Voigtlaender (2018); Imaizumi and Fukumizu (2019). In particular, an adaptive network enhancement method was proposed in Liu et al. (2022b) for the best least-squares approximation using two-layer neural networks. The adaptivity of neural networks to data distributions was considered in Zhang et al. (2023), where the concept of an effective Minkowski dimension was introduced and applied to anisotropic Gaussian distributions. The approximation error and generalization error for learning piecewise Hölder functions in $\mathbb{R}^d$ are developed in Petersen and Voigtlaender (2018) and Imaizumi and Fukumizu (2019), respectively. In the settings of Petersen and Voigtlaender (2018) and Imaizumi and Fukumizu (2019), each discontinuity boundary is parametrized by a $(d-1)$-dimensional Hölder function, which is called a horizon function. In this paper, we consider a function class based on nonlinear tree-based approximation, and provide approximation and generalization theories of deep neural networks for this function class, as well as several examples related with practical applications, which are not implied by existing works. For piecewise Hölder functions, our setting only assumes the boundary of each piece has Minkowski dimension $d-1$, which is more general than that considered in Petersen and Voigtlaender (2018); Imaizumi and Fukumizu (2019), see Section 4.3.3 and 4.5.3 for more detailed discussions.

Our paper is organized as follows: In Section 2, we introduce notations and concepts used in this paper. Tree based adaptive approximation and some examples are presented in Section 3. We present our main results, the adaptive approximation and generalization theories of deep neural networks in Section 4, and the proofs are deferred to Section 6. Our theories is validated by numerical experiments in Section 5. This paper is concluded in Section 7.

2. Notation and Preliminaries

In this section, we introduce our notation, some preliminary definitions and ReLU networks.

2.1 Notation

We use normal lower case letters to denote scalars, and bold lower case letters to denote vectors. For a vector $\mathbf{x} \in \mathbb{R}^d$, we use x_i to denote the i-th entry of $\mathbf{x}$. The standard 2-norm of $\mathbf{x}$ is $\|\mathbf{x}\|_2 = (\sum_{i=1}^d x_i^2)^{\frac{1}{2}}$. For a scalar $a > 0$, $\lfloor a \rfloor$ denotes the largest integer that is no larger than a, $\lceil a \rceil$ denotes the smallest integer that is no smaller than a. Let I be a set. We use χ_I to denote the indicator function on I such that $\chi_I(\mathbf{x}) = 1$ if $\mathbf{x} \in I$ and $\chi_I(\mathbf{x}) = 0$ if $\mathbf{x} \notin I$. The notation $\#I$ denotes the cardinality of I.

Denote the domain $X = [0,1]^d$. For a function $f : X \to \mathbb{R}$ and a multi-index $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_d]^\top$, $\partial^{\boldsymbol{\alpha}} f$ denotes $\partial^{|\boldsymbol{\alpha}|} f / \partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}$, where $|\boldsymbol{\alpha}| = \sum_{k=1}^d \alpha_k$. We denote $\boldsymbol{x}^{\boldsymbol{\alpha}} = x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_d^{\alpha_d}$. Let ρ be a measure on X. The L^2 norm of f with respect to the measure ρ is $\|f\|_{L^2(\rho)}^2 = \int_X |f(\mathbf{x})|^2 d\rho$. We say $f \in L^2(\rho)$ if $\|f\|_{L^2(\rho)}^2 < \infty$. We denote $\|f\|_{L^2(\rho(\Omega))}^2 = \int_\Omega |f(\mathbf{x})|^2 d\rho$ for any $\Omega \subset X$.

The notation of $f \lesssim g$ means that there exists a constant C independent of any variable upon which f and g depend, such that $f \leq Cg$; similarly for $\gtrsim$. $f \asymp g$ means that $f \lesssim g$ and $f \gtrsim g$.

2.2 Preliminaries

We define Hölder functions as follows:

Definition 1 (Hölder functions) *A function $f : X \to \mathbb{R}$ belongs to the Hölder space $\mathcal{H}^r(X)$ with a Hölder index $r > 0$, if*

$$\|f\|_{\mathcal{H}^r(X)} = \max_{|\boldsymbol{\alpha}| < \lceil r-1 \rceil} \sup_{\mathbf{x} \in X} |\partial^{\boldsymbol{\alpha}} f(\mathbf{x})| + \max_{|\boldsymbol{\alpha}| = \lceil r-1 \rceil} \sup_{\mathbf{x} \neq \mathbf{z} \in X} \frac{|\partial^{\boldsymbol{\alpha}} f(\mathbf{x}) - \partial^{\boldsymbol{\alpha}} f(\mathbf{z})|}{\|\mathbf{x} - \mathbf{z}\|_2^{r - \lceil r-1 \rceil}} < \infty. \tag{1}$$

For any r-Hölder function, it has partial derivatives up to $\lceil r-1 \rceil$-th order, and its $\lceil r-1 \rceil$-th order partial derivative is $(r - \lceil r-1 \rceil)$-Hölder continuous. Hölder functions and piecewise Hölder functions can be characterized using tree-based nonlinear approximation, see Section 3.4 for details.

Given a set in $\mathbb{R}^d$, its intrinsic dimension can be described by the Minkowksi dimension:

Definition 2 (Minkowski dimension) *Let $\Omega \subset [0,1]^d$. For any $\varepsilon > 0$, $\mathcal{N}(\varepsilon, \Omega, \|\cdot\|_\infty)$ denotes the fewest number of ε-balls that cover Ω in terms of $\|\cdot\|_\infty$. The (upper) Minkowski dimension of Ω is defined as*

$$d_M(\Omega) := \limsup_{\varepsilon \to 0^+} \frac{\log \mathcal{N}(\varepsilon, \Omega, \|\cdot\|_\infty)}{\log(1/\varepsilon)}.$$

We further define the Minkowski dimension constant of Ω as

$$c_M(\Omega) = \sup_{\varepsilon > 0} \mathcal{N}(\varepsilon, \Omega, \|\cdot\|_\infty) \varepsilon^{d_M(\Omega)}.$$

Such a constant is an upper bound on the rate of how $\mathcal{N}(\varepsilon, \Omega, \|\cdot\|_\infty)$ scales with $\varepsilon^{-d_M(\Omega)}$.

Consider covering Ω by Euclidean balls, Minkowski dimension measures how fast the number of balls increases as the ball radius goes to 0. A set Ω has Minkowksi dimension $d_M(\Omega)$ if $\mathcal{N}(\varepsilon, \Omega, \|\cdot\|_\infty) \approx C\varepsilon^{d_M(\Omega)}$ for some constant C. One can show that a d'-dimensional manifold has Minkowksi dimension no larger than d' (Nakada and Imaizumi, 2019, Lemma 9).

ReLU network. In this paper, we consider the standard feedforward neural networks defined in the form of

$$f_{\text{NN}}(\mathbf{x}) = W_L \cdot \text{ReLU}\big(W_{L-1} \cdots \text{ReLU}(W_1\mathbf{x} + \mathbf{b}_1) + \cdots + \mathbf{b}_{L-1}\big) + \mathbf{b}_L, \tag{2}$$

where $\mathbf{x}$ is a vector in $\mathbb{R}^d$, W_l's are weight matrices with $W_l \in \mathbb{R}^{c_l \times c_{l-1}}$ and c_l being the output dimension of the l-th layer, $\mathbf{b}_l \in \mathbb{R}^l$ are biases, and $\text{ReLU}(a) = \max\{a, 0\}$ denotes the rectified linear unit (ReLU), which is applied elementwisely. We define the network class as

$$\begin{aligned}\bar{\mathcal{F}}_{\text{NN}}(L, w, K, \kappa) = \big\{ & f_{\text{NN}} : \mathbb{R}^d \to \mathbb{R} | f_{\text{NN}}(\mathbf{x}) \text{ is in the form of (2) with } L \text{ layers,} \\ & \text{width bounded by } w,\ \|W_l\|_{\infty,\infty} \le \kappa, \|\mathbf{b}_l\|_\infty \le \kappa, \text{ and } \textstyle\sum_{l=1}^L \|W_l\|_0 + \|\mathbf{b}_l\|_0 \le K \big\},\end{aligned}$$

where $\|W\|_{\infty,\infty} = \max_{i,j} |W_{i,j}|$, $\|\mathbf{b}\|_\infty = \max_i |b_i|$ for any matrix W and vector $\mathbf{b}$, and $\|\cdot\|_0$ denotes the number of nonzero elements of its argument. The class $\bar{\mathcal{F}}_{\text{NN}}(L, w, K, \kappa)$ is the set of networks with L layers, width of each layer bounded by w, nonzero weight parameters no more than K and all weight parameters bounded by κ.

Our theory will focus on networks with bounded output. For any $M > 0$, define the projection operator for any scalar-valued function f as

$$\Pi_M(f)(\mathbf{x}) = \begin{cases} f(\mathbf{x}), & \text{if } |f(\mathbf{x})| \le M, \\ M, & \text{if } f(\mathbf{x}) > M, \\ -M, & \text{if } f(\mathbf{x}) < -M. \end{cases}$$

We define the function class

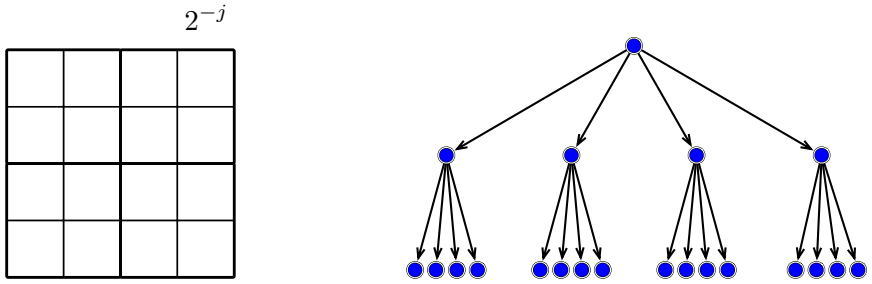
$$\mathcal{F}_{\text{NN}}(L, w, K, \kappa, M) = \left\{ f : \mathbb{R}^d \to \mathbb{R} | f = \Pi_M(g),\ g \in \bar{\mathcal{F}}_{\text{NN}}(L, w, K, \kappa) \right\}. \tag{3}$$

Note the projection Π_M can be realized by a two layer fixed ReLU network:

$$\Pi_M(f)(\mathbf{x}) = \text{ReLU}(2M - \text{ReLU}(M - f(\mathbf{x}))) - M.$$

3. Adaptive Approximation

This section is an introduction to tree-based nonlinear approximation and a function class whose regularity is defined through nonlinear approximation theory. In the classical tree-based nonlinear approximations (Binev et al., 2007, 2005; Cohen et al., 2001), piecewise polynomials are used to approximate the target function on an adaptive partition. For simplicity, we focus on the case that the function domain is $X = [0, 1]^d$. Let ρ be a probability measure on X and $f \in L^2(\rho)$. The multiscale dyadic partitions of X give rise to a tree structure. It is natural to consider nonlinear approximations based on this tree structure.



(a) Dyadic partition (b) Tree

Figure 1: The dyadic partition of the 2D unit cube $[0,1]^2$ and the associated tree.

We re-visit related definitions and tree-based nonlinear approximations in Subsection 3.1 and 3.2, respectively. We define the $\mathcal{A}^s_{\theta,\rho}$ function class, whose regularity is characterized through tree-based nonlinear approximations, in Subsection 3.3. Several examples of this function class are given in Subsection 3.4.

3.1 Related Definitions

We define dyadic cubes and related concepts that are used to define the tree-based nonlinear approximation.

Dyadic cubes. Let $\mathcal{C}_j = \{C_{j,k}\}_{k=1}^{2^{jd}}$ be the collection of dyadic subcubes of X of sidelength 2^{-j}. Here j denotes the scale of $C_{j,k}$ with a small j and k denotes the location. A small j represents the coarse scale, and a large j represents the fine scale. These dyadic cubes are naturally associated with a tree $\mathcal{T}$. Each node of this tree corresponds to a cube $C_{j,k}$. The dyadic partition of the 2D cube $[0,1]^2$ and its associated tree are illustrated in Figure 1. Every node $C_{j,k}$ at scale j has 2^d children at scale $j+1$. We denote the set of children of $C_{j,k}$ by $\mathcal{C}(C_{j,k})$.

Children and parents. When the node $C_{j,k}$ is a child of the node $C_{j-1,k'}$, we call $C_{j-1,k'}$ the parent of $C_{j,k}$, denoted by $\mathcal{P}(C_{j,k})$.

Subtree. A proper subtree $\mathcal{T}_0$ of $\mathcal{T}$ is a collection of nodes such that: (1) the root node X is in $\mathcal{T}_0$; (2) if $C_{j,k} \neq X$ is in $\mathcal{T}_0$, then its parent is also in $\mathcal{T}_0$. Given a proper subtree $\mathcal{T}_0$ of $\mathcal{T}$, the outer leaves of $\mathcal{T}_0$ contain all $C_{j,k} \in \mathcal{T}$ such that $C_{j,k} \notin \mathcal{T}_0$ but the parent of $C_{j,k}$ belongs to $\mathcal{T}_0$: $C_{j,k} \notin \mathcal{T}_0$ but $\mathcal{P}(C_{j,k}) \in \mathcal{T}_0$. The collection of the outer leaves of $\mathcal{T}_0$, denoted by $\Lambda = \Lambda(\mathcal{T}_0)$, forms a partition of X.

3.2 Nonlinear Approximation by Tree Structures

The tree-based nonlinear approximation generates an adaptive partition with a thresholding technique. In certain cases, this thresholding technique boils down to wavelet thresholding. Specifically, one defines a refinement quantity on each node of the tree, and then thresholds the tree to the smallest proper subtree containing all the nodes whose refinement quantity is above certain value. Adaptive partitions are given by the outer leaves of this proper subtree after thresholding.

We consider piecewise polynomial approximations of f by polynomials of degree θ, where θ is a nonnegative integer. Let $\mathcal{P}_\theta$ be the space of d-variable polynomials of degree no more

than θ. For any cube $C_{j,k}$, the best polynomial approximating f on $C_{j,k}$ is

$$p_{j,k} = p_{j,k}(f) = \operatorname*{argmin}_{p \in \mathcal{P}_\theta} \|(f-p)\chi_{C_{j,k}}\|_{L^2(\rho)}. \tag{4}$$

At a fixed scale j, f can be approximated by the piecewise polynomial $f_j = \sum_k p_{j,k}\chi_{C_{j,k}}$. Denote V_j as the space of θ-order piecewise polynomial functions on the partition $\cup_k C_{j,k}$. By definition, V_j is a linear subspace and $V_j \subset V_{j+1}$. We have $f_j \in V_j$ and f_j is the best approximation of f in V_j. Let $V_j^\perp$ be the orthogonal complement of V_j in V_{j+1}, and then

$$V_{j+1} = V_j \oplus V_j^\perp \quad \text{and} \quad V_j^\perp \perp V_{j'}^\perp \text{ if } j \neq j'.$$

When the node $C_{j,k}$ is refined to its children $\mathcal{C}(C_{j,k})$, the difference of the approximations between these two scales on $C_{j,k}$ is defined as

$$\psi_{j,k} = \psi_{j,k}(f) = \sum_{C_{j+1,k'} \in \mathcal{C}(C_{j,k})} p_{j+1,k'}(f)\chi_{C_{j+1,k'}} - p_{j,k}(f)\chi_{C_{j,k}}. \tag{5}$$

For $C_{0,1} = X$, we let $\psi_{0,1} = p_{0,1}$. Note that $\sum_k \psi_{j,k} \in V_j^\perp$ and therefore $\sum_k \psi_{j,k}$ and $\sum_{k'} \psi_{j',k'}$ are orthogonal if $j \neq j'$.

The refinement quantity on the node $C_{j,k}$ is defined as the norm of $\psi_{j,k}$:

$$\delta_{j,k} = \delta_{j,k}(f) = \|\psi_{j,k}\|_{L^2(\rho)}. \tag{6}$$

In the case piecewise constant approximations, i.e. $\theta = 0$, $\psi_{j,k}(f)$ corresponds to the Haar wavelet coefficient of f, and $\delta_{j,k}$ is the magnitude of the Haar wavelet coefficient.

The target function f can be decomposed as

$$f = \sum_{j \geq 0, k} \psi_{j,k}(f).$$

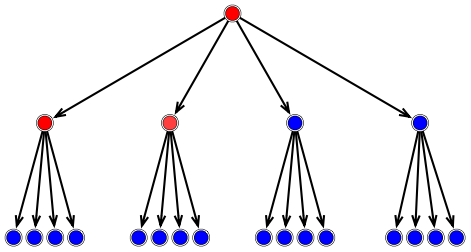
Due to the orthogonality of the $\psi_{j,k}$'s, we have

$$\|f\|^2_{L^2(\rho)} = \sum_{j \geq 0, k} [\delta_{j,k}(f)]^2.$$

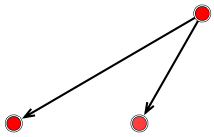
In the tree-based nonlinear approximation, one fixes a threshold value $\eta > 0$, and truncate $\mathcal{T}$ to $\mathcal{T}(f, \eta, \rho)$ – the smallest subtree that contains all $C_{j,k}$ with $\delta_{j,k}(f) > \eta$ under measure ρ. The collection of outer leaves of $\mathcal{T}(f, \eta, \rho)$, denoted by $\Lambda(f, \eta, \rho)$, gives rise to an adaptive partition. This truncation procedure is illustrated in Figure 2. In Figure 2, the red nodes have the refinement quantity above η, and then the master tree $\mathcal{T}$ is truncated to the smallest subtree containing the red nodes in (b). The outer leaves of this truncated tree are given by the green nodes in Figure 2 (c), and the corresponding adaptive partition is given in Figure 2 (d).

The piecewise polynomial approximation of f on this adaptive partition is

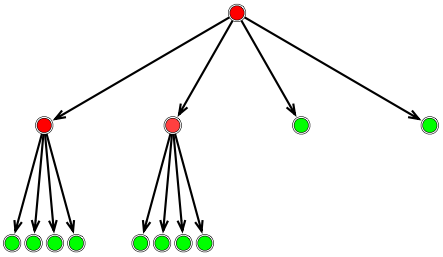
$$p_{\Lambda(f,\eta,\rho)} = \sum_{C_{j,k} \in \Lambda(f,\eta,\rho)} p_{j,k}(f)\chi_{C_{j,k}}. \tag{7}$$



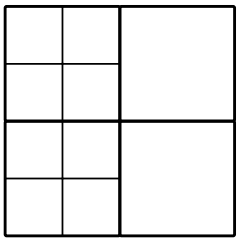
(a) Red nodes satisfy $\delta_{j,k} > \eta$



(b) The truncated tree



(c) Outer leaves given by the green nodes



(d) Dyadic partition

Figure 2: (a) For a fixed $\eta > 0$, the red nodes have the refinement quantity above η: $\delta_{j,k}(f) > \eta$. The master tree is then truncated to the smallest subtree containing the red nodes in (b). In (c), the outer leaves of the truncated tree are given by the green nodes. The corresponding adaptive partition is given in (d).

3.3 The $\mathcal{A}^s_{\theta,\rho}$ Class

In adaptive approximation, the regularity of f can be defined by the size of the tree $\#\mathcal{T}(f,\eta,\rho)$ ((2.19) in Binev et al. (2007)), leading to the $\mathcal{A}^s_{\theta,\rho}$ space as defined below.

Definition 3 (the $\mathcal{A}^s_{\theta,\rho}$ class) *For a fixed $s > 0$, a polynomial degree θ, we let the function class $\mathcal{A}^s_{\theta,\rho}$ be the collection of all $f \in L^2(X)$, such that*

$$|f|^m_{\mathcal{A}^s_{\theta,\rho}} = \sup_{\eta>0} \eta^m \#\mathcal{T}(f,\eta,\rho) < \infty, \quad \textit{with} \quad m = \frac{2}{2s+1}, \tag{8}$$

where $\mathcal{T}(f,\eta,\rho)$ is the truncated tree of approximating f with piecewise θ-th order polynomials with threshold η.

Unlike Hölder and Sobolev functions which have uniform regularities over the domain, the function class considered in this paper allows nonuniform regularities, such as discontinuity and piecewise smoothness. Piecewise smooth functions have wide applications in real world scenarios. For example, in human vision system, images are first perceived as cartoon-like functions, which are piecewise smooth functions with discontinuities at boundaries (Kutyniok and Labate, 2012). Piecewise smooth functions are also used in 3D surface modeling (Yu et al., 2022) and dynamical systems (Bernardo et al., 2008).

In Definition 3, the complexity of the adaptive approximation is measured by the cardinality of the truncated tree $\mathcal{T}(f,\eta,\rho)$. In fact, the cardinality of the adaptive partition

$\Lambda(f,\eta,\rho)$ is related with the cardinality of the truncated tree $\mathcal{T}(f,\eta,\rho)$ such that

$$\#\mathcal{T}(f,\eta,\rho) \le \#\Lambda(f,\eta,\rho) \le 2^d \#\mathcal{T}(f,\eta,\rho). \tag{9}$$

The lower bound follows from

$$\#\mathcal{T}(f,\eta,\rho) \le \sum_{k=1}^{\infty} \frac{\#\Lambda(f,\eta,\rho)}{(2^d)^k} = \frac{\#\Lambda(f,\eta,\rho)}{2^d(1-\frac{1}{2^d})} \le \#\Lambda(f,\eta,\rho).$$

The definition of $\mathcal{A}^s_{\theta,\rho}$ does not explicitly depend on the dimension d. The dimension d is actually hidden in the regularity parameter s (see Example 1a). This way of definition has the advantage of adapting to low-dimensional structures in the data distribution (see Example 5a).

When $f \in \mathcal{A}^s_{\theta,\rho}$, we have the approximation error

$$\|f - p_{\Lambda(f,\eta,\rho)}\|^2_{L^2(\rho)} \le C_s |f|^m_{\mathcal{A}^s_{\theta,\rho}} \eta^{2-m} \le C_s |f|^2_{\mathcal{A}^s_{\theta,\rho}} (\#\mathcal{T}(f,\eta,\rho))^{-2s}, \tag{10}$$

where

$$C_s = 2^m \sum_{\ell \ge 0} 2^{\ell(m-2)}, \quad \text{with } \ m = \frac{2}{2s+1}. \tag{11}$$

The approximation error in (10) is proved in Appendix A. The original proof can be found in Binev et al. (2007, 2005); Cohen et al. (2001).

3.4 Case Study of the $\mathcal{A}^s_{\theta,\rho}$ Function Class

The $\mathcal{A}^s_{\theta,\rho}$ class contains a large collection of functions, including Hölder functions, piecewise Hölder functions, functions which are irregular on a set of measure zero, and regular functions with distribution concentrated on a low-dimensional manifold. For some examples to be studied below, we make the following assumption on the measure ρ:

Assumption 1 *There exists a constant $C_\rho > 0$ such that any subset $S \subset X$ satisfies*

$$\rho(S) \le C_\rho |S|,$$

where $|S|$ is the Lebesgue measure of S.

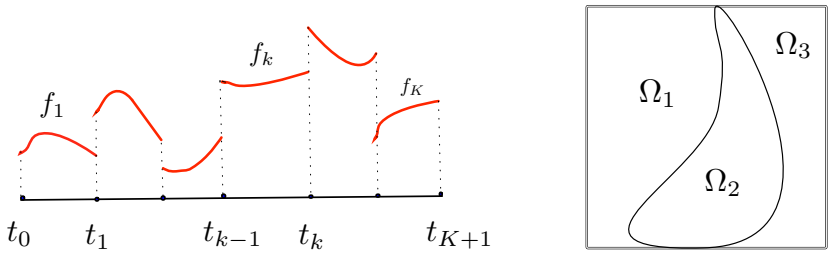
3.4.1 Hölder functions

Example 1a (Hölder functions) *Let $r > 0$. Under Assumption 1, the r-Hölder function class $\mathcal{H}^r(X)$ belongs to $\mathcal{A}^{r/d}_{\lceil r-1 \rceil,\rho}$. If $f \in \mathcal{H}^r(X)$, then we have $f \in \mathcal{A}^{r/d}_{\lceil r-1 \rceil,\rho}$. Furthermore, if $\|f\|_{\mathcal{H}^r(X)} \le 1$, then*

$$|f|_{\mathcal{A}^{r/d}_{\lceil r-1 \rceil,\rho}} \le C(r,d,C_\rho) \tag{12}$$

for some constant $C(r,d,C_\rho)$ depending on r, d and C_ρ in Assumption 1.

Example 1a is proved in Section 6.4.1. At the end of Example 1a, we assume $\|f\|_{\mathcal{H}^r(X)} \le 1$ without loss of generality. The same statement holds if $\|f\|_{\mathcal{H}^r(X)}$ is bounded by an absolute constant. Such a constant only changes the $|\cdot|_{\mathcal{A}^{r/d}_{\lceil r-1 \rceil,\rho}}$ bound in (12), i.e. $C(r,d,C_\rho)$ will



(a) 1D piecewise Hölder function

(b) 2D piecewise domain

Figure 3: (a) Example 2a: 1D piecewise Hölder function with K discontinuity points; (b) Example 3a: A 2D piecewise domain. The functions in Example 3a are r-Hölder in the interior of $\Omega_1, \Omega_2, \Omega_3$.

depends on this constant. In fact, the proof of Example 1a can be slightly revised to show that $\mathcal{H}^r(X) \subset \mathcal{A}_\theta^{r/d}$ for a fixed $\theta \geq \lceil r-1 \rceil$.

The neural network approximation theory for Hölder functions is given in Example 1b and the generalization theory is given in Example 1c.

3.4.2 Piecewise Hölder functions in 1D

Example 2a (Piecewise Hölder functions in 1D) *Let $d = 1$, $r > 0$ and K be a positive integer. Under Assumption 1, all bounded piecewise r-Hölder functions with K discontinuity points belong to $f \in \mathcal{A}^r_{\lceil r-1 \rceil, \rho}$. Specifically, let f be a piecewise r-Hölder function such that $f = \sum_{k=1}^{K+1} f_k \chi_{[t_{k-1}, t_k)}$, where $0 = t_0 < t_1 < \ldots < t_K < t_{K+1} = 1$. Each function $f_k : [t_{k-1}, t_k) \to \mathbb{R}$ is r-Hölder in (t_{k-1}, t_k), and f is discontinuous at $t_1, t_2, \ldots t_K$. Assume f is bounded such that $\|f\|_{L^\infty([0,1])} \leq 1$. In this case, we have $f \in \mathcal{A}^r_{\lceil r-1 \rceil, \rho}$. See Figure 3 (a) for an illustration of a piecewise Hölder function in 1D. Furthermore, if $\max_k \|f\|_{\mathcal{H}^r(t_k, t_{k+1})} \leq 1$, then*

$$|f|_{\mathcal{A}^r_{\lceil r-1 \rceil, \rho}} \leq C(r, d, C_\rho, K)$$

for some constant $C(r, d, C_\rho, K)$ depending on r, d, K and C_ρ in Assumption 1, and does not depend on specific t_k's.

Example 2a is proved in Section 6.4.2. Example 2a demonstrates that, for 1D bounded piecewise r-Hölder functions with a finite number of discontinuities, the overall regularity index is $s = r$ under Definition 3. In comparison with Example 1a, we prove that, a finite number of discontinuities in 1D does not affect the regularity index in Definition 3.

The neural network approximation theory for piecewise Hölder functions in 1D is given in Example 2b and the generalization theory is given in Example 2c.

3.4.3 Piecewise Hölder functions in multi-dimensions

In the next example, we will see that, for piecewise r-Hölder functions in multi-dimensions, the overall approximation error is dominated either by the approximation error in the

interior of each piece or by the error along the discontinuity. The overall regularity index s depends on r and d.

Example 3a (Piecewise Hölder functions in multi-dimensions) *Let $d \geq 2$, $r > 0$ and $\{\Omega_t\}_{t=1}^T$ be subsets of $[0,1]^d$ such that $\cup_{t=1}^T \Omega_t = [0,1]^d$ and the Ω_t's only overlap at their boundaries. Each Ω_t is a connected subset of $[0,1]^d$ and the union of their boundaries $\cup_t \partial\Omega_t$ has upper Minkowski dimension $d-1$. See Figure 3 (b) for an illustration of the Ω_t's. When ρ satisfies Assumption 1, all piecewise r-Hölder functions with discontinuity on $\cup_t \partial\Omega_t$ belong to*

$$\mathcal{A}^s_{\lceil r-1 \rceil, \rho}, \text{ where } s = \min\left\{\frac{r}{d}, \frac{1}{2(d-1)}\right\}. \tag{13}$$

Specifically, let f be a piecewise r-Hölder function such that $f = \sum_{t=1}^T f_t \chi_{\Omega_t}$ where χ_{Ω_t} is the indicator function on Ω_t. Each function $f_t : \Omega_t \to \mathbb{R}$ is r-Hölder in the interior of Ω_t: $f_t \in \mathcal{H}^r(\Omega_t^o)$ where Ω_t^o denotes the interior of Ω_t, and f is discontinuous at $\cup_t \partial\Omega_t$. Assume f is bounded such that $\|f\|_{L^\infty([0,1]^d)} \leq 1$. In this case, $f \in \mathcal{A}^s_{\lceil r-1 \rceil, \rho}$ with the s given in (13). Furthermore, if $\max_t \|f\|_{\mathcal{H}^r(\Omega_t^o)} \leq 1$, then

$$|f|_{\mathcal{A}^s_{\lceil r-1 \rceil, \rho}} \leq C(r, d, c_M(\cup_t \partial\Omega_t), C_\rho)$$

for some $C(r, d, c_M(\cup_t \partial\Omega_t), C_\rho)$ depending on $r, d, c_M(\cup_t \partial\Omega_t)$ (the Minkowski dimension constant of $\cup_t \partial\Omega_t$ defined in Definition 2) and C_ρ in Assumption 1.

Example 3a is proved in Section 6.4.3. Example 3a demonstrates that, for piecewise r-Hölder functions with discontinuity on a subset with upper Minkowski dimension $d-1$, the overall regularity index s has a phase transition. When $\frac{r}{d} \leq \frac{1}{2(d-1)}$, the approximation error is dominated by that in the interior of the Ω_t's. When $\frac{r}{d} > \frac{1}{2(d-1)}$, the approximation error is dominated by that around the boundary of the Ω_t's. As the result, the overall regularity index s is the minimum of $\frac{r}{d}$ and $\frac{1}{2(d-1)}$.

The neural network approximation theory for piecewise Hölder functions in multi-dimensions is given in Example 3b and the generalization theory is given in Example 3c.

Example 3a considers piecewise Hölder functions where all pieces have the same smoothness. With minor revisions of the proof of Example 3a, we can show that piecewise Hölder functions with different smoothness on each piece is also in $\mathcal{A}^s_{\theta,\rho}$ with a properly chosen s. Specifically, under the setting of Example 3a except we assume the f on Ω_k is a r_k-Hölder function for $k = 1, ..., T$. Denote $r = \min_k r_k$. Then we can show $f \in \mathcal{A}^s_{\theta,\rho}$ with $s = \min\left\{\frac{r}{d}, \frac{1}{2(d-1)}\right\}$, in which the smoothness parameter s is determined by the smallest r_k. This is because according to (8), the function class $\mathcal{A}^s_{\theta,\rho}$ is defined by taking the supremum over η. When the function f has different smoothness on each piece, the right-hand side of (8) is dominated by the piece with the smallest smoothness.

A related problem is to consider functions with continuously changing smoothness. However, the current definition of $\mathcal{A}^s_{\theta,\rho}$ class in Definition 3 cannot benefit from this setting. A more dedicated definition is needed. We defer the discussion to Appendix B.

3.4.4 Functions irregular on a set of measure zero

The definition of $\mathcal{A}^s_{\theta,\rho}$ is dependent on the measure ρ, since the refinement quantity $\delta_{j,k}$ is the L^2 norm with respect to ρ. This measure-dependent definition is not only adaptive to the regularity of f, but also adaptive to the distribution ρ. In the following example, we show that, the definition of $\mathcal{A}^s_{\theta,\rho}$ allows the function f to be irregular on a set of measure zero. For $\delta > 0$, Ω_δ denotes the set within δ distance to Ω such that $\Omega_\delta = \{x \in X : \text{dist}(x, \Omega) = \inf_{z\in\Omega} \|x - z\| \le \delta\}$.

Example 4a (Functions irregular on a set of measure zero) *Let $\delta > 0$, Ω be a subset of $X = [0,1]^d$ and $\Omega^{\complement} = X \setminus \Omega$. If f is an r-Hölder function on Ω_δ and $\rho(\Omega^{\complement}) = 0$, then $f \in \mathcal{A}^{r/d}_{\lceil r-1\rceil,\rho}$. Furthermore, if $\|f\|_{\mathcal{H}^r(\Omega_\delta)} \le 1$, then*

$$|f|_{\mathcal{A}^{r/d}_{\lceil r-1\rceil,\rho}} \le C(r, d, C_\rho)$$

for some constant $C(r, d, C_\rho)$ depending on r, d and C_ρ in Assumption 1.

Example 4a is proved in Section 6.4.4. In Example 4a, f is r-Hölder on Ω_δ, and can be irregular on a measure zero set. In comparison with Example 1a, such irregularity on a set of measure zero does not affect the smoothness parameter in Definition 3. In this example, we set Ω_δ to be a larger set than Ω, in order to avoid the discontinuity effect at the boundary of Ω.

The neural network approximation theory for functions in Example 4a is given in Example 4b and the generalization theory is given in Example 4c.

3.4.5 Hölder functions with distribution concentrated on a low-dimensional manifold

Since the definition of $\mathcal{A}^s_{\theta,\rho}$ is dependent on the probability measure ρ, this definition is also adaptive to lower-dimensional sets in X. We next consider a probability measure ρ concentrated on a d_{in}-dimensional manifold isometrically embedded in X.

Example 5a (Hölder functions with distribution concentrated on a low-dimensional manifold) *Let $r > 0$. Suppose $X = [0,1]^d$ can be decomposed to Ω and $\Omega^{\complement}$, i.e. $X = \Omega \cup \Omega^{\complement}$ where Ω is a compact d_{in}-dimensional Riemannian manifold isometrically embedded in X. Assume that $\rho(\Omega^{\complement}) = 0$, and ρ conditioned on Ω is the uniform distribution on Ω. If $f \in \mathcal{H}^r(X)$, then $f \in \mathcal{A}^{r/d_{\text{in}}}_{\lceil r-1\rceil,\rho}$. Furthermore, if $\|f\|_{\mathcal{H}^r(X)} \le 1$, then*

$$|f|_{\mathcal{A}^{r/d_{\text{in}}}_{\lceil r-1\rceil,\rho}} < C(r, d, d_{\text{in}}, \tau, |\Omega|)$$

with $C(r, d, d_{\text{in}}, \tau, |\Omega|)$ depending on $r, d, d_{\text{in}}, \tau$ and $|\Omega|$, where τ is the reach (Federer, 1959) of Ω and $|\Omega|$ is the surface area of Ω.

Example 5a is proved in Section 6.4.5. In this example, the function f is r-Hölder on X, but the measure ρ is supported on a lower dimensional manifold with intrinsic dimension d_{in}. The regularity index under Definition 3 is r/d_{in} instead of r/d.

4. Adaptive Approximation and Generalization Theory of Deep Neural Networks

This section contains our main results: approximation and generalization theories of deep ReLU networks for the $\mathcal{A}^s_{\theta,\rho}$ function class. We present some preliminaries in Subsection 4.1, the approximation theory in Subsection 4.2, case studies of the approximation error in Subsection 4.3, the generalization theory in Subsection 4.4, and case studies of the generalization error in Subsection 4.5.

4.1 Preliminaries

Each $C_{j,k}$ is a hypercube in the form of $\otimes_{\ell=1}^d [r_{\ell,j,k}, r_{\ell,j,k} + 2^{-j}]$, where $r_{\ell,j,k} \in [0,1]$ is a scalar and

$$\otimes_{\ell=1}^d [r_{\ell,j,k}, r_{\ell,j,k} + 2^{-j}] = [r_{1,j,k}, r_{1,j,k} + 2^{-j}] \times \cdots \times [r_{d,j,k}, r_{d,j,k} + 2^{-j}] \tag{14}$$

is a hypercube with edge length 2^{-j} in $\mathbb{R}^d$.

The collection of polynomials $\left((\mathbf{x} - \mathbf{r}_{j,k})/2^{-j}\right)^{\boldsymbol{\alpha}}$ form a basis for the space of d-variable polynomials of degree no more than θ. Let ρ be a measure on $[0,1]^d$ and $f \in L^2(\rho)$. The piecewise polynomial approximator $p_{j,k}$ for f can be written as

$$p_{j,k}(\mathbf{x}) = \sum_{|\boldsymbol{\alpha}| \le \theta} a_{\boldsymbol{\alpha}} \left(\frac{\mathbf{x} - \mathbf{r}_{j,k}}{2^{-j}} \right)^{\boldsymbol{\alpha}}, \tag{15}$$

where $\mathbf{r}_{j,k} = [r_{1,j,k}, ..., r_{d,j,k}]$.

In this paper, we focus on the set of functions with bounded coefficients in the piecewise polynomial approximation.

Assumption 2 *Let $s, R_{\mathcal{A}}, R, R_p > 0$, θ be a nonnegative integer, and ρ be a probability measure on $[0,1]^d$. We assume $f \in \mathcal{A}^s_{\theta,\rho}$ and*

(i) $|f|_{\mathcal{A}^s_{\theta,\rho}} \le R_{\mathcal{A}}$,

(ii) $\|f\|_{L^\infty([0,1]^d)} \le R$,

(iii) On every $C_{j,k}$, the polynomial approximator $p_{j,k}$ for f in the form of (15) satisfies $|a_{\boldsymbol{\alpha}}| \le R_p$ for all $\boldsymbol{\alpha}$ with $|\boldsymbol{\alpha}| \le \theta$.

By Assumption 2 (i) and (ii), f has a bounded $|\cdot|_{\mathcal{A}^s_{\theta,\rho}}$ quantity and L^∞ norm, which is a common assumption in nonparametric estimation theory (Györfi et al., 2002). Assumption 2 (iii) requires the polynomial coefficients in the best polynomial approximating f on every $C_{j,k}$ to be uniformly bounded by R_p. The following lemma shows that Assumption 2 (iii) can be implied from Assumption 2 (ii) when ρ is the Lebesgue measure on $X = [0,1]^d$.

Lemma 4 *Let $R > 0$, θ be a fixed nonnegative integer, and ρ be the Lebesgue measure on $X = [0,1]^d$. There exists a constant $R_p > 0$ depending on θ, d and R such that, for any function f on $[0,1]^d$ satisfying $\|f\|_{L^\infty(X)} \le R$, the $p_{j,k}$ in (4) has the form of (15) with $|a_{\boldsymbol{\alpha}}| \le R_p$, $\forall \boldsymbol{\alpha}$ with $|\boldsymbol{\alpha}| \le \theta$ and $\|p_{j,k}\|_{L^\infty(C_{j,k})} \le C R_p$ for some C depending on d and θ.*

Lemma 4 is proved in Appendix C. Lemma 4 implies that under the Lebesgue measure, for any $p_{j,k}$ in the form of (15), the coefficients $a_{\boldsymbol{\alpha}}$'s are uniformly bounded by a constant depending on θ, d, R, and is independent to the index (j,k). Thus Assumption 2(iii) holds.

4.2 Approximation Theory

Our approximation theory shows that deep neural networks give rise to universal approximations for functions in the $\mathcal{A}^s_{\theta,\rho}$ class under Assumption 2 if the network architecture is properly chosen.

Theorem 5 (Approximation) *Let $s, d, C_\rho, R_{\mathcal{A}}, R, R_p > 0$ and θ be a nonnegative integer. For any $\varepsilon > 0$, there is a ReLU network class $\mathcal{F} = \mathcal{F}_{\rm NN}(L, w, K, \kappa, M)$ with parameters*

$$L = O\left(\log \frac{1}{\varepsilon}\right),\ w = O(\varepsilon^{-\frac{1}{s}}),\ K = O\left(\varepsilon^{-\frac{1}{s}} \log \frac{1}{\varepsilon}\right),\ \kappa = O(\varepsilon^{-\max\{2, \frac{1}{s}\}}),\ M = R, \quad (16)$$

such that, for any ρ satisfying Assumption 1 and any $f \in \mathcal{A}^s_{\theta,\rho}$ satisfying Assumption 2, if the weight parameters of the network are properly chosen, the network yields a function $\widetilde{f} \in \mathcal{F}$ such that

$$\|\widetilde{f} - f\|^2_{L^2(\rho)} \le \varepsilon. \quad (17)$$

The constants hidden in $O(\cdot)$ depends on d (the dimension of the domain for f), C_ρ (in Assumption 1), θ (the polynomial order), s, R, R_p and $R_{\mathcal{A}}$ (in Assumption 2).

Theorem 5 is proved in Section 6.2. Theorem 5 demonstrates the universal approximation power of deep neural networks for the $\mathcal{A}^s_{\theta,\rho}$ function class. The parameters in (16) specifies the network architecture. To approximate a specific function $f \in \mathcal{A}^s_{\theta,\rho}$, there exist some proper weight parameters which give rise to a network function $\widetilde{f}$ to approximate f.

4.3 Case Studies of the Approximation Error

In this subsection, we apply Theorem 5 to the examples in Subsection 3.4, and derive the approximation theory for each example. In the following case studies, we need Assumptions 1, 2 (iii) and 3, but not Assumption 2 (i) - (ii). In each case, we have shown in Subsection 3.4 that Assumption 2 (i) holds: $f \in \mathcal{A}^s_{\theta,\rho}$ with a proper θ and the regularity index s depends on each specific case.

4.3.1 Hölder functions

Consider Hölder functions in Example 1a such that $\mathcal{H}^r \subset \mathcal{A}^{r/d}_{\lceil r-1\rceil,\rho}$. Applying Theorem 5 gives rise to the following neural network approximation theory for Hölder functions.

Example 1b (Hölder functions) *Let $r, d, C_\rho, R_p > 0$. For any $\varepsilon > 0$, there is a ReLU network class $\mathcal{F} = \mathcal{F}_{\rm NN}(L, w, K, \kappa, M)$ with parameters*

$$L = O\left(\log \frac{1}{\varepsilon}\right),\ w = O(\varepsilon^{-\frac{d}{r}}),\ K = O\left(\varepsilon^{-\frac{d}{r}} \log \frac{1}{\varepsilon}\right),\ \kappa = O(\varepsilon^{-\max\{2, \frac{d}{r}\}}),\ M = 1,$$

such that for any ρ satisfying Assumption 1 and any function $f \in \mathcal{H}^r(X)$ satisfying $\|f\|_{\mathcal{H}^r(X)} \le 1$ and Assumption 2 (iii), if the weight parameters of the network are properly chosen, the network yields a function $\widetilde{f} \in \mathcal{F}$ such that

$$\|\widetilde{f} - f\|^2_{L^2(\rho)} \le \varepsilon.$$

The constant hidden in $O(\cdot)$ depends on r, d, C_ρ, R_p.

Example 1b is a corollary of Theorem 5 with $s = r/d$. Note that Assumption 2(i) and (ii) are implied by the condition $\|f\|_{\mathcal{H}^r(X)} \leq 1$. In particular, $\|f\|_{\mathcal{H}^r(X)} \leq 1$ implies $\|f\|_{L^\infty(X)} \leq 1$ according to (1). Furthermore, If $\|f\|_{\mathcal{H}^r(X)} \leq 1$, then $|f|_{\mathcal{A}^{r/d}_{\lceil r-1\rceil,\rho}} \leq C(r,d,C_\rho)$ according to our argument in Example 1a.

In litertuare, the approximation theory of ReLU networks for Sobolev functions in $W^{r,\infty}$ has been established in Yarotsky (2017, Theorem 1). The proof in Yarotsky (2017, Theorem 1) can be applied to Hölder functions. Our network size in Example 1b is comparable to that in Yarotsky (2017, Theorem 1).

4.3.2 Piecewise Hölder functions in 1D

Considering 1D piecewise Hölder functions in Example 2a, we have the following approximation theory:

Example 2b (Piecewise Hölder functions in 1D) *Let $r, C_\rho, R_p, K > 0$. For any $\varepsilon > 0$, there is a ReLU network class $\mathcal{F} = \mathcal{F}_{\rm NN}(L, w, K, \kappa, M)$ with parameters*

$$L = O\left(\log \frac{1}{\varepsilon}\right),\ w = O(\varepsilon^{-\frac{1}{r}}),\ K = O\left(\varepsilon^{-\frac{1}{r}} \log \frac{1}{\varepsilon}\right),\ \kappa = O(\varepsilon^{-\max\{2,\frac{1}{r}\}}),\ M = 1, \quad (18)$$

such that for any ρ satisfing Assumption 1, and any piecewise r-Hölder function in the form of $f = \sum_{k=1}^{K+1} f_k \chi_{[t_{k-1},t_k)}$ in Example 2a satisfying $\|f\|_{L^\infty([0,1])} \leq 1$, $\max_k \|f_k\|_{\mathcal{H}^r(t_k,t_{k+1})} \leq 1$ and Assumption 2(iii), if the weight parameters of the network are properly chosen, the network yields a function $\widetilde{f} \in \mathcal{F}$ such that

$$\|\widetilde{f} - f\|^2_{L^2(\rho)} \leq \varepsilon.$$

The constant hidden in $O(\cdot)$ depends on r, C_ρ, R_p, K.

Example 2b is a corollary of Theorem 5 with $f \in \mathcal{A}^r_{\lceil r-1\rceil,\rho}$. Assumption 2 (i) is not explicitly enforced in Example 2b , but is implied from the condition of $\max_k \|f_k\|_{\mathcal{H}^r(t_k,t_{k+1})} \leq 1$ by Example 2a. Example 2b shows that, to achieve an ε approximation error for 1D functions with a finite number of discontinuities, the network size is comparable to that for 1D Hölder functions in Example 1b.

4.3.3 Piecewise Hölder function in multi-dimensions

Considering piecewise Hölder functions in multi-dimensions in Example 3a, we have the following approximation theory:

Example 3b (Piecewise Hölder functions in multi-dimensions) *Let $d \geq 2$, $r, C_\rho, R_p, T > 0$ and $\{\Omega_t\}_{t=1}^T$ be subsets of $[0,1]^d$ such that $\cup_{t=1}^T \Omega_t = [0,1]^d$ and the Ω_t's only overlap at their boundaries. Each Ω_t is a connected subset of $[0,1]^d$ and the union of their boundaries $\cup_{t=1}^T \partial\Omega_t$ has upper Minkowski dimension $d-1$. Denote the Minkowski dimension constant of $\cup_{t=1}^T \partial\Omega_t$ by $c_M(\cup_t \partial\Omega_t)$. For any $\varepsilon > 0$, there is a ReLU network class $\mathcal{F} = \mathcal{F}_{\rm NN}(L, w, K, \kappa, M)$ with parameters*

$$L = O\left(\log \frac{1}{\varepsilon}\right),\ w = O(\varepsilon^{-\max\left\{\frac{d}{r},2(d-1)\right\}}),\ K = O\left(\varepsilon^{-\max\left\{\frac{d}{r},2(d-1)\right\}} \log \frac{1}{\varepsilon}\right),$$

$$\kappa = O(\varepsilon^{-\max\{2,\frac{d}{r},2(d-1)\}}),\ M = 1,$$

such that for any ρ satisfying Assumption 1, and any piecewise r-Hölder function in the form of $f = \sum_{t=1}^{T} f_t \chi_{\Omega_t}$ in Example 3a satisfying $\|f\|_{L^\infty([0,1]^d)} \le 1$, $\max_t \|f_t\|_{\mathcal{H}^r(\Omega_t^\rho)} \le 1$ and Assumption 2(iii), if the weight parameters of the network are properly chosen, the network yields a function $\widetilde{f} \in \mathcal{F}$ such that

$$\|\widetilde{f} - f\|_{L^2(\rho)}^2 \le \varepsilon.$$

The constant hidden in $O(\cdot)$ depends on $r, d, c_M(\cup_t \partial\Omega_t), C_\rho, R_p$.

Example 3b is a corollary of Theorem 5 with $f \in \mathcal{A}^s_{\lceil r-1\rceil,\rho}$ for the s given in (13). Assumption 2 (i) is not explicitly enforced in Example 3b , but is implied from the condition of $\max_k \|f_k\|_{\mathcal{H}^r(t_k,t_{k+1})} \le 1$ by Example 3a. Example 3b implies that to approximate a piecewise Hölder function in Example 3a, the number of nonzero weight parameters is in the order of $\varepsilon^{-\max\{\frac{d}{r},2(d-1)\}} \log \frac{1}{\varepsilon}$. The discontinuity set in Example 3b has upper Minkowski dimension $d-1$, which is a weak assumption without additional regularity assumption or low-dimensional structures.

Neural network approximation theory for piecewise smooth functions has been considered in Petersen and Voigtlaender (2018). The setting in Petersen and Voigtlaender (2018) is similar but different from that of Example 3b. In Petersen and Voigtlaender (2018), the authors considered piecewise functions $f : [-1/2, 1/2]^d \to \mathbb{R}$, where the different "smooth regions" of f are separated by $\mathcal{H}^\beta$ hypersurfaces. If f is a piecewise r-Hölder function and the Hölder norm of f on each piece is bounded, it is shown in Petersen and Voigtlaender (2018, Corollary 3.7) that such f can be universally approximated by a ReLU network with at most $c\varepsilon^{-p(d-1)/\beta}$ weight parameters to guarantee an L^p approximation error ε. It is further shown in Petersen and Voigtlaender (2018, Theorem 4.2) that, to achieve an L^p approximation error ε, the optimal required number of weight parameters is lower bounded in the order of $\varepsilon^{-p(d-1)/\beta}/\log\frac{1}{\varepsilon}$. Our result in Example 3b is comparable to that of Petersen and Voigtlaender (2018) when $\beta = 1$ and $p = 2$. When the discontinuity hypersurface has higher order regularity, i.e. $\beta > 1$, the network size in Petersen and Voigtlaender (2018) is smaller/better than that in Example 3b since the higher order smoothness of the discontinuity hypersurface is exploited in the approximation theory. However, Example 3b imposes a weaker assumption on the discontinuity hypersurface. Example 3b requires the discontinuity hypersurface to have upper Minkowski dimension $d-1$, while the the discontinuity hypersurface in Petersen and Voigtlaender (2018, Definition 3.3) is the graph of a $\mathcal{H}^\beta$ function on $d-1$ coordinates. In this sense, Example 3b can be applied to a wider class of piecewise Hölder functions.

4.3.4 Functions irregular on a set of measure zero

Functions in the $\mathcal{A}^s_{\theta,\rho}$ class can be irregular on a set of measure zero, as in Example 4a. The approximation theory for functions in Example 4a is given below:

Example 4b (Functions irregular on a set of measure zero) *Let $r, d, C_\rho, R_p, \delta > 0$ and Ω be a subset of X. For any $\varepsilon > 0$, there is a ReLU network class $\mathcal{F} = \mathcal{F}_{\mathrm{NN}}(L, w, K, \kappa, M)$*

with parameters

$$L = O\left(\log \frac{1}{\varepsilon}\right),\ w = O(\varepsilon^{-\frac{d}{r}}),\ K = O\left(\varepsilon^{-\frac{d}{r}} \log \frac{1}{\varepsilon}\right),\ \kappa = O(\varepsilon^{-\max\{2, \frac{d}{r}\}}),\ M = 1.$$

For any ρ satisfying Assumption 1 and $\rho(\Omega^{\complement}) = 0$, and for any $f : X \to \mathbb{R}$ satisfying $\|f\|_{\mathcal{H}^r(\Omega_\delta)} \leq 1$ and Assumption 2(iii), if the weight parameters of the network are properly chosen, the network class yields a function $\widetilde{f} \in \mathcal{F}$ such that

$$\|\widetilde{f} - f\|_{L^2(\rho)}^2 \leq \varepsilon.$$

The constant hidden in $O(\cdot)$ depends on r, d, C_ρ, R_p.

Example 4b is a corollary of Theorem 5 with $f \in \mathcal{A}^{r/d}_{\lceil r-1 \rceil, \rho}$ and $|f|_{\mathcal{A}^{r/d}_{\lceil r-1 \rceil, \rho}} \leq C(r, d, C_\rho)$ by Example 4a. Example 4b shows that function irregularity on a set of measure zero does not affect the network size in approximation theory.

4.3.5 Hölder functions with distribution concentrated on a low-dimensional manifold

Theorem 5 cannot be directly applied to Example 5a, since Assumption 1 is violated when ρ is supported on a low-dimensional manifold. In literature, neural network approximation theory has been established in Chen et al. (2019a) for functions on a low-dimensional manifold, and in Nakada and Imaizumi (2020) for Hölder functions in $[0, 1]^D$ while the support of measure has a low Minkowski dimension. See Section 4.5.5 for a detailed discussion about the generalization error.

4.4 Generalization Error

Theorem 5 proves the existence of a neural network $\widetilde{f}$ to approximate f with an arbitrary accuracy ε, but it does not give an explicit method to find the weight parameters of $\widetilde{f}$. In practice, the weight parameters are learned from data through the empirical risk minimization. The generalization error of the empirical risk minimizer is analyzed in this subsection.

Suppose the training data set is $\mathcal{S} = \{\mathbf{x}_i, y_i\}_{i=1}^n$ where the $\mathbf{x}_i$'s are i.i.d. samples from ρ, and the y_i's have the form

$$y_i = f(\mathbf{x}_i) + \xi_i, \tag{19}$$

where the ξ_i's are i.i.d. noise, independently of the $\mathbf{x}_i$'s. In practice, one estimates the function f by minimizing the empirical mean squared risk

$$\widehat{f} = \operatorname*{argmin}_{f_{\mathrm{NN}} \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n |f_{\mathrm{NN}}(\mathbf{x}_i) - y_i|^2 \tag{20}$$

for some network class $\mathcal{F}$. The squared generalization error of $\widehat{f}$ is

$$\mathbb{E}_{\mathcal{S}} \mathbb{E}_{\mathbf{x} \sim \rho} |\widehat{f}(\mathbf{x}) - f(\mathbf{x})|^2,$$

where the expectation $\mathbb{E}_{\mathcal{S}}$ is taken over the joint the distribution of training data $\mathcal{S}$.

To establish an upper bound of the squared generalization error, we make the following assumption of noise.

Assumption 3 *Suppose the noise ξ is a sub-Gaussian random variable with mean 0 and variance proxy σ^2.*

Our second main theorem in this paper gives a generalization error bound of $\widehat{f}$.

Theorem 6 (Generalization error) *Let $\sigma, s, d, C_\rho, R_p, R_{\mathcal{A}}, R > 0$ and θ be a nonnegative integer. Suppose ρ satisfies Assumption 1, f satisfies Assumption 2, and Assumption 3 holds for noise. The training data $\mathcal{S}$ are sampled according to* (19)*. If the network class $\mathcal{F} = \mathcal{F}_{\rm NN}(L, w, K, \kappa, M)$ is set with parameters*

$$L = O(\log n),\ w = O(n^{\frac{1}{1+2s}}),\ K = O\left(n^{\frac{1}{1+2s}} \log n\right),\ \kappa = O\left(n^{\max\{\frac{2s}{1+2s}, \frac{1}{1+2s}\}}\right),\ M = R,$$

then the minimizer $\widehat{f}$ of (20) satisfies

$$\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\mathbf{x}\sim\rho}|\widehat{f}(\mathbf{x}) - f(\mathbf{x})|^2 \le C n^{-\frac{2s}{2s+1}} \log^3 n.$$

The constant C and the constant hidden in $O(\cdot)$ depend on σ (the variance proxy of noise in Assumption 3), d (the dimension of the domain for f), C_ρ (in Assumption 1), θ (the polynomial order), s, R, R_p and $R_{\mathcal{A}}$ (in Assumption 2).

Theorem 6 is proved in Section 6.3. Theorem 6 gives rise to a generalization error guarantee of deep neural networks for functions in the $\mathcal{A}^s_{\theta,\rho}$ class. A lower bound can be obtained by comparing the Hölder class and the $\mathcal{A}^{r/d}_{\theta,\rho}$ function class in some special cases. Let $X = [0,1]^d$ and ρ be the uniform distribution on X. As shown in Example 1a and the discussion after it, the r-Hölder function class $\mathcal{H}^r(X)$ belongs to $\mathcal{A}^{r/d}_{\theta,\rho}$ for $\theta \ge \lceil r - 1 \rceil$. We consider the function class $\mathcal{A}^s_{\theta,\rho}(X)$. Setting $r = sd$, we have $\mathcal{H}^r(X) \subset \mathcal{A}^s_{\theta,\rho}$ for $\theta \ge \lceil sd - 1 \rceil$. When ρ is the uniform distribution on X, it is known that the optimal rate for learning an r-Hölder function on X is of $O\left(n^{-\frac{2r}{2r+d}}\right)$ (Györfi et al., 2002, Theorem 3.2). Therefore, the optimal rate for learning a $\mathcal{A}^s_{\theta,\rho}$ functions is no faster than $O\left(n^{-\frac{2r}{2r+d}}\right) = O\left(n^{-\frac{2s}{2s+1}}\right)$. It demonstrates that our result in Theorem 6 is optimal up to a logarithmic factor for the $\mathcal{A}^s_{\theta,\rho}$ class when $\theta \ge \lceil sd - 1 \rceil$ and ρ is the uniform distribution. For a general distribution ρ and the general $\mathcal{A}^s_{\theta,\rho}$ function class, a more dedicated analysis is needed to derive a minimax rate. We leave it as our future work.

The network in Theorem 6 is a deep and under-parameterized network, as its depth scales with the number of samples n and the total number of weight parameters is smaller than n. For learning Hölder functions with uniform distribution ρ, our result achieves the minimax rate up to a logarithmic factor as shown in Example 1c below. Nonparametric regression by shallow (with one hidden layer) and over-parameterized networks for learning Hölder functions are shown to achieve the minimax in Yang and Zhou (2024b,a). In particular, the network in Yang and Zhou (2024a) is over-parameterized. Network construction and proof techniques in this paper and Yang and Zhou (2024b,a) are very different. Extending Theorem 6 to shallow and over-parameterized networks is not trivial and we leave it as our future work.

We will provide case studies in the following subsection.

4.5 Case Study of the Generalization Error

In this subsection, we apply Theorem 6 to the examples in Subsection 3.4, and derive the squared generalization error for each example. In the following case studies, we assume Assumptions 1, 2 (iii) and 3, but not Assumption 2 (i) and (ii). In each case, we have shown in Subsection 3.4 that Assumption 2 (i) holds: $f \in \mathcal{A}^s_{\theta,\rho}$ with a proper θ and the regularity index s depends on each specific case.

4.5.1 Hölder functions

Consider Hölder functions in Example 1a such that $\mathcal{H}^r \subset \mathcal{A}^{r/d}_{\lceil r-1 \rceil,\rho}$. Applying Theorem 6 gives rise to the following generalization error bound for Hölder functions.

Example 1c (Hölder functions) *Let $\sigma \geq 0, r, d, C_\rho, R_p > 0$. Suppose ρ satisfies Assumption 1, $f \in \mathcal{H}^r(X)$ satisfies $\|f\|_{\mathcal{H}^r(X)} \leq 1$ and Assumption 2(iii), and Assumptions 3 hold. The training data $\mathcal{S}$ are sampled according to* (19)*. If we set the network class $\mathcal{F}_{\mathrm{NN}}(L, p, K, \kappa, M)$ as*

$$L = O(\log n),\ p = O(n^{\frac{d}{2r+d}}),\ K = O\left(n^{\frac{d}{2r+d}} \log n\right),\ \kappa = O\left(n^{\max\left\{\frac{2r}{2r+d}, \frac{d}{2r+d}\right\}}\right),\ M = 1.$$

Then the empirical minimizer $\widehat{f}$ of (20) satisfies

$$\mathbb{E}_{\mathcal{S}} \mathbb{E}_{\mathbf{x}\sim\rho} |\widehat{f}(\mathbf{x}) - f(\mathbf{x})|^2 \leq C n^{-\frac{2r}{2r+d}} \log^3 n.$$

The constant C and the constant hidden in $O(\cdot)$ depend on $\sigma, r, d, C_\rho, R_p$.

Example 1c is a corollary of Theorem 6 with $s = r/d$. Our upper bound matches the rate in Schmidt-Hieber (2017, Theorem 1). When the distribution ρ is uniform, our bound is optimal up to a logarithmic factor in comparison with the minimax error given in Györfi et al. (2002, Theorem 3.2).

4.5.2 Piecewise Hölder functions in 1D

Considering 1D piecewise Hölder functions in Example 2a, we have the following generalization error bound:

Example 2c (Piecewise Hölder functions in 1D) *Let $\sigma \geq 0, r, d, C_\rho, R_p > 0$. Suppose ρ satisfies Assumption 1. Let f be an 1D piecewise r-Hölder function in the form of $f = \sum_{k=1}^{K+1} f_k \chi_{[t_{k-1}, t_k)}$ in Example 2a satisfying $\|f\|_{L^\infty([0,1])} \leq 1$, $\max_k \|f_k\|_{\mathcal{H}^r(t_k, t_{k+1})} \leq 1$ and Assumption 2(iii). Suppose Assumption 3 holds and the training data $\mathcal{S}$ are sampled according to* (19)*. Set the network class $\mathcal{F}_{\mathrm{NN}}(L, p, K, \kappa, M)$ as*

$$L = O(\log n),\ p = O(n^{\frac{1}{1+2r}}),\ K = O(n^{\frac{1}{1+2r}} \log n),\ \kappa = O(n^{\max\{\frac{2r}{1+2r}, \frac{1}{1+2r}\}}),\ M = 1. \quad (21)$$

Then the empirical minimizer $\widehat{f}$ of (20) satisfies

$$\mathbb{E}_{\mathcal{S}} \mathbb{E}_{\mathbf{x}\sim\rho} |\widehat{f}(\mathbf{x}) - f(\mathbf{x})|^2 \leq C n^{-\frac{2r}{2r+1}} \log^3 n. \quad (22)$$

The constant C and the constant hidden in $O(\cdot)$ depend on $\sigma, r, C_\rho, R_p, K$.

Example 2c shows that a finite number of discontinuities in 1D does not affect the rate of convergence of the generalization error.

4.5.3 Piecewise Hölder functions in multi-dimensions

Considering piecewise Hölder functions in multi-dimensions in Example 3a, we have the following generalization error bound:

Example 3c (Piecewise Hölder functions in multi-dimensions) *Let $\sigma \geq 0, r, d, C_\rho, R_p > 0$. Let $\{\Omega_t\}_{t=1}^T$ be subsets of $[0,1]^d$ such that $\cup_{t=1}^T \Omega_t = [0,1]^d$ and the Ω_t's only overlap at their boundaries. Each Ω_t is a connected subset of $[0,1]^d$ and the union of their boundaries $\cup_t \partial\Omega_t$ has upper Minkowski dimension $d-1$. Denote the Minkowski dimension constant of $\cup_{t=1}^T \Omega_t$ by $c_M(\cup_t \partial\Omega_t)$. Suppose ρ satisfies Assumption 1. Let f be a piecewise r-Hölder function in the form of $f = \sum_{t=1}^T f_t \chi_{\Omega_t}$ in Example 3a satisfying $\|f\|_{L^\infty([0,1]^d)} \leq 1$, $\max_t \|f_t\|_{\mathcal{H}^r(\Omega_t^o)} \leq 1$ and Assumption 2(iii). Suppose Assumption 3 holds and the training data $\mathcal{S}$ are sampled according to (19). Set the network class $\mathcal{F}_{\rm NN}(L, p, K, \kappa, M)$ as*

$$L = O(\log n),\ p = O(n^{\max\{\frac{d}{2r+d}, \frac{d-1}{d}\}}),\ K = O(n^{\max\{\frac{d}{2r+d}, \frac{d-1}{d}\}} \log n),$$
$$\kappa = O(n^{\max\left\{\min\{\frac{2r}{2r+d}, \frac{1}{d}\}, \max\{\frac{d}{2r+d}, \frac{d-1}{d}\}\right\}}),\ M = 1. \tag{23}$$

Then the empirical minimizer $\widehat{f}$ of (20) satisfies

$$\mathbb{E}_{\mathcal{S}} \mathbb{E}_{\mathbf{x}\sim\rho} |\widehat{f}(\mathbf{x}) - f(\mathbf{x})|^2 \leq C n^{-\min\{\frac{2r}{2r+d}, \frac{1}{d}\}} \log^3 n. \tag{24}$$

The constant C and the constant hidden in $O(\cdot)$ depend on $\sigma, r, d, c_M(\cup_t \partial\Omega_t), C_\rho, R_p$.

Example 3c shows that the convergence rate of the generalization error has a phase transition. When $\frac{r}{d} \leq \frac{1}{2(d-1)}$, the generalization error is dominated by that in the interior of the Ω_t's, so that the squared generalization error converges in the order of $n^{-\frac{2r}{2r+d}}$. When $\frac{r}{d} > \frac{1}{2(d-1)}$, the generalization error is dominated by that around the boundary of the Ω_t's, so that the squared generalization error converges in the order of $n^{-\frac{1}{d}}$. As a result, the overall rate of convergence is $n^{-\min\{\frac{2r}{2r+d}, \frac{1}{d}\}}$ up to log factors.

Under the setting of Petersen and Voigtlaender (2018) (discussed in Section 4.3.3 where different "smooth regions" of f are separated by $\mathcal{H}^\beta$ hypersurfaces), the generalization error for estimating piecewise r-Hölder function by ReLU network is proved in the order of $\max(n^{-\frac{2r}{2r+d}}, n^{-\frac{\beta}{\beta+d-1}})$ in Imaizumi and Fukumizu (2019). When $\beta = 1$, our rate matches the result in Imaizumi and Fukumizu (2019). When $\beta > 1$, the smoothness of boundaries is utilized in Imaizumi and Fukumizu (2019), leading to a better result than ours. Nevertheless, the setting considered in this paper assumes a weaker assumption on the discontinuous boundaries, which are not required to be hypersurfaces.

4.5.4 Functions irregular on a set of measure zero

Functions in the $\mathcal{A}_{\theta,\rho}^s$ class can be irregular on a set of measure zero, as in Example 4a. The generalization error for functions in Example 4a is given below:

Example 4c (Functions irregular on a set of measure zero) *Let $\sigma \geq 0, r, d, C_\rho, R_p > 0$. Let Ω be a subset of X and $\delta > 0$. Suppose ρ satisfies Assumption 1, f satisfies*

$\|f\|_{\mathcal{H}^r(\Omega_\delta)} \leq 1$ *and Assumption 2(iii), and Assumption 3 holds. Set the network class* $\mathcal{F}_{\mathrm{NN}}(L, w, K, \kappa, M)$ *as*

$$L = O(\log n),\ p = O(n^{\frac{d}{2r+d}}),\ K = O\left(n^{\frac{d}{2r+d}} \log n\right),\ \kappa = O\left(n^{\max\left\{\frac{2r}{2r+d}, \frac{d}{2r+d}\right\}}\right),\ M = 1.$$

Then the empirical minimizer $\widehat{f}$ *of (20) satisfies*

$$\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\mathbf{x}\sim\rho}|\widehat{f}(\mathbf{x}) - f(\mathbf{x})|^2 \leq Cn^{-\frac{2r}{2r+d}} \log^3 n.$$

The constant C *and the constant hidden in* $O(\cdot)$ *depend on* $\sigma, r, d, C_\rho, R_p$.

Example 4c shows that function irregularity on a set of measure zero does not affect the rate of convergence of the generalization error. Deep neural networks are adaptive to data distributions as well.

4.5.5 Hölder functions with distribution concentrated on a low-dimensional manifold

When the measure ρ is concentrated on a low-dimensional manifold as considered in Example 5a, Theorem 6 cannot be directly applied since Assumption 1 is not satisfied. Instead, a more dedicated network structure can be designed to develop approximation and generalization error analysis. The setting of Example 5a has been studied in Nakada and Imaizumi (2020), which assumes that the $\mathbf{x}_i$'s are sampled from a measure supported on a set with an upper Minkowski dimension d_{in}. If the target function f is an $r-$Hölder function on $[0,1]^d$ and if the network structure is properly set, the generalization error is in the order of $n^{-\frac{2r}{2r+d_{\mathrm{in}}}}$ up to a logarithmic factor (Nakada and Imaizumi, 2020). The result in Nakada and Imaizumi (2020) can be applied to the setting of Example 5a, since a d_{in} dimensional Riemannian manifold has the Minkowski dimension d_{in}.

Another related setting is that f is a r-Hölder function on a d_{in} dimensional Riemannian manifold embedded in $[0,1]^d$. This setting has been studied in Chen et al. (2019a) for approximation theory and in Chen et al. (2019b) for generalization theory by ReLU networks. In this setting, the squared generalization error converges in the order of $n^{-\frac{2r}{2r+d_{\mathrm{in}}}}$ up to a logarithmic factor.

5. Numerical Experiments

In this section, we perform numerical experiments on 1D piecewise smooth functions, which fit in Example 2a. The following functions are included in our experiments:

- Function with 1 discontinuity point,

$$f(x) = \begin{cases} \sin(2\pi x) + 1 & \text{when } 0 \leq x < \frac{1}{2}, \\ \sin(2\pi x) - 1 & \text{when } \frac{1}{2} \leq x < 1. \end{cases}$$

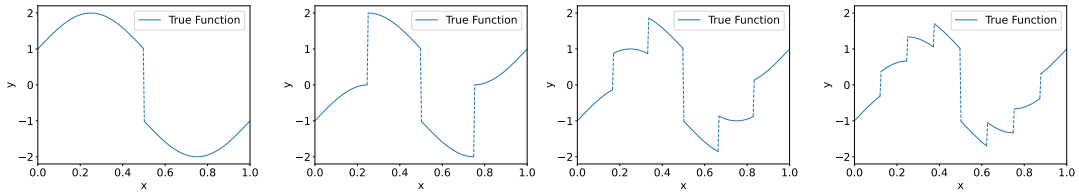


Figure 4: Functions with different numbers of discontinuity points.

- Function with 3 discontinuity points,

$$f(x) = \begin{cases} \sin(2\pi x) - 1 & \text{when } 0 \le x < \frac{1}{4}, \\ \sin(2\pi x) + 1 & \text{when } \frac{1}{4} \le x < \frac{1}{2}, \\ \sin(2\pi x) - 1 & \text{when } \frac{1}{2} \le x < \frac{3}{4}, \\ \sin(2\pi x) + 1 & \text{when } \frac{3}{4} \le x < 1. \end{cases}$$

- Function with 5 discontinuity points,

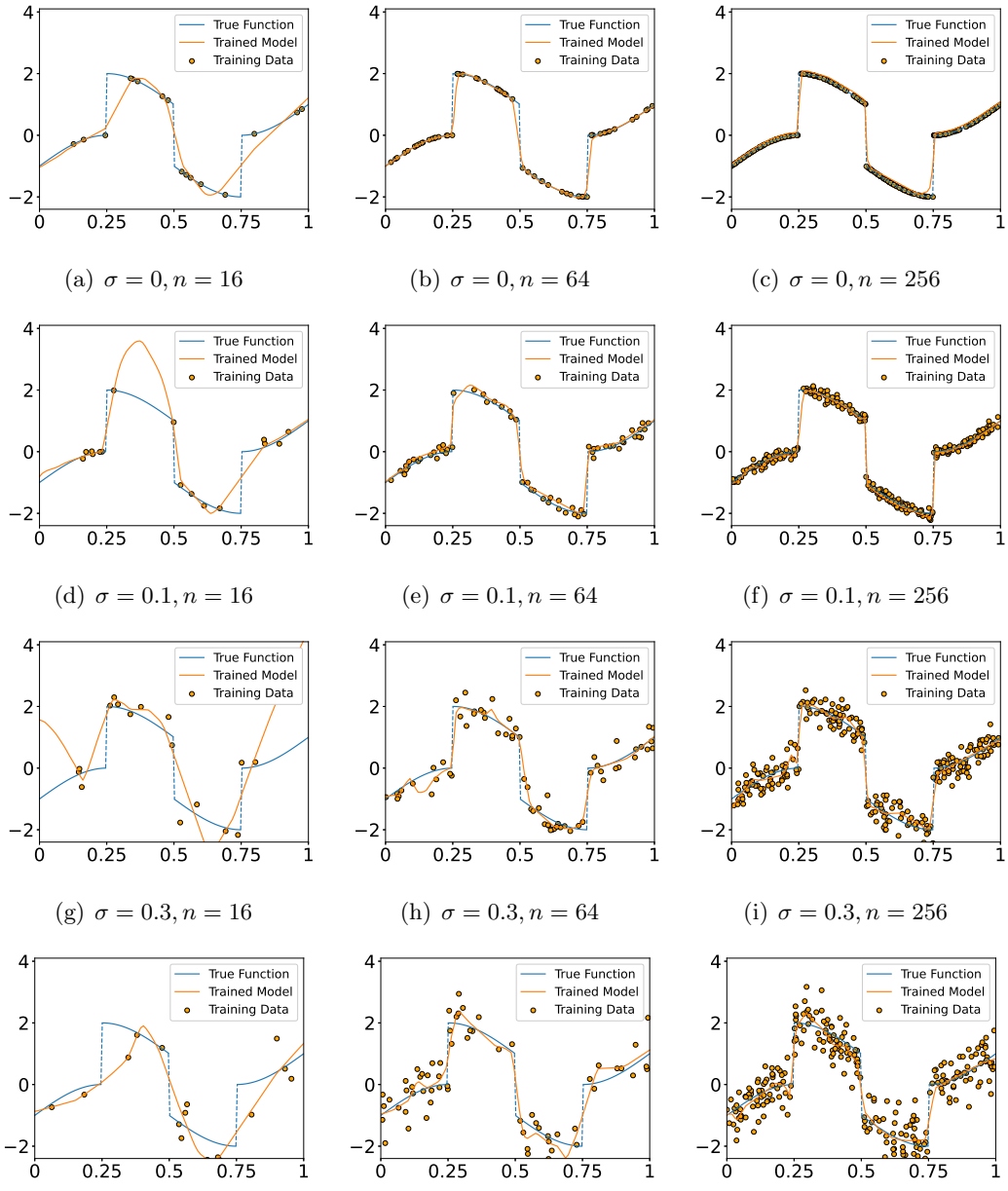
$$f(x) = \begin{cases} \sin(2\pi x) - 1 & \text{when } 0 \le x < \frac{1}{6}, \\ \sin(2\pi x) & \text{when } \frac{1}{6} \le x < \frac{1}{3}, \\ \sin(2\pi x) + 1 & \text{when } \frac{1}{3} \le x < \frac{1}{2}, \\ \sin(2\pi x) - 1 & \text{when } \frac{1}{2} \le x < 1, \\ \sin(2\pi x) & \text{when } \frac{2}{3} \le x < \frac{2}{3}, \\ \sin(2\pi x) + 1 & \text{when } \frac{5}{6} \le x < 1. \end{cases}$$

- Function with 7 discontinuity points,

$$f(x) = \begin{cases} \sin(2\pi x) - 1 & \text{when } 0 \le x < \frac{1}{8}, \\ \sin(2\pi x) - \frac{1}{3} & \text{when } \frac{1}{8} \le x < \frac{1}{4}, \\ \sin(2\pi x) + \frac{1}{3} & \text{when } \frac{1}{4} \le x < \frac{3}{8}, \\ \sin(2\pi x) + 1 & \text{when } \frac{3}{8} \le x < \frac{1}{2}, \\ \sin(2\pi x) - 1 & \text{when } \frac{1}{2} \le x < \frac{5}{8}, \\ \sin(2\pi x) - \frac{1}{3} & \text{when } \frac{5}{8} \le x < \frac{3}{4}, \\ \sin(2\pi x) + \frac{1}{3} & \text{when } \frac{3}{4} \le x < \frac{7}{8}, \\ \sin(2\pi x) + 1 & \text{when } \frac{7}{8} \le x < 1. \end{cases}$$

These functions are shown in the Figure 4. In each experiment, we sample n_{train} i.i.d. training samples $\{x_i, y_i\}_{i=1}^{n_{\text{train}}}$ according to the model in (19). Specifically, the x_i's are independently and uniformly sampled in $[0, 1]$, and $y_i = f(x_i) + \xi_i$ with $\xi_i \sim \mathcal{N}(0, \sigma^2)$ being a normal random variable with zero mean and standard deviation σ. Given the training data, we train a neural network through

$$\widehat{f} = \underset{f_{\text{NN}} \in \mathcal{F}}{\operatorname{argmin}} \frac{1}{n_{\text{train}}} \sum_{i=1}^{n_{\text{train}}} |f_{\text{NN}}(x_i) - y_i|^2, \tag{25}$$



(a) $\sigma = 0, n = 16$ (b) $\sigma = 0, n = 64$ (c) $\sigma = 0, n = 256$

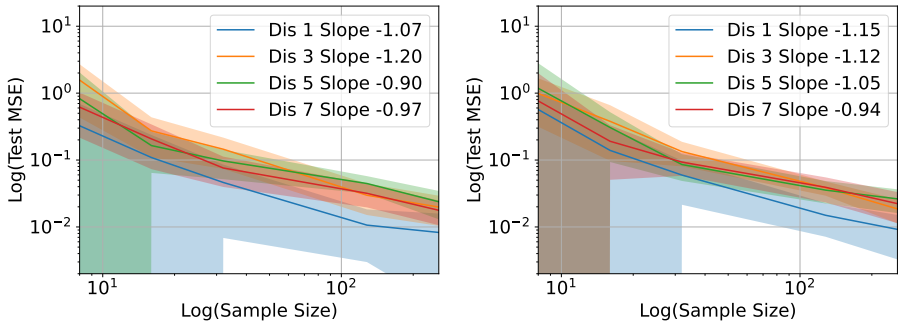
(d) $\sigma = 0.1, n = 16$ (e) $\sigma = 0.1, n = 64$ (f) $\sigma = 0.1, n = 256$

(g) $\sigma = 0.3, n = 16$ (h) $\sigma = 0.3, n = 64$ (i) $\sigma = 0.3, n = 256$

(j) $\sigma = 0.5, n = 16$ (k) $\sigma = 0.5, n = 64$ (l) $\sigma = 0.5, n = 256$

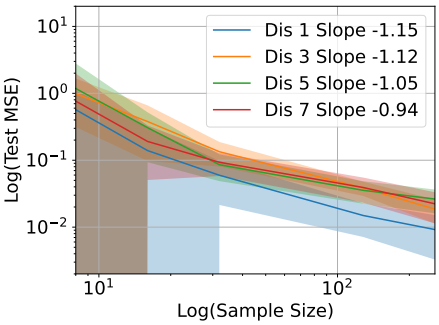
Figure 5: Trained model with $n_{\text{train}} = 16$ (1st column), 64 (2nd column), 256 (3rd column) when $\sigma = 0$ (1st row), 0.1 (2nd row), 0.3 (3rd row), and 0.5 (4th row).

where the ReLU neural network class $\mathcal{F}$ comprises four fully connected layer (64, 128, 64 neurons in each hidden layer). All weight and bias parameters are initialized with a uniform distribution, and we use Adam optimizer with learning rate 0.001 for training.

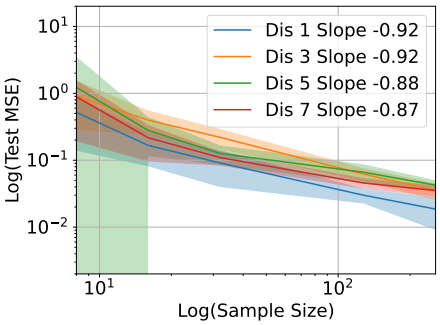


(a) $\sigma = 0$

(b) $\sigma = 0.1$



(c) $\sigma = 0.3$



(d) $\sigma = 0.5$

Figure 6: Test MSE versus n_{train} in log-log scale for the regression of functions with different number of discontinuity points (Dis) shown in Figure 4, when $\sigma = 0$(a), $\sigma = 0.1$(b), $\sigma = 0.3$(c) and $\sigma = 0.5$(d). We repeat 20 experiments for each setting. The curve represents the average test MSE in 20 experiments and the shade represents the standard deviation. A least-square fit of the curve gives rise to the slope in the legend.

Figure 5 shows the ground-truth, training data, and the trained model with different number of training data and noise levels. When σ is fixed, the trained model approaches the ground-truth function when n_{train} increases, which is consistent with Theorem 6.

The test mean squared error (MSE) is evaluated on the test samples $\{x_j, f(x_j)\}_{n_{\text{test}}}$

$$\text{Test MSE} = \frac{1}{n_{\text{test}}} \sum_{j=1}^{n_{\text{test}}} |\widehat{f}(x_j) - f(x_j)|^2$$

with $n_{\text{test}} = 10,000$. We use a large n_{test} in order to reduced the variance in the evaluation of the test MSE.

Figure 6 shows the Test MSE versus n_{train} in log-log scale for the regression of functions with different number of discontinuity points shown in Figure 4, when $\sigma = 0$ in (a), $\sigma = 0.1$ in (b), $\sigma = 0.3$ in (c) and $\sigma = 0.5$ in (d), respectively. We repeat 20 experiments for each setting. The curve represents the average test MSE in 20 experiments and the shade

represents the standard deviation. A least-square fit of the curve gives rise to the slope in the legend. These functions fit in Example 2c, which follows Test MSE $\lesssim {n_{\text{train}}}^{-\frac{2r}{2r+1}} \log^3 n$ for a large r since the functions are smooth except at the discontinuity points. Our Example 2c predicts the slope about -1 in the log-log plot of test MSE versus n_{train}, and the slope is not sensitive to the number of (finite) discontinuity points. The numerical slopes in Figure 6 are consistent with our theory.

6. Proof of Main Results

In this section, we present the proof of our main results. Some preliminaries for the proof are introduced in Subsection 6.1. Theorem 5 is proved in Subsection 6.2 and Theorem 6 is proved in Subsection 6.3.

6.1 Proof Preliminaries

We first introduce some preliminaries to be used in the proof.

6.1.1 Trapezoidal function and its neural network representation

Given an interval $[a,b] \subset [0,1]$ and $0 < \delta < b-a$, the function defined as

$$\psi_{[a,b]}(x) = \begin{cases} 0 & \text{if } x < a-\delta/2, \\ \frac{x-(a-\delta/2)}{\delta} & \text{if } a-\delta/2 \le x \le a+\delta/2, \\ 1 & \text{if } a+\delta/2 < x < b-\delta/2, \\ 1-\frac{x-(b-\delta/2)}{\delta} & \text{if } b-\delta/2 \le x \le b+\delta/2, \\ 0 & \text{if } x > b+\delta/2, \end{cases} \qquad \text{for} \quad a \neq 0,\ b \neq 1,$$

$$\psi_{[a,b]}(x) = \begin{cases} 1 & \text{if } a \le x < b-\delta/2, \\ 1-\frac{x-(b-\delta/2)}{\delta} & \text{if } b-\delta/2 \le x \le b+\delta/2, \\ 0 & \text{if } x > b+\delta/2, \end{cases} \qquad \text{for } a = 0,$$

$$\psi_{[a,b]}(x) = \begin{cases} 0 & \text{if } x < a-\delta/2, \\ \frac{x-(a-\delta/2)}{\delta} & \text{if } a-\delta/2 \le x \le a+\delta/2, \\ 1 & \text{if } a+\delta/2 < x \le b \end{cases} \qquad \text{for } b = 1 \tag{26}$$

is piecewise linear and supported on $[a-\delta/2, b+\delta/2]$ (or $[a, b+\delta/2]$ or $[a-\delta/2, b]$). In the rest of the proof, for simplicity, we only discuss the case for $0 < a < b < 1$. The case for $a = 0$ or $b = 1$ can be derived similarly. Function $\psi_{[a,b]}$ can be realized by the following ReLU network with 1 layer and width 4:

$$\widetilde{\psi}_{[a,b]}(x) = \frac{1}{\delta}\Big(\text{ReLU}(x-(a-\delta/2)) - \text{ReLU}(x-(a+\delta/2)) \\ - \text{ReLU}(x-(b-\delta/2) + \text{ReLU}(x-(b+\delta/2))\Big).$$

6.1.2 Multiplication operation and neural network approximation

The following lemma from Yarotsky (2017) shows that the product operation can be well approximated by a ReLU network.

Lemma 7 (Proposition 3 in Yarotsky (2017)) *For any $C > 0$ and $0 < \varepsilon < 1$. If $|x| \le C, |y| \le C$, there is a ReLU network, denoted by $\widetilde{\times}(\cdot,\cdot)$, such that*

$$|\widetilde{\times}(x,y) - xy| < \varepsilon,$$
$$\widetilde{\times}(x,0) = \widetilde{\times}(y,0) = 0,\ |\widetilde{\times}(x,y)| \le C^2.$$

Such a network has $O\left(\log \frac{1}{\varepsilon}\right)$ layers and parameters, where the constants hidden in O depends on C. The width of each layer is bounded by 6 and all parameters are bounded by $O(C^2)$, where the constant hidden in O is an absolute constant.

Furthermore, the following lemma shows that composition of products can be well approximated by a ReLU network (see a proof in Appendix D)

Lemma 8 *Let $\{a_i\}_{i=1}^N$ be a set of real numbers satisfying $|a_i| \le C$ for any i. For any $0 < \varepsilon < 1$, there exists a neural network $\widetilde{\Pi} \in \mathcal{F}_{\rm NN}(L, w, K, \kappa, M)$ such that*

$$|\widetilde{\Pi}(a_1, ..., a_N) - \prod_{i=1}^N a_i| \le N\varepsilon.$$

The network $\mathcal{F}_{\rm NN}(L, w, K, \kappa, M)$ has

$$L = O\left(N \log \frac{1}{\varepsilon}\right), w = N + 6, K = O\left(N \log \frac{1}{\varepsilon}\right), \kappa = O(C^N), M = C^N, \tag{27}$$

where the constant hidden in O of L and K depends on C, and κ is some absolute constant.

6.2 Proof of Theorem 5

Proof [Proof of Theorem 5] To prove Theorem 5, we first decompose the approximation error into two parts by applying the triangle inequality with the piecewise polynomial on the adaptive partition $p_{\Lambda(f,\eta,\rho)}$ defined in (7). The first part is the approximation error of f by $p_{\Lambda(f,\eta,\rho)}$, which can be bounded by (10). The second part is the network approximation error of $p_{\Lambda(f,\eta,\rho)}$. Then we show that $p_{\Lambda(f,\eta,\rho)}$ can be approximated by the given neural network with an arbitrary accuracy. Lastly, we estimate the total approximation error and quantify the network size. In the following we present the details of each step.

• **Decomposition of the approximation error.** For any $\widetilde{f}$ given by the network in (17), we decompose the error as

$$\|\widetilde{f} - f\|_{L^2(\rho)}^2 \le 2\|f - p_{\Lambda(f,\eta,\rho)}\|_{L^2(\rho)}^2 + 2\|\widetilde{f} - p_{\Lambda(f,\eta,\rho)}\|_{L^2(\rho)}^2, \tag{28}$$

where $\eta > 0$ is to be determined later. The first term in (28) can be bounded by (10) such that

$$2\|f - p_{\Lambda(f,\eta,\rho)}\|_{L^2(\rho)}^2 \le 2C_s R_{\mathcal{A}}^2 (\#\mathcal{T}(f,\eta,\rho))^{-2s}. \tag{29}$$

• **Bounding the second term in (28).** We next derive an upper bound for the second term in (28) by showing that $p_{\Lambda(f,\eta,\rho)}$ can be well approximated by a network $\widetilde{f}$. This part contains four steps:

Step 1 Estimate the finest scale of the truncated tree.

Step 2 Construct a partition of unity of X with respect to the truncated tree. Each element of the partition of unity is a network.

Step 3 Based on the partition of unity, construct a network to approximate $p_{j,k}$ on each cube.

Step 4 Estimate the approximation error.

In the following we disucss details of each step.
— **Step 1: Estimate the finest scale.** Denote the truncated tree and its outer leaves of $p_{\Lambda(f,\eta,\rho)}$ by $\mathcal{T}$ and Λ respectively for simplicity. Each $C_{j,k} \in \Lambda$ is a hypercube in the form of $\otimes_{\ell=1}^{d}[r_{\ell,j,k}, r_{\ell,j,k}+2^{-j}]$, where $r_{\ell,j,k} \in [0,1]$ are scalars and

$$\otimes_{\ell=1}^{d}[r_{\ell,j,k}, r_{\ell,j,k}+2^{-j}] = [r_{1,j,k}, r_{1,j,k}+2^{-j}] \times \cdots \times [r_{d,j,k}, r_{d,j,k}+2^{-j}]$$

is a hypercube with edge length 2^{-j} in $\mathbb{R}^d$.

We estimate the finest scale in Λ. Let $J > 0$ be the largest integer such that $C_{J,k} \in \Lambda$ for some k. In other words, 2^{-J} is the finest scale of the cubes in $\Lambda(f,\eta,\rho)$. Let $C_1 = R_{\mathcal{A}}^m$ so that

$$\sup_{\eta>0} \eta^m \#\mathcal{T}(f,\eta,\rho) \le R_{\mathcal{A}}^m = C_1,$$

with the m given in (8), which implies

$$\eta \le \left(\frac{C_1}{\#\mathcal{T}(f,\eta,\rho)}\right)^{\frac{1}{m}}. \tag{30}$$

For any $C_{j,k} \in \Lambda(f,\eta,\rho)$ and its parent $C_{j-1,k'}$, we have $\delta_{j-1,k'} > \eta$. Meanwhile, we have

$$\delta_{j-1,k'} \le 2C_\rho^{\frac{1}{2}} \|f\|_{L^\infty(X)} |C_{j-1,k'}| \le 2C_\rho^{\frac{1}{2}} R 2^{-(j-1)d/2}.$$

This implies

$$j < -\frac{2}{d}\log_2 \frac{\eta}{2C_\rho^{\frac{1}{2}} R} + 1. \tag{31}$$

Substituting (30) into (31) gives rise to

$$j \le \frac{2}{md}\log \#\mathcal{T}(f,\eta,\rho) - \frac{2}{md}\log C_1 + 1 + \frac{2}{d}\log_2(2C_\rho^{\frac{1}{2}} R) \le C_2 \log \#\mathcal{T}(f,\eta,\rho), \tag{32}$$

where C_2 is a constant depending on d, m, C_ρ and R. Since (32) holds for any $C_{j,k} \in \Lambda(f,\eta,\rho)$, we have

$$J \le C_2 \log \#\mathcal{T}(f,\eta,\rho). \tag{33}$$

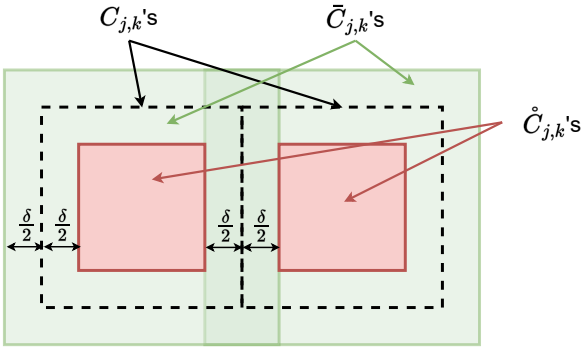


Figure 7: An illustration of the relation among $C_{j,k}, \mathring{C}_{j,k}$ and $\bar{C}_{j,k}$.

— **Step 2: Construct a partition of unity.** Let $0 < \delta \leq 2^{-(J+2)}$. For each $C_{j,k} \in \Lambda(f, \eta, \rho)$, we define two sets:

$$\begin{aligned} \mathring{C}_{j,k} &= \otimes_{\ell=1}^{d} [r_{\ell,j,k} + \delta/2, r_{\ell,j,k} + 2^{-j} - \delta/2] \cap X, \\ \bar{C}_{j,k} &= \left(\otimes_{\ell=1}^{d} [r_{\ell,j,k} - \delta/2, r_{\ell,j,k} + 2^{-j} + \delta/2] \cap X \right) \backslash \mathring{C}_{j,k}. \end{aligned} \tag{34}$$

The $\mathring{C}_{j,k}$ set is in the interior of $C_{j,k}$, with $\delta/2$ distance to the boundary of $C_{j,k}$. The $\bar{C}_{j,k}$ set contains the boundary of $C_{j,k}$. The relations of $\bar{C}_{j,k}, \mathring{C}_{j,k}$ and $C_{j,k}$ are illustrated in Figure 7.

For each $C_{j,k}$, we define the function

$$\phi_{j,k}(\mathbf{x}) = \prod_{\ell=1}^{d} \psi_{[r_{\ell,j,k}, r_{\ell,j,k}+2^{-j}]}(x_\ell),$$

where $\mathbf{x} = [x_1, ..., x_d]^\top$, and the ψ function is defined in (26). The function $\phi_{j,k}$ has the following properties:

1. $\phi_{j,k}$ is piecewise linear.
2. $\phi_{j,k}$ is supported on $\mathring{C}_{j,k} \cup \bar{C}_{j,k}$, and $\phi_{j,k}(\mathbf{x}) = 1$ when $\mathbf{x} \in \mathring{C}_{j,k}$.
3. The $\phi_{j,k}$'s form a partition of unity of X: $\sum_{C_{j,k} \in \Lambda} \phi_{j,k}(\mathbf{x}) = 1$ when $\mathbf{x} \in X = [0,1]^d$.

In this paper, we approximate $\phi_{j,k}(\mathbf{x})$ by

$$\widetilde{\phi}_{j,k}(\mathbf{x}) = \widetilde{\Pi}\left(\psi_{[r_{1,j,k}, r_{1,j,k}+2^{-j}]}(x_1), \psi_{[r_{2,j,k}, r_{2,j,k}+2^{-j}]}(x_2), \cdots, \psi_{[r_{d,j,k}, r_{d,j,k}+2^{-j}]}(x_d) \right),$$

where the network $\widetilde{\Pi}$ with d inputs is defined in Lemma 8 with accuracy $d\varepsilon_1$. We have $\widetilde{\phi}_{j,k} \in \mathcal{F}_1 = \mathcal{F}(L_1, w_1, K_1, \kappa_1, M_1)$ with

$$L_1 = O\left(d \log \frac{1}{\varepsilon_1} \right), w_1 = d + 6, K_1 = O\left(d \log \frac{1}{\varepsilon_1} \right), \kappa_1 = O\left(\frac{1}{\delta} \right), M_1 = 1.$$

By Lemma 8, we have

$$\sup_{\mathbf{x}\in[0,1]^d} |\widetilde{\phi}_{j,k}(\mathbf{x}) - \phi_{j,k}(\mathbf{x})| \leq d\varepsilon_1. \tag{35}$$

— **Step 3: Approximate $p_{j,k}$ on each cube.** According to (15), for each $C_{j,k} \in \Lambda(f,\eta,\rho)$, $p_{j,k}(f)$ is a polynomial of degree θ and is in the form of

$$p_{j,k} = \sum_{|\boldsymbol{\alpha}|\leq\theta} a_{\boldsymbol{\alpha}}(2^j(\mathbf{x}-\mathbf{r}_{j,k}))^{\boldsymbol{\alpha}}, \tag{36}$$

where $\mathbf{r}_{j,k} = [r_{1,j,k},...,r_{d,j,k}]^\top$.

We approximate $p_{j,k}$ by

$$\widetilde{p}_{j,k}(\mathbf{x}) = \sum_{|\boldsymbol{\alpha}|\leq\theta} a_{\boldsymbol{\alpha}}\widetilde{\Pi}(\underbrace{2^j(x_1-r_{1,j,k}),...,2^j(x_1-r_{1,j,k})}_{\alpha_1 \text{ times}},...,\underbrace{2^j(x_d-r_{d,j,k}),...,2^j(x_d-r_{d,j,k})}_{\alpha_d \text{ times}}),$$

where $\widetilde{\Pi}(\underbrace{2^j(x_1-r_{1,j,k}),...,2^j(x_1-r_{1,j,k})}_{\alpha_1 \text{ times}},...,\underbrace{2^j(x_d-r_{d,j,k}),...,2^j(x_d-r_{d,j,k})}_{\alpha_d \text{ times}})$ is the network approximation of $(2^j(\mathbf{x}-\mathbf{r}_{j,k}))^{\boldsymbol{\alpha}}$ with accuracy $\theta\varepsilon_1$, according to Lemma 8.

By Assumption 2(ii), there exists $R_p > 0$ so that $|a_{\boldsymbol{\alpha}}|$ is uniformly bounded by R_p for any $|\boldsymbol{\alpha}| \leq \theta$ for any $p_{j,k}$. By Lemma 8, we have

$$\begin{aligned}
&|\widetilde{p}_{j,k}(\mathbf{x}) - p_{j,k}(\mathbf{x})| \\
\leq &\sum_{|\boldsymbol{\alpha}|\leq\theta} a_{\boldsymbol{\alpha}} \left|\widetilde{\Pi}(\underbrace{2^j(x_1-r_{1,j,k}),...,2^j(x_1-r_{1,j,k})}_{\alpha_1 \text{ times}},...,\underbrace{2^j(x_d-r_{d,j,k}),...,2^j(x_d-r_{d,j,k})}_{\alpha_d \text{ times}}) - (2^j(\mathbf{x}-\mathbf{r}_{j,k}))^{\boldsymbol{\alpha}}\right| \\
\leq &C_3 R_p \theta\varepsilon_1
\end{aligned} \tag{37}$$

for some C_3 dpending on d,θ, and $\widetilde{p}_{j,k} \in \mathcal{F}_2 = \mathcal{F}(L_2,w_2,K_2,\kappa_2,M_2)$ with

$$L_2 = O(\theta\log\frac{1}{\varepsilon_1}),\ w_2 = O(\theta d^\theta),\ K_2 = O(\theta d^\theta\log\frac{1}{\varepsilon_1}),\ \kappa_2 = O(2^J),\ M_2 = R.$$

— **Step 4: Estimate the network approximation error for $p_{\Lambda(f,\eta,\rho)}$.** We approximate

$$p_{\Lambda(f,\eta,\rho)}(\mathbf{x}) = \sum_{C_{j,k}\in\Lambda(f,\eta,\rho)} p_{j,k}(\mathbf{x})$$

by

$$\widetilde{f}(\mathbf{x}) = \sum_{C_{j,k}\in\Lambda(f,\eta,\rho)} \widetilde{\times}(\widetilde{\phi}_{j,k}(\mathbf{x}),\widetilde{p}_{j,k}(\mathbf{x})), \tag{38}$$

where $\widetilde{\times}$ is the product network with accuracy ε_1, according to Lemma 7.

Denote $X_1 = \cup_{C_{j,k}\in\Lambda(f,\eta,\rho)}\mathring{C}_{j,k},\ X_2 = \cup_{C_{j,k}\in\Lambda(f,\eta,\rho)}\bar{C}_{j,k}$. The error is estimated as

$$\|\widetilde{f} - p_{\Lambda(f,\eta,\rho)}\|^2_{L^2(\rho(X))} = \|\widetilde{f} - p_{\Lambda(f,\eta,\rho)}\|^2_{L^2(\rho(X_1))} + \|\widetilde{f} - p_{\Lambda(f,\eta,\rho)}\|^2_{L^2(\rho(X_2))}. \tag{39}$$

For the first term in (39), we have

$$
\begin{aligned}
&\|\widetilde{f} - p_{\Lambda(f,\eta,\rho)}\|^2_{L^2(\rho(X_1))} \qquad (40)\\
=&\sum_{C_{j,k}\in\Lambda(f,\eta,\rho)} \int_{\mathring{C}_{j,k}} |\widetilde{\times}(\widetilde{\phi}_{j,k}(\mathbf{x}),\widetilde{p}_{j,k}(\mathbf{x})) - p_{j,k}(\mathbf{x})|^2 d\rho(\mathbf{x})\\
\leq&\sum_{C_{j,k}\in\Lambda(f,\eta,\rho)} \int_{\mathring{C}_{j,k}} 3\Big[|\widetilde{\times}(\widetilde{\phi}_{j,k}(\mathbf{x}),\widetilde{p}_{j,k}(\mathbf{x})) - \widetilde{\phi}_{j,k}(\mathbf{x})\widetilde{p}_{j,k}(\mathbf{x})|^2\\
&+ |\widetilde{\phi}_{j,k}(\mathbf{x})\widetilde{p}_{j,k}(\mathbf{x}) - \phi_{j,k}(\mathbf{x})\widetilde{p}_{j,k}(\mathbf{x})|^2 + |\phi_{j,k}(\mathbf{x})\widetilde{p}_{j,k}(\mathbf{x}) - p_{j,k}(\mathbf{x})|^2\Big] d\rho(\mathbf{x})\\
\leq&\sum_{C_{j,k}\in\Lambda(f,\eta,\rho)} \int_{\mathring{C}_{j,k}} 3\Big[\varepsilon_1^2 + R^2d^2\varepsilon_1^2 + |\widetilde{p}_{j,k}(\mathbf{x}) - p_{j,k}(\mathbf{x})|^2\Big] d\rho(\mathbf{x})\\
\leq&3(R^2d^2+1+C_3^2R_p^2\theta^2)\varepsilon_1^2 \sum_{C_{j,k}\in\Lambda(f,\eta,\rho)} \rho(\mathring{C}_{j,k})\\
\leq&3(R^2d^2+1+C_3^2R_p^2\theta^2)\varepsilon_1^2. \qquad (41)
\end{aligned}
$$

In the derivation above, we used $\phi_{j,k}(\boldsymbol{x}) = 1$ when $\mathbf{x} \in \mathring{C}_{j,k}$. Additionally, (35) is used in the second inequality, (37) is used in the third inequality.

We next derive an upper bound for the second term in (39) by bounding the volume of X_2. We define the common boundaries of a set of cubes being the outer leaves of a truncated tree as the set of points that belong to at least two cubes. We will use the following lemma to estimate the surface area for the common boundaries of $\Lambda(f,\eta,\rho)$ (see a proof in Appendix E).

Lemma 9 *Given a truncated tree $\mathcal{T}$ and its outer leaves $\Lambda_{\mathcal{T}}$, we denote the set of common boundaries of the subcubes in $\Lambda_{\mathcal{T}}$ by $\mathcal{B}(\Lambda_{\mathcal{T}})$. The surface area of $\mathcal{B}(\Lambda_{\mathcal{T}})$ in $\mathbb{R}^{d-1}$, denoted by $|\mathcal{B}(\Lambda_{\mathcal{T}})|$, satisfies*

$$|\mathcal{B}(\Lambda_{\mathcal{T}})| \leq 2^{d+1}d(\#\mathcal{T})^{1/d}. \qquad (42)$$

By Lemma 9 and Assumption 1, we have

$$\|\widetilde{f} - p_{\Lambda(f,\eta,\rho)}\|^2_{L^2(\rho(X_2))} \leq 4R^2\rho(X_2) \leq 4C_\rho R^2|X_2| \leq 4C_\rho R^2\delta|\mathcal{B}(\Lambda_{\mathcal{T}})| \leq 2^{d+3}C_\rho dR^2\delta(\#\mathcal{T})^{1/d}. \qquad (43)$$

Putting (41) and (43) together, we have

$$\|\widetilde{f} - p_{\Lambda(f,\eta,\rho)}\|^2_{L^2(\rho(X))} \leq 3(R^2d^2+1+C_3^2R_p^2\theta^2)\varepsilon_1^2 + 2^{d+3}C_\rho dR^2\delta(\#\mathcal{T})^{1/d}. \qquad (44)$$

• **Putting all terms together.** Set $\varepsilon_1 = (\#\mathcal{T})^{-s}, \delta = \min\{(\#\mathcal{T})^{-2s-1/d}, 2^{-\zeta-2}\}$, and substitute (44) and (29) into (28) gives rise to

$$
\begin{aligned}
\|\widetilde{f} - f\|^2_{L^2(X)} \leq&2C_sR_{\mathcal{A}}^2(\#\mathcal{T}(f,\eta,\rho))^{-2s} + 6(R^2d^2+1+C_3^2R_p^2\theta^2)\varepsilon_1^2 + 2^{d+3}C_\rho dR^2\delta\#\mathcal{T}^{1/d}\\
\leq&\left(2C_sR_{\mathcal{A}}^2 + 6(R^2d^2+1+C_3^2R_p^2\theta^2) + 2^{d+3}C_\rho Rd^2\right)(\#\mathcal{T}(f,\eta,\rho))^{-2s}. \qquad (45)
\end{aligned}
$$

We then quantify the network size of $\widetilde{f}$.

- $\widetilde{\times}$: The product network has depth $O(\log \frac{1}{\varepsilon_1})$, width 6, number of nonzero parameters $O(\log \frac{1}{\varepsilon_1})$, and all parameters are bounded by R^2.
- $\widetilde{\phi}_{j,k}$: Each $\widetilde{\phi}_{j,k}$ has depth $O(d \log \frac{1}{\varepsilon_1})$, width $O(1)$, number of parameters $O(d \log \frac{1}{\varepsilon_1})$, and all parameters are bounded by $O(\frac{1}{\delta})$.
- $\widetilde{p}_{j,k}$: Each $\widetilde{p}_{j,k}$ has depth $O(\theta \log \frac{1}{\varepsilon_1})$, width $O(\theta d^\theta)$, number of parameters $O(\theta d^\theta \log \frac{1}{\varepsilon_1})$, and all parameters are bounded by $O(2^\zeta)$.

Substituting the value of ε_1 and δ, we get $\widetilde{f} \in \mathcal{F}(L, w, K, \kappa, M)$ with

$$L = O(\log \#\mathcal{T}(f,\eta,\rho)),\ w = O(\#\mathcal{T}(f,\eta,\rho)),\ K = O(\#\mathcal{T}(f,\eta,\rho)\log \#\mathcal{T}(f,\eta,\rho)),$$
$$\kappa = O((\#\mathcal{T}(f,\eta,\rho))^{2s} + 2^\zeta),\ M = R. \tag{46}$$

Note that by (33) we have

$$2^\zeta < 2^{C_2} \cdot 2^{\log \#\mathcal{T}(f,\eta,\rho)} < 2^{C_2} \#\mathcal{T}(f,\eta,\rho). \tag{47}$$

We have $\kappa = O((\#\mathcal{T}(f,\eta,\rho))^{\max\{2s,1\}})$.

Setting

$$\#\mathcal{T}(f,\eta,\rho)) = \left(\frac{\sqrt{2C_s R_{\mathcal{A}}^2 + 6(R^2 d^2 + 1 + C_3^2 R_p^2 \theta^2) + 2^{d+3} C_\rho d R^2}}{\varepsilon} \right)^{\frac{1}{s}}$$

finishes the proof. ■

6.3 Proof of Theorem 6

Proof [Proof of Theorem 6] Let $\widehat{f}$ be the minimizer of (20). We decompose the squared generalization error as

$$\mathbb{E}_{\mathcal{S}} \mathbb{E}_{\mathbf{x}\sim\rho} |\widehat{f}(\mathbf{x}) - f(\mathbf{x})|^2 = \underbrace{2\mathbb{E}_{\mathcal{S}} \left[\frac{1}{n} \sum_{i=1}^n (\widehat{f}(\mathbf{x}_i) - f(\mathbf{x}_i))^2 \right]}_{\mathrm{T}_1} \tag{48}$$

$$+ \underbrace{\mathbb{E}_{\mathcal{S}} \left[\int_X (\widehat{f}(\mathbf{x}) - f(\mathbf{x}))^2 d\rho(\mathbf{x}) \right] - 2\mathbb{E}_{\mathcal{S}} \left[\frac{1}{n} \sum_{i=1}^n (\widehat{f}(\mathbf{x}_i) - f(\mathbf{x}_i))^2 d\rho(\mathbf{x}) \right]}_{\mathrm{T}_2}. \tag{49}$$

• **Bounding** T_1**.** We bound T_1 as

$$\mathrm{T}_1 = 2\mathbb{E}_{\mathcal{S}} \left[\frac{1}{n} \sum_{i=1}^n (\widehat{f}(\mathbf{x}_i) - y_i + \xi_i)^2 \right]$$

$$
\begin{aligned}
=&2\mathbb{E}_{\mathcal{S}}\left[\frac{1}{n}\sum_{i=1}^{n}(\widehat{f}(\mathbf{x}_i)-y_i)^2\right]+2\mathbb{E}_{\mathcal{S}}\left[\frac{1}{n}\sum_{i=1}^{n}\xi_i^2\right]+4\mathbb{E}_{\mathcal{S}}\left[\frac{1}{n}\sum_{i=1}^{n}\xi_i(\widehat{f}(\mathbf{x}_i)-y_i)\right]\\
=&2\mathbb{E}_{\mathcal{S}}\left[\inf_{f_{\mathrm{NN}}\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^{n}(f_{\mathrm{NN}}(\mathbf{x}_i)-y_i)^2\right)\right]+2\mathbb{E}_{\mathcal{S}}\left[\frac{1}{n}\sum_{i=1}^{n}\xi_i^2\right]+4\mathbb{E}_{\mathcal{S}}\left[\frac{1}{n}\sum_{i=1}^{n}\xi_i(\widehat{f}(\mathbf{x}_i)-f(\mathbf{x}_i)-\xi_i)\right]\\
=&2\mathbb{E}_{\mathcal{S}}\left[\inf_{f_{\mathrm{NN}}\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^{n}(f_{\mathrm{NN}}(\mathbf{x}_i)-y_i)^2\right)\right]-2\mathbb{E}_{\mathcal{S}}\left[\frac{1}{n}\sum_{i=1}^{n}\xi_i^2\right]+4\mathbb{E}_{\mathcal{S}}\left[\frac{1}{n}\sum_{i=1}^{n}\xi_i\widehat{f}(\mathbf{x}_i))\right]\\
\le&2\inf_{f_{\mathrm{NN}}\in\mathcal{F}}\mathbb{E}_{\mathcal{S}}\left[\frac{1}{n}\sum_{i=1}^{n}(f_{\mathrm{NN}}(\mathbf{x}_i)-y_i)^2\right]-2\mathbb{E}_{\mathcal{S}}\left[\frac{1}{n}\sum_{i=1}^{n}\xi_i^2\right]+4\mathbb{E}_{\mathcal{S}}\left[\frac{1}{n}\sum_{i=1}^{n}\xi_i\widehat{f}(\mathbf{x}_i))\right]\\
=&2\inf_{f_{\mathrm{NN}}\in\mathcal{F}}\mathbb{E}_{\mathcal{S}}\left[\frac{1}{n}\sum_{i=1}^{n}\left((f_{\mathrm{NN}}(\mathbf{x}_i)-f(\mathbf{x}_i)-\xi_i)^2-\xi_i^2\right)\right]+4\mathbb{E}_{\mathcal{S}}\left[\frac{1}{n}\sum_{i=1}^{n}\xi_i\widehat{f}(\mathbf{x}_i))\right]\\
=&2\inf_{f_{\mathrm{NN}}\in\mathcal{F}}\mathbb{E}_{\mathbf{x}\sim\rho}\left[(f_{\mathrm{NN}}(\mathbf{x})-f(\mathbf{x}))^2\right]+4\mathbb{E}_{\mathcal{S}}\left[\frac{1}{n}\sum_{i=1}^{n}\xi_i\widehat{f}(\mathbf{x}_i))\right]. \qquad (50)
\end{aligned}
$$

In (50), the first term is the network approximation error. The second term is a stochastic error arising from noise. By Theorem 5, we have an upper bound for the first term. Let $\varepsilon>0$. By Theorem 5, there exists a network architecture $\mathcal{F}=\mathcal{F}_{\mathrm{NN}}(L,w,K,\kappa,R)$ with

$$
L=O(\log\frac{1}{\varepsilon}),\ w=O(\varepsilon^{-\frac{1}{s}}),\ K=O(\varepsilon^{-\frac{1}{s}}\log\frac{1}{\varepsilon}),\ \kappa=O(\varepsilon^{-\max\{2,\frac{1}{s}\}}),\ M=R \qquad (51)
$$

so that there is a network function $\widetilde{f}$ with this architecture satisfying

$$
\|\widetilde{f}-f\|_{L^2(\rho)}\le\varepsilon, \qquad (52)
$$

where the constant hidden in O depends on $d,C_s,C_\rho,R,R_p,R_{\mathcal{A}},\theta$.

We thus have

$$
2\inf_{f_{\mathrm{NN}}\in\mathcal{F}}\mathbb{E}_{\mathbf{x}\sim\rho}\left[(f_{\mathrm{NN}}(\mathbf{x})-f(\mathbf{x}))^2\right]\le2\mathbb{E}_{\mathbf{x}\sim\rho}\left[(\widetilde{f}(\mathbf{x})-f(\mathbf{x}))^2\right]=2\|\widetilde{f}-f\|_{L^2(\rho)}^2\le2\varepsilon^2. \qquad (53)
$$

Let $\mathcal{N}(\delta,\mathcal{F},\|\cdot\|_{L^\infty(X)})$ be the covering number of $\mathcal{F}$ under the $\|\cdot\|_{L^\infty(X)}$ metric. Denote $\|f\|_n^2=\frac{1}{n}\sum_{i=1}^n(f(\mathbf{x}_i))^2$. The following lemma gives an upper bound for the second term in (50) (see a proof in Appendix F)

Lemma 10 *Under the conditions of Theorem 6, for any $\delta\in(0,1)$, we have*

$$
\mathbb{E}_{\mathcal{S}}\left[\frac{1}{n}\sum_{i=1}^{n}\xi_i\widehat{f}(\mathbf{x}_i))\right]\le2\sigma\left(\sqrt{\mathbb{E}_{\mathcal{S}}\left[\|\widehat{f}-f\|_n^2\right]}+\delta\right)\sqrt{\frac{2\log\mathcal{N}(\delta,\mathcal{F},\|\cdot\|_{L^\infty(X)})+3}{n}}+\delta\sigma. \qquad (54)
$$

Substituting (53) and (54) into (50) gives rise to

$$
\mathrm{T}_1=2\mathbb{E}_{\mathcal{S}}\left[\|\widehat{f}-f\|_n^2\right]
$$

$$\leq 2\varepsilon^2 + 8\sigma\left(\sqrt{\mathbb{E}_{\mathcal{S}}\left[\|\widehat{f}-f\|_n^2\right]}+\delta\right)\sqrt{\frac{2\log\mathcal{N}(\delta,\mathcal{F},\|\cdot\|_{L^\infty(X)})+3}{n}}+4\delta\sigma. \tag{55}$$

Let

$$\begin{aligned}
v &= \sqrt{\mathbb{E}_{\mathcal{S}}\left[\|\widehat{f}-f\|_n^2\right]},\\
a &= 2\sigma\sqrt{\frac{2\log\mathcal{N}(\delta,\mathcal{F},\|\cdot\|_{L^\infty(X)})+3}{n}},\\
b &= \varepsilon^2 + 4\delta\sigma\sqrt{\frac{2\log\mathcal{N}(\delta,\mathcal{F},\|\cdot\|_{L^\infty(X)})+3}{n}}+2\delta\sigma.
\end{aligned}$$

Relation (55) can be written as

$$v^2 \leq 2av + b,$$

which implies

$$v^2 \leq 4a^2 + 2b.$$

Thus we have

$$\begin{aligned}
\mathrm{T}_1 = 2v^2 \leq & 4\varepsilon^2 + \left(16\sqrt{\frac{2\log\mathcal{N}(\delta,\mathcal{F},\|\cdot\|_{L^\infty(X)})+3}{n}}+8\right)\delta\sigma\\
&+ 16\sigma^2\frac{2\log\mathcal{N}(\delta,\mathcal{F},\|\cdot\|_{L^\infty(X)})+3}{n}.
\end{aligned} \tag{56}$$

• **Bounding T_2.**

The following lemma gives an upper bound of T_2 (see a proof in Appendix G):

Lemma 11 *Under the condition of Theorem 6, we have*

$$\mathrm{T}_2 \leq \frac{35R^2}{n}\log\mathcal{N}\left(\frac{\delta}{4R},\mathcal{F},\|\cdot\|_{L^\infty(X)}\right)+6\delta. \tag{57}$$

• **Putting both terms together.** Substituting (56) and (57) into (49) gives rise to

$$\begin{aligned}
&\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\mathbf{x}\sim\rho}|\widehat{f}(\mathbf{x})-f(\mathbf{x})|^2\\
\leq & 4\varepsilon^2 + \left(16\sqrt{\frac{2\log\mathcal{N}(\delta,\mathcal{F},\|\cdot\|_{L^\infty(X)})+3}{n}}+8\right)\delta\sigma\\
&+ 16\sigma^2\frac{2\log\mathcal{N}(\delta,\mathcal{F},\|\cdot\|_{L^\infty(X)})+3}{n}+\frac{35R^2}{n}\log\mathcal{N}\left(\frac{\delta}{4R},\mathcal{F},\|\cdot\|_{L^\infty(X)}\right)+6\delta \qquad (58)\\
\leq & 4\varepsilon^2 + \left(16\sqrt{\frac{2\log\mathcal{N}\left(\frac{\delta}{4R},\mathcal{F},\|\cdot\|_{L^\infty(X)}\right)+3}{n}}+8\right)\delta\sigma\\
&+ \left(32\sigma^2+35R^2\right)\frac{\log\mathcal{N}(\frac{\delta}{4R},\mathcal{F},\|\cdot\|_{L^\infty(X)})+3}{n}+6\delta. \qquad (59)
\end{aligned}$$

The covering number of $\mathcal{F}$ can be bounded using network parameters, which is summarized in the following lemma:

Lemma 12 (Lemma 6 of Chen et al. (2019b)) *Let $\mathcal{F} = \mathcal{F}(L, w, K, \kappa, M)$ be a class of networks: $[0,1]^d \to [-M, M]$. For any $\delta > 0$, the δ-covering number of $\mathcal{F}_{\rm NN}$ is bounded by*

$$\mathcal{N}(\delta, \mathcal{F}, \|\cdot\|_{L^\infty(X)}) \le \left(\frac{2L^2(w+2)\kappa^L w^{L+1}}{\delta}\right)^K. \tag{60}$$

Substituting the network parameters in (51) into Lemma 12 gives

$$\log \mathcal{N}\left(\frac{\delta}{4R}, \mathcal{F}, \|\cdot\|_{L^\infty(X)}\right) \le C_1\left(\varepsilon^{-\frac{1}{s}}\log^3\varepsilon^{-1} + \log\varepsilon^{-1}\right), \tag{61}$$

where C_1 is some constant depending on $d, C_s, C_\rho, R, R_{\mathcal{A}}$ and θ.

Substituting (61) into (59) gives

$$\begin{aligned}
&\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\mathbf{x}\sim\rho}|\widehat{f}(\mathbf{x}) - f(\mathbf{x})|^2\\
\le& 4\varepsilon^2 + \left(16\sqrt{\frac{2C_1\left(\varepsilon^{-\frac{1}{s}}\log^3\varepsilon^{-1} + \log\delta^{-1}\right)+3}{n}} + 8\right)\delta\sigma\\
&+ \left(32\sigma^2 + 35R^2\right)\frac{C_1\left(\varepsilon^{-\frac{1}{s}}\log^3\varepsilon^{-1} + \log\delta^{-1}\right)+3}{n} + 6\delta.
\end{aligned} \tag{62}$$

Setting $\delta = 1/n$ and

$$\varepsilon = n^{-\frac{s}{2s+1}}$$

give rise to

$$\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\mathbf{x}\sim\rho}|\widehat{f}(\mathbf{x}) - f(\mathbf{x})|^2 \le C_2 n^{-\frac{2s}{1+2s}}\log^3 n, \tag{63}$$

for some C_2 depending on $\sigma, d, C_s, C_\rho, \theta, R, R_{\mathcal{A}}$.

The resulting network $\mathcal{F} = \mathcal{F}_{\rm NN}(L, w, K, \kappa, M)$ has parameters

$$L = O(\log n),\ w = O(n^{\frac{1}{2s+1}}),\ K = O(n^{\frac{1}{2s+1}}\log n),\ \kappa = O(n^{\max\{\frac{2s}{2s+1}, \frac{1}{1+2s}\}}),\ M = R. \tag{64}$$

■

6.4 Proof of the Examples in Section 3.4

6.4.1 Proof of Example 1a

Proof [Proof of Example 1a] We first estimate $\delta_{j,k}$ for every $C_{j,k}$. Let $\theta = \lceil r-1\rceil$ and $\mathbf{c}_{j,k}$ be the center of the cube $C_{j,k}$ (each coordinate of $\mathbf{c}_{j,k}$ is the midpoint of the corresponding side of $C_{j,k}$). Denote $\widetilde{p}_{j,k}^{(k)}$ be the kth order Taylor polynomial of f centered at $\mathbf{c}_{j,k}$. By analyzing the tail of the Taylor polynomial, we obtain that, for every $x \in C_{j,k}$,

$$|f(\mathbf{x}) - \widetilde{p}_{j,k}^{(\theta)}(\mathbf{x})| \le \frac{d^{\theta/2}\|f\|_{\mathcal{H}^r(X)}\|\mathbf{x}-\mathbf{c}_{j,k}\|^r}{\theta!} \le \frac{d^{\lceil r\rceil/2}\|f\|_{\mathcal{H}^r(X)}2^{-(j+1)r}}{\lceil r-1\rceil!}. \tag{65}$$

The proof of (65) is standard and can be found in Györfi et al. (2002, Lemma 11.1) and Liu and Liao (2024, Lemma 11). The point-wise error above implies the following L^2 approximation error of the $p_{j,k}$ in (4):

$$\begin{aligned}\|(f-p_{j,k})\chi_{C_{j,k}}\|_{L^2(\rho)} &\le \|(f-\widetilde{p}_{j,k}^{(\theta)})\chi_{C_{j,k}}\|_{L^2(\rho)}\\ &\le C_\rho^{\frac{1}{2}} 2^{-jd/2}\max_{x\in C_{j,k}}|f(x)-\widetilde{p}_{j,k}^{(\theta)}(x)|\\ &\le C_\rho^{\frac{1}{2}} 2^{-jd/2}\frac{d^{\lceil r\rceil/2}\|f\|_{\mathcal{H}^r(X)}2^{-(j+1)r}}{\lceil r-1\rceil!}. \end{aligned} \tag{66}$$

As a result, the refinement quantity $\delta_{j,k}$ satisfies

$$\delta_{j,k}(f)\le C_1 2^{-j(r+d/2)},\ \text{where } C_1=\frac{2^{1-r}C_\rho^{\frac{1}{2}}d^{\lceil r\rceil/2}\|f\|_{\mathcal{H}^r(X)}}{\lceil r-1\rceil!}. \tag{67}$$

Notice that C_1 depends on r,d,C_ρ and $\|f\|_{\mathcal{H}^r(X)}$. For any $\eta>0$, the nodes of $\mathcal{T}$ with $\delta_{j,k}>\eta$ satisfy $2^{-j}>(\eta/C_1)^{\frac{2}{2r+d}}$. The cardinality of $\mathcal{T}(f,\eta,\rho)$ satisfies

$$\begin{aligned}\#\mathcal{T}(f,\eta,\rho)&\lesssim 1+2^d+2^{2d}+\ldots+2^{jd}\ \text{with } 2^{-j}>(\eta/C_1)^{\frac{2}{2r+d}}\\ &\le \frac{2^d2^{jd}}{2^d-1}\le 2\cdot 2^{jd}\le 2\left(\frac{C_1}{\eta}\right)^{\frac{2d}{2r+d}}.\end{aligned}$$

Therefore, $\eta^{\frac{2}{2r/d+1}}\#\mathcal{T}(f,\eta,\rho)\le 2C_1^{\frac{2d}{2r+d}}$ for any $\eta>0$, so that $f\in\mathcal{A}^{r/d}_{\lceil r-1\rceil,\rho}$. Furthermore, since C_1 depends on r,d,C_ρ and $\|f\|_{\mathcal{H}^r(X)}$, if $\|f\|_{\mathcal{H}^r(X)}\le 1$, then $|f|_{\mathcal{A}^{r/d}_{\lceil r-1\rceil,\rho}}\le C(r,d,C_\rho)$ for some $C(r,d,C_\rho)$ depending on r,d and C_ρ.

■

6.4.2 Proof of Example 2a

Proof [Proof of Example 2a] We first estimate the $\delta_{j,k}(f)$ for every interval (1D cube) $C_{j,k}$. There are two types of intervals: the first type does not intersect with the discontinuities $\cup_k\{t_k\}$ and the second type has intersection with $\cup_k\{t_k\}$.
The first type (Type I): When $C_{j,k}\cap(\cup_k\{t_k\})=\emptyset$, we have

$$\delta_{j,k}(f)\le C_1 2^{-j(r+1/2)},\ \text{where } C_1=\frac{2^{1-r}C_\rho^{\frac{1}{2}}d^{\lceil r\rceil/2}\max_k\|f\|_{\mathcal{H}^r(t_k,t_{k+1})}}{\lceil r-1\rceil!},$$

according to (67).
The second type (Type II): When $C_{j,k}\cap(\cup_k\{t_k\})\neq\emptyset$, f is irregular on I. We have

$$\delta_{j,k}=\left\|\sum_{C_{j+1,k'}\in\mathcal{C}(C_{j,k})}p_{j+1,k'}(f)\chi_{C_{j+1,k'}}-p_{j,k}(f)\chi_{C_{j,k}}\right\|_{L^2(\rho)}$$

$$
\begin{aligned}
=&\left\| \sum_{C_{j+1,k'}\in\mathcal{C}(C_{j,k})} (p_{j+1,k'}-f)\chi_{C_{j+1,k'}} - (p_{j,k}-f)\chi_{C_{j,k}} \right\|_{L^2(\rho)}\\
\le& \sum_{C_{j+1,k'}\in\mathcal{C}(C_{j,k})} \left\|(p_{j+1,k'}-f)\chi_{C_{j+1,k'}}\right\|_{L^2(\rho)} + \left\|(p_{j,k}-f)\chi_{C_{j,k}}\right\|_{L^2(\rho)}\\
\le& \sum_{C_{j+1,k'}\in\mathcal{C}(C_{j,k})} \left\|(0-f)\chi_{C_{j+1,k'}}\right\|_{L^2(\rho)} + \left\|(0-f)\chi_{C_{j,k}}\right\|_{L^2(\rho)}\\
\le& \sum_{C_{j+1,k'}\in\mathcal{C}(C_{j,k})} \left\|f\chi_{C_{j+1,k'}}\right\|_{L^2(\rho)} + \left\|f\chi_{C_{j,k}}\right\|_{L^2(\rho)}\\
\le& 2\|f\|_{L^\infty([0,1])} C_\rho^{\frac{1}{2}} |C_{j,k}|^{\frac{1}{2}}\\
=& C_2 2^{-j/2},
\end{aligned} \tag{68}
$$

where $C_2 = 2C_\rho^{\frac{1}{2}}\|f\|_{L^\infty([0,1])}$.

For any $\eta > 0$, the master tree $\mathcal{T}$ is truncated to $\mathcal{T}(f,\eta,\rho)$. Consider the leaf node $C_{j,k} \in \mathcal{T}(f,\eta,\rho)$. The type-I leaf nodes satisfy $2^{-j} > (\eta/C_1)^{\frac{2}{2r+1}}$. There are at most 2^{j_1} leaf nodes of Type I where j_1 is the largest integer with $2^{-j_1} > (\eta/C_1)^{\frac{2}{2r+1}}$. The type-II leaf nodes in the truncated tree satisfy $2^{-j} > (\eta/C_2)^2$. Since there are at most K discontinuity points, there are at most K leaf nodes of Type II at scale j_2, where j_2 is the largest integer with $2^{-j_2} > (\eta/C_2)^2$, implying $j_2 < 2\log_2 \frac{C_2}{\eta}$. The cardinality of the outer leaf nodes of $\mathcal{T}(f,\eta,\rho)$ can be estimated as

$$
\begin{aligned}
\#\Lambda(f,\eta,\rho) &= \#\left[\text{Outer leaf nodes of } \mathcal{T}(f,\eta,\rho)\right]\\
&\le 2\cdot 2^{j_1} + 2j_2K\\
&\le 2\cdot 2^{j_1} + 4\log_2\frac{C_2}{\eta}K\\
&\le 2\left(\frac{C_1}{\eta}\right)^{\frac{2}{2r+1}} + 4K\log_2\frac{C_2}{\eta}\\
&\le 6\left(\frac{C}{\eta}\right)^{\frac{2}{2r+1}}
\end{aligned}
$$

when η is sufficiently small, where C is a constant depending on r, d, C_ρ, K and $\max_k \|f\|_{\mathcal{H}^r(t_k,t_{k+1})}$. Notice that

$$\#\mathcal{T}(f,\eta,\rho) \le \#\Lambda(f,\eta,\rho),$$

because of (9). Therefore, $\eta^{\frac{2}{2r+1}}\#\mathcal{T}(f,\eta,\rho) \le 6C^{\frac{2}{2r+1}}$ and we have $f \in \mathcal{A}^r_{\lceil r-1\rceil,\rho}$.

Furthermore, if $\max_k \|f\|_{\mathcal{H}^r(t_k,t_{k+1})} \le 1$, we have $|f|_{\mathcal{A}^r_{r-1}} \le C(r,d,C_\rho,K)$ for some $C(r,d,C_\rho,K)$ depending on r, d, C_ρ, K.

■

6.4.3 Proof of Example 3a

Proof [Proof of Example 3a] We first estimate $\delta_{j,k}(f)$ for every cube $C_{j,k}$. There are two types of cubes: the first type belongs to the interior of some Ω_t and the second type has intersection with some $\partial\Omega_t$ (the boundary of Ω_t).
The first type (Type I): When $C_{j,k} \subset \Omega_t^o$ for some t, we have

$$\delta_{j,k}(f) \le C_1 2^{-j(r+d/2)}, \text{ where } C_1 = \frac{2^{1-r} C_\rho^{\frac{1}{2}} d^{\lceil r \rceil/2} \max_t \|f\|_{\mathcal{H}^r(\Omega_t^o)}}{\lceil r-1 \rceil!},$$

according to (67).
The second type (Type II): When $C_{j,k} \cap \partial\Omega_t \neq \emptyset$ for some t, f is irregular on $C_{j,k}$. Similar to (68), we have

$$\delta_{j,k} \le C_2 2^{-jd/2},$$

where $C_2 = 2C_\rho^{\frac{1}{2}} \|f\|_{L^\infty(\Omega)}$.

For any $\eta > 0$, the master tree $\mathcal{T}$ is truncated to $\mathcal{T}(f,\eta,\rho)$. Consider the leaf node $C_{j,k} \in \mathcal{T}(f,\eta,\rho)$. The type-I leaf nodes satisfy $2^{-j} > (\eta/C_1)^{\frac{2}{2r+d}}$. There are at most $2^{j_1 d}$ leaf nodes of Type I where j_1 is the largest integer with $2^{-j_1} > (\eta/C_1)^{\frac{2}{2r+d}}$.

We next estimate the number of type-II leaf nodes. The type-II leaf nodes satisfy $2^{-j} > (\eta/C_2)^{\frac{2}{d}}$. Let j_2 be the largest integer satisfying $2^{-j_2} > (\eta/C_2)^{\frac{2}{d}}$, which implies $j_2 < \frac{2}{d}\log_2(C_\rho/\eta)$. We next count the number of dyadic cubes needed to cover $\cup_t \partial\Omega_t$, considering $\cup_t \partial\Omega_t$ has an upper Minkowski dimension $d-1$. Let $c_M(\cup_t\partial\Omega_t) = \sup_{\varepsilon>0} \mathcal{N}(\varepsilon, \cup_t\Omega_t, \|\cdot\|_\infty)\varepsilon^{d-1}$ be the Minkowski dimension constant of $\cup_t\Omega_t$. According to Definition 2, for each $j > 0$, there exists a collection of S cubes $\{V_k\}_{k=1}^S$ of edge length 2^{-j} covering $\cup_t\partial\Omega_t$ and $S \le c_M 2^{j(d-1)}$. Each V_k at most intersects with 2^{2d} dyadic cubes in the master tree $\mathcal{T}$ at scale j. Therefore, there are at most $c_M(\cup_t\partial\Omega_t)$ type-II nodes at scale j. In total, the number of type-II leaf node is no more than

$$\sum_{j=0}^{j_2} 2^{j(d-1)+2d} = c_M(\cup_t\partial\Omega_t) 2^{2d} \frac{2^{j_2(d-1)} \cdot 2^{d-1} - 1}{2^{d-1}-1} \le \bar{C} 2^{2d+1} 2^{j_2(d-1)} \tag{69}$$

for some $\bar{C}$ depending on c_M and d.

Finally, we count the outer leaf nodes of $\mathcal{T}(f,\eta,\rho)$. The cardinality of the outer leaf nodes of $\mathcal{T}(f,\eta,\rho)$ can be estimated as

$$\begin{aligned}
\#\Lambda(f,\eta,\rho) &= \#\left[\text{Outer leaf nodes of } \mathcal{T}(f,\eta,\rho)\right] \\
&\le 2^d \cdot 2^{j_1 d} + 2^d \cdot \bar{C} 2^{2d+1} 2^{j_2(d-1)} \\
&< 2^d \left(\frac{C_1}{\eta}\right)^{\frac{2d}{2r+d}} + 2^d \cdot \bar{C} 2^{2d+1} \left(\frac{C_2}{\eta}\right)^{\frac{2(d-1)}{d}} \\
&\le \widetilde{C} \eta^{-\max\left\{\frac{2d}{2r+d}, \frac{2(d-1)}{d}\right\}}
\end{aligned}$$

for some $\widetilde{C}$ depending on $r, d, c_M(\cup_t\partial\Omega_t), C_\rho$ and $\max_t \|f\|_{\mathcal{H}^r(\Omega_t^o)}$. Notice that

$$\#\mathcal{T}(f,\eta,\rho) \le \#\Lambda(f,\eta,\rho).$$

We have $\eta^{\max\left\{\frac{2d}{2r+d},\frac{2(d-1)}{d}\right\}}\#\mathcal{T}(f,\eta,\rho)\le\widetilde{C}$. Thus $f\in\mathcal{A}^s_{\lceil r-1\rceil,\rho}$ with $s=\min\left\{\frac{r}{d},\frac{1}{2(d-1)}\right\}$.

Furthermore, if $\max_t\|f\|_{\mathcal{H}^r(\Omega_t^o)}\le 1$, then

$$|f|_{\mathcal{A}^s_{\lceil r-1\rceil,\rho}}\le C(r,d,c_M(\cup_t\partial\Omega_t),C_\rho)$$

for some $C(r,d,c_M(\cup_t\partial\Omega_t),C_\rho)$ depending on $r,d,c_M(\cup_t\partial\Omega_t)$ and C_ρ. ■

6.4.4 Proof of Example 4a

Proof [Proof of Example 4a] We first estimate $\delta_{j,k}(f)$ for every cube $C_{j,k}$. There are two types of cubes: the first type belongs to Ω_δ and the second type has intersection with $\Omega_\delta^{\complement}$.
The first type (Type I): When $C_{j,k}\subset\Omega_\delta$, we have

$$\delta_{j,k}\le C_1 2^{-j(r+d/2)},$$

with the C_1 given in (67).
The second type (Type II): When $C_{j,k}\cap\Omega_\delta^{\complement}\neq\emptyset$, f may be irregular on $C_{j,k}$ but $\delta_{j,k}=0$ since $C_{j,k}\subset\Omega^{\complement}$, when j is sufficiently large.

For sufficiently small $\eta>0$, the master tree $\mathcal{T}$ is truncated to $\mathcal{T}(f,\eta,\rho)$. The size of the tree is dominated by the nodes within Ω_δ. Therefore, $\eta^{\frac{2}{2r/d+1}}\#\mathcal{T}(f,\eta,\rho)\le 2C^{\frac{2d}{2r+d}}$ and $f\in\mathcal{A}^{r/d}_{\lceil r-1\rceil,\rho}$. Furthermore, since C depends on r,d,C_ρ and $\|f\|_{\mathcal{H}^r(\Omega_\delta)}$, if $\|f\|_{\mathcal{H}^r(\Omega_\delta)}\le 1$, then $|f|_{\mathcal{A}^{r/d}_{\lceil r-1\rceil,\rho}}\le C(r,d,C_\rho)$ for some $C(r,d,C_\rho)$ depending on r,d and C_ρ.

■

6.4.5 Proof of Example 5a

Proof [Proof of Example 5a] We first estimate $\delta_{j,k}(f)$ for every cube $C_{j,k}$. There are two types of cubes: the first type intersects with Ω and the second type has no intersection with Ω.
The first type: When $I\cap\Omega\neq\emptyset$, thanks to (65), we have

$$\delta^2_{j,k}(f)\le\int_{C_{j,k}\cap\Omega}\left(\frac{d^{\lceil r\rceil/2}\|f\|_{\mathcal{H}^r(X)}2^{-(j+1)r}}{\lceil r-1\rceil!}\right)^2 d\rho=C_1^2 2^{-2jr}\rho(C_{j,k}\cap\Omega),$$

where $C_1=d^{\lceil r\rceil/2}\|f\|_{\mathcal{H}^r(X)}2^{-r}/\lceil r-1\rceil!$. We next estimate $\rho(C_{j,k}\cap\Omega)$. Since Ω is a compact d_{in}-dimensional Riemannian manifold isometrically embedded in X, Ω has a positive reach $\tau>0$ (Thäle, 2008, Proposition 14). Each $C_{j,k}$ is a d-dimensional cube of side length 2^{-j}, and therefore $C_{j,k}\cap\Omega$ is contained in an Euclidean ball of diameter $\sqrt{d}2^{-j}$. We denote ρ_Ω as the conditional measure of ρ on Ω. According to Maggioni et al. (2016, Lemma 19), when j is sufficiently small such that $\sqrt{d}2^{-j}<\tau/8$,

$$\rho_\Omega(C_{j,k}\cap\Omega)\le\left(1+\left(\frac{2\cdot\sqrt{d}2^{-j}}{\tau-2\cdot\sqrt{d}2^{-j}}\right)^2\right)^{\frac{d}{2}}\frac{\text{Vol}(B_{\sqrt{d}2^{-j}}(\mathbf{0}_{d_{\text{in}}}))}{|\Omega|}\le C_2^2 2^{-jd_{\text{in}}},\qquad(70)$$

where $B_{\sqrt{d}2^{-j}}(\mathbf{0}_{d_{\text{in}}})$ denotes the Euclidean ball of radius $\sqrt{d}2^{-j}$ centered at origin in $\mathbb{R}^{d_{\text{in}}}$, $|\Omega|$ is the surface area of Ω, and C_2 is a constant depending on d, τ and $|\Omega|$. (70) implies that

$$\delta_{j,k}(f) \leq C_1 C_2 2^{-j(r+\frac{d_{\text{in}}}{2})}.$$

The second type: When $C_{j,k} \cap \Omega = \emptyset$, $\rho_X(C_{j,k}) = 0$ and then $\delta_{j,k}(f) = 0$.

For any $\eta > 0$, the master tree $\mathcal{T}$ is truncated to $\mathcal{T}(f, \eta, \rho)$. The size of the tree is dominated by the nodes intersecting Ω. The Type-I leaf nodes with $\delta_{j,k}(f) > \eta$ satisfy $2^{-j} \gtrsim \eta^{\frac{2}{2r+d_{\text{in}}}}$. At scale j, there are at most $O(2^{jd_{\text{in}}})$ Type-I leaf nodes. The cardinality of $\mathcal{T}(f, \eta, \rho)$ satisfies

$$\#\mathcal{T}(f, \eta, \rho) \lesssim \eta^{-\frac{2d_{\text{in}}}{2r+d_{\text{in}}}}.$$

Therefore, $\sup_{\eta>0} \eta^{\frac{2d_{\text{in}}}{2r+d_{\text{in}}}} \#\mathcal{T}(f, \eta, \rho) < \infty$, so that $f \in \mathcal{A}^{r/d_{\text{in}}}_{\lceil r-1 \rceil, \rho}$. Furthermore, if $\|f\|_{\mathcal{H}^r(X)} \leq 1$, we have $|f|_{\mathcal{A}^{r/d_{\text{in}}}_{\lceil r-1 \rceil, \rho}} < C(r, d, d_{\text{in}}, \tau, |\Omega|)$ with $C(r, d, d_{\text{in}}, \tau, |\Omega|)$ depending on $r, d, d_{\text{in}}, \tau$ and $|\Omega|$.

■

7. Conclusion

In this paper, we establish approximation and generalization theories for a large function class which is defined by nonlinear tree-based approximation theory. Such a function class allows the regularity of the function to vary at different locations and scales. It covers common function classes, such as Hölder functions and discontinuous functions, such as piecewise Hölder functions. Our theory shows that deep neural networks are adaptive to nonuniform regularity of functions and data distributions at different locations and scales by minimizing the empirical risk in (20). This is different from classical nonparametric methods (such as adaptive splines, kernel regression, tree-based methods etc.) for learning functions with varying regularity, which requires a careful selection of some adaptive parameters, such as the location of knots in adaptive splines, kernel bandwidth in kernel regression, and adaptive tree structure in tree-based methods.

When deep learning is used for regression, different network architectures can give rise to very different results. The success of deep learning relies on the optimization algorithm, initialization and a proper choice of network architecture. We will leave the computational study as our future research.

Acknowledgments

Hao Liu was partially supported by the National Natural Science Foundation of China 12201530, HKRGC ECS 22302123, HKRGC GRF 12301925 and Guangdong and Hong Kong Universities "1+1+1" Joint Research Collaboration Scheme UICR0800008-24. Wenjing Liao

was partially supported by the National Science Foundation under the NSF DMS 2145167 and the U.S. Department of Energy under the DOE SC0024348.

Appendix

Appendix A. Proof of the Approximation Error in (10)

The approximation error in (10) can be proved as follows:

$$\begin{aligned}
\|f - p_{\Lambda(f,\eta,\rho)}\|^2_{L^2(\rho)} =& \sum_{C_{j,k}\notin\mathcal{T}(f,\eta,\rho)} \|\psi_{j,k}(f)\|^2_{L^2(\rho)} \\
=& \sum_{\ell\geq 0} \sum_{C_{j,k}\in\mathcal{T}(f,2^{-(\ell+1)}\eta,\rho)\setminus\mathcal{T}(f,2^{-\ell}\eta)} \|\psi_{j,k}(f)\|^2_{L^2(\rho)} \\
\leq& \sum_{\ell\geq 0} (2^{-\ell}\eta)^2 \#[\mathcal{T}(f,2^{-(\ell+1)}\eta,\rho)] \\
\leq& \sum_{\ell\geq 0} (2^{-\ell}\eta)^2 |f|^m_{\mathcal{A}^s_{\theta,\rho}} [2^{-(\ell+1)}\eta]^{-m} \\
=& [2^m \sum_{\ell\geq 0} 2^{\ell(m-2)}] |f|^m_{\mathcal{A}^s_{\theta,\rho}} \eta^{2-m} \\
=& C_s |f|^m_{\mathcal{A}^s_{\theta,\rho}} \eta^{2-m} \\
\leq& C_s |f|^2_{\mathcal{A}^s_{\theta,\rho}} (\#\mathcal{T}(f,\eta,\rho))^{-2s},
\end{aligned}$$

where the first equality follows from the definition of $\psi_{j,k}$ in (5), the first inequality follows from the fact that $\psi_{j,k}(f)$ is supported on $C_{j,k}$ and $2^{-(\ell+1)}\eta < \|\psi_{j,k}(f)\|^2_{L^2(\rho)} \leq 2^{-\ell}\eta$ for each $C_{j,k} \in \mathcal{T}(f, 2^{-(\ell+1)}\eta) \setminus \mathcal{T}(f, 2^{-\ell}\eta)$, the second inequality follows from Definition 3, and the last inequality holds since $\eta^{2-m} \leq |f|^{2-m}_{\mathcal{A}^s_{\theta,\rho}} (\#\mathcal{T}(f,\eta,\rho))^{-\frac{2-m}{m}} = |f|^{2-m}_{\mathcal{A}^s_{\theta,\rho}} (\#\mathcal{T}(f,\eta,\rho))^{-2s}$ by the definition in (8).

Appendix B. Discussion on functions with continuously varying smoothness

As discussed in Section 3.4.3, the regularity of piecewise Hölder functions with various smoothness in different pieces is dominated by the piece with the worst smoothness. A related problem is to consider Hölder functions with continuously varying smoothness.

Let us consider the one dimensional example for $\mathcal{H}^r([0,1])$ when $r(x) = 1+3x,\ x \in [0,1]$. As x varies from 0 to 1, the regularity index changes from 1 to 4. If we use the lowest smoothness to construct the tree, according to Example 1a, we have

$$\#\mathcal{T}(f,\eta,\rho) = \widetilde{C}_1 \eta^{-2/3}. \tag{71}$$

where $\widetilde{C}_1$ is a constant independent of η.

Now let us use the varying smoothness to derive a tighter bound. For any $\eta > 0$, according to the proof of Example 1a in Section 6.4.1, given a regularity r, we need j to be

the largest integer such that $2^{-j} \gtrsim \eta^{\frac{2}{2r+1}}$. Roughly speaking, we have

$$j \approx \frac{2}{2r+1} \log_2 \eta^{-1}.$$

Define

$$j(x) = \frac{2}{2r(x)+1} \log_2 \eta^{-1}. \tag{72}$$

As x varies from 0 to 1, $j(x)$ varies from $\frac{2}{3}\log_2 \eta^{-1}$ to $\frac{2}{9}\log_2 \eta^{-1}$. Note that j must be some integer. We next derive the interval length for each j, denoted by $p(j)$, which is the length of the interval needed for leaves with scale 2^{-j}. Substituting $r(x) = 1 + 3x$ into (72) gives rise to

$$j(x) = \frac{2}{3+6x} \log_2 \eta^{-1},$$

implying that

$$x = \frac{1}{6}(\frac{2}{j} \log_2 \eta^{-1} - 3).$$

We have

$$p(j) = x(j) - x(j-1) = \frac{1}{3j(j-1)} \log_2 \eta^{-1}.$$

The number of leaf nodes can be calculated as

$$\begin{aligned}
\#\Lambda(f,\eta,\rho) = \sum_{j=\frac{2}{9}\log_2 \eta^{-1}}^{\frac{2}{3}\log_2 \eta^{-1}} 2^j p(j) &= \sum_{j=\frac{2}{9}\log_2 \eta^{-1}}^{\frac{2}{3}\log_2 \eta^{-1}} 2^j \frac{1}{3j(j-1)} \log_2 \eta^{-1} \\
&\leq \frac{1}{6} \sum_{j=\frac{2}{9}\log_2 \eta^{-1}}^{\frac{2}{3}\log_2 \eta^{-1}} 2^j j^{-2} \log_2 \eta^{-1} \\
&\leq \widetilde{C}_2 \frac{2^{\frac{2}{3}\log_2 \eta^{-1}}}{\log_2^2 \eta^{-1}} \log_2 \eta^{-1} \\
&= \widetilde{C}_2 \frac{\eta^{-2/3}}{\log_2 \eta^{-1}}
\end{aligned}$$

for some $\widetilde{C}_2 > 0$. We thus have

$$\#\mathcal{T}(f,\eta,\rho) = \widetilde{C}_3 \frac{\eta^{-2/3}}{\log_2 \eta^{-1}} \tag{73}$$

for some $\widetilde{C}_3 > 0$.

By utilizing the varying smoothness, the tree size in (71) is improved by a logrithmic factor to that in (73). As $\eta \to 0$, the bound in (73) and the one in (71) are both dominated by the order of $\eta^{-2/3}$. If we consider the additional logrithmic factor, the tree size in (73) is still significantly smaller than the one in (71), which demonstrates the advantage of adaptive approximation.

Appendix C. Proof of Lemma 4

Proof [Proof of Lemma 4]

Let $\mathcal{R} = \{\mathbf{x}^{\boldsymbol{\alpha}} : |\boldsymbol{\alpha}| = \alpha_1 + \ldots + \alpha_d \leq \theta\}$, and $n_p = \#\mathcal{R}$ be the cardinality of $\mathcal{R}$. Denote $\widetilde{\Omega} = [0,1]^d$. We first index the elements in $\mathcal{R}$ according to $|\boldsymbol{\alpha}|$ in the non-decreasing order. One can obtain a set of orthonormal polynomials on $\widetilde{\Omega}$ from $\mathcal{R}$ by the Gram–Schmidt process. This set of polynomials forms an orthonormal basis for polynomials on $\widetilde{\Omega}$ with degree no more than θ. Denote this orthonormal set of polynomials by $\{\widetilde{\phi}_\ell\}_{\ell=1}^{n_p}$. Each $\widetilde{\phi}_\ell$ can be written as

$$\widetilde{\phi}_\ell(\mathbf{x}) = \sum_{|\boldsymbol{\alpha}|\leq\theta} \widetilde{b}_{\ell,\boldsymbol{\alpha}} \mathbf{x}^{\boldsymbol{\alpha}} \tag{74}$$

for some $\{\widetilde{b}_{\ell,\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha}}$. There exists a constant C_1 only depending on θ and d so that

$$|\widetilde{b}_{\ell,\boldsymbol{\alpha}}| \leq C_1 \quad \forall \ell = 1, ..., n_p \text{ and } |\boldsymbol{\alpha}| \leq \theta. \tag{75}$$

For the simplicity of notation, we denote $\Omega = C_{j,k} = [\mathbf{a}, \boldsymbol{b}]$ with $\mathbf{a} = [a_1, ..., a_d], \boldsymbol{b} = [b_1, ..., b_d]$ and $b_1 - a_1 = \cdots = b_d - a_d = h = 2^{-j}$. The idea of this proof is to obtain a set of orthonormal basis on Ω from (74), where each basis is a linear combination of monomials of $\left(\frac{\mathbf{x}-\mathbf{a}}{h}\right)$. Then the coefficients of $p_{j,k}$ can be expressed as inner product between f and each basis. Let

$$\phi_\ell(\mathbf{x}) = \frac{1}{h^{d/2}} \widetilde{\phi}_\ell \left(\frac{\mathbf{x}-\mathbf{a}}{h}\right). \tag{76}$$

The ϕ_ℓ's form a set of orthonormal polynomials on Ω, since

$$\begin{aligned}
&\langle \phi_{\ell_1}, \phi_{\ell_2} \rangle \\
=&\int_\Omega \phi_{\ell_1}(\mathbf{x}) \phi_{\ell_2}(\mathbf{x}) d\mathbf{x} \\
=&\frac{1}{h^d} \int_\Omega \widetilde{\phi}_{\ell_1} \left(\left(\frac{\mathbf{x}-\mathbf{a}}{h}\right) \right) \widetilde{\phi}_{\ell_2} \left(\left(\frac{\mathbf{x}-\mathbf{a}}{h}\right) \right) d\mathbf{x} \\
=&\frac{h^d}{h^d} \int_\Omega \widetilde{\phi}_{\ell_1} \left(\left(\frac{\mathbf{x}-\mathbf{a}}{h}\right) \right) \widetilde{\phi}_{\ell_2} \left(\left(\frac{\mathbf{x}-\mathbf{a}}{h}\right) \right) d\left(\frac{\mathbf{x}-\mathbf{a}}{h}\right) \\
=&\int_{\widetilde{\Omega}} \widetilde{\phi}_{\ell_1}(\mathbf{x}) \, \widetilde{\phi}_{\ell_2}(\mathbf{x}) \, d\mathbf{x} \\
=&\begin{cases} 1 \text{ if } \ell_1 = \ell_2, \\ 0 \text{ otherwise.} \end{cases}
\end{aligned}$$

Thus $\{\phi_\ell\}_{\ell=1}^{n_p}$ form an orthonormal basis for polynomials with degree no more than θ on Ω. The $p_{j,k}$ in (4) has the form

$$p_{j,k} = \sum_{\ell=1}^{n_p} c_\ell \phi_\ell \quad \text{with} \quad c_\ell = \int_\Omega f(\mathbf{x}) \phi_\ell(\mathbf{x}) d\mathbf{x}. \tag{77}$$

Using Hölder's inequality, we have

$$|c_\ell| = |\langle f, \phi_\ell \rangle| \leq \|f\|_{L^2(\Omega)} \|\phi_\ell\|_{L^2(\Omega)} \leq R\sqrt{|\Omega|} \leq Rh^{d/2}. \tag{78}$$

Substituting (76) and (74) into (77) gives rise to

$$p_{j,k}(\mathbf{x}) = \sum_{\ell=1}^{n_p} c_k \sum_{|\boldsymbol{\alpha}|\leq\theta} \frac{\widetilde{b}_{k,\boldsymbol{\alpha}}}{h^{d/2}} \left(\frac{\mathbf{x}-\mathbf{a}}{h}\right)^{\boldsymbol{\alpha}} = \sum_{|\boldsymbol{\alpha}|\leq\theta} \left(\sum_{\ell=1}^{n_p} \frac{c_\ell \widetilde{b}_{\ell,\boldsymbol{\alpha}}}{h^{d/2}}\right) \left(\frac{\mathbf{x}-\mathbf{a}}{h}\right)^{\boldsymbol{\alpha}},$$

implying that

$$a_{\boldsymbol{\alpha}} = \sum_{\ell=1}^{n_p} \frac{c_\ell \widetilde{b}_{\ell,\boldsymbol{\alpha}}}{h^{d/2}}.$$

Putting (75) and (78) together, we have

$$|a_{\boldsymbol{\alpha}}| \leq \sum_{k=1}^{n_p} \frac{|c_k||\widetilde{b}_{k,\boldsymbol{\alpha}}|}{h^{d/2}} \leq \sum_{k=1}^{n_p} \frac{C_1 R h^{d/2}}{h^{d/2}} = C_1 n_p R, \tag{79}$$

where C_1, as defined in (75), is a constant depending on θ.

Furthermore, since $\| \left(\frac{\mathbf{x}-\mathbf{a}}{h}\right)^{\boldsymbol{\alpha}} \|_\infty \leq 1$, we have

$$\|p_{j,k}\|_{L^\infty(\Omega)} \leq C_1 n_p^2 R.$$

■

Appendix D. Proof of Lemma 8

Proof [Proof of Lemma 8] Denote the product $\times(a,b) = a \times b$. Let $\widetilde{\times}(\cdot,\cdot)$ be the network specified in Lemma 8 with accuracy ε. We construct

$$\widetilde{\Pi}(a_1, ..., a_N) = \widetilde{\times}(a_1, \widetilde{\times}(a_2, \widetilde{\times}(\cdots, \widetilde{\times}(a_{N-1}, a_N)\cdots))) \tag{80}$$

to approximate the multiplication operation $\prod_{i=1}^N a_i = a_1 \times a_2 \times \cdots \times a_N$. The approximation error can be bounded as

$$\begin{aligned}
&|\widetilde{\Pi}(a_1, ..., a_N) - \prod_{i=1}^N a_i| \\
=&|\widetilde{\times}(a_1, \widetilde{\times}(a_2, \widetilde{\times}(\cdots, \widetilde{\times}(a_{N-1}, a_N)\cdots))) - \times(a_1, \times(a_2, \times(\cdots, \times(a_{N-1}, a_N)\cdots)))| \\
\leq&|\widetilde{\times}(a_1, \widetilde{\times}(a_2, \widetilde{\times}(\cdots, \widetilde{\times}(a_{N-1}, a_N)\cdots))) - \times(a_1, \widetilde{\times}(a_2, \widetilde{\times}(\cdots, \widetilde{\times}(a_{N-1}, a_N)\cdots)))| \\
&+ |\times(a_1, \widetilde{\times}(a_2, \widetilde{\times}(\cdots, \widetilde{\times}(a_{N-1}, a_N)\cdots))) - \times(a_1, \times(a_2, \widetilde{\times}(\cdots, \widetilde{\times}(a_{N-1}, a_N)\cdots)))| \\
&+ \cdots \\
&+ |\times(a_1, \times(a_2, \times(\cdots, \widetilde{\times}(a_{N-1}, a_N)\cdots))) - \times(a_1, \times(a_2, \times(\cdots, \times(a_{N-1}, a_N)\cdots)))| \\
\leq&N\varepsilon.
\end{aligned} \tag{81}$$

The network size is specified in (27). ■

Appendix E. Proof of Lemma 9

Proof [Proof of Lemma 9]

Let j^* be the smallest integer so that $2^{dj^*} \geq \#\Lambda_{\mathcal{T}}$. Based on $\mathcal{T}$, we first construct a $\mathcal{T}'$ so that $\#\Lambda_{\mathcal{T}'} = \#\Lambda_{\mathcal{T}}$ and $j \leq j^*$ for any $C_{j,k} \in \Lambda_{\mathcal{T}'}$ by the following procedure.

Note that if there exists $C_{j_1,k} \in \Lambda_{\mathcal{T}}$ with $j_1 > j^*$, there must be a $C_{j_2,k'} \in \Lambda_{\mathcal{T}}$ with $j_2 < j^*$. Otherwise, we must have $\#\Lambda_{\mathcal{T}} > 2^{dj^*}$, contradicting to the definition of j^*. Let $C_{j_1,k}$ be a subcube in the finest scale of $\Lambda_{\mathcal{T}}$, and $C_{j_2,k'}$ be a subcube in the coarsest scale of $\Lambda_{\mathcal{T}}$. Suppose $j_1 > j^*$. We have $j_1 - j_2 \geq 2$. Denote the set of children and the parent of $C_{j,k}$ by $\mathcal{C}(C_{j,k})$ and $\mathcal{P}(C_{j,k})$, respectively. Since $C_{j_1,k}$ is at the finest scale, we have $\mathcal{C}(\mathcal{P}(C_{j_1,k})) \subset \Lambda_{\mathcal{T}}$. By replacing $\mathcal{C}(\mathcal{P}(C_{j_1,k}))$ by $\mathcal{P}(C_{j_1,k})$ and $C_{j_2,k'}$ by $\mathcal{C}(C_{j_2,k'})$, we obtain a new tree $\mathcal{T}_1$ with $\#\Lambda_{\mathcal{T}_1} = \#\Lambda_{\mathcal{T}} \leq 2^{dj^*}$. Note that the subcubes in $\mathcal{C}(\mathcal{P}(C_{j_1,k}))$ have side length 2^{-j_1}, $C_{j_2,k'}$ has side length 2^{-j_2}. Since $j_1 - j_2 \geq 2$, we have

$$|\mathcal{B}(\Lambda_{\mathcal{T}_1})| - |\mathcal{B}(\Lambda_{\mathcal{T}})| = d2^{-j_2(d-1)} - d2^{-j_1(d-1)} > 0, \tag{82}$$

implying that $|\mathcal{B}(\Lambda_{\mathcal{T}})| < |\mathcal{B}(\Lambda_{\mathcal{T}_1})|$. Replace $\mathcal{T}$ by $\mathcal{T}_1$ and repeat the above procedure, we can generate a set of trees $\{\mathcal{T}_m\}_{m=1}^M$ for some $M > 0$ until $j \leq j^*$ for all $C_{j,k} \in \Lambda_{\mathcal{T}_M}$. We have

$$|\mathcal{B}(\Lambda_{\mathcal{T}})| < |\mathcal{B}(\Lambda_{\mathcal{T}_1})| < \cdots < |\mathcal{B}(\Lambda_{\mathcal{T}_M})|.$$

All leaf nodes of $\mathcal{T}_M$ are at scale no larger than $j^* - 1$. For any leaf node $C_{j,k}$ of $\mathcal{T}_M$ with $j < j^* - 1$, we partition it into its children. Repeat this process until all leaf nodes of the tree is at scale $j^* - 1$, and denote the tree by $\mathcal{T}^*$. Note that doing so only creates additional common boundaries, thus $\mathcal{B}(\Lambda_{\mathcal{T}_M}) \subset \mathcal{B}(\Lambda_{\mathcal{T}^*})$, and we have

$$|\mathcal{B}(\Lambda_{\mathcal{T}})| < |\mathcal{B}(\Lambda_{\mathcal{T}_M})| < |\mathcal{B}(\Lambda_{\mathcal{T}^*})|.$$

We next compute $|\mathcal{B}(\Lambda_{\mathcal{T}^*})|$. Note that $\mathcal{B}(\Lambda_{\mathcal{T}^*})$ can be generated sequentially by slicing each cube at scale j for $j = 0, ..., j^* - 1$. When $C_{j-1,k}$ is sliced to get cubes at scale j, d hyper-surfaces with area $2^{-(j-1)(d-1)}$ are created as common boundaries. There are in total $2^{d(j-1)}$ cubes at scale j. Thus we compute $|\mathcal{B}(\Lambda_{\mathcal{T}^*})|$ as:

$$\begin{aligned}
|\mathcal{B}(\Lambda_{\mathcal{T}^*})| &= \sum_{j=1}^{j^*} d2^{-(j-1)(d-1)}2^{d(j-1)} \\
&= \sum_{j=1}^{j^*} d2^{j-1} = d(2^{j^*} - 1) \leq d2^{j^*} \leq 2^d d(\#\Lambda_{\mathcal{T}})^{1/d} \leq 2^{d+1} d(\#\mathcal{T})^{1/d},
\end{aligned} \tag{83}$$

where we used $\#\Lambda_{\mathcal{T}} \leq 2^{dj^*} \leq 2^d \#\Lambda_{\mathcal{T}}$ in the second inequality according to the definition of j^*. ■

Appendix F. Proof of Lemma 10

Proof [Proof of Lemma 10] Let $\mathcal{F}^* = \{f_{\text{NN},j}\}_{j=1}^{\mathcal{N}(\delta,\mathcal{F},\|\cdot\|_{L^\infty(X)})}$ be a δ cover of $\mathcal{F}$. There exists $f_{\text{NN}}^* \in \mathcal{F}^*$ satisfying $\|f_{\text{NN}}^* - \widehat{f}\|_{L^\infty(X)} \leq \delta$.

We have

$$\begin{aligned}
\mathbb{E}_{\mathcal{S}}\left[\frac{1}{n}\sum_{i=1}^{n}\xi_i\widehat{f}(\mathbf{x}_i))\right] =& \mathbb{E}_{\mathcal{S}}\left[\frac{1}{n}\sum_{i=1}^{n}\xi_i(\widehat{f}(\mathbf{x}_i)-f^*_{\mathrm{NN}}(\mathbf{x}_i)+f^*_{\mathrm{NN}}(\mathbf{x}_i)-f(\mathbf{x}_i))\right]\\
\leq& \mathbb{E}_{\mathcal{S}}\left[\frac{1}{n}\sum_{i=1}^{n}\xi_i(f^*_{\mathrm{NN}}(\mathbf{x}_i)-f(\mathbf{x}_i))\right]+\mathbb{E}_{\mathcal{S}}\left[\frac{1}{n}\sum_{i=1}^{n}|\xi_i|\left|\widehat{f}(\mathbf{x}_i)-f^*_{\mathrm{NN}}(\mathbf{x}_i)\right|\right]\\
\leq& \mathbb{E}_{\mathcal{S}}\left[\frac{\|f^*_{\mathrm{NN}}-f\|_n}{\sqrt{n}}\frac{\sum_{i=1}^{n}\xi_i\left(f^*_{\mathrm{NN}}(\mathbf{x}_i)-f^*_{\mathrm{NN}}(\mathbf{x}_i)\right)}{\sqrt{n}\|f^*_{\mathrm{NN}}-f\|_n}\right]+\delta\sigma\\
\leq& \sqrt{2}\mathbb{E}_{\mathcal{S}}\left[\frac{\|\widehat{f}-f\|_n+\delta}{\sqrt{n}}\left|\frac{\sum_{i=1}^{n}\xi_i\left(f^*_{\mathrm{NN}}(\mathbf{x}_i)-f^*_{\mathrm{NN}}(\mathbf{x}_i)\right)}{\sqrt{n}\|f^*_{\mathrm{NN}}-f\|_n}\right|\right]+\delta\sigma. \qquad (84)
\end{aligned}$$

In (84), the first inequality follows from Cauchy-Schwarz inequality, the second inequality holds by Jensen's inequality and

$$\begin{aligned}
\mathbb{E}_{\mathcal{S}}\left[\frac{1}{n}\sum_{i=1}^{n}|\xi_i|\left|\widehat{f}(\mathbf{x}_i)-f^*_{\mathrm{NN}}(\mathbf{x}_i)\right|\right] \leq& \mathbb{E}_{\mathcal{S}}\left[\frac{1}{n}\sum_{i=1}^{n}|\xi_i|\left\|\widehat{f}(\mathbf{x}_i)-f^*_{\mathrm{NN}}(\mathbf{x}_i)\right\|_{L^\infty(X)}\right]\\
\leq& \delta\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\mathcal{S}}\left[\sqrt{\xi_i^2}\right]\\
\leq& \delta\frac{1}{n}\sum_{i=1}^{n}\sqrt{\mathbb{E}_{\mathcal{S}}\left[\xi_i^2\right]}\\
\leq& \delta\frac{1}{n}\sum_{i=1}^{n}\sqrt{\sigma^2}\\
=& \delta\sigma, \qquad (85)
\end{aligned}$$

the last inequality holds since

$$\begin{aligned}
\|f^*_{\mathrm{NN}}-f\|_n =& \sqrt{\frac{1}{n}\sum_{i=1}^{n}(f^*_{\mathrm{NN}}(\mathbf{x}_i)-\widehat{f}(\mathbf{x}_i)+\widehat{f}(\mathbf{x}_i)-f(\mathbf{x}_i))^2}\\
\leq& \sqrt{\frac{2}{n}\sum_{i=1}^{n}\left[(f^*_{\mathrm{NN}}(\mathbf{x}_i)-\widehat{f}(\mathbf{x}_i))^2+(\widehat{f}(\mathbf{x}_i)-f(\mathbf{x}_i))^2\right]}\\
\leq& \sqrt{\frac{2}{n}\sum_{i=1}^{n}\left[\delta^2+(\widehat{f}(\mathbf{x}_i)-f(\mathbf{x}_i))^2\right]}\\
\leq& \sqrt{2}\|\widehat{f}-f\|_n+\sqrt{2}\delta. \qquad (86)
\end{aligned}$$

Denote $z_j=\frac{\sum_{i=1}^{n}\xi_i(\widehat{f}(\mathbf{x}_i)-f^*_{\mathrm{NN}}(\mathbf{x}_i))}{\sqrt{n}\|f^*_{\mathrm{NN}}-f\|_n}$. The first term in (84) can be bounded as

$$\sqrt{2}\mathbb{E}_{\mathcal{S}}\left[\frac{\|\widehat{f}-f\|_n+\delta}{\sqrt{n}}\left|\frac{\sum_{i=1}^{n}\xi_i(\widehat{f}(\mathbf{x}_i)-f^*_{\mathrm{NN}}(\mathbf{x}_i))}{\sqrt{n}\|f^*_{\mathrm{NN}}-f\|_n}\right|\right]$$

$$
\begin{aligned}
\leq & \sqrt{2}\mathbb{E}_{\mathcal{S}}\left[\frac{\|\widehat{f}-f\|_n+\delta}{\sqrt{n}}\max_j|z_j|\right]\\
= & \sqrt{2}\mathbb{E}_{\mathcal{S}}\left[\frac{\|\widehat{f}-f\|_n}{\sqrt{n}}\max_j|z_j|+\frac{\delta}{\sqrt{n}}\max_j|z_j|\right]\\
= & \sqrt{2}\mathbb{E}_{\mathcal{S}}\left[\sqrt{\frac{1}{n}\|\widehat{f}-f\|_n^2}\sqrt{\max_j|z_j|^2}+\frac{\delta}{\sqrt{n}}\sqrt{\max_j|z_j|^2}\right]\\
\leq & \sqrt{2}\sqrt{\frac{1}{n}\mathbb{E}_{\mathcal{S}}\left[\|\widehat{f}-f\|_n^2\right]}\sqrt{\mathbb{E}_{\mathcal{S}}\left[\max_j|z_j|^2\right]}+\frac{\delta}{\sqrt{n}}\sqrt{\mathbb{E}_{\mathcal{S}}\left[\max_j|z_j|^2\right]}\\
= & \sqrt{2}\left(\sqrt{\frac{1}{n}\mathbb{E}_{\mathcal{S}}\left[\|\widehat{f}-f\|_n^2\right]}+\frac{\delta}{\sqrt{n}}\right)\sqrt{\mathbb{E}_{\mathcal{S}}\left[\max_j|z_j|^2\right]},
\end{aligned}
\tag{87}
$$

where the second inequality comes from Jensen's inequality and Cauchy-Schwarz inequality.

For given $\{\mathbf{x}_i\}_{i=1}^n$, z_j is a sub-Gaussian variable with variance proxy σ^2.

For any $t>0$, we have

$$
\begin{aligned}
\mathbb{E}_{\mathcal{S}}\left[\max_j|z_j|^2|\mathbf{x}_1,...,\mathbf{x}_n\right] = & \frac{1}{t}\log\exp\left(t\mathbb{E}_{\mathcal{S}}\left[\max_j|z_j|^2|\mathbf{x}_1,...,\mathbf{x}_n\right]\right)\\
\leq & \frac{1}{t}\log\mathbb{E}_{\mathcal{S}}\left[\exp\left(t\max_j|z_j|^2\right)|\mathbf{x}_1,...,\mathbf{x}_n\right]\\
\leq & \frac{1}{t}\log\mathbb{E}_{\mathcal{S}}\left[\sum_j\exp\left(t|z_j|^2\right)|\mathbf{x}_1,...,\mathbf{x}_n\right]\\
\leq & \frac{1}{t}\log\mathcal{N}(\delta,\mathcal{F},\|\cdot\|_{L^\infty(X)})+\frac{1}{t}\log\mathbb{E}_{\mathcal{S}}\left[\exp(t|z_1|^2)|\mathbf{x}_1,...,\mathbf{x}_n\right].
\end{aligned}
\tag{88}
$$

Since z_1 is a sub-Gaussian variable with parameter σ, we have

$$
\begin{aligned}
\mathbb{E}_{\mathcal{S}}\left[\exp(t|z_1|^2)|\mathbf{x}_1,...,\mathbf{x}_n\right] = & 1+\sum_{k=1}^{\infty}\frac{t^k\mathbb{E}_{\mathcal{S}}\left[z_1^{2k}|\mathbf{x}_1,...,\mathbf{x}_n\right]}{k!}\\
= & 1+\sum_{k=1}^{\infty}\frac{t^k}{k!}\int_0^{\infty}\mathbb{P}\left(|z_1|\geq\lambda^{\frac{1}{2k}}|\mathbf{x}_1,...,\mathbf{x}_n\right)d\lambda\\
\leq & 1+2\sum_{k=1}^{\infty}\frac{t^k}{k!}\int_0^{\infty}\exp\left(-\frac{\lambda^{1/k}}{2\sigma^2}\right)d\lambda\\
= & 1+\sum_{k=1}^{\infty}\frac{2k(2t\sigma^2)^k}{k!}\Gamma_G(k)\\
= & 1+2\sum_{k=1}^{\infty}(2t\sigma^2)^k,
\end{aligned}
\tag{89}
$$

where Γ_G represents the Gamma function. Set $t = (4\sigma^2)^{-1}$, we have

$$\begin{aligned}\mathbb{E}_{\mathcal{S}}\left[\max_j |z_j|^2 | \mathbf{x}_1, ..., \mathbf{x}_n\right] \leq & 4\sigma^2 \log \mathcal{N}(\delta, \mathcal{F}, \|\cdot\|_{L^\infty(X)}) + 4\sigma^2 \log 3 \\ \leq & 4\sigma^2 \log \mathcal{N}(\delta, \mathcal{F}, \|\cdot\|_{L^\infty(X)}) + 6\sigma^2. \end{aligned} \tag{90}$$

Combining (84), (87) and (90) proves the lemma. ■

Appendix G. Proof of Lemma 11

Proof [Proof of Lemma 11] Denote $\widehat{g}(\mathbf{x}) = (\widehat{f}(\mathbf{x}) - f(\mathbf{x}))^2$. We have $\|\widehat{g}\|_{L^\infty(X)} \leq 4R^2$. The term T_2 can be written as

$$\begin{aligned}\mathrm{T}_2 =& \mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\mathbf{x}\sim\rho}[\widehat{g}(\mathbf{x})|\mathcal{S}] - \frac{2}{n}\sum_{i=1}^n \widehat{g}(\mathbf{x}_i)\right] \\ =& 2\mathbb{E}_{\mathcal{S}}\left[\frac{1}{2}\mathbb{E}_{\mathbf{x}\sim\rho}[\widehat{g}(\mathbf{x})|\mathcal{S}] - \frac{1}{n}\sum_{i=1}^n \widehat{g}(\mathbf{x}_i)\right] \\ =& 2\mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\mathbf{x}\sim\rho}[\widehat{g}(\mathbf{x})|\mathcal{S}] - \frac{1}{n}\sum_{i=1}^n \widehat{g}(\mathbf{x}_i) - \frac{1}{2}\mathbb{E}_{\mathbf{x}\sim\rho}[\widehat{g}(\mathbf{x})|\mathcal{S}]\right]. \end{aligned} \tag{91}$$

A lower bound of $\frac{1}{2}\mathbb{E}_{\mathbf{x}\sim\rho}[\widehat{g}(\mathbf{x})|\mathcal{S}]$ is derived as

$$\mathbb{E}_{\mathbf{x}\sim\rho_X}[\widehat{g}(\mathbf{x})|\mathcal{S}] = \mathbb{E}_{\mathbf{x}\sim\rho}\left[\frac{4R^2}{4R^2}\widehat{g}(\mathbf{x})|\mathcal{S}\right] \geq \frac{1}{4R^2}\mathbb{E}_{\mathbf{x}\sim\rho}[\widehat{g}^2(\mathbf{x})|\mathcal{S}]. \tag{92}$$

Substituting (92) into (91) gives rise to

$$\mathrm{T}_2 \leq 2\mathbb{E}_{\mathcal{S}}\left[\mathbb{E}_{\mathbf{x}\sim\rho}[\widehat{g}(\mathbf{x})|\mathcal{S}] - \frac{1}{n}\sum_{i=1}^n \widehat{g}(\mathbf{x}_i) - \frac{1}{8R^2}\mathbb{E}_{\mathbf{x}\sim\rho}[\widehat{g}^2(\mathbf{x})|\mathcal{S}]\right]. \tag{93}$$

Define the set

$$\mathcal{R} = \{g : g(\mathbf{x}) = (f_{\mathrm{NN}}(\mathbf{x}) - f(\mathbf{x})) \text{ for } f_{\mathrm{NN}} \in \mathcal{F}\}. \tag{94}$$

Denote $\mathcal{S}' = \{\mathbf{x}_i'\}_{i=1}^n$ be an independent copy of $\mathcal{S}$. We have

$$\begin{aligned}\mathrm{T}_2 \leq & 2\mathbb{E}_{\mathcal{S}}\left[\sup_{g\in\mathcal{R}}\left(\mathbb{E}_{\mathcal{S}'}\left[\frac{1}{n}\sum_{i=1}^n g(\mathbf{x}_i')\right] - \frac{1}{n}\sum_{i=1}^n g(\mathbf{x}_i) - \frac{1}{8R^2}\mathbb{E}_{\mathcal{S}'}\left[\frac{1}{n}\sum_{i=1}^n g^2(\mathbf{x}_i')\right]\right)\right] \\ \leq & 2\mathbb{E}_{\mathcal{S}}\left[\sup_{g\in\mathcal{R}}\left(\mathbb{E}_{\mathcal{S}'}\left[\frac{1}{n}\sum_{i=1}^n \left(g(\mathbf{x}_i') - g(\mathbf{x}_i)\right)\right] - \frac{1}{16R^2}\mathbb{E}_{\mathcal{S},\mathcal{S}'}\left[\frac{1}{n}\sum_{i=1}^n \left(g^2(\mathbf{x}_i) + g^2(\mathbf{x}_i')\right)\right]\right)\right] \\ \leq & 2\mathbb{E}_{\mathcal{S},\mathcal{S}'}\left[\sup_{g\in\mathcal{R}}\left(\frac{1}{n}\sum_{i=1}^n\left(\left(g(\mathbf{x}_i) - g(\mathbf{x}_i')\right) - \frac{1}{16R^2}\mathbb{E}_{\mathcal{S},\mathcal{S}'}\left[g^2(\mathbf{x}_i) + g^2(\mathbf{x}_i')\right]\right)\right)\right]. \end{aligned} \tag{95}$$

Let $\mathcal{R}^* = \left\{g_j^*\right\}_{j=1}^{\mathcal{N}(\delta,\mathcal{R},\|\cdot\|_{L^\infty(X)})}$ be a δ-cover of $\mathcal{R}$. For any $g \in \mathcal{R}$, there exists $g^* \in \mathcal{R}^*$ such that $\|g - g^*\|_{L^\infty(X)} \le \delta$.

We bound (95) using $\mathcal{R}^*$. The first term in (95) can be bounded as

$$\begin{aligned} g(\mathbf{x}_i) - g(\mathbf{x}_i') =& g(\mathbf{x}_i) - g^*(\mathbf{x}_i) + g^*(\mathbf{x}_i) - g^*(\mathbf{x}_i) + g^*(\mathbf{x}_i') - g(\mathbf{x}_i') \\ =& \left(g(\mathbf{x}_i) - g^*(\mathbf{x}_i)\right) + \left(g^*(\mathbf{x}_i) - g^*(\mathbf{x}_i')\right) + \left(g^*(\mathbf{x}_i') - g(\mathbf{x}_i')\right) \\ \le& \left(g^*(\mathbf{x}_i) - g^*(\mathbf{x}_i')\right) + 2\delta. \end{aligned} \tag{96}$$

We then lower bound $g^2(\mathbf{x}_i) + g^2(\mathbf{x}_i')$ as

$$\begin{aligned} & g^2(\mathbf{x}_i) + g^2(\mathbf{x}_i') \\ =& \left(g^2(\mathbf{x}_i) - (g^*)^2(\mathbf{x}_i)\right) + \left((g^*)^2(\mathbf{x}_i) - (g^*)^2(\mathbf{x}_i')\right) + \left((g^*)^2(\mathbf{x}_i') - g^2(\mathbf{x}_i')\right) \\ \ge& (g^*)^2(\mathbf{x}_i) + (g^*)^2(\mathbf{x}_i') - |g(\mathbf{x}_i) - g^*(\mathbf{x}_i)||g(\mathbf{x}_i) + g^*(\mathbf{x}_i)| - |g^*(\mathbf{x}_i') - g(\mathbf{x}_i')||g^*(\mathbf{x}_i') + g(\mathbf{x}_i')| \\ \ge& (g^*)^2(\mathbf{x}_i) + (g^*)^2(\mathbf{x}_i') - 16R^2\delta. \end{aligned} \tag{97}$$

Substituting (96) and (97) into (95) gives rise to

$$\begin{aligned} \mathrm{T}_2 \le& 2\mathbb{E}_{\mathcal{S},\mathcal{S}'}\left[\sup_{g^*\in\mathcal{R}^*}\left(\frac{1}{n}\sum_{i=1}^n\left(\left(g^*(\mathbf{x}_i) - g^*(\mathbf{x}_i')\right) - \frac{1}{16R^2}\mathbb{E}_{\mathcal{S},\mathcal{S}'}\left[(g^*)^2(\mathbf{x}_i) + (g^*)^2(\mathbf{x}_i')\right]\right)\right)\right] + 6\delta \\ =& 2\mathbb{E}_{\mathcal{S},\mathcal{S}'}\left[\max_{j}\left(\frac{1}{n}\sum_{i=1}^n\left(\left(g^*(\mathbf{x}_i) - g^*(\mathbf{x}_i')\right) - \frac{1}{16R^2}\mathbb{E}_{\mathcal{S},\mathcal{S}'}\left[(g^*)^2(\mathbf{x}_i) + (g^*)^2(\mathbf{x}_i')\right]\right)\right)\right] + 6\delta. \end{aligned} \tag{98}$$

Denote $h_j(\mathbf{x}_i, \mathbf{x}_i') = g_j^*(\mathbf{x}_i) - g_j^*(\mathbf{x}_i')$. We have

$$\begin{aligned} \mathbb{E}_{\mathcal{S},\mathcal{S}'}\left[h_j(\mathbf{x}_i, \mathbf{x}_i')\right] =& 0, \\ \mathrm{Var}\left[h_j(\mathbf{x}_i, \mathbf{x}_i'\right] =& \mathbb{E}_{\mathcal{S},\mathcal{S}'}\left[h_j^2(\mathbf{x}_i, \mathbf{x}_i')\right] \\ =& \mathbb{E}_{\mathcal{S},\mathcal{S}'}\left[\left(g_j^*(\mathbf{x}_i) - g_j^*(\mathbf{x}_i')\right)^2\right] \\ \le& 2\mathbb{E}_{\mathcal{S},\mathcal{S}'}\left[(g_j^*)^2(\mathbf{x}_i) + (g_j^*)^2(\mathbf{x}_i')\right]. \end{aligned}$$

Thus T_2 is bounded as

$$\begin{aligned} & \mathrm{T}_2 \le \widetilde{\mathrm{T}}_2 + 6\delta \\ & \text{with } \widetilde{\mathrm{T}}_2 = 2\mathbb{E}_{\mathcal{S},\mathcal{S}'}\left[\max_j\left(\frac{1}{n}\sum_{i=1}^n\left(h_j(\mathbf{x}_i, \mathbf{x}_i') - \frac{1}{32R^2}\mathrm{Var}[h_j(\mathbf{x}_i, \mathbf{x}_i')]\right)\right)\right]. \end{aligned} \tag{99}$$

Note that $\|h_j(\mathbf{x}_i, \mathbf{x}_i')\|_{L^\infty(X\times X)} \le 4R^2$. We next study the moment generating function of h_j. For any $0 < t < \frac{3}{4R^2}$, we have

$$\mathbb{E}_{\mathcal{S},\mathcal{S}'}\left[\exp(th_j(\mathbf{x}_i, \mathbf{x}_i')\right] = \mathbb{E}_{\mathcal{S},\mathcal{S}'}\left[1 + th_j(\mathbf{x}_i, \mathbf{x}_i') + \sum_{k=2}^\infty \frac{t^k h_j^k(\mathbf{x}_i, \mathbf{x}_i')}{k!}\right]$$

$$
\begin{aligned}
&\leq \mathbb{E}_{\mathcal{S},\mathcal{S}'}\left[1+th_j(\mathbf{x}-i,\mathbf{x}_i')+\sum_{k=2}^{\infty}\frac{(4R^2)^{k-2}t^k h_j^{k-2}(\mathbf{x}_i,\mathbf{x}_i')}{2\times 3^{k-2}}\right]\\
&= \mathbb{E}_{\mathcal{S},\mathcal{S}'}\left[1+th_j(\mathbf{x}-i,\mathbf{x}_i')+\frac{t^2h_j^2(\mathbf{x}_i,\mathbf{x}_i')}{2}\sum_{k=2}^{\infty}\frac{(4R^2)^{k-2}t^{k-2}}{3^{k-2}}\right]\\
&= \mathbb{E}_{\mathcal{S},\mathcal{S}'}\left[1+th_j(\mathbf{x}-i,\mathbf{x}_i')+\frac{t^2h_j^2(\mathbf{x}_i,\mathbf{x}_i')}{2}\frac{1}{1-4R^2t/3}\right]\\
&= 1+t^2\mathrm{Var}\left[h_j(\mathbf{x}_i,\mathbf{x}_i')\right]\frac{1}{2-8R^2t/3}\\
&\leq \exp\left(\mathrm{Var}\left[h_j(\mathbf{x}_i,\mathbf{x}_i')\right]\frac{3t^2}{6-8R^2t}\right), \qquad (100)
\end{aligned}
$$

where the last inequality comes from $1+x\leq\exp(x)$ for $x\geq 0$.

For $0<\frac{t}{n}<\frac{3}{4R^2}$, we have

$$
\begin{aligned}
&\exp\left(\frac{t\widetilde{\mathrm{T}}_2}{2}\right)\\
&= \exp\left(t\mathbb{E}_{\mathcal{S},\mathcal{S}'}\left[\max_j\left(\frac{1}{n}\sum_{i=1}^{n}h_j(\mathbf{x}_i,\mathbf{x}_i')-\frac{1}{32R^2}\frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}[h_j(\mathbf{x}_i,\mathbf{x}_i')]\right)\right]\right)\\
&\leq \mathbb{E}_{\mathcal{S},\mathcal{S}'}\left[\exp\left(t\max_j\left(\frac{1}{n}\sum_{i=1}^{n}h_j(\mathbf{x}_i,\mathbf{x}_i')-\frac{1}{32R^2}\frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}[h_j(\mathbf{x}_i,\mathbf{x}_i')]\right)\right)\right]\\
&\leq \mathbb{E}_{\mathcal{S},\mathcal{S}'}\left[\sum_j\exp\left(\left(\frac{t}{n}\sum_{i=1}^{n}h_j(\mathbf{x}_i,\mathbf{x}_i')-\frac{1}{32R^2}\frac{t}{n}\sum_{i=1}^{n}\mathrm{Var}[h_j(\mathbf{x}_i,\mathbf{x}_i')]\right)\right)\right]\\
&\leq \sum_j\exp\left(\left(\sum_{i=1}^{n}\frac{3t^2/n^2}{6-8R^3t/n}\mathrm{Var}[h_j(\mathbf{x}_i,\mathbf{x}_i')]-\frac{1}{32R^2}\frac{t}{n}\sum_{i=1}^{n}\mathrm{Var}[h_j(\mathbf{x}_i,\mathbf{x}_i')]\right)\right)\\
&= \sum_j\exp\left(\left(\sum_{i=1}^{n}\frac{t}{n}\mathrm{Var}[h_j(\mathbf{x}_i,\mathbf{x}_i')]\left(\frac{3t^2/n^2}{6-8R^3t/n}-\frac{1}{32R^2}\right)\right)\right), \qquad (101)
\end{aligned}
$$

where the first inequality follows from Jesen's inequality, the third inequality uses (100).

Setting

$$\frac{3t^2/n^2}{6-8R^3t/n}-\frac{1}{32R^2}=0$$

gives rise to $t=\frac{3n}{52R^2}<\frac{3n}{4R^2}$. Substitute the choice of t into (101) gives

$$\frac{t\widetilde{\mathrm{T}}_2}{2}\leq\log\sum_j\exp(0)=\log\mathcal{N}(\delta,\mathcal{R},\|\cdot\|_{L^\infty(X)}).$$

Therefore, we have

$$\widetilde{\mathrm{T}}_2\leq\frac{2}{t}\log\mathcal{N}(\delta,\mathcal{R},\|\cdot\|_{L^\infty(X)})=\frac{104R^2}{3n}\log\mathcal{N}(\delta,\mathcal{R},\|\cdot\|_{L^\infty(X)})$$

and

$$T_2 \leq \frac{104R^2}{3n} \log \mathcal{N}(\delta, \mathcal{R}, \|\cdot\|_{L^\infty(X)}) + 6\delta \leq \frac{35R^2}{n} \log \mathcal{N}(\delta, \mathcal{R}, \|\cdot\|_{L^\infty(X)}) + 6\delta.$$

We then derive a relation between the covering number of $\mathcal{F}_2$ and $\mathcal{R}$. For any $g, g' \in \mathcal{R}$, we have

$$g(\mathbf{x}) = (f_{\text{NN}}(\mathbf{x}) - f(\mathbf{x}))^2,\ g'(\mathbf{x}) = (f'_{\text{NN}}(\mathbf{x}) - f(\mathbf{x}))^2$$

for some $f_{\text{NN}}, f'_{\text{NN}} \in \mathcal{F}$. We have

$$\begin{aligned}
\|g - g'\|_\infty =& \sup_{\mathbf{x}} \left|(f_{\text{NN}}(\mathbf{x}) - f(\mathbf{x}))^2 - (f'_{\text{NN}}(\mathbf{x}) - f(\mathbf{x}))^2\right| \\
=& \sup_{\mathbf{x}} \left|(f_{\text{NN}}(\mathbf{x}) - f'_{\text{NN}}(\mathbf{x}))(f_{\text{NN}}(\mathbf{x}) + f'_{\text{NN}}(\mathbf{x}) - 2f(\mathbf{x}))\right| \\
\leq& \sup_{\mathbf{x}} \left|f_{\text{NN}}(\mathbf{x}) - f'_{\text{NN}}(\mathbf{x})\right| \left|f_{\text{NN}}(\mathbf{x}) + f'_{\text{NN}}(\mathbf{x}) - 2f(\mathbf{x})\right| \\
\leq& 4R\|f_{\text{NN}} - f'_{\text{NN}}\|_{L^\infty(X)}.
\end{aligned}$$

As a result, we have

$$\mathcal{N}(\delta, \mathcal{R}, \|\cdot\|_{L^\infty(X)}) \leq \mathcal{N}\left(\frac{\delta}{4R}, \mathcal{F}, \|\cdot\|_{L^\infty(X)}\right)$$

and the lemma is proved. ■

References

Babak Alipanahi, Andrew Delong, Matthew T. Weirauch, and Brendan J. Frey. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature Biotechnology*, 33:831–838, 2015.

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

Mario Bernardo, Chris Budd, Alan Richard Champneys, and Piotr Kowalczyk. *Piecewise-smooth dynamical systems: theory and applications*, volume 163. Springer Science & Business Media, 2008.

Peter Binev, Albert Cohen, Wolfgang Dahmen, Ronald DeVore, Vladimir Temlyakov, and Peter Bartlett. Universal algorithms for learning theory part I: Piecewise constant functions. *Journal of Machine Learning Research*, 6(9), 2005.

Peter Binev, Albert Cohen, Wolfgang Dahmen, and Ronald DeVore. Universal algorithms for learning theory part II: Piecewise polynomial functions. *Constructive Approximation*, 26(2):127–152, 2007.

T Tony Cai. Minimax and adaptive inference in nonparametric function estimation. *Statistical Science*, 27(1):31–50, 2012.

Minshuo Chen, Haoming Jiang, Wenjing Liao, and Tuo Zhao. Efficient approximation of deep ReLU networks for functions on low dimensional manifolds. In *Advances in Neural Information Processing Systems*, pages 8172–8182, 2019a.

Minshuo Chen, Haoming Jiang, Wenjing Liao, and Tuo Zhao. Nonparametric regression on low-dimensional manifolds using deep ReLU networks: Function approximation and statistical recovery. *arXiv preprint arXiv:1908.01842*, 2019b.

Charles K Chui and Xin Li. Approximation by ridge functions and neural networks with one hidden layer. *Journal of Approximation Theory*, 70(2):131–141, 1992.

Junyoung Chung, Sungjin Ahn, and Yoshua Bengio. Hierarchical multiscale recurrent neural networks. *arXiv preprint arXiv:1609.01704*, 2016.

Alexander Cloninger and Timo Klock. ReLU nets adapt to intrinsic dimensionality beyond the target domain. *arXiv preprint arXiv:2008.02545*, 2020.

Albert Cohen, Wolfgang Dahmen, Ingrid Daubechies, and Ronald DeVore. Tree approximation and optimal encoding. *Applied and Computational Harmonic Analysis*, 11(2): 192–226, 2001.

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

Ingrid Daubechies. *Ten Lectures on Wavelets*. SIAM, 1992.

Ingrid Daubechies, Ronald DeVore, Simon Foucart, Boris Hanin, and Guergana Petrova. Nonlinear approximation and (deep) ReLU networks. *Constructive Approximation*, 55 (1):127–172, 2022.

DGT Denison, BK Mallick, and AFM Smith. Automatic bayesian curve fitting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(2):333–350, 1998.

Ronald DeVore, Boris Hanin, and Guergana Petrova. Neural network approximation. *Acta Numerica*, 30:327–444, 2021.

Ronald A DeVore. Nonlinear approximation. *Acta Numerica*, 7:51–150, 1998.

David L Donoho and Iain M Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994.

David L Donoho and Iain M Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, 90(432):1200–1224, 1995.

David L Donoho and Iain M Johnstone. Minimax estimation via wavelet shrinkage. *The Annals of Statistics*, 26(3):879–921, 1998.

David L Donoho, Iain M Johnstone, Gérard Kerkyacharian, and Dominique Picard. Wavelet shrinkage: asymptopia? *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(2):301–337, 1995.

Jianqing Fan and Irene Gijbels. *Local Polynomial Modelling and Its Applications: Monographs on Statistics and Applied Probability 66*, volume 66. CRC Press, 1996.

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

JH Fridedman. Multivariate adaptive regression splines (with discussion). *Ann Stat*, 19(1): 79–141, 1991.

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

Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6645–6649. IEEE, 2013.

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

László Györfi, Michael Kohler, Adam Krzyzak, Harro Walk, et al. *A Distribution-free Theory of Nonparametric Regression*, volume 1. Springer, 2002.

Eldad Haber, Lars Ruthotto, Elliot Holtham, and Seong-Hwan Jun. Learning across scales—multiscale methods for convolution neural networks. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Boris Hanin. Universal function approximation by deep neural nets with bounded width and ReLU activations. *arXiv preprint arXiv:1708.02691*, 2017.

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

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

Masaaki Imaizumi and Kenji Fukumizu. Deep neural networks learn non-smooth functions effectively. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 869–878. PMLR, 2019.

Bunpei Irie and Sei Miyake. Capabilities of three-layered perceptrons. In *IEEE International Conference on Neural Networks*, volume 1, page 218, 1988.

Seonghyun Jeong and Veronika Rockova. The art of BART: Minimax optimality over nonhomogeneous smoothness in high dimension. *Journal of Machine Learning Research*, 24(337):1–65, 2023.

Fei Jiang, Yong Jiang, Hui Zhi, Yi Dong, Hao Li, Sufeng Ma, Yilong Wang, Qiang Dong, Haipeng Shen, and Yongjun Wang. Artificial intelligence in healthcare: past, present and future. *Stroke and vascular neurology*, 2(4):230–243, 2017.

David LB Jupp. Approximation to data by splines with free knots. *SIAM Journal on Numerical Analysis*, 15(2):328–343, 1978.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

Gitta Kutyniok and Demetrio Labate. Introduction to shearlets. *Shearlets: Multiscale analysis for multivariate data*, pages 1–38, 2012.

Moshe Leshno, Vladimir Ya Lin, Allan Pinkus, and Shimon Schocken. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Networks*, 6(6):861–867, 1993.

Hao Liu and Wenjing Liao. Learning functions varying along a central subspace. *SIAM Journal on Mathematics of Data Science*, 6(2):343–371, 2024.

Hao Liu, Minshuo Chen, Tuo Zhao, and Wenjing Liao. Besov function approximation and binary classification on low-dimensional manifolds using convolutional residual networks. In *International Conference on Machine Learning*, pages 6770–6780. PMLR, 2021.

Hao Liu, Minshuo Chen, Siawpeng Er, Wenjing Liao, Tong Zhang, and Tuo Zhao. Benefits of overparameterized convolutional residual networks: Function approximation under smoothness constraint. In *International Conference on Machine Learning*, pages 13669–13703. PMLR, 2022a.

Min Liu, Zhiqiang Cai, and Jingshuang Chen. Adaptive two-layer ReLU neural network: I. Best least-squares approximation. *Computers & Mathematics with Applications*, 113: 34–44, 2022b.

Ziyue Liu and Wensheng Guo. Data driven adaptive spline smoothing. *Statistica Sinica*, pages 1143–1163, 2010.

Zhou Lu, Hongming Pu, Feicheng Wang, Zhiqiang Hu, and Liwei Wang. The expressive power of neural networks: A view from the width. In *Advances in Neural Information Processing Systems*, pages 6231–6239, 2017.

Zhen Luo and Grace Wahba. Hybrid adaptive splines. *Journal of the American Statistical Association*, 92(437):107–116, 1997.

Mauro Maggioni, Stanislav Minsker, and Nate Strawn. Multiscale dictionary learning: non-asymptotic bounds and robustness. *The Journal of Machine Learning Research*, 17(1): 43–93, 2016.

Stéphane Mallat. *A Wavelet Tour of Signal Processing*. Elsevier, 1999.

Tong Mao and Ding-Xuan Zhou. Rates of approximation by ReLU shallow neural networks. *Journal of Complexity*, 79:101784, 2023.

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

Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T Dudley. Deep learning for healthcare: review, opportunities and challenges. *Briefings in Bioinformatics*, 19(6):1236–1246, 2018.

Hans-Georg Muller and Ulrich Stadtmuller. Variable bandwidth kernel estimators of regression curves. *The Annals of Statistics*, pages 182–201, 1987.

Ryumei Nakada and Masaaki Imaizumi. Adaptive approximation and estimation of deep neural network with intrinsic dimensionality. *arXiv preprint arXiv:1907.02177*, 2019.

Ryumei Nakada and Masaaki Imaizumi. Adaptive approximation and generalization of deep neural network with intrinsic dimensionality. *J. Mach. Learn. Res.*, 21(174):1–38, 2020.

Kenta Oono and Taiji Suzuki. Approximation and non-parametric estimation of ResNet-type convolutional neural networks. In *International Conference on Machine Learning*, pages 4922–4931, 2019.

Philipp Petersen and Felix Voigtlaender. Optimal approximation of piecewise smooth functions using deep ReLU neural networks. *Neural Networks*, 108:296–330, 2018.

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

Alexandre Pintore, Paul Speckman, and Chris C Holmes. Spatially adaptive smoothing splines. *Biometrika*, 93(1):113–125, 2006.

David Ruppert and Raymond J Carroll. Theory & methods: Spatially-adaptive penalties for spline fitting. *Australian & New Zealand Journal of Statistics*, 42(2):205–223, 2000.

Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with ReLU activation function. *arXiv preprint arXiv:1708.06633*, 2017.

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

Jonathan W Siegel and Jinchao Xu. Sharp bounds on the approximation rates, metric entropy, and n-widths of shallow neural networks. *Foundations of Computational Mathematics*, 24(2):481–537, 2024.

Michael Smith and Robert Kohn. Nonparametric regression using bayesian variable selection. *Journal of Econometrics*, 75(2):317–343, 1996.

Taiji Suzuki. Adaptivity of deep ReLU network for learning in Besov and mixed smooth Besov spaces: optimal rate and curse of dimensionality. *arXiv preprint arXiv:1810.08033*, 2018.

Christoph Thäle. 50 years sets with positive reach–a survey. *Surveys in Mathematics and its Applications*, 3:123–165, 2008.

Ryan J Tibshirani. Adaptive piecewise polynomial estimation via trend filtering1. *The Annals of Statistics*, 42(1):285–323, 2014.

G Wahba. In Discussion of 'Wavelet shrinkage: asymptopia?' by D. L. Donoho, I. M. Johnstone, G. Kerkyacharian & D. Picard. *J. R. Statist. Soc. B*, 57:360–361, 1995.

Xiao Wang, Pang Du, and Jinglai Shen. Smoothing splines with varying smoothing parameter. *Biometrika*, 100(4):955–970, 2013.

Sally A Wood, Wenxin Jiang, and Martin Tanner. Bayesian mixture of splines for spatially adaptive nonparametric regression. *Biometrika*, 89(3):513–528, 2002.

Yunfei Yang and Ding-Xuan Zhou. Nonparametric regression using over-parameterized shallow ReLU neural networks. *Journal of Machine Learning Research*, 25:1–35, 2024a.

Yunfei Yang and Ding-Xuan Zhou. Optimal rates of approximation by shallow ReLU^k neural networks and applications to nonparametric regression. *Constructive Approximation*, pages 1–32, 2024b.

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

Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. Recent trends in deep learning based natural language processing. *IEEE Computational Intelligence Magazine*, 13(3):55–75, 2018.

Emilie Yu, Rahul Arora, J Andreas Baerentzen, Karan Singh, and Adrien Bousseau. Piecewise-smooth surface fitting onto unstructured 3d sketches. *ACM Transactions on Graphics (TOG)*, 41(4):1–16, 2022.

Zixuan Zhang, Minshuo Chen, Mengdi Wang, Wenjing Liao, and Tuo Zhao. Effective Minkowski dimension of deep nonparametric regression: function approximation and statistical theories. In *International Conference on Machine Learning*, pages 40911–40931. PMLR, 2023.

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

Jian Zhou and Olga G Troyanskaya. Predicting effects of noncoding variants with deep learning–based sequence model. *Nature Methods*, 12(10):931–934, 2015.