Journal of Machine Learning Research 21 (2020) 1-38 Submitted 1/20; Revised 7/20; Published 9/20

Adaptive Approximation and Generalization of Deep Neural Network with Intrinsic Dimensionality

Ryumei Nakada RYUMEI.NAKADA@RUTGERS.EDU
Department of Statistics, Rutgers University, USA
The University of Tokyo, Japan

Masaaki Imaizumi IMAIZUMI@G.ECC.U-TOKYO.AC.JP
Komaba Institute for Science, The University of Tokyo, Japan
Center for Advanced Intelligence Project, RIKEN, Japan

Editor: Rina Foygel Barber

Abstract

In this study, we prove that an intrinsic low dimensionality of covariates is the main factor that determines the performance of deep neural networks (DNNs). DNNs generally provide outstanding empirical performance. Hence, numerous studies have actively investigated the theoretical properties of DNNs to understand their underlying mechanisms. In particular, the behavior of DNNs in terms of high-dimensional data is one of the most critical questions. However, this issue has not been sufficiently investigated from the aspect of covariates, although high-dimensional data have practically low intrinsic dimensionality. In this study, we derive bounds for an approximation error and a generalization error regarding DNNs with intrinsically low dimensional covariates. We apply the notion of the Minkowski dimension and develop a novel proof technique. Consequently, we show that convergence rates of the errors by DNNs do not depend on the nominal high dimensionality of data, but on its lower intrinsic dimension. We further prove that the rate is optimal in the minimax sense. We identify an advantage of DNNs by showing that DNNs can handle a broader class of intrinsic low dimensional data than other adaptive estimators. Finally, we conduct a numerical simulation to validate the theoretical results.

Keywords: Deep Learning, Deep Neural Network, Generalization Analysis, Intrinsic Dimension, Minimax Optimal Rate.

1. Introduction

Deep neural networks (DNNs) (LeCun et al., 2015; Goodfellow et al., 2016) have attracted considerable attention as statistical modeling for deep learning, owing to favorable outcomes of deep learning in various applications (Collobert and Weston, 2008; He et al., 2016). We often observe that prediction and estimation by DNNs can achieve higher accuracy than that by several standard conventional methods, such as kernel methods (Schmidhuber, 2015). To understand the performance of DNNs and effectively exploit them, numerous studies have extensively investigated their theoretical aspects, such as the approximation power of DNNs (Yarotsky, 2017; Arora et al., 2018; Bartlett et al., 2017; Schmidt-Hieber, 2020).

A nonparametric regression problem is one of the standard frameworks for investigating the mechanisms of DNNs (Bauer and Kohler, 2019; Schmidt-Hieber, 2020; Kohler et al., 2019; Imaizumi and Fukumizu, 2019; Suzuki, 2019; Schmidt-Hieber, 2019). Suppose we have

a set of n observations $\{(X_i, Y_i)\}_{i=1}^n \subset [0,1]^D \times \mathbb{R}$ which is independently and identically generated from the regression model

$$Y_i = f_0(X_i) + \xi_i,\ X_i \sim \mu,\quad i = 1, ..., n, \tag{1}$$

where the covariates X_i marginally follow a probability measure μ, and ξ_i is an independent and identically distributed (i.i.d.) Gaussian noise that is independent of X_i, $E[\xi_i] = 0$ and $E[\xi_i^2] = \sigma^2$, where $\sigma > 0$. The aim of this study is to estimate the unknown function $f_0 : [0,1]^D \to \mathbb{R}$ by an estimator $\widehat{f}$ applied to DNNs. To measure a performance of DNNs, we consider the following value:

$$\|\widehat{f} - f_0\|_{L^2(\mu)}^2 = E_{X\sim\mu}[(\widehat{f}(X) - f_0(X))^2], \tag{2}$$

with the marginal measure μ. The error is also known as a generalization error and is generally used to evaluate the performance of DNNs.

The curse of dimensionality is one of the most significant problems with the DNN framework, in which the theoretical performance of DNNs deteriorates as data dimensionality increases. For the regression problem, the generalization error of DNNs is on the order $\widetilde{O}(n^{-2\beta/(2\beta+D)})$, where $\beta > 0$ is the degree of smoothness of f_0. The rate is known to be optimal in the minimax sense in a typical setting (Schmidt-Hieber, 2020). Since D tends to be very large in machine learning applications (e.g., the number of pixels in images), the theoretical generalization error decreases very slowly as n increases. Thus, the theoretical bound for the error is quite loose for describing the empirical performance of DNNs. To avoid the slow convergence rate, numerous studies have investigated specific structures of functions as a target of approximation when using DNNs. One typical approach is to consider various types of smoothness or spectral distribution; many general notions of smoothness are investigated (Barron, 1993, 1994; Barron and Klusowski, 2018; Montanelli and Du, 2017; Suzuki, 2019; Schmidt-Hieber, 2020). Alternatively, it is also common to introduce a specific form of f_0. For example, when f_0 takes on the form of conventional statistical models (e.g., the generalized single index or multivariate adaptive regression splines (MARS)), the convergence rates of DNNs evidently improve (Bauer and Kohler, 2019; Kohler et al., 2019). Additionally, when f_0 has a functional form involving manifolds, the convergence rate of DNNs depends on the manifold dimensions (Schmidt-Hieber, 2019).

In contrast to the studies focused on f_0, the behavior of the measure μ of covariates has not been well studied despite several significant motivations for the investigation of μ exist. First, we frequently observe that high-dimensional data have an implicit structure such as lying around low dimensional sets (e.g., manifolds) in practice (Tenenbaum et al., 2000; Belkin and Niyogi, 2003). We numerically confirm that several well-known real data have approximately 30 intrinsic dimensions, while their nominal dimensions are approximately $1,000$ (see Section 2.1). Since the low intrinsic dimensionality is well known in the field of machine learning, several well-designed methodologies use this empirical fact (Masci et al., 2015; Arjovsky and Bottou, 2017). Second, the literature on nonlinear dimension reduction indicates that the low intrinsic dimensionality of covariates can be a crucial factor in overcoming the curse of dimensionality. Several estimators, such as kernel methods and the Gaussian process regression, have been shown to achieve a fast convergence rate depending only on their intrinsic dimensionality (Bickel and Li, 2007; Kpotufe, 2011; Kpotufe and

Garg, 2013; Yang and Dunson, 2016). Despite these motivations, connecting DNNs with an intrinsic dimension of data is a nontrivial task; thus, investigating this property remains an important open question.

In this study, we investigate the performance of DNNs regarding D-dimensional data, which have a d-dimensional intrinsic structure such that $d < D$. To describe the intrinsic dimensionality of data, we apply the notion of *Minkowski dimension*. Moreover, we develop a proof technique to adapt DNNs to the intrinsic low dimensional structure. Consequently, we derive the rates of the approximation and generalization errors in DNNs, which depend only on d and β, but not on D, as summarized in Table 1. In summary, we prove that the convergence speed of DNNs is independent of the nominal dimension of data, but does on their intrinsic dimension. We also prove that the derived rate is optimal in the minimax sense and, finally, verify the theoretical results using numerical experiments.

	Approximation error	Generalization error
Existing	$\mathcal{O}(W^{-\beta/D})$	$\widetilde{\mathcal{O}}(n^{-2\beta/(2\beta+D)})$
Ours (d-Minkowski dim.)	$\mathcal{O}(W^{-\beta/d})$	$\widetilde{\mathcal{O}}(n^{-2\beta/(2\beta+d)})$

Table 1: Derived rates of approximation and generalization errors by DNNs with W parameters and n observations. $\beta > 0$ denotes the smoothness of the generating function, and D is dimension of the observations, and d is an intrinsic dimension of μ. In general, $d \leq D$ holds.

Our results describe an advantage inherent in DNNs compared to several other methods. That is, we demonstrate that DNNs can achieve a fast convergence rate over a broader class of data distributions. Some adaptive methods, such as the kernel and Gaussian process estimators, can achieve a fast convergence rate with the intrinsic dimension. However, they can achieve a fast rate only when data are on smooth manifolds or are generated from doubling measures. In contrast, we show that DNNs can achieve a fast rate over a broader class of μ, such as data on highly non-smooth fractal sets. This advantage is due to the use of the Minkowski dimension, which can cover a broader class of distributions.

As a technical contribution of this study, we develop a proof for DNNs with an optimal partition of hypercubes in the domain to handle the Minkowski dimension. To evaluate an error regarding the dimension, we have to follow the two steps: (i) divide the domain of f^0 into hypercubes and (ii) combine sub-neural networks in each of the hypercubes. However, a naive combination makes the depth (number of layers) of DNNs diverge because of the accumulation of errors of adjacent hypercubes. In our proof, to avoid this problem, we develop a particular set of partitions for a set of hypercubes and then unify the sub-neural networks within each of the partitions. Using this technique, we can avoid the accumulation of errors to achieve the desired convergence rate.

We summarize the contributions of this study as follows:

1. We prove that DNNs can avoid the curse of dimensionality by adapting to the intrinsic low dimensionality of the data with the Minkowski dimension.

2. We present a relative advantage of DNNs, in which a fast convergence rate with a broader class of distributions can be achieved compared to other methods that are also adaptive to an intrinsic dimension.
3. As proof, we derive rates of approximation and generalization errors for D-dimensional data, which have d intrinsic dimension, and demonstrate that the rate is minimax optimal.

The remainder of this study is organized as follows. Section 2 discusses the notion of intrinsic dimensionality and defines the Minkowski dimension. Section 3 shows an upper bound for the approximation error of DNNs. Section 4 provides the upper and lower bounds of the generalization error of DNNs. Section 5 compares our main results with several related studies involving DNNs and other methods. Section 6 provides experimental evidence to support the theoretical results. Finally, Section 7 summarizes our conclusions. The appendix includes a full version of the proof.

1.1 Basic Notation

For an integer z, $[z] := \{1, 2, ..., z\}$ is a set of positive integers no greater than z. For a vector $b \in \mathbb{R}^d$, $\|b\|_q := (\sum_{j=1,...,d} b_j^q)^{1/q}$ is a q-norm for $q \in [0, \infty]$. For a measure μ, the support of μ is written as $\mathrm{Supp}(\mu)$. For a function $g : \mathbb{R}^D \to \mathbb{R}$, $\|g\|_{L^p(\mu)} := (\int g^p \, \mathrm{d}\mu)^{1/p}$ is the $L^p(\mu)$ norm, with a probability measure μ. $\widetilde{O}(\cdot)$ is the Landau's big O, ignoring a logarithmic factor. With $\varepsilon > 0$, $\mathcal{N}(\Omega, \varepsilon)$ is the fewest number of ε-balls that cover Ω in terms of $\|\cdot\|_\infty$. For a measure μ, $\mathrm{Supp}(\mu)$ denotes the support for μ.

2. Intrinsic Low Dimensionality of Covariates

2.1 Empirical Motivation

As the motivation for considering low intrinsic dimensionality, we provide an empirical analysis of real datasets, such as handwritten letter images using the modified National Institute of Standards and Technology (MNIST) dataset (LeCun et al., 2015) and object images using the Canadian Institute for Advanced Research (v) dataset (Krizhevsky and Hinton, 2009). Since the data are images, their nominal dimension D is equal to the number of pixels in each image. We apply several dimension estimators, such as the local principal component analysis (LPCA) (Fukunaga and Olsen, 1971; Bruske and Sommer, 1998), the method with maximum likelihood method (ML) (Haro et al., 2008), and the expected simplex skewness (ESS) (Johnsson et al., 2015) to estimate the intrinsic dimensions of $30,000$ samples from each of the datasets. The results in Table 2 indicate that the estimated intrinsic dimensions are significantly less than D. Although the definitions of intrinsic dimensions are not standard, the results provide motivation to conduct an in-depth investigation on low intrinsic dimensionality.

2.2 Notion of Intrinsic Dimensionality

We introduce the notion of dimensionality in this study. Although there are numerous definitions for dimensionality, we employ the following general notion.

Data set	D	Intrinsic Dimension d LPCA	ML	ESS
MNIST	784	37	13.12	29.41
CIFAR-10	1024	9	25.84	27.99

Table 2: Estimated intrinsic dimensions of the MNIST and CIFAR-10 datasets. The dimensions are estimated from $30,000$ sub-samples from the original datasets.

Definition 1 (Minkowski Dimension). *The (upper) Minkowski dimension of a set $E \subset [0,1]^D$ is defined as*

$$\dim_{\mathrm{M}} E := \inf\left\{d^* \geq 0 \mid \limsup_{\varepsilon \downarrow 0} \mathcal{N}(E,\varepsilon)\varepsilon^{d^*} = 0\right\}.$$

The Minkowski dimension measures how the number of covering balls for E is affected by the radius of the balls. Since the dimension does not depend on smoothness, it can measure the dimensionality of highly non-smooth sets, such as fractal sets (e.g., Koch curve). Figure 1 shows an image of how the covering balls measure the Minkowski dimension of E.

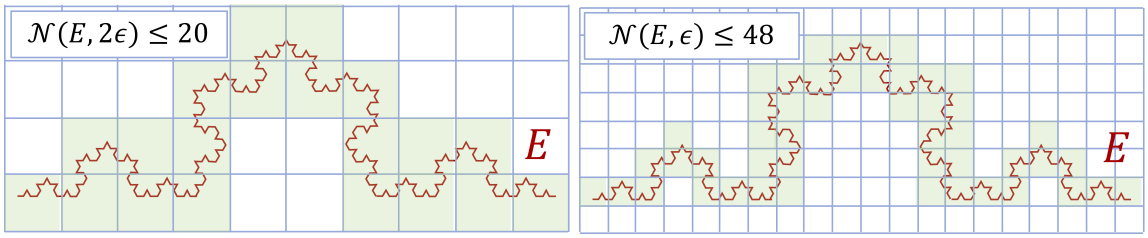

Figure 1: The Koch curve E (red lines) and covering max-balls (green squares) for E. The Minkowski dimension of E is $d = \log 4/\log 3 \approx 1.26$, while E is a subset of $\mathbb{R}^D$ with $D = 2$.

Relationship to Other Dimensions: The Minkowski dimension can describe a broader class of low dimensional sets compared to several other dimensionalities. For example, the notion of *manifold dimension* describes the dimensionality of sets with smooth structures and is one of the most common notions used to describe an intrinsic dimensionality (Bickel and Li, 2007; Yang and Dunson, 2016). While manifold dimensions are valid only for smooth sets such as circles, we can apply the Minkowski dimension to sets without such restriction. Consequently, for a set $E \subset [0,1]^D$, we show that

$$\{E \mid \dim_M E \leq d\} \supset \{E \mid E \text{ is a } d\text{-dimensional manifold}\},$$

holds (see Lemma 9 in Section A.2). In addition, the notion of *regularity dimensions* is used for an intrinsic dimensionality (Kpotufe, 2011; Kpotufe and Garg, 2013). Similar to the manifold case, the Minkowski dimension is a more general notion than the regularity dimension (see Lemma 12 in Section A.2).

3. Approximation Results

3.1 Preparation

To investigate an approximation power of DNNs, we provide a rigorous formulation. Here, we consider DNNs with a rectified linear unit (ReLU) activation function $\rho(x_1, \dots, x_p) := (\max\{x_1, 0\}, \dots, \max\{x_p, 0\})$.

Definition 2 (Deep Neural Networks)**.** *Let L be a number of layers. For each $\ell \in [L] \cup \{0\}$, $p_\ell \in \mathbb{N}$ be a number of nodes for each layer ℓ, and $A_\ell \in \mathbb{R}^{p_\ell \times p_{\ell-1}}$ and $b_\ell \in \mathbb{R}^{p_\ell}$ be a parameter matrix and vector. Let $\rho_{b_\ell} := \rho(\cdot + b_\ell)$ be a shifted ReLU activation. Then, the realization of a neural network architecture $\Phi := ((A_L, b_L), \dots, (A_1, b_1))$ is denoted as $R(\Phi) : \mathbb{R}^{p_0} \to \mathbb{R}^{p_L}$, which has a form*

$$R(\Phi)(x) = A_L \rho_{b_{L-1}} \circ \cdots \circ A_2 \rho_{b_1}(A_1 x) + b_L, \text{ for } x \in [0,1]^{p_0}. \tag{3}$$

$R(\Phi)(x)_i$ denotes the i-th output of $R(\Phi)(x)$.

Further, we define a set of realizations of DNNs with several restrictions. Specifically, we bound a number of layers, parameters, and its scale. For each Φ, the number of layers of Φ is written as $L(\Phi)$, a number of parameters of Φ is $W(\Phi) := \sum_{\ell=1}^{L} \|b_\ell\|_0 + \|\mathrm{vec}(A_\ell)\|_0$, and the scale of parameters of Φ is $B(\Phi) = \max_{\ell=1,\dots,L} \max\{\|b_\ell\|_\infty, \|\mathrm{vec}(A_\ell)\|_\infty\}$.

Definition 3 (Functional Set by DNNs)**.** *With a tuple (W', L', B'), a functional set by DNNs is defined as follows:*

$$\mathcal{F}(W', L', B') = \left\{ R(\Phi) : [0,1]^{p_0} \to \mathbb{R}^{p_L} \mid L(\Phi) \le L', W(\Phi) \le W', B(\Phi) \le B' \right\}.$$

To discuss the approximation power of DNNs, we define the *Hölder space* as a family of smooth functions. For a function $f : \mathbb{R}^D \to \mathbb{R}$, $\partial_d f(x)$ is a partial derivative with respect to a d-th component, and $\partial^\alpha f := \partial_1^{\alpha_1} \cdots \partial_D^{\alpha_D} f$ using multi-index $\alpha = (\alpha_1, ..., \alpha_D)$. For $z \in \mathbb{R}$, $\lfloor z \rfloor$ denotes the largest integer that is less than z.

Definition 4 (Hölder space)**.** *Let $\beta > 0$ be a degree of smoothness. For $f : [0,1]^D \to \mathbb{R}$, the Hölder norm is defined as*

$$\|f\|_{\mathcal{H}(\beta,[0,1]^D)} := \max_{\alpha:\|\alpha\|_1 < \lfloor\beta\rfloor} \sup_{x \in [0,1]^D} |\partial^\alpha f(x)| + \max_{\alpha:\|\alpha\|_1 = \lfloor\beta\rfloor} \sup_{x,x' \in [0,1]^D, x \neq x'} \frac{|\partial^\alpha f(x) - \partial^\alpha f(x')|}{\|x - x'\|_\infty^{\beta - \lfloor\beta\rfloor}}.$$

Then, the Hölder space on $[0,1]^D$ is defined as

$$\mathcal{H}(\beta, [0,1]^D) = \left\{ f \in C^{\lfloor\beta\rfloor}([0,1]^D) \middle| \|f\|_{\mathcal{H}(\beta,[0,1]^D)} < \infty \right\}.$$

Also, $\mathcal{H}(\beta, [0,1]^D, M) = \left\{ f \in \mathcal{H}(\beta, [0,1]^D) \middle| \|f\|_{\mathcal{H}(\beta,[0,1]^D)} \le M \right\}$ denotes the M-radius closed ball in $\mathcal{H}(\beta, [0,1]^D)$.

3.2 Approximation with Low Dimensionality

We evaluate how well DNNs approximate a function $f_0 \in \mathcal{H}(\beta, [0,1]^D)$ with a probability measure μ whose support has a Minkowski dimension less than d. That is, we measure an approximation error using the norm $\|\cdot\|_{L^\infty(\mu)}$ with μ as a base measure.

Theorem 5 (Approximation with Minkowski dimension)**.** *Suppose $d > \dim_{\mathrm{M}} \mathrm{Supp}(\mu)$ holds with $d < D$. For $\beta, M > 0$, we define $r := 2 + \lfloor (1 + \lfloor \beta \rfloor)/(2d) \rfloor$, $c_{\beta,D,d} := 384\beta(11 + (1+\beta)/d)(36r + 83 + 6 \cdot 4^{r+2}) \cdot 6^{d/(1+\lfloor \beta \rfloor)}$, $c_\mu := \sup_{\varepsilon > 0} \mathcal{N}(\mathrm{Supp}\,\mu, \varepsilon)\varepsilon^d$, and constants $C_1, C_2 > 0$ such as*

$$C_1 \leq \Big(4c_\mu(50D + 17 + 8D^{2+\lfloor \beta \rfloor} c_{\beta,D,d})D^d (3M)^{d/\beta} + D^{2+\lfloor \beta \rfloor} c_{\beta,D,d} 2^{d/\beta + 5}\Big),$$
$$C_2 \leq 11 + 6D + (11 + (1+\beta)/d)(1 + (1 \vee \lceil \log_2 \beta \rceil)).$$

Also, for $\varepsilon > 0$, we consider a triple (W, L, B) with

$$W = C_1 \varepsilon^{-d/\beta},\ L = C_2,\ \text{and } B = O(\varepsilon^{-s}). \tag{4}$$

Then, for sufficiently small $\varepsilon_0 > 0$, any $\varepsilon \in (0, \varepsilon_0)$ and any $f_0 \in \mathcal{H}(\beta, D, M)$, we obtain

$$\inf_{R(\Psi) \in \mathcal{F}(W,L,B)} \|R(\Psi) - f_0\|_{L^\infty(\mu)} \leq \varepsilon.$$

The following corollary summarizes the result in Theorem 5.

Corollary 6 (Approximation Rate)**.** *With the triple (W, L, B) as defined in Theorem 5, an existing $R(\Psi) \in \mathcal{F}(W, L, B)$ satisfies for sufficiently large W,*

$$\|R(\Psi) - f_0\|_{L^\infty(\mu)} = \mathcal{O}(W^{-\beta/d}).$$

The result indicates that the order $\mathcal{O}(W^{-\beta/d})$ depends only on d and β, but not on D. That is, the approximation rate behaves as if the data are d-dimensional, although they are nominally D-dimensional.

Additionally, the results of Theorem 2 suggest the following intuitions. First, a finite number of layers is sufficient to achieve the convergence rate, because L does not diverge with small ε. Second, the constant terms C_1 and C_2 in Theorem 5 depend on D polynomially. To achieve the results, we develop an additional proof technique to achieve it with for low intrinsic dimensionality.

Proof Outline of Theorem 5: Let $\mathcal{I}$ be a minimum set of hypercubes of side length γ covering $\mathrm{Supp}(\mu)$. We partition $\mathcal{I}$ into $\mathcal{I}_1, \ldots, \mathcal{I}_K$ such that each subset $\mathcal{I}_k$ consists of hypercubes separated by γ from each other. For $I \in \mathcal{I}$, let $R(\Phi_I)$ be a trapezoid-type approximator represented by a neural network Φ_I that approximates $f_I \mathbb{1}_I$ (the green curve in the right panel in Figure 2), where f_I is the Taylor expansion of f_0 around any point in I. We note that for any $\mathcal{I}_k$, support of $R(\Phi_I)$ for $I \in \mathcal{I}_k$ are disjoint. Then, we define a neural network Φ to realize $R(\Phi) = \max_{1 \leq k \leq K} \sum_{I' \in \mathcal{I}_k} R(\Phi_{I'})$ as our novel approximator. For $x \in I$ for some $I \in \mathcal{I}$, we have

$$|R(\Phi)(x) - f_0(x)| \leq \underbrace{\max_{I' \in \Xi(I)} |R(\Phi_{I'})(x) - f_{I'}(x)|}_{T_1} + \underbrace{\max_{I' \in \Xi(I)} |f_{I'}(x) - f_0(x)|}_{T_2}$$

where $\Xi(I)$ denotes the set of hypercubes neighbouring I, including I itself. The inequality holds because $R(\Phi_{I'})(x) = 0$ holds for all $I' \notin \Xi(I)$ and thus $\sum_{I' \in \mathcal{I}_k} R(\Phi_{I'})(x) = R(\Phi_{I''})(x)$ for some $I'' \in \mathcal{I}_k \cap \Xi(I)$. To bound the term T_1, we evaluate the trapezoid approximation using the *sawtooth* approximation (Telgarsky, 2016) for the Taylor polynomials. T_2 is evaluated by the Taylor approximation for $\mathcal{H}(\beta, [0,1]^D, M)$. Regarding the effect of the partition $\mathcal{I}_k$, we limit K by a constant depending only on D. Hence, we show that the number of layers can be finite. Since $\dim_{\mathrm{M}} \mathrm{Supp}(\mu) \leq d$ holds, there are $O(\gamma^{-d})$ hypercubes used to approximate f_0 on $\mathrm{Supp}(\mu)$. An image of the entire procedure is provided in Figure 2.

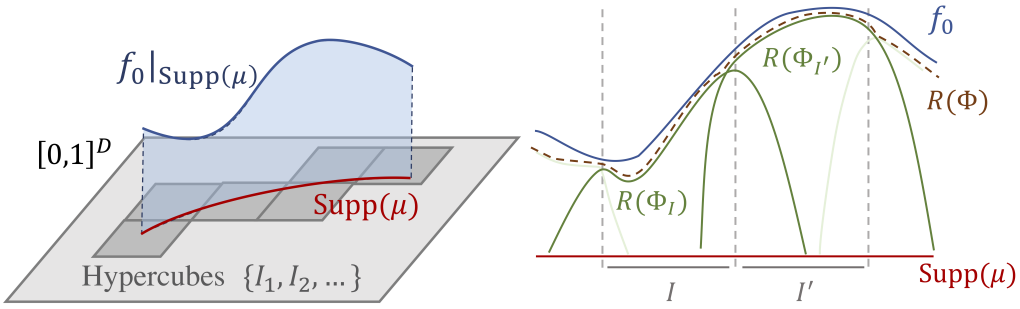

Figure 2: An illustration of our proof. [Left] $\mathrm{Supp}(\mu)$ (the red curve) in $[0,1]^D$, where f_0 (the blue curve) is restricted to $\mathrm{Supp}(\mu)$. The hypercubes (the gray squares) $\mathcal{I} = \{I_1, I_2, ...\}$ cover $\mathrm{Supp}(\mu)$. DNNs approximate f_0 within each of the hypercubes. [Right] Within hypercubes $I, I' \in \mathcal{I}$, the DNNs produce a function that approximates f_0 locally. Using the trapezoid-type approximators $R(\Phi_I)$ and $R(\Phi_{I'})$ (the green curves), we define the approximator $R(\Phi)$ (the brown curve) as the maximum of the trapezoid approximators.

4. Generalization Results

We investigate the generalization error of DNNs using a nonparametric regression problem. Suppose we have a set of n observations $\{(X_i, Y_i)\}_{i=1}^n$ from the regression model (1) with $f_0 \in \mathcal{H}(\beta, [0,1]^D, M)$, the marginal measure μ, and Gaussian noise ξ_i. From the observations, we introduce an estimator for f_0. The estimator $\widehat{f}$ is defined as $\widehat{f}(x) := \max\{-C_B, \min\{C_B, \widetilde{f}(x)\}\}$, where

$$\widetilde{f} \in \mathrm{argmin}_{f \in \mathcal{F}(W,L,B)} \sum_{i=1}^{n} (Y_i - f(X_i))^2. \tag{5}$$

$C_B > 0$ denotes a threshold for the clipped estimator $\widetilde{f}$. We note that calculating $\widehat{f}$ is not straightforward because the loss function in (5) is non-convex. However, we employ the estimator $\widehat{f}$, because we aim to investigate the generalization error in terms of n, which is independent of the difficulty of optimization. We can obtain an approximated version of $\widehat{f}$ using various optimization techniques, such as multiple initializations or Bayesian optimization.

4.1 Generalization Error with Low Dimensionality

We provide a generalization error of $\widehat{f}$ for the case when $\mathrm{Supp}(\mu)$ has a low Minkowski dimension.

Theorem 7 (Generalization with Minkowski dimension)**.** *Fix any $f_0 \in \mathcal{H}(\beta, [0,1]^D, M)$ and suppose $d > \dim_{\mathrm{M}} \mathrm{Supp}(\mu)$. Set a triple (W, L, B) with the constants C_1, C_2 and s appearing in Theorem 5 as $W = C_1 n^{d/(2\beta+d)}$, $L = C_2$, and $B = O(n^{2\beta s/(2\beta+d)} \log n)$. Then, there exists a constant $C = C(c_\mu, \beta, D, d, M, \sigma)$ such that*

$$\|\widehat{f} - f_0\|_{L^2(\mu)}^2 \leq C n^{-2\beta/(2\beta+d)} (1 + \log n)^2$$

holds with probability at least $1 - 2\exp\left(-n^{d/(2\beta+d)}\right)$ for any $n \geq N$ and $C_B \geq \|f_0\|_{L^\infty(\mu)}$ with a sufficiently large N.

The derived generalization error is on the order of $\widetilde{O}(n^{-2\beta/(2\beta+d)})$, which is independent from D. That is, we show that the convergence rate of DNNs is determined by the intrinsic dimension d of $\mathrm{Supp}(\mu)$, which is much faster than the existing rate $\widetilde{O}(n^{-2\beta/(2\beta+D)})$ (Schmidt-Hieber, 2020) without low dimensionality. We note that the order of parameters (W, L, B) is not affected by D; however, it does depend on d. Moreover, using DNNs for estimation requires $\mathcal{O}(1)$ layers to achieve the desired rate based on our approximation technique presented in Theorem 5.

Proof Outline of Theorem 7: First, we decompose the empirical loss into two terms, which are analogous to the bias and variance. Following the definition of $\widehat{f}$ as (5), a simple calculation leads to

$$\|\widehat{f} - f_0\|_n^2 \leq \underbrace{\|f - f_0\|_n^2}_{=:T_B} + \underbrace{\frac{2}{n} \sum_{i=1} \xi_i (\widehat{f}(X_i) - f(X_i))}_{:=T_V},$$

for any $f \in \mathcal{F}(W, L, B)$. We define an empirical norm as $\|f\|_n^2 := n^{-1} \sum_{i=1}^n f(X_i)^2$.

The first term T_B, which is analogous to an approximation bias, is evaluated using an approximation power of $\mathcal{F}(W, L, B)$. We apply Theorem 5 and bound the term. The second term, T_V, which describes the variance of the estimator, is evaluated using the technique of the empirical process theory (van der Vaart and Wellner, 1996). By using the notion of the local Rademacher complexity and concentration inequalities (Giné and Nickl, 2016), we bound T_V by an integrated covering number of $\mathcal{F}(W, L, B)$. Further, we derive a bound for the covering number using the parameters (W, L, B); thus, we can evaluate T_V in terms of the parameters. By combining the results for T_B and T_V and selecting proper values for (W, L, B), we obtain the claimed result.

4.2 Minimax Optimal Rate with Low Dimensionality

We prove the optimality of the obtained rate in Theorem 7 by deriving the minimax error of the estimation problem. To this end, we consider a probability measure μ with $\dim_{\mathrm{M}} \mathrm{Supp}(\mu) \leq d$. Then, we obtain the following minimax lower bound.

Theorem 8 (Minimax Rate with Low Dimensionality). *Let $\mathcal{P}_d$ be a set of probability measures on $[0,1]^D$ satisfying $\dim_{\mathrm{M}} \mathrm{Supp}(\mu) \leq d$. Then, there exists a constant $C' > 0$ such that*

$$\inf_{\check{f}} \sup_{(f_0,\mu)\in\mathcal{H}(\beta,[0,1]^D,M)\times\mathcal{P}_d} \|\check{f} - f_0\|^2_{L^2(\mu)} \geq C' n^{-2\beta/(2\beta+d)}, \tag{6}$$

where $\check{f}$ is any arbitrary estimator for f_0.

That is, any estimator provides an error $\Omega(n^{-2\beta/(2\beta+d)})$ in a worst case scenario; therefore, it is regarded as a theoretical limit of efficiency. Since the rate in Theorem 7 corresponds to the rate up to logarithmic factors, our rate almost achieves the minimax optimality.

5. Comparison with Related Studies

5.1 Nonparametric Analysis for DNNs

Setting		Error	
f_0	μ	Approximation	Estimation
Hölder / Sobolev		$\widetilde{O}(W^{-\beta/D})$	$\widetilde{O}(n^{-2\beta/(2\beta+D)})$
Barron		$\widetilde{O}(W^{-1/2})$	$\widetilde{O}(n^{-1})$
Mixed Smooth		$\widetilde{O}(W^{-\gamma})$	$\widetilde{O}(n^{-2\gamma/(2\gamma+1)})$
Hölder	**d-dimensional**	$O(W^{-\beta/d})$	$\widetilde{O}(n^{-2\beta/(2\beta+d)})$

Table 3: Comparison of the derived rates of approximation and generalization errors with a non-parametric class of target functions. W denotes a number of parameters in DNNs, and n is the number of observations. D is the dimension of X, β is the smoothness of f_0, γ is the index of mixed smoothness, and d is an intrinsic dimension of X.

Many studies have investigated the approximation and estimation performance of DNNs and some (Yarotsky, 2017; Schmidt-Hieber, 2020) clarify the performance of DNNs with ReLU activations when f_0 is in the Hölder space. They show that this performance is $\widetilde{O}(W^{-\beta/D})$ for approximation and $\widetilde{O}(n^{-2\beta/(2\beta+D)})$ for estimation. Other studies (Montanelli and Du, 2017; Suzuki, 2019) consider a different functional class with *mixed smoothness* for f_0 and then obtain a novel convergence rate that depends on its particular smoothness index γ. Additionally, other studies have investigated more specific structures of f_0. If we can decompose f_0 into a composition of feature maps, the convergence rate depends on the dimensionality of the feature space (Petersen and Voigtlaender, 2018). As a close analog to this study, the DNNs' error converges faster when f_0 is associated with a manifold structure (Schmidt-Hieber, 2019). Bach (2017) derives an error for a class of Lipschitz-continuous functions, which is considered as a special case of the Hölder space with $\beta = 1$.

A large number of research have been conducted to achieve faster convergence rates by introducing specific structures. Bach (2017) showed that D-independent convergence rates can be achieved when f_0 has a parametric structure such as a general additive model or a

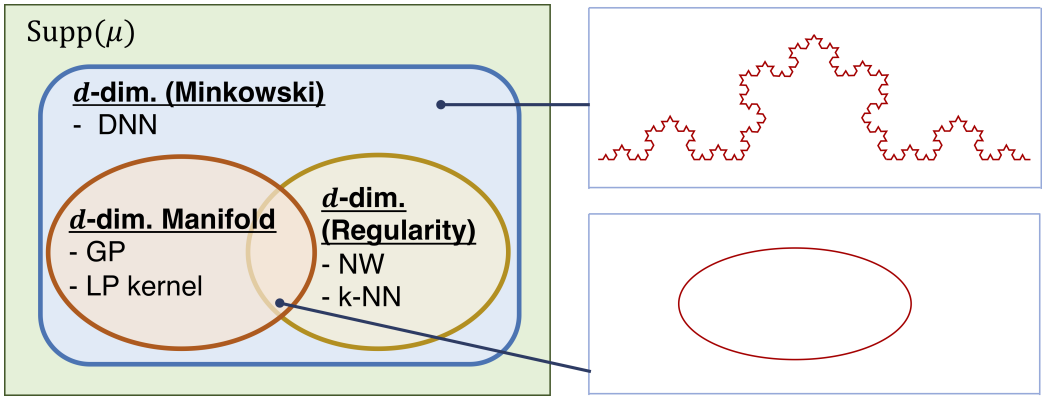

Figure 3: [Left] Various low dimensional sets as Supp(μ), and corresponding regression methods that can obtain the optimal rate on each of the sets. The Minkowski dimension can describe a wider class of low dimensional sets. [Right] The top right is the Koch curve, which is a low dimensional set in terms of the Minkowski dimension. The bottom right is an ellipse representing a smooth manifold. It is also a low dimensional set in terms of the Minkowski and the regularity dimension.

single index model. Similarly, when f_0 has the form of a generalized hierarchical version of a single index model (Bauer and Kohler, 2019) or a form with multivariate adaptive regression splines (MARS) (Kohler et al., 2019), we can obtain a faster convergence rate with DNNs. Although these rates are fast, it is unlikely that f_0 has such specific parametric model structures in practice. As a non-parametric attempt to obtain a faster convergence rate, a classical approach (Barron, 1993, 1994) considers a restricted functional class (referred to as the *Barron class* in this study) for f_0, and achieving a very fast rate: $\widetilde{O}(W^{-1/2})$ for approximation, and $\widetilde{O}(n^{-1})$ for estimation. The Barron class is non-parametric and has no model constraints, but it has constraints on differentiability through spectral conditions. Given that the class requires higher differentiability when D is large, a more flexible functional class is required.

Our study assumes that f_0 is an element of the Hölder space, and that μ has an intrinsic low dimensional structure; that is, Supp(μ) is d-dimensional in the Minkowski or the manifold sense. The Hölder class allows the smoothness of the function to be determined separately of D. Thus, we can investigate the effects of d and D independently. We obtain the approximation rate $O(W^{-\beta/d})$ and the estimation rate $\widetilde{O}(n^{-2\beta/(2\beta+d)})$. Because d is much less than D empirically (as mentioned in Section 2.1), the rate can alleviate the curse of dimensionality caused by a large D. To the best of our knowledge, this is the first study proving that errors in DNNs converge faster with data having general intrinsic low dimensionality.

5.2 Other Adaptive Methods with Intrinsic Low Dimensionality

Except for DNNs, several nonparametric estimators can obtain a convergence rate that is adaptive to the intrinsic dimension of a distribution d. The local polynomial kernel (LP kernel) regression (Bickel and Li, 2007) and the Gaussian process (GP) regression (Yang and Dunson, 2016) can achieve the rate $\mathcal{O}(n^{-2\beta/(2\beta+d')})$, where $\mathrm{Supp}(\mu)$ is a d'-dimensional manifold with $\beta = 2$. Similarly, the k-nearest neighbor (k-NN) regression (Kpotufe, 2011) and the Nadaraya-Watson (NW) kernel regression (Kpotufe and Garg, 2013) can achieve the rate with d', when μ has a *regularity dimension* d', that is less general than the Minkowski dimension (Lemma 12 in the supplementary material).

We show that DNNs with finite layers can obtain the fast convergence rate of $\mathcal{O}(n^{-2\beta/(2\beta+d)})$ over a broader scope of cases compared to the existing adaptive methods. Theorem 7 indicates that DNNs with $L = \mathcal{O}(1)$ can obtain the rate when $\dim_M \mathrm{Supp}(\mu) < d$, which is less restrictive than the settings with manifolds and regularity dimensions. Intuitively, DNNs can obtain a fast adaptive rate even when $\mathrm{Supp}(\mu)$ does not have a smooth structure such as that of manifolds. Figure 3 presents an overview of the results.

6. Simulation

6.1 Estimation by DNNs with Different d

We calculate the generalization errors of DNNs using synthetic data. We set the true function as $f_0(x) := (D-1)^{-1}\sum_{i=1}^{D-1} x_i x_{i+1} + D^{-1}\sum_{i=1}^{D} 2\sin(2\pi x_i)\mathbb{1}_{\{x_i \le 0.5\}} + D^{-1}\sum_{i=1}^{D}(4\pi(\sqrt{2}-1)^{-1}(x_i - 2^{-1/2})^2 - \pi(\sqrt{2}-1))\mathbb{1}_{\{x_i > 0.5\}}$, which belongs to $\mathcal{H}(\beta, [0,1]^D)$ with $\beta = 2$. We set μ as a uniform measure on a d-dimensional sphere embedded in $[0,1]^D$, and also set a noise ξ_i as a Gaussian variable with zero mean and variance $\sigma^2 = 0.1$. We generate n pairs of (X_i, Y_i) from the regression model (1) and learn the estimator (5). For the learning process, a DNN architecture with four layers and the ReLU activation function are employed, and each layer has D units except the output layer. For optimization, we employ Adam (Kingma and Ba, 2015) with the following hyper-parameters; 0.001 learning rate and $(\beta_1, \beta_2) = (0.9, 0.999)$.

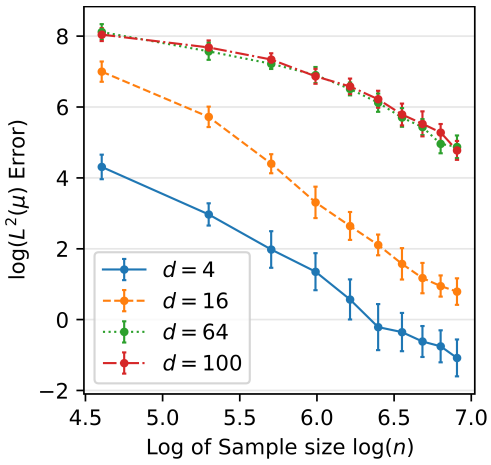

Figure 4: Simulated generalization errors of DNNs with $D = 128$ and $d \in \{4, 16, 64, 100\}$. The error bars show the standard deviation from the replication.

We set the nominal dimension as $D = 128$ and consider different numbers of samples $n \in \{100, 200, ..., 1000\}$ and intrinsic dimensions $d \in \{4, 16, 64, 100\}$. We measure the generalization errors using validation data in terms of the $L^2(\mu)$-norm. We replicate the learning procedure 100 times with different initial weights for the parameters of neural networks from a standard normal distribution.

We plot the generalization errors in log against $\log n$ in Figure 4. The slope of the curve corresponds to the convergence rate of the errors as the figure is double logarithmic. From

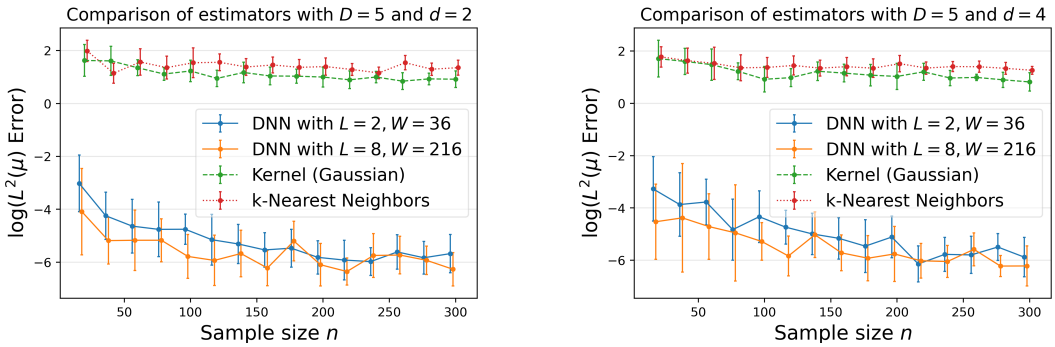

Figure 5: Simulated generalization errors for NW with Gaussian kernel, k-NN regression, and DNNs (2 layers and 8 layers). The error bars shows the standard deviation of the 10 replication. The left panel shows the case with $D = 5$ and $d = 2$, and the right panel is for $D = 5$ and $d = 4$.

the results, we observe the following two findings: (i) The error is lower with a fewer d, and (ii) The convergence rates with $d \in \{4, 16\}$ are faster than those with $d \in \{64, 100\}$.

6.2 Comparison with the Other Estimators

We compare the performances of DNNs with existing methods, such as k-NN method and the NW kernel method. We note that they can achieve a rate whose exponent depends on d.

We set the true function as $f_0(x) := (1/D) \sum_{i=1}^{D} x_i^2 \mathbb{1}_{\{x_i \leq 0.5\}} + (-x_i + 3/4)\mathbb{1}_{\{x_i > 0.5\}}$, which belongs to $\mathcal{H}(\beta, [0,1]^D)$ with $\beta = 1$. Let μ be the uniform measure on a union of d-dimensional $\ell^{1/2}$ ball and ℓ^2 ball embedded in $[0,1]^D$. We note that the support of μ is not a smooth manifold. We set $n \in \{20, 40, ..., 300\}$, and also consider two configurations $(D, d) = (5, 2)$ and $(D, d) = (5, 4)$. For each sample size, we replicate the estimation 10 times with different initial weights from a standard normal distribution. The learning procedure of DNNs is the same as that described in Section 6.1. For k-NN, its hyper-parameter k is selected from [50]. Fir the NW kernel method, we employ a Gaussian kernel whose bandwidth is selected from $\{0.10, 0.11, \ldots, 1.00\}$. We select all the hyper-parameters based on cross-validation.

We plot the simulated generalization error with the validation data and the other methods based on n in Figure 5. The results indicate that DNNs outperform the other estimators. Since Supp(μ) is not a smooth manifold, this is likely to affect the dominance observed for the DNNs.

6.3 Real Data Analysis

We compare the performance of DNNs using the modified National Institute of Standards and Technology (MNIST) dataset. The dataset contains 784-dimensional grayscale images of handwritten digits.

For the experiment, we contaminate the data with d-dimensional Gaussian noise using $d \in \{0, 25, 50, 75, 100\}$ with variance 0.01. We set $n \in \{100, 200, 300, 400, 500\}$. Considering that the task is classification, we employ the soft-max activation function in the last layer

and measure the error based on the $\|\cdot\|_{L^2}$-norm. We replicate the setting 10 times, and discard two replications with the first and second-largest test errors to eliminate the effect of the difficulty involved with non-convex optimization using DNNs. All the other settings inherit those of Section 6.1.

d	0	25	50	75	100
Convergence Rate	-0.20	-0.19	-0.17	-0.18	-0.16

Table 4: Estimated convergence rate with the MNIST data. d denotes an intrinsic dimension of the contaminated Gaussian noise. The convergence rate is estimated by regressing the logarithm of test error on $\log n$.

Table 4 shows the convergence rates of the generalization error for each value of d. The rate is estimated by using the least square method for the logarithm of the test error versus the logarithm of the sample size. The result indicates that the convergence gets slower as d increases.

7. Conclusion

We theoretically elucidate that the intrinsic low dimensionality of data mainly determines the performance of deep learning. To show the result, we introduce the notion of the Minkowski dimension and then derive the rates of approximation and generalization errors, which only depend on the intrinsic dimension and are independent of the large nominal dimension. Additionally, we find that DNNs can achieve the convergence rate over a broader class of data than several conventional methods. Our results provided evidence of the inherent convergence advantage of deep learning over other models with respect to data with low intrinsic dimension.

Acknowledgments

We would like to thank the editor for valuable comments and suggestions. We also acknowledge support for this project from JSPS KAKENHI (18K18114, 18K19793 and 20J21717) and JST Presto.

Appendix A. Supportive Discussion

A.1 Additional Notation and Definition

A closed ball in $\mathbb{R}^D$ with its center x and radius r with norm $\|\cdot\|$ is described as $\overline{B}^D(x,r) := \{x' \in \mathbb{R}^D \mid \|x - x'\| \leq r\}$. An open ball is similarly defined as $B^D(x,r)$. A closed ball in $\mathbb{R}^D$ with its center x and radius r with the ℓ^p-norm is described as $\overline{B}_p^D(x,r)$. An open version of the ball is B_p^D. For a set Ω, $\mathbb{1}_\Omega(\cdot)$ is an indicator function such that $\mathbb{1}_\Omega(x) = 1$ if $x \in \Omega$, and $\mathbb{1}_\Omega(x) = 0$ otherwise. We write the set of non-negative real numbers as

$\mathbb{R}_{\geq}$. For sequences $\{a_n\}_n$ and $\{b_n\}_n$, $a_n \lesssim b_n$ denotes $a_n \leq Cb_n$ with a finite constant $C > 0$ for all n. With $\varepsilon > 0$, $\mathcal{N}_2(\Omega, \varepsilon)$ is a smallest number of ε-balls, which cover Ω in terms of $\|\cdot\|_2$. Rigorously, a support of a probability measure ν on a set $\mathcal{X}$ is defined as $\mathrm{Supp}(\nu) \coloneqq \{x \in \mathcal{X} \mid V \in \mathcal{N}_x \Rightarrow \nu(V) > 0\}$, where $\mathcal{N}_x$ is a set of open neighborhoods with its center $x \in \mathcal{X}$. We define a standard distance $d(A, B)$ of two sets A and B by

$$d(A, B) \coloneqq \inf\{\|x - y\| \mid x \in A, y \in B\}.$$

A.2 Other Notions for Dimensionality

We show a relation between the Minkowski dimension and the dimension of manifolds. The notion of *manifolds* is common for analyzing low dimensionality of data (Belkin and Niyogi, 2003; Niyogi et al., 2008; Genovese et al., 2012). The Minkowski dimension can describe the dimensionality of manifolds.

Lemma 9. *Let $\mathcal{M}$ be a compact d-dimensional manifold in $[0,1]^D$. Assume $\mathcal{M} = \bigcup_{k=1}^K \mathcal{M}_k \subset [0,1]^D$ for $K \in \mathbb{N}$. Also assume that for any $1 \leq k \leq K$, there exists an onto and continuously differentiable map $\psi_k : [0,1]^{d_k} \to \mathcal{M}_k$ each of which has the input dimension $d_k \in \mathbb{N}$. Then, $\dim_{\mathrm{M}} \mathcal{M} \leq \max_{1 \leq k \leq K} d_k$.*

Proof of Lemma 9 We first assume that the statement holds with $K = 1$. Then, we investigate the case with general K.

Suppose the lemma is correct for $K = 1$. Take $d_k^* > d_k$. Since $\dim_{\mathrm{M}} \mathcal{M}_k < d_k^*$, there exists a constant $C > 0$ such that for any $\varepsilon > 0$, existing a finite set $F_\varepsilon^k \subset [0,1]^D$ satisfies the followings:

1. $\mathcal{M}_k \subset \bigcup_{x \in F_\varepsilon^k} B_2^D(x, \varepsilon)$,
2. $\mathrm{card}(F_\varepsilon^k) \leq C_k \varepsilon^{-d_k^*}$.

Let $F_\varepsilon \coloneqq \bigcup_{k=1}^K F_\varepsilon^k$, then we have $\mathcal{M} \subset \bigcup_{x \in F_\varepsilon} B_2^D(x, \varepsilon)$ and

$$\mathcal{N}_2(\mathcal{M}, \varepsilon) \leq \mathrm{card}(F_\varepsilon) \leq \left(\sum_{k=1}^K C_k\right) \varepsilon^{-\max_k d_k^*}.$$

For $d^* > d \coloneqq \max_k d_k$, we choose $d_k^* > d_k$ such that $\max_k d_k^* < d^*$ holds. Then, it yields $\limsup_{\varepsilon \downarrow 0} \mathcal{N}_2(\mathcal{M}, \varepsilon) \varepsilon^{d^*} = 0$. So the proof is reduced to the case of $K = 1$.

We investigate the case $K = 1$. For brevity, we omit the subscript k and write $\psi = (\psi_1, ..., \psi_D)$. Recall that ψ_i is continuously differentiable. We also define

$$L_i \coloneqq \max_{x \in [0,1]^D} \sqrt{\sum_{j=1}^d |\partial \psi_i'(x)/\partial x_j|^2}.$$

Applying the mean-value theorem to ψ_i yields $|\psi_i(z) - \psi_i(w)| \leq L_i \|z - w\|_2$. By the Lipschitz continuity of $\psi = (\psi_1, \ldots, \psi_D)$, for any z, w, $\|\psi(z) - \psi(w)\|_2 \leq \sqrt{D} L \|z - w\|_2$ where $L \coloneqq \max_i L_i$. Fix any $\varepsilon > 0$ and $d \in \mathbb{N}$. Recall that there exists a constant $C > 0$ so that for any $\delta > 0$, an existing finite set $F_\delta \subset [0,1]^d$ (see Example 27.1 in Shalev-Shwartz and Ben-David (2014)) satisfies

1. $\mathrm{card}(F_\delta) \leq C\delta^{-d}$,
2. $[0,1]^d \subset \bigcup_{y \in F_\delta} B_2^d(y, \delta)$.

Choosing $\delta = \varepsilon/(\sqrt{D}L)$ yields

$$\mathcal{M} \subset \psi\left(\bigcup_{y \in F_\delta} B_2^d(y,\delta) \cap [0,1]^d\right) \subset \bigcup_{y \in F_\delta} \psi\Big(B_2^d(y,\delta) \cap [0,1]^d\Big) \subset \bigcup_{y \in F_\delta} B_2^D(\psi(y), \varepsilon)$$

where the last inclusion follows from the Lipschitz continuity of ψ. Since we have

$$\mathrm{card}(F_\delta) \leq C(\sqrt{D}L)^d \varepsilon^{-d},$$

we obtain the conclusion. □

Additionally, we explain the notion of a *doubling measure*, which is an alternative way to describe an intrinsic dimension. It is employed in several studies (Kpotufe, 2011; Kpotufe and Garg, 2013).

Definition 10 (Doubling Measure)**.** *A probability measure ν on $\mathcal{X}$ is called a doubling measure, if there exists a constant $C > 0$ such as*

$$\nu(\overline{B}^D(x, 2r) \cap \mathcal{X}) \leq C\nu(\overline{B}^D(x, r) \cap \mathcal{X}),$$

for all $x \in \mathrm{Supp}(\nu)$ and $r > 0$.

Then, we can define a dimensionality by the regularity property of doubling measures.

Definition 11 (Regularity Dimension)**.** *For a doubling measure ν, the (upper) regularity dimension* $\dim_{\mathrm{R}} \nu$ *is defined by the infimum of $d^* > 0$ such that there exists a constant $C_\nu > 0$ satisfying*

$$\frac{\nu(\overline{B}^D(x, r) \cap \mathcal{X})}{\nu(\overline{B}^D(x, \varepsilon r) \cap \mathcal{X})} \leq C_\nu \varepsilon^{-d^*},$$

for all $x \in \mathrm{Supp}(\nu)$, $\varepsilon \in (0,1)$ and $r > 0$.

There also exists a relation between the *Minkowski dimension* and the *regularity dimension*. Intuitively, when $\mathcal{X} \subset \mathbb{R}^D$, a measure ν with $\dim_{\mathrm{R}} \nu = d$ behaves as if a domain of ν is $\mathbb{R}^d$, as shown in Figure 2. The regularity dimension of ν can evaluate the Minkowski dimension of $\mathrm{Supp}(\nu)$.

Lemma 12 (Lemma 3.4 in Käenmäki et al. (2013))**.** *Let $(\mathcal{X}, \mathcal{B}, \nu)$ be a probability space with $\mathcal{X} \subset \mathbb{R}^D$ is bounded. Suppose ν is a doubling measure. Then,*

$$\dim_{\mathrm{M}} \mathrm{Supp}(\nu) \leq \dim_{\mathrm{R}} \nu.$$

Appendix B. Proof of Main Results

B.1 About Theorem 5

Our proof strategy is an extended version of those of the previous studies (Yarotsky, 2017; Petersen and Voigtlaender, 2018). The previous studies employ a simultaneous approximation of Taylor polynomials multiplied by approximated indicator functions for each disjoint hypercube. However, their approach fails with the concentrated measure of covariates. To avoid the problem, we develop a novel way to unite the approximators by the disjoint subsets and the max operation as described in the proof outlined in Section 3.

B.1.1 Sub-Neural Networks as Preparation

Before a central part of the proof, for convenience, we define several sub-neural networks: (i) concatenation of neural networks, (ii) parallelization of neural networks, (iii) parallelization of neural networks with different inputs, (iv) an approximated identity function, and (v) a max function.

(i) Concatenation of Neural Networks: Given two neural networks Φ^1 and Φ^2, we aim to construct a network Φ such that $R(\Phi) = R(\Phi^2) \circ R(\Phi^1)$, which is possible because ReLU activation function has the property $\rho(x) - \rho(-x) = x$. Write $\Phi^1 = ((A^1_{L_1}, b^1_{L_1}), \ldots, (A^1_1, b^1_1))$ and $\Phi^2 = ((A^2_{L_1}, b^2_{L_1}), \ldots, (A^2_1, b^2_1))$. We define parameter matrices and vectors as

$$\widetilde{A}^2_1 \coloneqq \begin{pmatrix} A^2_1 \\ -A^2_1 \end{pmatrix}, \widetilde{b}^2_1 \coloneqq \begin{pmatrix} b^2_1 \\ -b^2_1 \end{pmatrix}, \widetilde{A}^1_{L_1} \coloneqq \begin{pmatrix} A^1_{L_1} & -A^1_{L_1} \end{pmatrix}.$$

Then, the concatenation of neural networks Φ^1 and Φ^2 is defined as

$$\Phi^2 \odot \Phi^1 \coloneqq \big((A^2_{L_2}, b^2_{L_2}), \ldots, (A^2_2, b^2_2), (\widetilde{A}^2_1, \widetilde{b}^2_1), (\widetilde{A}^1_{L_1}, b^1_{L_1}), (A^1_{L_1-1}, b^1_{L_1-1}), \ldots, (A^1_1, b^1_1).\big)$$

It is easy to show the following relations:

1. $W(\Phi^2 \odot \Phi^1) \leq 2W(\Phi^2) + 2W(\Phi^1)$,
2. $L(\Phi^2 \odot \Phi^1) = L(\Phi^2) + L(\Phi^1)$,
3. $B(\Phi^2 \odot \Phi^1) = \max\big\{B(\Phi^2), B(\Phi^1)\big\}$.

By repeating the discussion, we obtain the concatenation of k neural networks. Some properties of the concatenation are summarized in the following remark.

Remark 13 (Concatenation)**.** *For any neural networks $\Phi^1, ..., \Phi^k$, we have the following inequalities:*

$$W(\Phi^k \odot \cdots \odot \Phi^1) \leq 2\sum_{i=1}^{k} W(\Phi^i),$$
$$L(\Phi^k \odot \cdots \odot \Phi^1) = \sum_{i=1}^{k} L(\Phi^i),$$
$$B(\Phi^k \odot \cdots \odot \Phi^1) = \max_{1 \leq i \leq K} B(\Phi^i).$$

(ii) Parallelization of Neural Networks: We define the parallelization of multiple neural networks. Let $\Phi^i = ((A^i_{L_i}, b^i_{L_i}), \dots, (A^i_1, b^i_1))$, be neural networks with a d_i-dimensional input and an m_i-dimensional output. When the input dimension $d_i = d$ for all i, we can define the parallelization of neural networks.

Suppose the number of layers of the network is the same for all Φ^i, Write $\Phi^{i'} = ((A^i_L, b^i_L), \dots, (A^i_1, b^i_1))$ and define parameters as

$$A_1 \coloneqq \begin{pmatrix} {A^1_1}^\top & {A^2_1}^\top & \cdots & {A^K_1}^\top \end{pmatrix}^\top, \text{ and } b_1 \coloneqq \begin{pmatrix} {b^1_1}^\top & \cdots & {b^K_1}^\top \end{pmatrix}^\top.$$

Also, for $\ell \geq 2$, we define parameters as

$$A_\ell \coloneqq \begin{pmatrix} A^1_\ell & O & \dots & O \\ O & A^2_\ell & \dots & O \\ \vdots & \vdots & \ddots & \vdots \\ O & O & \dots & A^K_\ell \end{pmatrix}, \text{ and } b_\ell \coloneqq \begin{pmatrix} b^1_\ell \\ \vdots \\ b^K_\ell \end{pmatrix}.$$

Then, we define the parallelization of networks $(\Phi^i)_{i=1}^K$ with d-dimensional input and $\sum_{i=1}^K m_i =: m$-dimensional output as

$$[\Phi^1, \Phi^2, \dots, \Phi^K] \coloneqq ((A_L, b_L), \dots, (A_1, b_1)).$$

Combined with the previous results, we can show the following result.

Remark 14 (Parallelization)**.** *Let $\Phi^i = ((A^i_{L_i}, b^i_{L_i}), \dots, (A^i_1, b^i_1))$ be the neural network with d-dimensional input, m_i-dimensional output and the same number of layers L for $i \in \{1, \dots, K\}$. Then, we obtain the followings:*

$$\begin{aligned} W([\Phi^1, \Phi^2, \dots, \Phi^K]) &= \sum_{i=1}^K W(\Phi^i), \\ L([\Phi^1, \Phi^2, \dots, \Phi^K]) &= L, \\ B([\Phi^1, \Phi^2, \dots, \Phi^K]) &= \max_{1 \leq i \leq K} B(\Phi^i). \end{aligned}$$

(iii) Parallelization of Neural Networks with Different Input: We define the parallelization of multiple neural networks without sharing the input. Under this case, the networks share the number of layers only. Let $\Phi^i = ((A^i_L, b^i_L), \dots, (A^i_1, b^i_1))$ be a neural network with a d_i-dimensional input and an m_i-dimensional output. Define $m \coloneqq \sum_{i=1}^K m_i$ and $d \coloneqq \sum_{i=1}^K d_i$. We construct a neural network version of a function $\mathbb{R}^d \to \mathbb{R}^m$

$$((x_{1,1}, \dots x_{1,d_1}), \dots, (x_{K,1}, \dots, x_{K,d_K}))^\top \mapsto \begin{pmatrix} R(\Phi_1)(x_{1,1}, \dots, x_{1,d_1}) \\ \vdots \\ R(\Phi_K)(x_{K,1}, \dots, x_{K,d_K}) \end{pmatrix}.$$

Write $\Phi^i = ((A^i_L, b^i_L), \dots, (A^i_1, b^i_1))$ and we define parameters as

$$A_\ell \coloneqq \begin{pmatrix} A^1_\ell & O & \dots & O \\ O & A^2_\ell & \dots & O \\ \vdots & \vdots & \ddots & \vdots \\ O & O & \dots & A^K_\ell \end{pmatrix}, \text{ and } b_\ell \coloneqq \begin{pmatrix} b^1_\ell \\ \vdots \\ b^K_\ell \end{pmatrix},$$

We define the parallelization of networks $(\Phi^i)_{i=1}^K$ with d-dimensional input and m-dimensional output as

$$\langle \Phi^1, \Phi^2, \dots, \Phi^K \rangle \coloneqq ((A_L, b_L), \dots, (A_1, b_1)).$$

We obtain the following remark:

Remark 15 (Parallelization with Different Input)**.** *Let $\Phi^i = ((A^i_{L_i}, b^i_{L_i}), \dots, (A^i_1, b^i_1))$, $i \in [K]$ be neural networks with d_i-dimensional input and m_i-dimensional output with the same number of layers L. Then, we obtain the following relations:*

$$\begin{aligned} W\big(\langle \Phi^1, \Phi^2, \dots, \Phi^K \rangle\big) &= \sum_{i=1}^K W(\Phi_i), \\ L\big(\langle \Phi^1, \Phi^2, \dots, \Phi^K \rangle\big) &= L, \\ B\big(\langle \Phi^1, \Phi^2, \dots, \Phi^K \rangle\big) &= \max_{1 \le i \le K} B(\Phi^i). \end{aligned}$$

(iv) Identity Function: We define a neural network that approximates an identity function. For $L \ge 2$, let $\Phi^{Id}_{D,L} : \mathbb{R}^D \to \mathbb{R}^D$ be a neural network of identity function

$$\Phi^{Id}_{D,L} \coloneqq \left(((I_D \ -I_D), 0), \underbrace{(I_{2D}, 0), \dots, (I_{2D}, 0)}_{L-2 \text{ times}}, \left(\begin{pmatrix} I_D \\ -I_D \end{pmatrix}, 0 \right) \right),$$

where $I_p \in \mathbb{R}^{p \times p}$ is the identity matrix. For $L = 1$, let $\Phi^{Id}_{D,L} = ((I_D, 0))$. Then its realization is the identity function $\mathbb{R}^p \ni x \mapsto x$. We can see $W(\Phi^{Id}_{D,L}) = 2DL$, $L(\Phi^{Id}_{D,L}) = L$ and $B(\Phi^{Id}_{D,L}) = 1$ for all $L \ge 2$.

(v) Max function: We define a neural network that works as a max function. We implement $\max : \mathbb{R}^s_{\ge} \to \mathbb{R}$ by a neural network. Let $t = \lceil \log_2 s \rceil$. Note that $\log_2 s \le t < \log_2 s + 1$ and $s \le 2^t < 2s$. With the ReLU activation function, the max function $(x_1, x_2) \mapsto \max(x_1, x_2)$ for non-negative inputs can be easily implemented by a ReLU neural network

$$\Phi^{\max,2} \coloneqq \left(((1,1), 0), \left(\begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \end{pmatrix} \right) \right),$$

due to the identity $\max(x_1, x_2) = \rho(x_2 - x_1) + \rho(x_1)$ for $x_1, x_2 \ge 0$.

Next, we extend $\Phi^{\max,2}$ to take multiple inputs. Since $2^t = s$ does not always hold, we need a following dummy network to make $\Phi^{\max,s}$ take s-dimensional input. Define a dummy network $\Phi^{\text{dummy}} \coloneqq \left(\begin{pmatrix} I_s \\ O_{(2^t - s) \times s} \end{pmatrix}, 0_{2^t} \right)$. Then, a max function with s dimensional inputs is defined by

$$\Phi^{\max,s} \coloneqq \Phi^{\max,2} \odot \langle \underbrace{\Phi^{\max,2}, \Phi^{\max,2}}_{2^1} \rangle \odot \cdots \odot \langle \underbrace{\Phi^{\max,2}, \dots, \Phi^{\max,2}}_{2^t} \rangle \odot \Phi^{\text{dummy}}.$$

Then, from Remark 13 and Remark 15, we can verify the following relations:

1. $W(\Phi^{\max,s}) \le 2(2^0 \times 5 + 2^1 \times 5 + \cdots + 2^t \times 5) + 2s = 42s$,
2. $L(\Phi^{\max,s}) = 2(t+1) + 1 < 2\log_2 s + 3$,
3. $B(\Phi^{\max,s}) = 1$.

B.1.2 Approximation for Smooth Functions by Neural Networks

We here investigate an approximation property of neural networks for several types of functions. Namely, we provide several lemmas for the following functions: (i) general smooth functions, (ii) smooth functions on hypercubes, and (iii) smooth functions with finite layers.

(i) Approximation for Smooth Functions: We discuss several approximation for $f_0 \in \mathcal{H}(\beta, [0,1]^D, M)$ by DNNs. We review several existing lemmas, and also provide a novel approximation result. To begin with, we cite the following lemma for convenience.

Lemma 16 (Lemma A.8 in Petersen and Voigtlaender (2018)). *Fix any $f \in \mathcal{H}(\beta, D, M)$ and $\bar{x} \in [0,1]^D$. Let $\bar{f}(x)$ be the Taylor polynomial of degree $\lfloor \beta \rfloor$ of f around $\bar{x}$, namely,*

$$\bar{f}(x) \coloneqq \sum_{|\alpha| \leq \lfloor \beta \rfloor} \frac{\partial^\alpha f(\bar{x})}{\alpha!} (x - \bar{x})^\alpha.$$

Then, $\left| f(x) - \bar{f}(x) \right| \leq D^\beta M \|x - \bar{x}\|^\beta$ holds for any $x \in [0,1]^D$.

Note that there exists some constant $\overline{C} = \overline{C}(\beta, D, M)$ such that $\sup_{|\alpha| \leq \lfloor \beta \rfloor} |\partial^\alpha f(\bar{x})/\alpha!| \leq \overline{C} M$ for $f \in \mathcal{H}(\beta, [0,1]^D, M)$.

Also, we cite the following lemma, which describes an approximation for a multiplication function.

Lemma 17 (Lemma A.4 in Petersen and Voigtlaender (2018)). *Fix any $b > 0$. There are constants $s^{\mathrm{mul}} = s^{\mathrm{mul}}(\beta) \in \mathbb{N}$, $c_1^{\mathrm{mul}} = c_1^{\mathrm{mul}}(\beta, D, b)$, $c_2^{\mathrm{mul}} = c_2^{\mathrm{mul}}(\beta, D, b)$ and $\varepsilon_0^{\mathrm{mul}} = \varepsilon_0^{\mathrm{mul}}(\beta, D, b)$ such that for any $\varepsilon \in (0, \varepsilon_0^{\mathrm{mul}})$ and $\alpha \leq \lfloor \beta \rfloor$, there is a neural network $\Phi_\varepsilon^{\mathrm{mul}}$ with D-dimensional input and 1-dimensional output satisfying the following inequalities:*

1. $\sup_{x \in [0,1]^D} \left| R(\Phi_\varepsilon^{\mathrm{mul}})(x) - x^\alpha \right| \leq \varepsilon$,
2. $W(\Phi_\varepsilon^{\mathrm{mul}}) \leq c_1^{\mathrm{mul}} \varepsilon^{-D/b}$,
3. $L(\Phi_\varepsilon^{\mathrm{mul}}) \leq (1 + \lceil \log_2 \lfloor \beta \rfloor \rceil)(11 + b/D)$,
4. $B(\Phi_\varepsilon^{\mathrm{mul}}) \leq c_2^{\mathrm{mul}} \varepsilon^{-s^{\mathrm{mul}}}$.

The constant c_1^{mul} is upper-bounded by $384\beta(36r + 83 + 6 \cdot 4^{r+2}) \cdot 6^{D/b}$ for $r \coloneqq 1 + \lfloor (b+1)/(2D) \rfloor$.

Using this lemma, we construct a neural network approximating multiple taylor polynomials in each output.

Based on Lemma 17 and Lemma A.5 in Petersen and Voigtlaender (2018), we develop an approximation result for an m-dimensional multiple output neural network Φ. Here, we write $R(\Phi) = (R(\Phi)_1, \ldots, R(\Phi)_m)$.

Lemma 18 (Simultaneous approximation of multiple Taylor polynomials). *Fix any $m \in \mathbb{N}$. Let $\{c_{\lambda,\alpha}\} \subset [-B, B]$ for $1 \leq \lambda \leq m$. Let $(x_\lambda)_{\lambda=1}^m \subset [0,1]^D$. Then there exist constants $c_1^{\mathrm{pol}} = c_1^{\mathrm{pol}}(\beta, D, d, B)$, $c_2^{\mathrm{pol}} = c_2^{\mathrm{pol}}(\beta, D, d, B)$, $s_1^{\mathrm{pol}} = s_1^{\mathrm{pol}}(\beta, D, d, B)$ and $\varepsilon_0^{\mathrm{pol}} = \varepsilon_0^{\mathrm{pol}}(\beta, D, d)$ such that for any $\varepsilon \in (0, \varepsilon_0^{\mathrm{pol}})$, there is a neural network $\Phi_\varepsilon^{\mathrm{pol}}$ which satisfies the followings:*

1. $\max_{\lambda=1,\ldots,m} \sup_{x \in [0,1]^D} \left| R(\Phi_\varepsilon^{\mathrm{pol}})_\lambda(x) - \sum_{|\alpha| < \beta} c_{\lambda,\alpha} (x - x_\lambda)^\alpha \right| \leq \varepsilon$,

2. $W(\Phi_\varepsilon^{\mathrm{pol}}) \le c_1^{\mathrm{pol}}(\varepsilon^{-d/\beta} + m)$,
3. $L(\Phi_\varepsilon^{\mathrm{pol}}) \le 1 + (1 + \lceil \log_2 \beta \rceil)(11 + (1+\beta)/d)$,
4. $B(\Phi_\varepsilon^{\mathrm{pol}}) \le c_2^{\mathrm{pol}}\, \varepsilon^{-s_1^{\mathrm{pol}}}$.

Proof of Lemma 18 Firstly, we rewrite the target polynomial $\sum_{|\alpha|<\beta} c_{\lambda,\alpha}(x - x_\lambda)^\alpha$. By the binomial theorem (Folland, 2013) for example, we have

$$(x - x_\lambda)^\alpha = \sum_{\gamma \le \alpha} \binom{\alpha}{\gamma} (-x_\lambda)^{\alpha-\gamma} x^\gamma.$$

Then, we evaluate the polynomial as

$$\begin{aligned}
\sum_{|\alpha|<\beta} c_{\lambda,\alpha}(x - x_\lambda)^\alpha &= \sum_{|\alpha| \le \lfloor\beta\rfloor} \left\{ \sum_{\gamma \le \alpha} c_{\lambda,\alpha} \binom{\alpha}{\gamma} (-x_\lambda)^{\alpha-\gamma} x^\gamma \right\} \\
&= \sum_{|\gamma| \le \lfloor\beta\rfloor} \left\{ \sum_{\gamma \le \alpha, |\alpha| \le \lfloor\beta\rfloor} c_{\lambda,\alpha} \binom{\alpha}{\gamma} (-x_\lambda)^{\alpha-\gamma} \right\} x^\gamma \\
&=: \sum_{|\gamma| \le \lfloor\beta\rfloor} \widetilde{c}_{\lambda,\gamma} x^\gamma.
\end{aligned}$$

Note that $|\widetilde{c}_{\lambda,\gamma}| \le cM$ for all γ with $|\gamma| \le \lfloor\beta\rfloor$ where $c = c(\beta, D)$ is a constant. In fact since $\binom{\alpha}{\gamma} \le \alpha_1^{\gamma_1} \dots \alpha_D^{\gamma_D}$, we can bound $|\widetilde{c}_{\lambda,\gamma}|$ by

$$\begin{aligned}
|\widetilde{c}_{\lambda,\gamma}| &\le \sup_{x \in [0,1]^D} \left| \sum_{|\alpha| \le \lfloor\beta\rfloor} \sum_{\gamma \le \alpha} c_{\lambda,\alpha} \binom{\alpha}{\gamma} (-x_\lambda)^{\alpha-\gamma} \right| \\
&\le M \sum_{|\alpha| \le \lfloor\beta\rfloor} \sum_{\gamma \le \alpha} \alpha_1^{\gamma_1} \dots \alpha_D^{\gamma_D} \\
&\le M \sum_{|\alpha| \le \lfloor\beta\rfloor} (\alpha_1 + 1) \dots (\alpha_D + 1) \alpha_1^{\alpha_1} \dots \alpha_D^{\alpha_D} \\
&\le M D^{\lfloor\beta\rfloor + 1} (1 + \lfloor\beta\rfloor)^{D(1+\lfloor\beta\rfloor)}.
\end{aligned}$$

Secondly, we define an explicit neural network to approximate the polynomial. Write $\{\gamma \mid |\gamma| \le \lfloor\beta\rfloor\} = \{\gamma_1, \dots, \gamma_K\}$ for some $K = K(\beta)$. Let $\Phi_\varepsilon^{\mathrm{tlr},\lambda} := ((\widetilde{c}_{\lambda,\gamma_1}, \dots, \widetilde{c}_{\lambda,\gamma_K}), 0)$. Define $\varepsilon_0^{\mathrm{pol}} := \varepsilon_0^{\mathrm{mul}}/(cKM)$. The number of parameters in Lemma 17 has the exponential decay with exponent $-D/b$. In order to moderate the exponent, we define the neural network $\Phi_{\varepsilon/cKM}^{\mathrm{mul},\gamma_k}$ as the one constructed by Lemma 17 with substitution $\varepsilon \leftarrow \varepsilon/(cKM)$, $b \leftarrow (1+\lfloor\beta\rfloor)D/d$ and $\alpha \leftarrow \gamma_k$ for $k \in [K]$. Then, there exist constants $c_3^{\mathrm{pol}} = c_3^{\mathrm{pol}}(\beta, D, d, M)$ and $c_4^{\mathrm{pol}} = c_4^{\mathrm{pol}}(\beta, D, d, M)$ such that

1. $W(\Phi_{\varepsilon/cKM}^{\mathrm{mul},\gamma_k}) \le c_3^{\mathrm{pol}}\, \varepsilon^{-d/(1+\lfloor\beta\rfloor)}$,
2. $L(\Phi_{\varepsilon/cKM}^{\mathrm{mul},\gamma_k}) \le (1 + \lceil \log_2 \beta \rceil)(11 + (1+\beta)/d)$,

3. $B(\Phi^{\mathrm{mul},\gamma_k}_{\varepsilon/cKM}) \leq c_4^{\mathrm{pol}} \varepsilon^{-s_2^{\mathrm{mul}}}$,

holds for all $k \in [K]$. Note that since $K \leq D^0 + D^1 + \cdots + D^{\lfloor\beta\rfloor} \leq (1+\lfloor\beta\rfloor)D^{\lfloor\beta\rfloor+1}$, it is easily shown that

$$
\begin{aligned}
c_3^{\mathrm{pol}} &\leq c_1^{\mathrm{mul}} D^{2d} (1+\lfloor\beta\rfloor)^{Dd+d/(1+\lfloor\beta\rfloor)} M^{d/(1+\lfloor\beta\rfloor)}, \\
c_4^{\mathrm{pol}} &\leq c_2^{\mathrm{mul}} D^{s^{\mathrm{mul}} d/(1+\lfloor\beta\rfloor)+s^{\mathrm{mul}}(1+\lfloor\beta\rfloor)} (cM)^{s^{\mathrm{mul}}} (1+\lfloor\beta\rfloor)^{s^{\mathrm{mul}}+D(1+\lfloor\beta\rfloor)s^{\mathrm{mul}}}.
\end{aligned}
$$

Finally, we define a concatenated and parallelized neural networks as

$$
\begin{aligned}
\Phi_\varepsilon^{\mathrm{pol},1} &\coloneqq [\Phi^{\mathrm{mul},\gamma_1}_{\varepsilon/cKM}, \Phi^{\mathrm{mul},\gamma_2}_{\varepsilon/cKM}, \ldots, \Phi^{\mathrm{mul},\gamma_K}_{\varepsilon/cKM}], \\
\Phi_\varepsilon^{\mathrm{pol},2} &\coloneqq [\Phi_\varepsilon^{\mathrm{tlr},1}, \Phi_\varepsilon^{\mathrm{tlr},2}, \ldots, \Phi_\varepsilon^{\mathrm{tlr},m}], \\
\Phi_\varepsilon^{\mathrm{pol}} &\coloneqq \Phi_\varepsilon^{\mathrm{pol},2} \odot \Phi_\varepsilon^{\mathrm{pol},1}.
\end{aligned}
$$

Then, we can simply obtain the error bound as

$$
\sup_{x\in[0,1]^D} \left| \sum_{|\gamma|\leq\lfloor\beta\rfloor} \widetilde{c}_{\lambda,\gamma} x^\gamma - R(\Phi_\varepsilon^{\mathrm{pol}})_\lambda(x) \right| \leq cMK \frac{\varepsilon}{cMK} = \varepsilon.
$$

About parameters of the network $\Phi_\varepsilon^{\mathrm{pol}}$, the result in Remark 13 and 14 shows the following inequalities:

$$
\begin{aligned}
W(\Phi_\varepsilon^{\mathrm{pol}}) &\leq 2W(\Phi_\varepsilon^{\mathrm{pol},2}) + 2W(\Phi_\varepsilon^{\mathrm{pol},1}) \\
&\leq 2Km + 2(2Kc_3^{\mathrm{pol}} \varepsilon^{-d/(1+\lfloor\beta\rfloor)} + 4KD(1+\lceil\log_2\beta\rceil)(11+(1+\beta)/d), \\
L(\Phi_\varepsilon^{\mathrm{pol}}) &= L(\Phi_\varepsilon^{\mathrm{pol},2}) + L(\Phi_\varepsilon^{\mathrm{pol},1}) \leq 1 + (1+\lceil\log_2\beta\rceil)(11+(1+\beta)/d), \\
B(\Phi_\varepsilon^{\mathrm{pol}}) &\leq \max\left\{ cM, c_4^{\mathrm{pol}} \varepsilon^{-s_2^{\mathrm{pol}}} \right\}.
\end{aligned}
$$

Then, we obtain the statement. □

(ii) Approximation for Function on a Hypercube: We investigate simultaneous approximation of functions on several hypercubes, namely, we approximate set of functions from a set $\{(f_0 + M + 1)\mathbb{1}_I \mid I \in \mathcal{I}\}$. To make the functions positive, we define $f_1 \coloneqq f_0 + M + 1$. Notice that $f_1 \in \mathcal{H}(\beta, [0,1]^D, 2M+1)$ and $1 \leq f_1(x) \leq 2M+1$ for any $x \in [0,1]^D$. Let $\mathcal{I}$ be a minimum γ-covering of $\mathrm{Supp}(\mu)$. Consequently, $\mathcal{I}$ is regarded as an ordered set and is accompanied by an index set $\Lambda = [\mathrm{card}\,\mathcal{I}]$. We define a bijective map $\psi : \mathcal{I} \to \Lambda$ which returns a corresponding index in Λ of $I \in \mathcal{I}$. Also, let $\Xi : \mathcal{I} \to 2^\Lambda$ be a set-function defined by $\Xi(I) = \{I' \in \mathcal{I} \mid (I \oplus 3\gamma/2) \cap I' \neq \phi\}$.

We first approximate Taylor polynomials of Hölder class functions. For any fixed $I \in \mathcal{I}$, we define its center of I as $(\iota_1, \ldots, \iota_D)$. Define a neural network $\Phi_\gamma^{\mathrm{cut},I} = (2M+2, 0) \odot (A_\ell^2, -D) \odot [(A_1^1, b_1^1), \ldots, (A_D^1, b_D^1)]$ where $A_\ell^1, b_\ell^1, A_\ell^2$ are defined by the following parameters:

$$
A_\ell^1 \coloneqq \begin{pmatrix} e_\ell^\top & e_\ell^\top & e_\ell^\top & e_\ell^\top \\ 0 & 0 & 0 & 0 \end{pmatrix}^\top, \quad b_\ell^1 \coloneqq \begin{pmatrix} -\iota_\ell + \gamma & -\iota_\ell + \gamma/2 & -\iota_\ell - \gamma/2 & -\iota_\ell - \gamma \end{pmatrix},
$$

and

$$A_\ell^2 \coloneqq (\underbrace{2/\gamma, -2/\gamma, -2/\gamma, 2/\gamma, 2/\gamma, -2/\gamma, -2/\gamma, 2/\gamma, \ldots, 2/\gamma, -2/\gamma, -2/\gamma, 2/\gamma}_{4D}, 1/(2M+2)).$$

Then, a function by the neural network $R(\Phi_\gamma^{\mathrm{cut},I})[0,1]^D \times \mathbb{R}_{\geq} \to \mathbb{R}_{\geq}$ has the following form

$$R(\Phi_\gamma^{\mathrm{cut},I}) = (2M+2)\rho\left(\sum_{\ell=1}^{D} \mathbb{1}_\gamma^{I,\ell}(x_\ell) + \frac{y}{2M+2} - D\right), \tag{7}$$

where $\mathbb{1}_\gamma^{I,\ell} : \mathbb{R} \to [0,1]$ is the approximated indicator function with the form:

$$\mathbb{1}_\gamma^{I,\ell}(z) = \begin{cases} 0 & \text{if } z \leq \iota_\ell - \gamma, \\ \frac{z-(\iota_\ell-\gamma)}{\gamma/2} & \text{if } \iota_\ell - \gamma < z \leq \iota_\ell - \frac{\gamma}{2}, \\ 1 & \text{if } \iota_\ell - \frac{\gamma}{2} < z \leq \iota_\ell + \frac{\gamma}{2}, \\ \frac{(\iota_\ell+\gamma)-z}{\gamma/2} & \text{if } \iota_\ell + \frac{\gamma}{2} < z \leq \iota_\ell + \gamma, \\ 0 & \text{if } \iota_\ell + \gamma < z. \end{cases}$$

Then, we can claim that the function $R(\Phi_\gamma^{\mathrm{cut},I})(x,y)$ approximates a function $(x,y) \mapsto y\mathbb{1}_I(x)$. Its properties are summarized in the following remark:

Remark 19. *For any $y \in [0, 2M+2]$, $R(\Phi_\gamma^{\mathrm{cut},I})(x,y) = y$ holds for any $x \in I$. Also $R(\Phi_\gamma^{\mathrm{cut},I})(x,y) \leq y$ holds for any $x \in I \oplus \gamma/2$. Also, $R(\Phi_\gamma^{\mathrm{cut},I})(x,y) = 0$ holds for $x \notin \mathrm{I} \oplus \gamma/2$ and any y. Furthermore, we obtain the following properties:*

1. $W(\Phi_\gamma^{\mathrm{cut},I}) = 24D + 6$,
2. $L(\Phi_\gamma^{\mathrm{cut},I}) = 3$,
3. $B(\Phi_\gamma^{\mathrm{cut},I}) \leq \max\{1, 2M+2, 1/(2M+2), D, 1+\gamma, 2/\gamma\}$.

Then, we define a neural network to approximate f_I, which is a $\varepsilon/2$-accuracy Taylor polynomial of f_1. For any $I \in \mathcal{I}$ and any point $x_I \in I$, take $f_I(x)$ as a Taylor polynomial function as Lemma 16 with setting $\bar{x} \leftarrow x_I$ and $f \leftarrow f_1$. For a fixed $\varepsilon \in (0, \varepsilon_0^{\mathrm{pol}}/2)$, let $\Phi_{\varepsilon/2}^{\mathrm{pol}}$ be a neural network constructed in Lemma 18 with $\varepsilon \leftarrow \varepsilon/2$, $m \leftarrow \operatorname{card}\mathcal{I}$, $(x_\lambda)_{\lambda=1}^m \leftarrow (x_{\psi^{-1}(\lambda)})_{\lambda=1}^{\operatorname{card}\mathcal{I}}$, $(c_{\lambda,\alpha})_{\lambda=1}^m \leftarrow (\partial^\alpha f(x_{\psi^{-1}(\lambda)})/\alpha!)_{\lambda=1}^{\operatorname{card}\mathcal{I}}$ and $B \leftarrow \overline{C}(2M+1)$, where $\overline{C} = \overline{C}(\beta, D, 2M+1)$ appearing in Lemma 16. Then, we obtain

$$\sup_{I\in\mathcal{I}} \sup_{x\in[0,1]^D} \left| f_I(x) - R(\Phi_{\varepsilon/2}^{\mathrm{pol}})_{\psi(I)}(x) \right| \leq \frac{\varepsilon}{2}. \tag{8}$$

Also, we construct a neural network to aggregate the outputs of $\Phi_{\varepsilon/2}^{\mathrm{pol}}$. Let us define a neural network $\Phi^{\mathrm{filter},i} : \mathbb{R}^{D+m} \to \mathbb{R}^{D+1}$ which picks up the first D inputs and $D+i$-th input as

$$\Phi^{\mathrm{filter},i} \coloneqq \left(\begin{pmatrix} I_D & e_i^\top \\ O_D & e_i^\top \end{pmatrix}, 0_{D+1} \right).$$

Then, we define

$$\Phi^{\mathrm{simul}}_{\varepsilon/2} \coloneqq [\Phi^{\mathrm{cut},\psi^{-1}(1)}_{\gamma} \odot \Phi^{\mathrm{filter},1}, \ldots, \Phi^{\mathrm{cut},\psi^{-1}(\mathrm{card}\,\mathcal{I})}_{\gamma} \odot \Phi^{\mathrm{filter},\mathrm{card}\,\mathcal{I}}] \odot [\Phi^{Id}_{D,L}, \Phi^{\mathrm{pol}}_{\varepsilon/2}] \tag{9}$$

where $\Phi^{Id}_{D,L}$ is the neural network version of the identity function $\mathbb{R}^D \to \mathbb{R}^D$ with the number of layers $L = L(\Phi^{\mathrm{pol}}_{\varepsilon/2})$.

(iii) Approximation of Smooth Functions with Finite Layers: We develop a Taylor polynomial approximation for f_0 on hypercubes $\mathcal{I}$ with finite layers. To avoid divergence of the number of layers of DNNs, we provide novel techniques for neural networks. We first need the following lemma.

Lemma 20. *Let $\mathcal{I}$ be a minimum γ-covering of* $\mathrm{Supp}(\mu)$. *Then, there exists a disjoint partition $\{\mathcal{I}_i\}_{i=1}^{5^D}$ of $\mathcal{I}$ such that $\mathcal{I} = \bigcup_{i=1}^{5^D} \mathcal{I}_i$ and $d(I_j, I_k) \geq \gamma$ hold for any $I_j \neq I_k \in \mathcal{I}_i$ if* $\mathrm{card}(\mathcal{I}_i) \geq 2$.

Proof of Lemma 20 We construct a partition explicitly. Generate a sequence $(\mathcal{I}_i)_{i\geq 1}$ inductively by the following procedure. Firstly, let $\mathcal{I}_i = \{\}$ for all $i \geq 1$. Starting from $i = 1$, repeat the following procedures for each i. Choose any $I \in \left\{ I' \in \mathcal{I} \setminus \bigcup_{\ell=1}^{i} \mathcal{I}_\ell \mid \min_{I'' \in \mathcal{I}_i} d(I', I'') \geq \gamma \right\}$ and let $\mathcal{I}_i \leftarrow \mathcal{I}_i \cup \{I\}$ until we fail to take I.

We now prove $\mathcal{I}_{5^D+1} = \emptyset$ by contradiction. Suppose $\mathcal{I}_{5^D+1} \neq \emptyset$. Then also $\mathcal{I}_i \neq \emptyset$ for $1 \leq i \leq 5^D$. Take any $I \in \mathcal{I}_{5^D+1}$. By construction, we can always take $I_i \in \mathcal{I}_i$ such that $d(I, I_i) < \gamma$ for any $i \in [5^D]$. The set $\{I, I_1, \ldots, I_{5^D}\}$ is covered by $I \oplus 2\gamma$. But $I \oplus 2\gamma$ can be covered by 5^D hypercubes with diameter γ, which contradicts the fact that $\mathcal{I}$ is a minimum covering.

Since $\mathcal{I}_{5^D+1} = \emptyset$, $\mathcal{I}_i = \emptyset$ holds for all $i > 5^D$. □

B.1.3 Proof of Approximation Error bound

Proof of Theorem 5 Let $\mathcal{I}$ be a minimum γ-covering of $\mathrm{Supp}(\mu)$. By Lemma 20, $\mathcal{I}$ can be partitioned into $\mathcal{I}_1, \ldots, \mathcal{I}_{5^D}$ such that $\mathcal{I} = \bigcup_{i=1}^{5^D} \mathcal{I}_i$ and for all $i \in [5^D]$, $d(I_j, I_k) \geq \gamma$ for any $I_j, I_k \in \mathcal{I}_i$ satisfying $I_j \neq I_k$.

We define a neural network that summate the output of $\Phi^{\mathrm{simul}}_{\varepsilon/2}$ in each partition $\mathcal{I}_i$. We provide parameters $A^{\mathrm{sum}}_{ij} \coloneqq \mathbb{1}(\psi^{-1}(j) \in \mathcal{I}_i)$ and $A^{\mathrm{sum}} \coloneqq (A^{\mathrm{sum}}_{ij})_{i,j} \in \mathbb{R}^{5^D \times \mathrm{card}\,\mathcal{I}}$. Then, we define a neural network $\Phi^{\mathrm{sum}} \coloneqq (A^{\mathrm{sum}}, 0_{5^D})$. The function by Φ^{sum} has the following representation:

$$R(\Phi^{\mathrm{sum}})(x_1, \ldots, x_{\mathrm{card}\,\mathcal{I}}) = \left(\sum_{I \in \mathcal{I}_1} x_{\psi(I)}, \ldots, \sum_{I \in \mathcal{I}_{5^D}} x_{\psi(I)} \right).$$

Then, we construct a neural network $\Phi^{f_1}_{\varepsilon}$ to approximate $f_1 \coloneqq f_0 + M + 1$. Let $\Phi^{\mathrm{simul}}_{\varepsilon/2}$ as defined in (9). Define a neural network $\Phi^{f_1}_{\varepsilon}$ as $\Phi^{f_1}_{\varepsilon} \coloneqq \Phi^{\max,5^D} \odot \Phi^{\mathrm{sum}} \odot \Phi^{\mathrm{simul}}_{\varepsilon/2}$. We obtain a function with a form $R(\Phi^{f_1}_{\varepsilon}) = \max_{i \in [5^D]} \sum_{I \in \mathcal{I}_i} R(\Phi^{\mathrm{simul}}_{\varepsilon/2})_{\psi(I)}$.

Next, we bound an approximation error of $\Phi^{f_1}_{\varepsilon}$. When $x \in I$ for some $I \in \mathcal{I}$,

$$R(\Phi^{f_1}_{\varepsilon})(x) = \max_{I' \in \Xi(I)} R(\Phi^{\mathrm{simul}}_{\varepsilon/2})_{\psi(I')}(x)$$

$$\leq \max_{I' \in \Xi(I)} R(\Phi^{\mathrm{pol},\varepsilon/2})_{\psi(I')}(x),$$

where we used the fact that

$$\max_{i \in [5^D]} \sum_{I' \in \mathcal{I}_i} R(\Phi^{\mathrm{simul}}_{\varepsilon/2})_{\psi(I')}(x) = \max_{I' \in \Xi(I)} R(\Phi^{\mathrm{simul}}_{\varepsilon/2})_{\psi(I')}(x).$$

In other words, when computing $R(\Phi^{f_1}_{\varepsilon})(x)$ we only have to take maximum over the outputs of $R(\Phi^{\mathrm{simul}}_{\varepsilon/2})(x)$ related to hypercubes near x. This follows from the fact that $R(\Phi^{f_1}_{\varepsilon})_{\psi(I')}(x) = 0$ for $I' \notin \Xi(I)$ and $d(I', I'') > \gamma$ holds for $I' \neq I'' \in \mathcal{I}_i$ for all i. The last inequality follows by construction of $\Phi^{\mathrm{simul}}_{\varepsilon/2}$. For a further parameter tuning, we set $\gamma = D^{-1}(3M)^{-1/\beta}\varepsilon^{1/\beta}$.

Given $\varepsilon \in (0, 1)$, we can ensure $0 \leq R(\Phi^{\mathrm{simul}}_{\varepsilon/2})_{\psi(I)}(x) \leq 2M + 2$ for all $I \in \mathcal{I}$ by Remark 19, since $R(\Phi^{\mathrm{simul}}_{\varepsilon/2})_{\psi(I)}$ approximates f_I which is a $\varepsilon/2$-accuracy Taylor polynomial of $f_1 \in [1, 2M + 1]$. The error is bounded as

$$\begin{aligned}
&\left|R(\Phi^{f_1}_{\varepsilon})(x) - f_1(x)\right| \\
&= \max\left\{\max_{I' \in \Xi(I)} R(\Phi^{\mathrm{simul}}_{\varepsilon/2})_{\psi(I')}(x) - f_1(x), f_1(x) - \max_{I' \in \Xi(I)} R(\Phi^{\mathrm{simul}}_{\varepsilon/2})_{\psi(I')}(x)\right\} \\
&\leq \max\left\{\max_{I' \in \Xi(I)} R(\Phi^{\mathrm{pol},\varepsilon/2})_{\psi(I')}(x) - f_1(x), f_1(x) - R(\Phi^{\mathrm{pol},\varepsilon/2})_{\psi(I)}(x)\right\} \\
&\leq \max_{I' \in \Xi(I)} \left|R(\Phi^{\mathrm{pol},\varepsilon/2})_{\psi(I')}(x) - f_1(x)\right| \\
&\leq \max_{I' \in \Xi(I)} \left|R(\Phi^{\mathrm{pol},\varepsilon/2})_{\psi(I')}(x) - f_{I'}(x)\right| + \max_{I' \in \Xi(I)} |f_{I'}(x) - f_1(x)| \\
&\leq \frac{\varepsilon}{2} + D^{\beta} M \left(\frac{3\gamma}{2}\right)^{\beta} = \varepsilon,
\end{aligned}$$

where the second last inequality follows from Lemma 16 because $f_{I'}$ for $I' \in \Xi(I)$ is a Taylor polynomial around some $x_{I'} \in I'$ satisfying $\|x_{I'} - x_I\| \leq 3\gamma/2$. The last inequality follows from (8) and Lemma 16.

From the result for f_1, we provide an approximation for f_0. To the end, let us define a neural network $\Phi^{\mathrm{mod},M}$ as $\Phi^{\mathrm{mod},M} \coloneqq (-1, M) \odot (-1, 2M) \odot (1, -1)$. Its realization is $R(\Phi^{\mathrm{mod},M})(x) = \min(\max(1, x), 2M + 1) - (M + 1)$ for any $x \in \mathbb{R}$. By Remark 13 and 14, the following properties holds:

1. $W(\Phi^{\mathrm{mod},M}) = 12$,
2. $L(\Phi^{\mathrm{mod},M}) = 3$,
3. $B(\Phi^{\mathrm{mod},M}) \leq \max\{2M, 1\}$.

Then, we define $\Phi^{f_0}_{\varepsilon} \coloneqq \Phi^{\mathrm{mod},M} \odot \Phi^{f_1}_{\varepsilon}$. Then, an approximation error by $\Phi^{f_0}_{\varepsilon}$ is bounded as

$$\sup_{x \in \mathrm{Supp}(\mu)} \left|R(\Psi^{f_0}_{\varepsilon})(x) - f_0(x)\right|$$

$$
\begin{aligned}
&= \sup_{x \in \mathrm{Supp}(\mu)} \left| \min(\max(1, R(\Phi_\varepsilon^{f_1})(x), 2M+1) - (f_0(x) + M + 1) \right| \\
&\leq \sup_{x \in \mathrm{Supp}(\mu)} \left| R(\Phi_\varepsilon^{f_1})(x) - f_1(x) \right| \\
&\leq \varepsilon.
\end{aligned}
$$

Here, note that $\mathrm{card}\, \mathcal{I} \leq c_\mu \gamma^{-d}$. Combined with Remark 13 and 14, $\Psi_\varepsilon^{f_0}$ has the following properties:

$$
\begin{aligned}
W(\Psi_\varepsilon^{f_0}) &\leq 2W(\Phi^{\mathrm{mod},M}) + 2W(\Phi^{\max,5^D}) \\
&\quad + 2(\mathrm{card}\, \mathcal{I}) W(\Phi_\gamma^{\mathrm{cut},\psi^{-1}(1)} \odot \Phi^{\mathrm{filter},1}) + 2W(\Phi_{D,L}^{Id}) + 2W(\Phi_{\varepsilon/2}^{\mathrm{pol}}) \\
&\leq 2((50D+17)c_\mu D^d (3M)^{d/\beta} + 2D(11 + (1+\beta)/d) c_1^{\mathrm{pol}} (2^{d/\beta} + C D^d (3M)^{d/\beta})) \varepsilon^{-d/\beta} \\
&\quad + 2(12 + 42 \times 5^D + 2D + 2D(11 + (1+\beta)/d)(1 + \lceil \log_2 \beta \rceil)), \\
L(\Psi_\varepsilon^{f_0}) &= L(\Phi^{\mathrm{mod},M}) + L(\Phi^{\max,5^D}) + L(\Phi_\gamma^{\mathrm{cut},\psi^{-1}(1)}) + L(\Phi^{\mathrm{filter},1}) + L(\Phi_{\varepsilon/2}^{\mathrm{pol}}) \\
&\leq 11 + 2D \log_2 5 + (11 + (1+\beta)/d)(1 + \lceil \log_2 \beta \rceil), \\
B(\Psi_\varepsilon^{f_0}) &\leq \max \left\{ 1, 2M+2, 1/(2M+2), D, 1+\gamma, 2/\gamma, c_4^{\mathrm{pol}} \varepsilon^{-s_2^{\mathrm{pol}}} \right\}.
\end{aligned}
$$

By adjusting several constants, we obtain the statement. □

B.2 Proof of Bound for Generalization Error

The proof of Theorem 7 follows proof techniques developed by several studies (Suzuki, 2017; Schmidt-Hieber, 2020; Imaizumi and Fukumizu, 2019) with some adaptation for our setting.

Proof of Theorem 7 Without loss of generality, we can assume Theorem 5 holds for any approximation accuracy $\varepsilon \in (0, 1)$. In the beginning, we apply the optimal condition of $\widehat{f}$ and derive a basic inequality. Recall that $\widehat{f}$ is defined as $\widehat{f}(x) = \max\{-C_B, \min\{C_B, \widetilde{f}(x)\}\}$, where

$$
\widetilde{f} \in \mathrm{argmin}_{f \in \mathcal{F}(W,L,B)} \sum_{i=1}^n (Y_i - f(X_i))^2. \tag{10}
$$

Then, it is easily seen that $\|\widehat{f} - f_0\|_{L^2(\mu)} \leq \|\widetilde{f} - f_0\|_{L^2(\mu)}$. Hence we regard $\widehat{f}$ as the unclipped estimator $\widetilde{f}$ without loss of generality. By definition of $\widehat{f}$, $\|Y - \widehat{f}\|_n^2 \leq \|Y - f\|_n^2$ for any $f \in \mathcal{F}(W, L, D)$. By substituting $Y_i = f_0(X_i) + \xi_i$, we obtain the base inequality as

$$
\|\widehat{f} - f_0\|_n^2 \leq \|f - f_0\|_n^2 + \frac{2}{n} \sum_{i=1}^n \xi_i \Big(\widehat{f}(X_i) - f(X_i) \Big). \tag{11}
$$

To bound the two terms in (11), we provide a neural network $\Psi_\varepsilon^{f_0}$ as constructed in the proof of Theorem 5 for approximating f_0. Specifically, we set a triple (W, L, B) as in Theorem 5 with accuracy $\varepsilon \leftarrow n^{d/(2\beta+d)}$. We define $f^* = R(\Psi_\varepsilon^{f_0})$. Note that by construction, $|\widehat{f}| \leq C_B$ and $|f^*| \leq M$.

We divide the proof into following 3 steps.

Step 1. Derive an upper bound of $\|\widehat{f} - f^*\|^2_{L^2(\mu)}$ using its empirical counterpart $\|\widehat{f} - f^*\|^2_n$.

Step 2. Evaluate the variance term $(1/n)\sum_{i=1}^n \xi_i\Big(\widehat{f}(X_i) - f(X_i)\Big)$.

Step 3. Combine the results of step 1 and step 2.

Step 1. Upper Bound of $\|\widehat{f} - f^*\|^2_{L^2(\mu)}$: We prepare an evaluation of the entropy number bound. Let $\mathcal{N}(\varepsilon, \mathcal{F}, \|\cdot\|)$ be the minimum ε-covering number of $\mathcal{F}$ by a norm $\|\cdot\|$. Similar results are well-known (e.g., Anthony and Bartlett (2009); Schmidt-Hieber (2020)). However, our setting, such as a parameter bound, is slightly different from those of the studies. Hence, we provide the following lemma and its full proof.

Lemma 21 (Covering entropy bound for $\mathcal{F}$)**.** *Let $\mathcal{F} = \mathcal{F}(W, L, B)$ be a space of neural networks with the number of nonzero weights, the number of layers, and the maximum absolute value of weights bounded by W, L and B respectively. Then,*

$$\log \mathcal{N}\Big(\varepsilon, \mathcal{F}(W, L, B), \|\cdot\|_{L^\infty(\mu)}\Big) \leq W \log\left(\frac{2LB^L(W+1)^L}{\varepsilon}\right).$$

Before presenting proof of Lemma 21, we need the following preliminary result, which makes it possible to regard neural networks in $\mathcal{F}(W, L, B)$ share the same dimensional parameter space.

Lemma 22. *Let $\mathcal{F}(W, L, B)$ be a class of neural networks. Define*

$$S_B(p, q) \coloneqq \big\{(A, b) \mid A \in [-B, B]^{p\times q}, b \in [-B, B]^p\big\},$$
$$\mathcal{G}(W, L, B) \coloneqq S_B(1, W) \times S_B(W, W) \times \cdots \times S_B(W, D).$$

Then there exists a map $Q : \mathcal{F}(W, L, B) \to \mathcal{G}(W, L, B)$ such that

$$R(\Phi)(x) = A_L^Q \rho_{b_{L-1}^Q} \circ \cdots \circ A_2^Q \rho_{b_1^Q}(A_1^Q x) + b_L^Q,$$

where $((A_L^Q, b_L^Q), \ldots, (A_1^Q, b_1^Q)) = Q(R(\Phi)) \in \mathcal{G}(W, L, B)$.

Proof of Lemma 22 Take any $R(\Phi) \in \mathcal{F}(W, L, B)$. Write $\Phi = ((A_L, b_L), \ldots, (A_1, b_1))$ and assume $A_l \in \mathbb{R}^{p_l \times p_{l-1}}$ and $b_l \in \mathbb{R}^{p_l}$. Consider (A_{l-1}, b_{l-1}) for $l = 2, \ldots, L$. Since the number of nonzero parameters are bounded by W, the number of nonzero parameters in $A_{l-1}x + b_{l-1}$ for any $x \in \mathbb{R}^{p_{l-2}}$ is at most W. For $p_{l-1} > W$, without loss of generality, we can assume the W-th, $\ldots$, p_l-th element of $A_{l-1}x + b_{l-1}$ are 0. Let $A'_{l-1} \in \mathbb{R}^{W\times p_{l-1}}$ be the upper-left part of A_{l-1} and $A'_l \in \mathbb{R}^{p_l \times W}$ be the upper-left part of A_l. Also let $b'_{l-1} \in \mathbb{R}^{p_{l-1}}$ be the first W elements of b_{l-1}. Then $A'_l(A'_{l-1}x + b'_{l-1}) = A_l(A_{l-1}x + b_{l-1})$. For $p_{l-1} < W$, we can simply extend A_l, A_{l-1}, b_{l-1} to be in $\mathbb{R}^{p_l\times W}, \mathbb{R}^{W\times p_{l-2}}, \mathbb{R}^W$, respectively. Applying this procedure multiple times yields the conclusion. □

Proof of Lemma 21 Firstly, consider neural networks $\Phi = Q(R(\Phi)) = ((A_L, b_L), \ldots, (A_1, b_1))$ and $\Phi' = Q(R(\Phi')) = ((A'_L, b'_L), \ldots, (A'_1, b'_1))$, such that for each $l \in [L]$, (A'_l, b'_l) has elements at most ε apart from (A_l, b_l). Let us write $(A_l, b_l) = ((a^l_{ij})_{ij}, (b^l_i)_i)$ and $(A'_l, b'_l) = ((a'^l_{ij})_{ij}, (b'^l_i)_i)$, then define functions in internal layers as

$$h^l(x) \coloneqq (h^l_1(x), \ldots, h^l_{p_l}(x))^\top \coloneqq A_l x + b_l,$$

$$h'^l(x) \coloneqq (h'^l_1(x), \ldots, h'^l_{p_l}(x))^\top \coloneqq A'_l x + b'_l,$$

and

$$g^l(x) \coloneqq (g^l_1(x), \ldots, g^l_{p_l}(x))^\top \coloneqq h^l(x) - h'^l(x).$$

For any $E \geq 0$, we can bound supremums of the functions as

$$\begin{aligned}\sup_{x \in [-E,E]^{p_{l-1}}} \left| g^l_i(x) \right| &\leq \sum_{j=1}^{p_{l-1}} \left| a_{ij} - a'_{ij} \right| |x_j| + \left| b_i - b'_i \right| \\ &\leq p_{l-1} \varepsilon E + \varepsilon \leq (WE+1)\varepsilon \leq (W+1)E\varepsilon. \end{aligned} \tag{12}$$

Also, we have

$$\begin{aligned}\sup_{x \in [-E,E]^{p_{l-1}}} \left| h^l_i(x) \right| &\leq \sum_{j=1}^{p_{l-1}} |a_{ij}| |x_j| + |b_i| \\ &\leq p_{l-1} BE + B \leq (WE+1)B \leq (W+1)EB. \end{aligned} \tag{13}$$

Since a Lipschitz constant of the ReLU actiavtion function is 1 for each coordinate, we can apply (13) repeatedly for $(h^1_1, \ldots, h^1_D), \ldots, (h^L_1, \ldots, h^L_{p_L})$. Then, we obtain the bound for $\sup_{x \in [0,1]^D} |R(\Phi) - R(\Phi')|$ as

$$\begin{aligned}&\sup_{x \in [0,1]^D} \left| R(\Phi) - R(\Phi') \right| \\ &= \left| h^L \circ \rho \circ h^{L-1} \circ \rho \circ \cdots \circ \rho \circ h^2 \circ \rho \circ h^1(x) - h'^L \circ \rho \circ h'^{L-1} \circ \rho \circ \cdots \circ \rho \circ h'^2 \circ \rho \circ h'^1(x) \right| \\ &\leq \left| h^L \circ \rho \circ h^{L-1} \circ \rho \circ \cdots \circ \rho \circ h^2 \circ \rho \circ h^1(x) - h^L \circ \rho \circ h^{L-1} \circ \rho \circ \cdots \circ \rho \circ h^2 \circ \rho \circ h'^1(x) \right| \\ &+ \left| h^L \circ \rho \circ h^{L-1} \circ \rho \circ \cdots \circ \rho \circ h^2 \circ \rho \circ h'^1(x) - h^L \circ \rho \circ h^{L-1} \circ \rho \circ \cdots \circ \rho \circ h'^2 \circ \rho \circ h'^1(x) \right| \\ &\vdots \\ &+ \left| h^L \circ \rho \circ h'^{L-1} \circ \rho \circ \cdots \circ \rho \circ h'^2 \circ \rho \circ h'^1(x) - h'^L \circ \rho \circ h'^{L-1} \circ \rho \circ \cdots \circ \rho \circ h'^2 \circ \rho \circ h'^1(x) \right| \\ &\leq L(W+1)^L B^{L-1} \varepsilon. \end{aligned}$$

Note that E in (12) and (13) is bounded by $B^\ell \leq B^L$ for any $\ell = 1, ..., L$. Then, we discretize W parameters with $\varepsilon / L(W+1)^L B^{L-1}$ grid size. Thus we obtain the covering number bound in the statement. □

Next, we bound the term $\|\widehat{f} - f^*\|^2_{L^2(\mu)}$. Let us define $N = \mathcal{N}(\delta, \mathcal{F}(W,L,B), \|\cdot\|_{L^\infty(\mu)})$, and also $\{f_1, \ldots, f_N\}$ be a set of centers of the minimal δ-cover of $\mathcal{F}(W,L,B)$ with $\|\cdot\|_{L^\infty(\mu)}$ norm. Without loss of generality, we can assume $|f_j| \leq M$ for all $j \in [N]$. Take any random $f_{\widehat{j}} \in \{f_1, \ldots, f_N\}$ so that $\|\widehat{f} - f_{\widehat{j}}\|_{L^\infty(\mu)} \leq \delta$. By the triangle inequality, we have

$$\|\widehat{f} - f^*\|^2_{L^2(\mu)} \leq 2\|\widehat{f} - f_{\widehat{j}}\|^2_{L^2(\mu)} + 2\|f_{\widehat{j}} - f^*\|^2_{L^2(\mu)} \leq 2\delta^2 + 2\|f_{\widehat{j}} - f^*\|^2_{L^2(\mu)}.$$

We bound the term $\|f_j - f^*\|^2_{L^2(\mu)}$ uniformly for all $j \in [N]$ in order to bound the random quantity $\|f_{\widehat{j}} - f^*\|^2_{L^2(\mu)}$. Firstly, from Bernstein's inequality, for independent and identically distributed random variables Z_i satisfying $|Z_i| \leq c$ and $E[Z_i] = 0$, it holds that

$$P(|\bar{Z}| \geq u) \leq \exp\left(-\frac{nu^2}{2\tau^2 + 2cu/3}\right)$$

for any $u > 0$, where $\tau^2 \coloneqq \mathrm{Var}(Z_i)$. Substitute $u \leftarrow \max\left\{v, (1/2)\|f_j - f^*\|^2_{L^2(\mu)}\right\}$, $Z_i \leftarrow (f_j(X_i) - f^*(X_i))^2 - E[(f_j(X_i) - f^*(X_i))^2]$ and $c \leftarrow 8M^2$. Notice that

$$\begin{aligned}\tau^2 &= E\Big[\big((f_j(X_i) - f^*(X_i))^2 - E[(f_j(X_i) - f^*(X_i))^2]\big)^2\Big] \\ &\leq 4M^2\|f_j - f^*\|^2_{L^2(\mu)} \leq 8M^2 u,\end{aligned}$$

holds. Then, for fixed j, we bound the tail probability of $\|f_j - f^*\|_n$ as

$$P\Big(\|f_j - f^*\|^2_{L^2(\mu)} \geq \|f_j - f^*\|^2_n + u\Big) \leq \exp\left(-\frac{3nv}{64M^2}\right).$$

By the uniform bound argument, $\|f_j - f^*\|^2_{L^2(\mu)} \geq \|f_j - f^*\|^2_n + u$ holds for all $j \in [N]$ with probability at most $N\exp\big(-3nv/(64M^2)\big)$. Substitute $v \leftarrow 64M^2(n^{d/(2\beta+d)} + \log N)/(3n)$ together with the trivial inequality $u \leq v + (1/2)\|f_j - f^*\|^2_{L^2(\mu)}$ leads to the following inequality

$$\|f_j - f^*\|^2_n + u \leq \|f_j - f^*\|^2_n + \frac{64M^2 n^{-2\beta/(2\beta+d)}}{3} + \frac{64M^2 \log N}{3n} + \frac{1}{2}\|f_j - f^*\|^2_{L^2(\mu)}.$$

Hence, for all $j \in [N]$, the following inequality

$$\|f_j - f^*\|^2_{L^2(\mu)} \leq 2\|f_j - f^*\|^2_n + \frac{128M^2 n^{-2\beta/(2\beta+d)}}{3} + \frac{128M^2 \log N}{3n}$$

holds with probability at least $1 - \exp\big(-n^{d/(2\beta+d)}\big)$.

Back to the inequality $\|\widehat{f} - f^*\|^2_{L^2(\mu)} \leq 2\delta^2 + 2\|f_{\widehat{j}} - f^*\|^2_{L^2(\mu)}$ with $\delta \leftarrow n^{-\beta/(2\beta+d)}$ and with Lemma 21, we obtain

$$\begin{aligned}&\|\widehat{f} - f^*\|^2_{L^2(\mu)} \\ &\leq 2n^{-2\beta/(2\beta+d)} + 4\|f_{\widehat{j}} - f^*\|^2_n + \frac{2^8 M^2 n^{-2\beta/(2\beta+d)}}{3} + \frac{2^8 M^2 \log N}{3n} \\ &\leq \left(10 + \frac{2^8 M^2}{3}\right) n^{-2\beta/(2\beta+d)} + 8\|\widehat{f} - f^*\|^2_n + \frac{2^8 M^2 W}{3n}\log\Big(2n^{\beta/(2\beta+d)} L(W+1)^L B^L\Big),\end{aligned} \tag{14}$$

with probability at least $1 - \exp\big(-n^{d/(2\beta+d)}\big)$.

Step 2. Evaluate Variance: Let $\mathcal{G}_\delta \coloneqq \{g \mid g \coloneqq f - f', \|g\|_{L^\infty(\mu)} \leq \delta, f, f' \in \mathcal{F}\}$. Given the observed variables $X_1, \ldots, X_n$, we regard $(1/n)\sum_{i=1}^n \xi_i g(X_i)$ as a stochastic process

indexed by $g \in \mathcal{G}_\delta$. By the Gaussian concentration inequality(Theorem 2.5.8 in (Giné and Nickl, 2016)),

$$P\left(\sup_{g\in\mathcal{G}_\delta}\left|\frac{1}{n}\sum_{i=1}^n \xi_i g(X_i)\right| \geq E\left[\sup_{g\in\mathcal{G}_\delta}\left|\frac{1}{n}\sum_{i=1}^n \xi_i g(X_i)\right|\right] + r_2\right) \leq \exp\left\{-\frac{nr_2^2}{2\sigma^2\delta^2}\right\}. \tag{15}$$

Also, by the covering entropy bound in (Giné and Nickl, 2016) combined with the inequality $\log \mathcal{N}(\varepsilon, \mathcal{G}_\delta, \|\cdot\|_n) \leq 2\log \mathcal{N}(\varepsilon, \mathcal{F}, \|\cdot\|_{L^\infty(\mu)})$, we obtain

$$\begin{aligned} E\left[\sup_{g\in\mathcal{G}_\delta}\left|\frac{1}{n}\sum_{i=1}^n \xi_i g(X_i)\right|\right] &\leq \frac{4\sqrt{2}\sigma}{\sqrt{n}}\int_0^{2\delta}\sqrt{\log(2\mathcal{N}(\varepsilon,\mathcal{G},\|\cdot\|_n))}\,\mathrm{d}\varepsilon \\ &\leq \frac{4\sigma\sqrt{2W}\delta}{\sqrt{n}}\log\left(\frac{L(W+1)^L B^L}{\delta}+1\right). \end{aligned} \tag{16}$$

Finally, combining (15) and (16) yields that

$$\begin{aligned} \sup_{g\in\mathcal{G}_\delta}\left|\frac{1}{n}\sum_{i=1}^n \xi_i g(X_i)\right| &\leq \frac{4\sigma\sqrt{2W}\delta}{\sqrt{n}}\log\left(\frac{L(W+1)^L B^L}{\delta}+1\right) + r_2 \\ &\leq \frac{1}{128}\delta^2 + 2^{11}\sigma^2\frac{W}{n}\left(\log\left(\frac{L(W+1)^L B^L}{\delta}+1\right)\right)^2 + r_2, \end{aligned}$$

with probability at least $1-\exp\bigl(-nr_2^2/(2\sigma^2\delta^2)\bigr)$. Here, the last inequality follows from the inequality $xy \leq (1/32)x^2 + 16y^2$. We substitute $r_2 \leftarrow 2^{-7}\delta^2$, then we have

$$\sup_{g\in\mathcal{G}_\delta}\left|\frac{1}{n}\sum_{i=1}^n \xi_i g(X_i)\right| \leq \frac{1}{64}\delta^2 + 2^{11}\sigma^2\frac{W}{n}\left(\log\left(\frac{L(W+1)^L B^L}{\delta}+1\right)\right)^2, \tag{17}$$

with probability $1-\exp\bigl(-n\delta^2/(2^{-13}\sigma^2)\bigr)$.

We bound $\|\widehat{f}-f^*\|_n^2$ with substituting $\delta \leftarrow \max\{2^{15}\sigma^2 n^{-\beta/(2\beta+d)}, 2\|\widehat{f}-f_0\|_n\}$. Here, we consider the following two cases. Firstly, suppose that $\|\widehat{f}-f^*\|_n \leq \delta$ holds. Then, we obtain

$$\begin{aligned} &\|\widehat{f}-f^*\|_n^2 \\ &\leq 2\|\widehat{f}-f_0\|_n^2 + 2\|f^*-f_0\|_n^2 \\ &\leq 4\|f^*-f_0\|_n^2 + 4\sup_{g\in\mathcal{G}_\delta}\left|\frac{1}{n}\sum_{i=1}^n \xi_i g(X_i)\right| \\ &\leq 4\|f^*-f_0\|_n^2 + 2^{13}\sigma^2\frac{W}{n}\left(\log\left(\frac{L(W+1)^L B^L}{\delta}+1\right)\right)^2 + \frac{\delta^2}{16} \\ &\leq 4\|f^*-f_0\|_n^2 + 2^{13}\sigma^2\frac{W}{n}\left(\log\left(L(W+1)^L B^L n^{\beta/(2\beta+d)}+1\right)\right)^2 + 2^{26}\sigma^4 n^{-2\beta/(2\beta+d)} \\ &\quad + \frac{1}{2}\|\widehat{f}-f^*\|_n^2 + \frac{1}{2}\|f^*-f_0\|_n^2, \end{aligned}$$

where the second inequality we apply (11) with $f \leftarrow f^*$. Therefore,

$$\begin{aligned}&\|\widehat{f} - f^*\|_n^2 \\ &\leq 9\|f^* - f_0\|_n^2 + 2^{14}\sigma^2\frac{W}{n}\Big(\log\Big(L(W+1)^L B^L n^{\beta/(2\beta+d)} + 1\Big)\Big)^2 + 2^{27}\sigma^4 n^{-2\beta/(2\beta+d)}. \quad (18)\end{aligned}$$

Secondly, suppose that $\|\widehat{f} - f^*\|_n \geq \delta$, namely, $2\|\widehat{f} - f_0\|_n \leq \|\widehat{f} - f^*\|_n$ holds. Then, we obtain

$$\|\widehat{f} - f^*\|_n^2 \leq 2\|\widehat{f} - f_0\|_n^2 + 2\|f^* - f_0\|_n^2 \leq \frac{1}{2}\|\widehat{f} - f^*\|_n^2 + 2\|f^* - f_0\|_n^2.$$

Therefore, $\|\widehat{f} - f^*\|_n^2 \leq 4\|f^* - f_0\|_n^2$. Hence the inequality (18) holds.

Step 3. Combine the Results: From the conclusion of (14) in Step 1 and (18) in Step 2, we obtain

$$\begin{aligned}\|\widehat{f} - f^*\|_{L^2(\mu)}^2 \leq &\left(2^{30}\sigma^4 + 10 + \frac{2^8 M^2}{3}\right) n^{-2\beta/(2\beta+d)} + 72\|f^* - f_0\|_n^2 \\ &+ 2^{17}\sigma^2\frac{W}{n}\Big(\log\Big(L(W+1)^L B^L n^{\beta/(2\beta+d)} + 1\Big)\Big)^2 \\ &+ \frac{2^8 M^2 W}{3n}\log\Big(2n^{\beta/(2\beta+d)} L(W+1)^L B^L\Big),\end{aligned}$$

with probability at least $1 - 2\exp\left(-n^{d/(2\beta+d)}\right)$.

Under the choice of triples (W, L, B) so that $\|f^* - f_0\|_{L^\infty(\mu)} \leq n^{-\beta/(2\beta+d)}$, the terms inside log are polynomial to n. With the inequality $\|\widehat{f} - f_0\|_{L^2}^2 \leq 2\|\widehat{f} - f^*\|_{L^2(\mu)}^2 + 2\|f^* - f_0\|_{L^2(\mu)}^2$, we conclude that there exists a constant $C_1 = C_1(c_\mu, \beta, D, d, M, \sigma)$ such as

$$\|\widehat{f} - f_0\|_{L^2(\mu)}^2 \leq C_1 n^{-2\beta/(2\beta+d)}(1 + \log n)^2,$$

with probability at least $1 - 2\exp\left(-n^{d/(2\beta+d)}\right)$. □

B.3 Proof for Minimax Rate of Generalization Error

In this proof, we obtain the statement by evaluating a covering number with intrinsic dimensionality and employing the minimax rate result by Yang et al. (1999). We write the packing number of class $\mathcal{F}$ with norm $\|\cdot\|$ as $\mathcal{S}(\varepsilon, \mathcal{F}, \|\cdot\|)$, which is the maximum size of ε-packing of $\mathcal{F}$.

Lemma 23 (Proposition 1 in Yang and Barron (1999)). *Let $\mathcal{F}$ be any class of functions f with $\sup_{f\in\mathcal{F}}|f| < \infty$. For the regression model $Y_i = f_0(X_i) + \xi_i$, assume X and ε are independent, where $X_i \sim \mu$ and $\xi_i \sim N(0, \sigma^2)$. Let ε_n be the solution of $\varepsilon_n^2 = \log \mathcal{S}(\varepsilon_n, \mathcal{F}, \|\cdot\|_{L^2(\mu)})/n$. Then, we have*

$$\inf_{\widehat{f}} \sup_{f_0\in\mathcal{F}} \|\widehat{f} - f_0\|_{L^2(\mu)} = \Theta(\varepsilon_n),$$

where $\widehat{f}$ is any estimator based on n independent and identically distributed observations $(X_1, Y_1), \ldots, (X_n, Y_n)$.

To apply Lemma 23 to $\mathcal{H}(\beta, [0,1]^D, M)$, we need to evaluate the covering entropy number of the smooth function class $\mathcal{H}(\beta, [0,1]^D, M)$. For a tight evaluation of the covering entropy of the $\mathcal{H}(\beta, [0,1]^D, M)$, we introduce the following condition. Roughly speaking, this condition states that some minimal ϵ-cover can be grouped into moderate number of subgroups in which covers are neighbouring.

Definition 24 (Concentration Condition)**.** *A set $E \subset \mathbb{R}^K$ satisfies the Concentration Condition, when there exist constants $C_1 > 0$ and $C_2 > 0$ such that for any $\varepsilon > 0$, some ε-cover $\left\{B_\infty^K(x_i, \varepsilon)\right\}_{i=1}^T$ of E satisfies the following properties: there exists a map $g : \{x_1, \ldots, x_T\} \to [U]$ for some $U \in \mathbb{N}$ such that for all $j \in [U]$, and for all $X \in 2^{g^{-1}(j)} \setminus \left\{\emptyset, g^{-1}(j)\right\}$, some $y = y \in g^{-1}(j) \setminus X$ satisfy $\min_{x \in X} \|x - y\|_\infty \leq \varepsilon$. Also $T \leq C_1 \mathcal{N}(\varepsilon, E, \|\cdot\|_\infty)$ and $U \log(1/\varepsilon) \leq C_2 \mathcal{N}(\varepsilon, E, \|\cdot\|_\infty)$ hold.*

In short, this condition requires an existence of some nearly minimal ε-cover of E that can be grouped into properly concentrated parts. To make clear this condition, we introduce the following lemma.

Lemma 25. *Assume $\mathcal{M}$ is a compact d-dimensional manifold in $[0,1]^D$, namely, assume $\mathcal{M} = \bigcup_{k=1}^K \mathcal{M}_k \subset [0,1]^D$ for some $K \in \mathbb{N}$. Also assume for any $1 \leq k \leq K$, there exists an onto and continuously differentiable map $\psi_k : [0,1]^{d_k} \to \mathcal{M}_k$ each of which has the input dimension $d_k \in \mathbb{N}$. Then, $\mathcal{M}$ satisfies the Concentration Condition.*

Proof of Lemma 25 If $\mathcal{M}_k \subset [0,1]^D$ satisfy the Concentration Condition, it is easily shown that $\bigcup_{k=1}^K \mathcal{M}_k$ satisfy the Concentration Condition. So the problem is reduced to showing that for any $k \in [K]$, $\mathcal{M}_k$ satisfies the Concentration Condition.

Fix any $k \in [K]$. For simplicity, we omit the subscript k from ψ_k, d_k and $\mathcal{M}_k$. Write $\psi = (\psi_1, \ldots, \psi_D)$. Define $L_i \coloneqq \max_{x \in [0,1]^D} \sqrt{\sum_{j=1}^d |\partial \psi_i'(x)/\partial x_j|^2}$. Applying the mean-value theorem to ψ_i along with Cauchy-Schwartz inequality yields $|\psi_i(x) - \psi_i(y)| \leq L_i \|x - y\|_2$ for any $x, y \in [0,1]^d$. By the Lipschitz continuity of $\psi = (\psi_1, \ldots, \psi_D)$, for any $z, w \in [0,1]^d$, $\|\psi(x) - \psi(y)\|_\infty \leq \sqrt{D} L \|x - y\|_\infty$ where $L \coloneqq \max_i L_i$.

Note that $[0,1]^d$ satisfies Concentration Condition, since for any $\delta > 0$, the δ-cover $\left\{B_\infty^D(x_i', \delta)\right\}_{i=1}^T$ constructed by expanding the minimal $\delta/2$-cover $\left\{B_\infty^D(x_i', \delta/2)\right\}_{i=1}^T$ always satisfy the property of the condition with $U = 1$.

Fix any $\varepsilon > 0$. Since $[0,1]^d$ satisfies Concentration condition, we can take an $\varepsilon/(\sqrt{D}L)$-cover $\left\{B_\infty^d(x_i, \varepsilon/(\sqrt{D}L))\right\}_{i=1}^T$ of $[0,1]^d$ so that for any $X \in 2^{\{x_1, \ldots, x_T\}} \setminus \{\emptyset, \{x_1, \ldots, x_T\}\}$, there exists some $y \in \{x_1, \ldots, x_T\} \setminus X$ such that $\min_{x \in X} \|x - y\|_\infty \leq \varepsilon/(\sqrt{D}L)$. Let $\mathcal{C} \coloneqq \left\{B_\infty^D(\psi(x_1), \varepsilon), \ldots, B_\infty^D(\psi(x_T), \varepsilon)\right\}$. We first verify that $\mathcal{C}$ is a ε-cover of $\mathcal{M}$. Since ψ is onto, for any $z \in \mathcal{M}$, there exists some $x \in [0,1]^d$ such that $z = \psi(x)$. For this x, some $y \in \{x_1, \ldots, x_T\}$ satisfies $\|x - y\|_\infty \leq \varepsilon/(\sqrt{D}L)$. Thus, we obtain

$$\|z - \psi(y)\|_\infty = \|\psi(x) - \psi(y)\|_\infty \leq \sqrt{D} L \|x - y\|_\infty \leq \sqrt{D} L \frac{\varepsilon}{\sqrt{D}L} = \epsilon.$$

This verifies that $\mathcal{C}$ is an ε-cover of $\mathcal{M}$.

Take any $X' \in 2^{\{\psi(x_1), \ldots, \psi(x_T)\}} \setminus \{\emptyset, \{\psi(x_1), \ldots, \psi(x_T)\}\}$. Write $X' = \{\psi(x_{j_1}), \ldots, \psi(x_{j_t})\}$. By assumption, there exists some $y \in \{x_1, \ldots, x_T\} \setminus \psi^{-1}(X')$ such that $\min_{x \in \psi^{-1}(X')} \|x - y\|_\infty \leq$

$\varepsilon/(\sqrt{D}L)$ holds. Hence, for this y, the following holds:

$$\min_{\psi(x)\in X'} \|\psi(x)-\psi(y)\|_\infty \le \sqrt{D}L \min_{x\in\psi^{-1}(X')} \|x-y\|_\infty \le \sqrt{D}L\frac{\varepsilon}{\sqrt{D}L} = \varepsilon.$$

This concludes the proof. □

Theorem 8 is a direct consequence of the following lemma.

Lemma 26 (Minimax optimal rate under Concentration Condition)**.** *Let μ be a probability measure on $[0,1]^D$. Assume $\mathcal{N}(\varepsilon, \mathrm{Supp}(\mu), \|\cdot\|_\infty) = \Theta(\varepsilon^{-d})$ for some $d > 0$. Also assume that $\mathrm{Supp}(\mu)$ satisfy Concentration Condition. Then, the following holds:*

$$\inf_{\widehat{f}} \sup_{f_0\in\mathcal{F}} \|\widehat{f} - f_0\|_{L^2(\mu)} = \Theta(\varepsilon^{-\beta/(2\beta+d)}).$$

Proof of Lemma 26 This proof contains the following two steps: (i) derive a lower bound of $\mathcal{N}(\varepsilon, \mathcal{H}(\beta,[0,1]^D,M), \|\cdot\|_{L^\infty(\mu)})$, (ii) derive an upper bound of $\mathcal{N}(\varepsilon, \mathcal{H}(\beta,[0,1]^D,M), \|\cdot\|_{L^\infty(\mu)})$, then (iii) apply Lemma 23.

Step (i): The lower bound. For the lower bound of $\mathcal{N}(\varepsilon, \mathcal{H}(\beta,[0,1]^D,M), \|\cdot\|_{L^\infty(\mu)})$, we basically follow (Wainwright, 2019). We first construct a packing $\left\{f_\gamma \mid \gamma\in\{-1,1\}^S\right\}$ for some $S\in\mathbb{N}$. Define

$$\phi(y) \coloneqq \begin{cases} c2^{2\beta D}\prod_{j=1}^D (1/2-y_j)^\beta(1/2+y_j)^\beta & \text{if } y\in[-1/2,1/2]^D,\\ 0 & \text{if } y\notin[-1/2,1/2]^D,\end{cases}$$

where $c = c(\beta, D, M)$ is chosen small enough so that $\phi\in\mathcal{H}(\beta,[0,1]^D,M)$ holds.

For any $\varepsilon>0$, set $\delta = (\varepsilon/2c)^{1/\beta}$. Consider $\delta/2$-packing of $\mathrm{Supp}(\mu)$ as $\{x_i\}_{i=1}^S \subset \mathrm{Supp}(\mu)$. Recall that $\mathcal{N}(\delta, \mathrm{Supp}(\mu), \|\cdot\|_\infty) \le S \le \mathcal{N}(\delta/2, \mathrm{Supp}(\mu), \|\cdot\|_\infty)$.

For each $\gamma\in\{-1,1\}^S$, define the following

$$f_\gamma(x) = \sum_{i=1}^S \gamma_i\delta^\beta\phi\left(\frac{x-x_i}{\delta}\right).$$

If $\gamma\neq\gamma'$, then for $x\in(x_{i1}-\delta/2, x_{i1}+\delta/2)\times\cdots\times(x_{iD}-\delta/2, x_{iD}+\delta/2)$, the following holds:

$$\left|f_\gamma(x) - f_{\gamma'}(x)\right| = 2\delta^\beta\phi\left(\frac{x-x_i}{\delta}\right).$$

Setting $x\leftarrow x_i$ yields

$$\left|f_\gamma(x) - f_{\gamma'}(x)\right| = 2\delta^\beta c = \varepsilon.$$

Since $\left\{f_\gamma \mid \gamma\in\{-1,1\}^S\right\}$ is an ε-packing of $\mathcal{H}(\beta,[0,1]^D,M)$, we obtain the lower bound of the covering number of $\mathcal{H}(\beta,[0,1]^D,M)$ as

$$\begin{aligned}\log\mathcal{N}(\varepsilon, \mathcal{H}(\beta,[0,1]^D,M), \|\cdot\|_{L^\infty(\mu)}) &\ge \log 2^S\\ &\ge \mathcal{N}(2(\varepsilon/2c)^{1/\beta}, [0,1]^D, \|\cdot\|_\infty)\log 2\\ &\gtrsim \varepsilon^{-d/\beta},\end{aligned}$$

where the last inequality follows from the assumption $\mathcal{N}(\varepsilon, \mathrm{Supp}(\mu), \|\cdot\|_\infty) = \Theta(\varepsilon^{-d})$.

Step (ii): The upper bound. For the upper bound of $\mathcal{N}(\varepsilon, \mathcal{H}(\beta, [0,1]^D, M), \|\cdot\|_{L^\infty(\mu)})$. We modify the Theorem 2.7.1 in van der Vaart and Wellner (1996).

For a preparation, we define several notions to form a covering set with its cardinality $\mathcal{N}(\varepsilon, \mathcal{H}(\beta, [0,1]^D, M), \|\cdot\|_{L^\infty(\mu)})$. Take the minimal δ-cover $(x_i)_{i=1}^T \subset \mathrm{Supp}(\mu)$, where $T = \mathcal{N}(\delta, \mathrm{Supp}(\mu), \|\cdot\|_\infty)$. Note that $T = \Theta(\delta^{-d})$ holds by the setting. For a multi-index $k = (k_1, \ldots, k_D)$ with $k \leq \beta$, define the operators A_k, B_k as

$$A_k f \coloneqq (\lfloor D^k f(x_1)/\delta^{\beta-|k|} \rfloor, \ldots, \lfloor D^k f(x_T)/\delta^{\beta-|k|} \rfloor), \text{ and } B_k f \coloneqq \delta^{\beta-|k|} A_k f.$$

If $A_k f = A_k g$ for all k with $|k| \leq \beta$, then $\|f - g\|_{L^\infty(\mu)} \lesssim \varepsilon$. For each f, we define a matrix Af to provide a covering set as follows:

$$Af \coloneqq \begin{pmatrix} A_{0,0,\ldots,0} f & A_{1,0,\ldots,0} f & A_{0,1,\ldots,0} f & \cdots & A_{0,0,\ldots,\beta} f \end{pmatrix}^\top \in \mathbb{R}^{r \times T},$$

where r is a combination of the multi-index, we can bound the number of row r of Af as $r \leq \binom{D}{0} + \binom{D}{1} + \cdots + \binom{D}{\beta} \leq (\beta+1)^D$. Since $\left|D^k f(x)\right| \leq M$ for all $x \in [0,1]^D$, each element in $A_k f$ takes at most $2M/\delta^{\beta-|k|} + 1 \leq 2M\delta^{-\beta} + 1$ values.

Moreover, we define a smooth approximation of $D^k f(x_i)$ and evaluate its approximation error. Suppose $\|x_i - x_{i'}\|_\infty \leq \delta$ for some i, i'. Since $D^k f(x_i) = \sum_{|k|+|l| \leq \lfloor \beta \rfloor} D^{k+l} f(x_{i'}) \frac{(x_i - x_{i'})^l}{l!} + R$ with $|R| \lesssim \|x_i - x_{i'}\|_\infty^{\beta-|k|}$, we bound the value as

$$\begin{aligned}
&\left| D^k f(x_i) - \sum_{|k|+|l| \leq \lfloor \beta \rfloor} B_{k+l} f(x_{i'}) \frac{(x_i - x_{i'})^l}{l!} \right| \\
&\lesssim \sum_{|k|+|l| \leq \lfloor \beta \rfloor} \left| D^{k+l} f(x_i) - B_{k+l} f(x_{i'}) \right| \frac{(x_i - x_{i'})^l}{l!} + \delta^{\beta-|k|} \\
&\leq \sum_{|k|+|l| \leq \lfloor \beta \rfloor} \delta^{\beta-|k|-|l|} \frac{\delta^{|l|}}{l!} + \delta^{\beta-|k|} \\
&\lesssim \delta^{\beta-|k|}.
\end{aligned}$$

Given the i'-th column, i-th column ranges over $\Theta(\delta^{\beta-|k|}/\delta^{\beta-|k|}) = \Theta(1)$.

Now, we bound the covering number by using the notions stated above. By assumption, $\mathrm{Supp}(\mu)$ satisfy the Concentration Condition. Thus there exist disjoint sets $X_1, \ldots, X_U$ such that $X \coloneqq \{x_1, \ldots, x_T\} = \bigcup_{u=1}^U X_u$ and that $X_u = \{x_1, \ldots, x_{T_u}\}$ with $\|x_{i+1} - x_i\|_\infty \leq \delta$ for any $i = 1, \ldots, T_u - 1$. Thus

$$\mathrm{card}\left\{Af \mid f \in \mathcal{H}(\beta, [0,1]^D, M)\right\} \leq (2M\delta^{-\beta} + 1)^{U(\beta+1)^D} C^{T-U},$$

holds for some constant $C > 0$. Substitute $\delta \leftarrow \varepsilon^{1/\beta}$, we obtain

$$\log\left(\mathrm{card}\left\{Af \mid f \in \mathcal{H}(\beta, [0,1]^D, M)\right\}\right) \lesssim \max\left\{U \log\left(\frac{1}{\varepsilon}\right), T - U\right\}.$$

Since $U\log(1/\varepsilon) = \mathcal{O}(T)$, we obtain

$$\log \mathcal{N}\Big(\varepsilon, \mathcal{H}(\beta, [0,1]^D, M), \|\cdot\|_{L^\infty(\mu)}\Big) \lesssim \varepsilon^{-d/\beta}.$$

Step (iii): Combine the results. It is ready to obtain the statement. Since it is shown that $\log \mathcal{N}\Big(\varepsilon, \mathcal{H}(\beta, [0,1]^D, M), \|\cdot\|_{L^\infty(\mu)}\Big) = \Theta(\varepsilon^{-d/\beta})$, applying Lemma 23 together with $\mathcal{S}(2\varepsilon, \mathcal{H}(\beta, [0,1]^D, M), M), \|\cdot\|_{L^\infty(\mu)}) \leq \mathcal{N}(\varepsilon, \mathcal{H}(\beta, [0,1]^D, M), M), \|\cdot\|_{L^\infty(\mu)}) \leq \mathcal{S}(\varepsilon, \mathcal{H}(\beta, [0,1]^D, M), M), \|\cdot\|_{L^\infty(\mu)})$ yields the statement. □

Proof of Theorem 8 By a direct application of Lemma 26, we obtain the statement. □

References

Martin Anthony and Peter L Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 2009.

Martin Arjovsky and Léon Bottou. Towards principled methods for training generative adversarial networks. In *International Conference on Learning Representations*, 2017.

Sanjeev Arora, Rong Ge, Behnam Neyshabur, and Yi Zhang. Stronger generalization bounds for deep nets via a compression approach. In *International Conference on Machine Learning*, pages 254–263, 2018.

Francis Bach. Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research*, 18(1):629–681, 2017.

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

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

Andrew R Barron and Jason M Klusowski. Approximation and estimation for high-dimensional deep learning networks. *arXiv preprint arXiv:1809.03090*, 2018.

Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*, pages 6240–6249, 2017.

Benedikt Bauer and Michael Kohler. On deep learning as a remedy for the curse of dimensionality in nonparametric regression. *Annals of Statistics*, 47(4):2261–2285, 2019.

Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003.

Peter J Bickel and Bo Li. Local polynomial regression on unknown manifolds. In *Complex Datasets and Inverse Problems*, pages 177–186. 2007.

Jörg Bruske and Gerald Sommer. Intrinsic dimensionality estimation with optimally topology preserving maps. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(5): 572–575, 1998.

Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *International Conference on Machine Learning*, pages 160–167, 2008.

Gerald B Folland. *Real Analysis: Modern Techniques and Their Applications*. John Wiley & Sons, 2013.

Keinosuke Fukunaga and David R Olsen. An algorithm for finding intrinsic dimensionality of data. *IEEE Transactions on Computers*, 100(2):176–183, 1971.

Christopher Genovese, Marco Perone-Pacifico, Isabella Verdinelli, and Larry Wasserman. Minimax manifold estimation. *Journal of Machine Learning Research*, 13:1263–1291, 2012.

Evarist Giné and Richard Nickl. *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge University Press, 2016.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep Learning*. MIT Press Cambridge, 2016.

Gloria Haro, Gregory Randall, and Guillermo Sapiro. Translated poisson mixture model for stratification learning. *International Journal of Computer Vision*, 80(3):358–374, 2008.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

Masaaki Imaizumi and Kenji Fukumizu. Deep neural networks learn non-smooth functions effectively. In *Artificial Intelligence and Statistics*, pages 869–878, 2019.

Kerstin Johnsson, Charlotte Soneson, and Magnus Fontes. Low bias local intrinsic dimension estimation from expected simplex skewness. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(1):196–202, 2015.

Antti Käenmäki, Juha Lehrbäck, and Matti Vuorinen. Dimensions, whitney covers, and tubular neighborhoods. *Indiana University Mathematics Journal*, 62(6):1861–1889, 2013.

Diederick P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Michael Kohler, Adam Krzyzak, and Sophie Langer. Estimation of a function of low local dimensionality by deep neural networks. *arXiv preprint arXiv:1908.11140*, 2019.

Samory Kpotufe. k-nn regression adapts to local intrinsic dimension. In *Advances in Neural Information Processing Systems*, pages 729–737, 2011.

Samory Kpotufe and Vikas Garg. Adaptivity to local smoothness and dimension in kernel regression. In *Advances in Neural Information Processing Systems*, pages 3075–3083, 2013.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. 2009.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436, 2015.

Jonathan Masci, Davide Boscaini, Michael Bronstein, and Pierre Vandergheynst. Geodesic convolutional neural networks on riemannian manifolds. In *IEEE International Conference on Computer Vision Workshops*, pages 37–45, 2015.

Hadrien Montanelli and Qiang Du. Deep relu networks lessen the curse of dimensionality. *arXiv preprint arXiv:1712.08688*, 2017.

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

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

Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 61: 85–117, 2015.

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

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

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

Taiji Suzuki. Fast learning rate of deep learning via a kernel perspective. In *Artificial Intelligence and Statistics*, 2017.

Taiji Suzuki. Adaptivity of deep relu network for learning in besov and mixed smooth besov spaces: optimal rate and curse of dimensionality. In *International Conference on Learning Representations*, 2019.

Matus Telgarsky. Benefits of depth in neural networks. In *Conference On Learning Theory*, volume 49, 2016.

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

Aad W van der Vaart and Jon A Wellner. *Weak Convergence and Empirical Processes*. Springer, 1996.

Martin J Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.

Yuhong Yang, Andrew Barron, et al. Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, 27(5):1564–1599, 1999.

Yun Yang and David B Dunson. Bayesian manifold regression. *Annals of Statistics*, 44(2): 876–905, 2016.

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