Deep learning is adaptive to intrinsic dimensionality of model smoothness in anisotropic Besov space

Taiji Suzuki
Department of Mathematical Informatics, The University of Tokyo, Tokyo, Japan
RIKEN Center for Advanced Intelligence Project, Tokyo, Japan
taiji@mist.i.u-tokyo.ac.jp

Atsushi Nitanda
Kyushu Institute of Technology, Fukuoka, Japan
RIKEN Center for Advanced Intelligence Project, Tokyo, Japan
nitanda@ai.kyutech.ac.jp

Abstract

Deep learning has exhibited superior performance for various tasks, especially for high-dimensional datasets, such as images. To understand this property, we investigate the approximation and estimation ability of deep learning on *anisotropic Besov spaces*. The anisotropic Besov space is characterized by direction-dependent smoothness and includes several function classes that have been investigated thus far. We demonstrate that the approximation error and estimation error of deep learning only depend on the average value of the smoothness parameters in all directions. Consequently, the curse of dimensionality can be avoided if the smoothness of the target function is highly anisotropic. Unlike existing studies, our analysis does not require a low-dimensional structure of the input data. We also investigate the minimax optimality of deep learning and compare its performance with that of the kernel method (more generally, linear estimators). The results show that deep learning has better dependence on the input dimensionality if the target function possesses anisotropic smoothness, and it achieves an adaptive rate for functions with spatially inhomogeneous smoothness.

1 Introduction

Based on the recent literature pertaining to machine learning, deep learning has exhibited superior performance in several tasks such as image recognition (Krizhevsky et al., 2012), natural language processing (Devlin et al., 2018), and image synthesis (Radford et al., 2015). In particular, its superiority is remarkable for complicated and high-dimensional data like images. This is mainly due to its high flexibility and superior feature-extraction ability for effectively extracting the intrinsic structure of data. Its theoretical analysis also has been extensively developed considering several aspects such as expressive ability, optimization, and generalization error.

Amongst representation ability analysis of deep neural networks such as universal approximation ability (Cybenko, 1989; Hornik, 1991; Sonoda & Murata, 2017), approximation theory of deep neural networks on typical function classes such as Hölder, Sobolev, and Besov spaces have been extensively studied. In particular, analyses of deep neural networks with the ReLU activation (Nair & Hinton, 2010; Glorot et al., 2011) have been recently developed. Schmidt-Hieber (2020) showed that the deep learning with ReLU activations can achieve the minimax optimal estimation accuracy to estimate composite functions in Hölder spaces by using the approximation theory of Yarotsky (2017). Suzuki (2019) generalized this analysis to those on the *Besov space* and the *mixed smooth*

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

Table 1: Relationship between existing research and our work. indicates the smoothness of the target function, d is the dimensionality of input x, D is the dimensionality of a low-dimensional structure on which the data are distributed, and $\tilde{}$ is the average smoothness of an anisotropic Besov space (Eq. (1)).

Function class	Hölder	Besov	mixed smooth Besov	Hölder on a low-dimensional set	anisotropic Besov
Author	Schmidt-Hieber (2020)	Suzuki (2019)	Suzuki (2019)	Nakada & Imaizumi (2020); Schmidt-Hieber (2019); Chen et al. (2019)	This work
Estimation error	$O(n^{-\frac{2}{2+d}})$	$O(n^{-\frac{2}{2+d}})$	$O\left(n^{-\frac{2}{2+1}} \log(n)^{\frac{2(d-1)(u+)}{1+2}}\right)$	$O(n^{-\frac{2}{2+D}})$	$O(n^{-\frac{2}{2+1}})$

Besov space by utilizing the techniques developed in approximation theories (Temlyakov, 1993; DeVore, 1998). It was shown that deep learning can achieve an *adaptive approximation* error rate that is faster than that of (non-adaptive) linear approximation methods (DeVore & Popov, 1988; DeVore et al., 1993; Dũng, 2011b), and it outperforms any linear estimators (including kernel ridge regression) in terms of the minimax optimal rate.

From these analyses, one can see that the approximation errors and estimation errors are strongly influenced by two factors, i.e., the *smoothness* of the target function and the *dimensionality* of the input (see Table 1). In particular, they suffer from the *curse of dimensionality*, which is unavoidable. However, these analyses are about the worst case errors and do not exploit specific intrinsic properties of the true distributions. For example, practically encountered data usually possess low intrinsic dimensionality, i.e., data are distributed on a low dimensional sub-manifold of the input space (Tenenbaum et al., 2000; Belkin & Niyogi, 2003). Recently, Nakada & Imaizumi (2020); Schmidt-Hieber (2019); Chen et al. (2019); Chen et al. (2019) have shown that deep ReLU network has adaptivity to the intrinsic dimensionality of data and can avoid curse of dimensionality if the intrinsic dimensionality is small. However, one drawback is that they assumed *exact* low dimensionality of the input data. This could be a strong assumption because practically observed data are always noisy, and injecting noise immediately destroys the low-dimensional structure. Therefore, we consider another direction in this paper. In terms of curse of dimensionality, Suzuki (2019) showed that deep learning can alleviate the curse of dimensionality to estimate functions in a so called mixed smooth Besov space (m-Besov). However, m-Besov space assumes strong smoothness toward *all* directions uniformly and does not include the ordinary Besov space as a special case. Moreover, the convergence rate includes heavy poly-log term which is not negligible (see Table 1).

In practice, one of the typically expected properties of a true function on high-dimensional data is that it is invariant against perturbations of an input in some specific directions (Figure 1). For example, in image-recognition tasks, the target function must be invariant against the spatial shift of an input image, which is utilized by data-augmentation techniques (Simard et al., 2003; Krizhevsky et al., 2012). In this paper, we investigate the approximation and estimation abilities of deep learning on *anisotropic Besov spaces* (Nikol'skii, 1975; Vybiral, 2006; Triebel, 2011) (also called dominated mixed-smooth Besov spaces). An anisotropic Besov space is a set of functions that have "direction-dependent" smoothness, whereas ordinary function spaces such as Hölder, Sobolev, and Besov spaces assume isotropic smoothness that is uniform in all directions. We consider a composition of functions included in an anisotropic Besov space, including several existing settings as special cases; it includes analyses of the Hölder space Schmidt-Hieber (2020) and Besov space Suzuki (2019), as well as the low-dimensional sub-manifold setting (Nakada & Imaizumi, 2020; Schmidt-Hieber, 2019; Chen et al., 2019; Chen et al., 2019)[1]. By considering such a space, we can show that deep learning can alleviate curse of dimensionality if the smoothness in each direction is highly anisotropic. Interestingly, any linear estimator (including kernel ridge regression) has worse dependence on the dimensionality than deep learning. Our contributions can be summarized as follows:

- We consider a situation in which the target function is included in a class of anisotropic Besov spaces and show that deep learning can avoid the curse of dimensionality *even if the input data*

[1]We would like to remark that the analysis of Nakada & Imaizumi (2020) does not require smoothness of the embedded manifold that is not covered in this paper.

do not lie on a low-dimensional manifold. Moreover, deep learning can achieve the optimal adaptive approximation error rate and minimax optimal estimation error rate.

- We compare deep learning with general linear estimators (including kernel methods) and show that deep learning has better dependence on the input dimensionality than linear estimators.

2 Problem setting and the model

In this section, we describe the problem setting considered in this work. We consider the following nonparametric regression model:

$$y_i = f^o(x_i) + \varepsilon_i \quad (i = 1, \dots, n),$$

where x_i is generated from a probability distribution P_X on $[0,1]^d$, $\varepsilon_i \sim N(0, \sigma^2)$, and the data $D_n = (x_i, y_i)_{i=1}^n$ are independently identically distributed. f^o is the true function that we want to estimate. We are interested in the mean squared estimation error of an estimator $\widehat{f}$: $\mathbb{E}_{D_n}[\|\widehat{f} - f^o\|_{L^2(P_X)}^2]$, where $\mathbb{E}_{D_n}[\cdot]$ indicates the expectation with respect to the training data D_n. We consider a least-squares estimator in the deep neural network model as $\widehat{f}$ (see Eq. (5)) and discuss its optimality. More specifically, we investigate how the "intrinsic dimensionality" of data affects the estimation accuracy of deep learning. For this purpose, we consider an *anisotropic Besov space* as a model of the target function.

2.1 Anisotropic Besov space

In this section, we introduce the anisotropic Besov which was investigated as the model of the true function in this paper. Throughout this paper, we set the domain of the input to $\Omega = [0,1]^d$. For a function $f : \Omega \to \mathbb{R}$, let $\|f\|_p := \|f\|_{L^p(\Omega)} := (\int_\Omega |f|^p dx)^{1/p}$ for $0 < p < \infty$. For $p = \infty$, we define $\|f\|_\infty := \|f\|_{L^\infty(\Omega)} := \sup_{x \in \Omega} |f(x)|$. For $\alpha \in \mathbb{R}^d_{++}$, let $|\alpha| = \sum_{j=1}^d |\alpha_j|$[2].

For a function $f : \mathbb{R}^d \to \mathbb{R}$, we define the rth difference of f in the direction $h \in \mathbb{R}^d$ as

$$\Delta_h^r(f)(x) := \Delta_h^1(\Delta_h^{r-1}(f))(x), \qquad \Delta_h^0(f)(x) := f(x),$$

for $x \in \Omega$ with $x + rh \in \Omega$, otherwise, let $\Delta_h^r(f)(x) = 0$.

Definition 1. *For a function $f \in L^p(\Omega)$ where $p \in (0, \infty]$, the r-th modulus of smoothness of f is defined by $w_{r,p}(f, t) = \sup_{h \in \mathbb{R}^d : |h_i| \le t_i} \|\Delta_h^r(f)\|_p$, for $t = (t_1, \dots, t_d)$, $t_i > 0$.*

With this modulus of smoothness, we define the anisotropic Besov space $B_{p,q}(\Omega)$ for $\alpha = (\alpha_1, \dots, \alpha_d)^\top \in \mathbb{R}^d_{++}$ as follows.

Definition 2 (Anisotropic Besov space ($B_{p,q}(\Omega)$))**.** *For $0 < p, q \le \infty$, $\alpha = (\alpha_1, \dots, \alpha_d)^\top \in \mathbb{R}^d_{++}$, $r := \max_i \lfloor \alpha_i \rfloor + 1$, let the seminorm $|\cdot|_{B_{p,q}}$ be*

$$|f|_{B_{p,q}} := \begin{cases} \left(\sum_{k=0}^{\infty} [2^{k\alpha} w_{r,p}(f, (2^{-k\alpha_1}, \dots, 2^{-k\alpha_d}))]^q \right)^{1/q} & (q < \infty), \\ \sup_{k \ge 0} 2^{k\alpha} w_{r,p}(f, (2^{-k\alpha_1}, \dots, 2^{-k\alpha_d})) & (q = \infty). \end{cases}$$

The norm of the anisotropic Besov space $B_{p,q}(\Omega)$ is defined by $\|f\|_{B_{p,q}} := \|f\|_p + |f|_{B_{p,q}}$, and $B_{p,q}(\Omega) = \{f \in L^p(\Omega) \mid \|f\|_{B_{p,q}} < \infty\}$.

Roughly speaking α represents the smoothness in each direction. If α_i is large, then a function in $B_{p,q}$ is smooth to the ith coordinate direction, otherwise, it is non-smooth to that direction. p is also an important quantity that controls the *spatial inhomogeneity* of the smoothness. If $\alpha_1 = \alpha_2 = \cdots = \alpha_d$, then the definition is equivalent to the usual Besov space (DeVore & Popov, 1988; DeVore et al., 1993). Suzuki (2019) analyzed curse of dimensionality of deep learning through a so-called *mixed smooth Besov* (m-Besov) space which imposes a stronger condition toward all directions uniformly.

[2] We let $\mathbb{N} := \{1, 2, 3, \dots\}$, $\mathbb{Z}_+ := \{0, 1, 2, 3, \dots\}$, $\mathbb{Z}_+^d := \{(z_1, \dots, z_d) \mid z_i \in \mathbb{Z}_+\}$, $\mathbb{R}_+ := \{x \ge 0 \mid x \in \mathbb{R}\}$, and $\mathbb{R}_{++} := \{x > 0 \mid x \in \mathbb{R}\}$. We let $[N] := \{1, \dots, N\}$ for $N \in \mathbb{N}$.

Particularly, it imposes stronger smoothness toward non-coordinate axis directions. Moreover, m-Besov space does *not* include the vanilla Besov space as a special case and thus cannot capture the situation that we consider in this paper.

Throughout this paper, for given $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d)^\top \in \mathbb{R}^d_{++}$, we write $\underline{\alpha} := \min_j \alpha_j$ (smallest smoothness) and $\overline{\alpha} := \max_j \alpha_j$ (largest smoothness). The approximation error of a function in anisotropic Besov spaces is characterized by the harmonic mean of $(\alpha_j)_{j=1}^d$, which corresponds to the average smoothness, and thus we define

$$\tilde{\alpha} := \left(\sum_{j=1}^d 1/\alpha_j \right)^{-1}. \tag{1}$$

The Besov space is closely related to other function spaces such as Hölder space. Let $\partial^{\boldsymbol{\alpha}} f(x) = \frac{\partial^{\alpha_1 + \cdots + \alpha_d} f}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}}(x)$.

Definition 3 (Hölder space ($C(\mathcal{X})$)). *For a smoothness paraemter $\alpha \in \mathbb{R}_{++}$ with $\alpha \notin \mathbb{N}$, consider an m times differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ where $m = \lfloor \alpha \rfloor$ (the largest integer less than α), and let the norm of the Hölder space $C(\mathcal{X})$ be $\|f\|_{C} := \max_{|\boldsymbol{\alpha}| \le m} \|\partial^{\boldsymbol{\alpha}} f\|_\infty + \max_{|\boldsymbol{\alpha}| = m} \sup_{x,y \in \mathcal{X}} \frac{|\partial^{\boldsymbol{\alpha}} f(x) - \partial^{\boldsymbol{\alpha}} f(y)|}{\|x - y\|_\infty^{\alpha - m}}$. Then, ($\alpha$-)Hölder space $C(\mathcal{X})$ is defined as $C(\mathcal{X}) = \{f \mid \|f\|_C < \infty\}$.*

Let $C^0(\mathcal{X})$ be the set of continuous functions equipped with L^∞-norm: $C^0(\mathcal{X}) := \{f : \mathcal{X} \to \mathbb{R} \mid f \text{ is continuous and } \|f\|_\infty < \infty\}$. These function spaces are closely related to each other.

Proposition 1 (Triebel (2011)). *There exist the following relations between the spaces:*

1. *For $\boldsymbol{\alpha} = (\alpha_0, \ldots, \alpha_0)^\top \in \mathbb{R}^d$ with $\alpha_0 \notin \mathbb{N}$, it holds that $C^0(\mathcal{X}) = B_{\infty,\infty}(\mathcal{X})$.*
2. *For $0 < p_1, p_2, q \le \infty$, $\alpha_1 \ne \alpha_2$ and $\boldsymbol{\alpha} \in \mathbb{R}^d_{++}$ with $\tilde{\alpha} > (1/p_1 - 1/p_2)_+$[3], it holds that*[4] *$B_{p_1,q}(\mathcal{X}) \hookrightarrow B_{p_2,q}(\mathcal{X})$ for $\alpha = \alpha_1 - (1/p_1 - 1/p_2)_+ = \tilde{\alpha}$.*
3. *For $0 < p, q_1, q_2 \le \infty$, $\alpha_1 < \alpha_2$, and $\boldsymbol{\alpha} \in \mathbb{R}^d_{++}$, it holds that $B_{p,q_1} \hookrightarrow B_{p,q_2}$. In particular, with properties 1 and 2, if $\tilde{\alpha} > 1/p$, it holds that $B_{p,q}(\mathcal{X}) \hookrightarrow C_b(\mathcal{X})$ where $\alpha = \alpha_1 - (1/p - \tilde{\alpha})$.*
4. *For $0 < p, q \le \infty$, α_1 and $\boldsymbol{\alpha} \in \mathbb{R}^d_{++}$, if $\tilde{\alpha} > 1/p$, then $B_{p,q}(\mathcal{X}) \hookrightarrow C^0(\mathcal{X})$.*

This result is basically proven by Triebel (2011). For completeness, we provide its derivation in Appendix D. If the average smoothness $\tilde{\alpha}$ is sufficiently large ($\tilde{\alpha} > 1/p$), then the functions in $B_{p,q}$ are continuous; however, if it is small ($\tilde{\alpha} < 1/p$), then they are no longer continuous. Small p indicates spatially inhomogeneous smoothness; thus, spikes and jumps appear (see Donoho & Johnstone (1998) for this perspective, from the viewpoint of wavelet analysis).

2.2 Model of the true function

As a model of the true function f°, we consider two types of models: *Affien composition model* and *deep composition model*. For a Banach space H, we let $U(H)$ be the unit ball of H.

(a) Affine composition model: The first model we introduced is a very naive model which is just a composition of an affine transformation and a function in the anisotropic Besov space:

$$H_{\mathrm{a}} := \{h(Ax + b) \mid h \in U(B_{p,q}([0,1]^{d'})),\ A \in \mathbb{R}^{d' \times d},\ b \in \mathbb{R}^{b} \text{ s.t. } Ax + b \in [0,1]^{d'} (\forall x \in \mathcal{X})\};$$

where we assume $d' \le d$. Here, we assumed that the affine transformation has an appropriate scaling such that $Ax + b$ is included in the domain of h for all $x \in \mathcal{X}$. This is a quite naive model but provides an instructive example to understand how the estimation error of deep learning behaves under the anisotropic setting.

(b) Deep composition model: The *deep composition model* generalizes the affine composition model to a composition of nonlinear functions. Let $m_1 = d, m_{L+1} = 1, m_\ell$ be the dimension of the

[3] Here, we let $(x)_+ := \max\{x, 0\}$.

[4] The symbol $\hookrightarrow$ means continuous embedding.

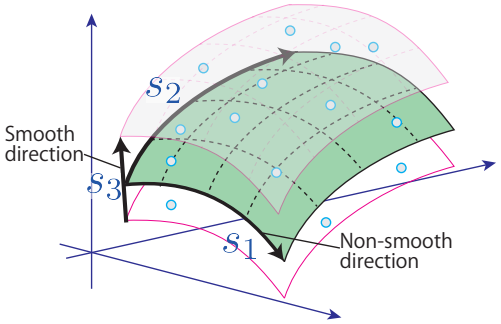

Figure 1: Near low dimensional data distribution with anisotropic smoothness of the target function. The target function has less smoothness ($s_1; s_2$) toward the first two coordinates on the manifold while it is almost constant toward the third coordinate (large s_3).

`th layer, and let $^{(`)}$ 2 $\mathrm{R}^{m_`}_{++}$ be the smoothness parameter in the `th layer. The deep composition model is defined as

$$H_{\mathrm{deep}} := fh_H \qquad h_1(x) \; j \; h_` : [0;1]^{m_`} \; ! \; [0;1]^{m_{`+1}}; \; h_{`;k} \; 2 \; U(B^{(`)}_{p;q}([0;1]^{m_`})) \; (8k \; 2 \; [m_{`+1}])g:$$

Here, the interval $[0;1]$ can be replaced by another compact interval, such as $[a_`; b_`]$, but this difference can be absorbed by changing a scaling factor. The assumption $kh_{`;k}k_{B^{(`)}_{p;q}}$ 1 can also be relaxed, but we do not pursue that direction due to presentation simplicity. This model includes the affine composition model as a special case. However, it requires a stronger assumption to properly evaluate the estimation error on this model.

Examples The model we have introduced includes some instructive examples as listed below:

(a) Linear projection Schmidt-Hieber (2020) analyzed estimation of the following model by deep learning: $f^o(x) = g(w^\top x)$ where $g \; 2 \; C \; ([0;1])$ and $w \; 2 \; \mathrm{R}^d$. In this example, the function f^o varies along only one direction, w. Apparently, this is an example of the affine composition model.

(b) Distribution on low dimensional smooth manifold Assume that the input x is distributed on a low-dimensional smooth manifold embedded in , and the smoothness of the true function f^o is anisotropic along a coordinate direction on the manifold. We suppose that the low dimensional manifold is d^*-dimensional and d^* d. In this situation, the true function can be written as $f^o(x) = h(\; (x))$ where $: \mathrm{R}^d \; ! \; \mathrm{R}^{d^*}$ is a map that returns the coordinate of x on the manifold and h is an element in an anisotropic Besov space on R^{d^*}. This situation appears if data is distributed on a low-dimensional sub-manifold of and the target function is invariant against noise injection to some direction on the manifold at each input point x (Figure 1 illustrates this situation). One typical example of this situation is a function invariant with data augmentation (Simard et al., 2003; Krizhevsky et al., 2012). Even if the noise injection destroys low dimensionality of the data distribution (i.e., $d^* = d$), an anisotropic smoothness of the target function eases the curse of dimensionality as analyzed below, which is quite different from existing works (Yang & Dunson, 2016; Bickel & Li, 2007; Nakada & Imaizumi, 2020; Schmidt-Hieber, 2019; Chen et al., 2019; Chen et al., 2019).

Related work Here, we introduce some more related work and discuss their relation to our analysis. The statistical analysis on an anisotropic Besov space can be back to Ibragimov & Khas'minskii (1984) who considered density estimation, where the density is assumed to be included in an anisotropic Sobolev space with p 2, and derived the minimax optimal rate $n^{-\bar{r}=(2\bar{r}+1)}$ with respect to L^r-norm. Nyssbaum (1983, 1987) analyzed a nonparametric regression problem on an anisotropic Besov space. Following these results, several studied have been conducted in the literature pertaining to nonparametric statistics, such as nonlinear kernel estimator Kerkyacharian et al. (2001), adaptive confidence band construction Hoffman & Lepski (2002), optimal aggregation Gaiffas & Lecue (2011), Gaussian process estimator Bhattacharya et al. (2011, 2014), and kernel ridge regression Hang & Steinwart (2018). Basically, these studies investigated estimation problems in which the target function is in anisotropic Besov spaces, but the composition models considered in this paper have not been analyzed. Hoffman & Lepski (2002); Bhattacharya et al. (2011) considered a dimension reduction model; that is, the target function is dependent on only a few variables of x, but they did not deal with more general models, such as the affine/deep composition models. The nonparametric regression problems where the input data are distributed on a low-dimensional smooth manifold has been studied as a "manifold regression" Yang & Dunson (2016); Bickel & Li

(2007); Yang & Tokdar (2015). Such a model can be considered as a specific example of the deep composition model. In this sense, our analysis is a significant extension of these analyses.

3 Approximation error analysis

Here, we consider approximating the true function f^o via a deep neural network and derive the approximation error. As the activation function, we consider the ReLU activation denoted by $(x) = \max\{x; 0\}$ $(x \in \mathrm{R})$. For a vector x, (x) is operated in an element-wise manner. The model of neural networks with height L, width W, sparsity constraint S, and norm constraint B as $(L; W; S; B) := \{f(W^{(L)} \ (\ ) + b^{(L)}) \quad (W^{(1)}x + b^{(1)}) \mid W^{(L)} \in \mathrm{R}^{1 \ W}; b^{(L)} \in \mathrm{R}; W^{(1)} \in \mathrm{R}^{W \ d}; b^{(1)} \in \mathrm{R}^{W}; W^{(\ell)} \in \mathrm{R}^{W \ W}; b^{(\ell)} \in \mathrm{R}^{W} (1 < \ell < L); \sum_{\ell=1}^{L} (\|W^{(\ell)}\|_0 + \|b^{(\ell)}\|_0) \le S; \max_\ell \|W^{(\ell)}\|_{1} \vee \|b^{(\ell)}\|_{1} \le B\}$, where $\|\ \|_0$ is the ℓ_0-norm of the matrix (the number of non-zero elements of the matrix), and $\|\ \|_{1}$ is the ℓ_{1}-norm of the matrix (maximum of the absolute values of the elements). The sparsity constraint and norm bounds are required to obtain the near-optimal rate of the estimation error. To evaluate the accuracy of the deep neural network model in approximating target functions, we define the worst-case approximation error as

$$R_r(F; H) := \sup_{f \in H} \inf_{f \in F} \|f - f\|_{L^r(\ )};$$

where F is the set of functions used for approximation, and H is the set of target functions.

Proposition 2 (Approximation ability for anisotropic Besov space)**.** *Suppose that* $0 < p; q; r \le \infty$ *and* $\in \mathrm{R}^d_{++}$ *satisfy the following condition:* $\tilde{} > (1/p - 1/r)_+$. *Assume that* $m \in \mathrm{N}$ *satisfies* $0 < \bar{} < \min(m; m - 1 + 1/p)$. *Let* $= (1/p - 1/r)_+$, $= (\tilde{} \ )/(2\ )$ *and* $W_0(d) := 6dm(m+2) + 2d$. *Then, for* $N \in \mathrm{N}$, *we can bound the approximation error as*

$$R_r(\ (L_1; W_1; S_1; B_1); U(B_{p;q}(\ ))) \ . \ N^{\tilde{}};$$

by setting

$$L_1(d) := 3 + 2d\log_2\left(\frac{3^{d \vee m}}{c_{(d;m)}}\right) + 5ed\log_2(d \vee m)e; \quad W_1(d) := NW_0; \tag{2}$$

$$S_1(d) := [(L - 1)W_0^2 + 1]N; \quad B_1(d) := O(N^{d(1 + \ ^{1})(1/p \ \tilde{})_+}); \tag{3}$$

for $= N^{\tilde{}} \log(N)^{\ 1}$ *and a constant* $c_{(d;m)}$ *depending only on* d *and* m.

The proof of this proposition is provided in Appendix B. The rate $N^{\tilde{}}$ is the optimal *adaptive* approximation error rate that can be achieved by a model with N parameters (the difference between adaptive and non-adaptive methods is explained in the discussion below). Note that this is an approximation error in an oracle setting and no sample complexity appears here. We notice that we can avoid the *curse of dimensionality* if the average smoothness $\tilde{}$ is small. This means that if the target function is non-smooth in only a few directions and smooth in other directions, we can avoid the curse of dimensionality. In contrast, if we consider an isotropic Besov space where $_1 = \ = \ _d (= \bar{})$, then $\tilde{} = \bar{} = d$, which directly depends on the dimensionality d, and we need an exponentially large number of parameters in this situation to achieve $$-accuracy. Therefore, the anisotropic smoothness has a significant impact on the approximation error rate. The assumption $\tilde{} > (1/p - 1/r)_+$ ensures the L_r-integrability of the target function, and the inequality (without equality) admits a near-optimal wavelet approximation of the target function in terms of L_r-norm.

Using this evaluation as a basic tool, we can obtain the approximation error for the deep composition models. We can also obtain the approximation error for the affine composition models, but it is almost identical to Proposition 2. Therefore, we defer it to Appendix A.

Theorem 1 (Deep composition model)**.** *Assume that* $\tilde{}^{(\ell)} > 1/p$ *for all* $\ell = 1; \dots; H$. *Then, the estimation error on the deep composition model is bounded as*

$$R_1(\ (L; W; S; B); H_{\mathrm{deep}}) \ . \ \max_{\ell \in [H]} N^{\tilde{}^{(\ell)}}; \tag{4}$$

where $\tilde{}^{(\ell)} = \tilde{}^{(\ell)} \prod_{k=\ell+1}^{H} [(\_^{(k)} \ 1/p) \wedge 1]$, *and* $L = \sum_{\ell=1}^{H} (L_1(m_\ell) + 1); W = \max_\ell (W_1(m_\ell) \_ m_{\ell+1}); S = \sum_{\ell=1}^{H} (S_1(m_\ell) + 3m_{\ell+1}); B = \max_\ell B_1(m_\ell)$:

The proof can be found in Appendix B.1. Since the model is more general than the vanilla anisotropic Besov space, we require a stronger assumption $\tilde{\beta}(\ell) > 1/p$ on $\tilde{\beta}(\ell)$ than the condition in Proposition 2. This is because we need to bound the Hölder smoothness of the remaining layers to bound the influence of the approximation error in the internal layers to the entire function. Hölder smoothness is ensured according to the embedding property under this condition (Proposition 1). This Hölder smoothness assumption affects the approximation error rate. The convergence rate $\tilde{\beta}(\ell)$ in Eq. (4) is different from that in Eq. (8). This is because the approximation error in the internal layers are propagated through the remaining layers with Hölder smoothness and its amplitude is controlled by the Hölder smoothness.

Approximation error by non-adaptive method The approximation error obtained in the previous section is called an adaptive error rate in the literature regarding approximation theory (DeVore, 1998). If we fix N bases beforehand and approximate the target function by a linear combination of the N bases (which is called the non-adaptive method), then we *cannot* achieve the adaptive error rate obtained in the previous section. Roughly speaking, the approximation error of non-adaptive methods is lower bounded by $N^{-(\frac{1}{p} - \frac{1}{\min\{2,r\}})_+}$ (Myronyuk, 2015, 2016, 2017), which is slower than the approximation error rate of deep neural networks especially for small p.

4 Estimation error analysis

In this section, we analyze the accuracy of deep learning in estimating a function in compositions of anisotropic Besov spaces. We consider a least-squares estimator in the deep neural network model:

$$\widehat{f} = \operatorname{argmin}_{f: f \in \mathcal{F}(L,W,S,B)} \sum_{i=1}^{n} (y_i - f(x_i))^2 \tag{5}$$

where f is the *clipping* of f defined by $f = \min\{\max\{f, -F\}, F\}$ for a constant $F > 0$ which is realized by ReLU units. The network parameters (L, W, S, B) should be specified appropriately as indicated in Theorems 2 and 3. In practice, these parameters can be specified by cross validation. Indeed, we can theoretically show that cross validation can provide the appropriate choice of these parameters in compensation to an additional $\log(n)$-factor in the estimation error bound. This estimator can be seen as a sparsely regularized estimator because there are constraints on S. In terms of optimization, this requires a combinatorial optimization, but we do not pursue the computational aspect. The estimation error that we derive in this section can involve the optimization error, but for simplicity, we only demonstrate the estimation error of the *ideal* situation where there is no optimization error.

Affine composition model The following theorem provides an upper bound of the estimation error for the affine composition model.

Theorem 2. *Assume the same condition as in Theorem 6; in particular, suppose* $0 < p, q \leq 1$ *and* $\tilde{\beta} > (1/p - 1/2)_+$ *for* $\tilde{\beta} \in \mathbb{R}^{d^*}_{+,+}$. *Moreover, we assume that the distribution* P_X *has a density* p_X *such that* $\|p_X\|_\infty \leq R$ *for a constant* $R > 0$. *If* $f^o \in \mathcal{H}_a \cap L^1([0,1]^{d^*})$, *and* $\|f^o\|_1 \leq F$ *for* $F \geq 1$; *then, letting* $(W, L, S, B) = (L_1(d^*), W_1(d^*), S_1(d^*), (d^*C + 1)B_1(d^*))$ *as in Theorem 6 with* $N \asymp n^{\frac{1}{2\tilde{\beta}+1}}$, *we obtain*

$$\mathbb{E}_{D_n}[\|f^o - \widehat{f}\|^2_{L^2(P_X)}] \lesssim n^{-\frac{2\tilde{\beta}}{2\tilde{\beta}+1}} \log(n)^3;$$

where $\mathbb{E}_{D_n}[\cdot]$ *indicates the expectation with respect to the training data* D_n.

The proof is provided in Appendix C. We will show that the convergence rate $n^{-2\tilde{\beta}/(2\tilde{\beta}+1)}$ is minimax optimal in Section 5 (see also Kerkyacharian & Picard (1992); Donoho et al. (1996); Donoho & Johnstone (1998); Giné & Nickl (2015) for ordinary Besov spaces). The L^1-norm constraint $\|f^o\|_1 \leq F$ is used to derive a uniform bound on the discrepancy between the population and the empirical L^2-norm. Without this condition, the convergence rate could be slower.

Deep composition model For the deep composition model, we obtain the following convergence rate. This is an extension of Theorem 2 but requires a stronger assumption on the smoothness.

Theorem 3. *Suppose that* $0 < p, q \leq 1$ *and* $\tilde{\beta}(\ell) > 1/p$ *for all* $\ell \in [H]$. *If* $f^o \in \mathcal{H}_{\text{deep}} \cap L^1([0,1]^{d^*})$, *and* $\|f\|_1 \leq F$ *for* $F \geq 1$, *then we obtain*

$$\mathbb{E}_{D_n}[\|f^o - \widehat{f}\|^2_{L^2(P_X)}] \lesssim \max_{\ell \in [H]} n^{-2\tilde{\beta}^*(\ell)/(2\tilde{\beta}^*(\ell)+1)} \log(n)^3;$$

where $\tilde{\beta}^{(\ell)}$ *is defined in Theorem 1, and* (L, W, S, B) *is as given in Theorem 1 with* $N \asymp \max_{\ell \in [L]} n^{\frac{1}{2\tilde{\beta}^{(\ell)}+1}}$.

The proof is provided in Appendix C. We will show that this is also minimax optimal in Theorem 4. Because of the Hölder continuity, the convergence rate becomes slower than the affine composition model (that is, $\tilde{\beta}^{(\ell)} \leq \tilde{\beta}^{(\ell)}$). However, this slower rate is unavoidable in terms of the minimax optimal rate. Schmidt-Hieber (2020) analyzed the same situation for the Hölder class which corresponds to $\beta_1^{(\ell)} = \cdots = \beta_d^{(\ell)} (\beta^{\ell})$ and $p = q = 1$. Our analysis far extends their analysis to the setting of anisotropic Besov spaces in which the parameters $\beta^{(\ell)}, p, q$ have much more freedom.

From these two bounds (Theorems 2 and 3), we can see that as the smoothness $\tilde{\beta}$ becomes large, the convergence rates faster. If the target function is included in the isotropic Besov space with smoothness $\beta_1 = \cdots = \beta_d (= \beta)$, then the estimation error becomes

(Isotropic Besov) $\qquad n^{-2\beta/(2\beta+d)}.$

In the exponent, the dimensionality d appears, which causes the curse of dimensionality. In contrast, if the target function is in the anisotropic Besov space, and the smoothness in each direction is sufficiently imbalanced such that $\tilde{\beta}$ does not depend on d, our obtained rate

(Anisotropic Besov) $\qquad n^{-2\tilde{\beta}/(2\tilde{\beta}+1)}$

avoids the curse of dimensionality. For high-dimensional settings, there would be several redundant directions in which the true function does not change. Deep learning is adaptive to this redundancy and achieves a better estimation error. However, in Section 6, we prove that linear estimators are affected by the dimensionality more strongly than deep learning. This indicates the superiority of deep learning.

5 Minimax optimal rate

Here, we show that the estimation error rate, that we have presented, of deep learning achieves the *minimax optimal rate*. Roughly speaking the minimax optimal risk on a model F of the true function is the smallest worst case error over all estimators:

$$R(F) := \inf_{\widehat{f}} \sup_{f^* \in F} \mathbb{E}_{D_n}[\|\widehat{f} - f^*\|^2_{L^2(P_X)}],$$

where $\widehat{f}$ runs over all estimators. The convergence rate of the minimax optimal risk is referred to as minimax optimal rate. We obtain the following minimax optimal rate for anisotropic Besov spaces.

Theorem 4. **(a) Affine composition model:** *For* $0 < p, q \leq \infty$ *and* $\beta \in \mathbb{R}^d_{+,+}$, *assume that* $\tilde{\beta} > \max\{1/p - 1/2, 1/p - 1, 0\}$. *Then, the minimax optimal risk of the affine composition model is lower bounded as* $R(H_a) \gtrsim n^{-\frac{2\tilde{\beta}}{2\tilde{\beta}+1}}$. **(b) Deep composition model:** *For* $0 < p, q \leq \infty$ *and* $\beta^{(\ell)} \in \mathbb{R}^d_{+,+}$ $(\ell = 1, \ldots, H)$, *assume that* $\tilde{\beta}^{(\ell)} > 1/p$. *Let* $\gamma > 0$ *be arbitrarily small for* $q < \infty$, *and let* $\gamma = 0$ *for* $q = 0$. *Let* $\tilde{\beta}^{(\ell)} = \tilde{\beta}^{(\ell)} \prod_{k=\ell+1}^{H} [(\beta_*^{(k)} \wedge 1)p + \gamma) \wedge 1]$, *and* $\tilde{\beta}^* := \min_{\ell} \tilde{\beta}^{(\ell)}$. *Then, the minimax optimal risk of the deep composition model is lower bounded as* $R(H_{\mathrm{deep}}) \gtrsim n^{-\frac{2\tilde{\beta}^*}{2\tilde{\beta}^*+1}}$.

The proof is provided in Appendix E (see also Ibragimov & Khas'minskii (1984); Nyssbaum (1987)). From this theorem, we can see that the estimation error of deep learning shown in Theorems 2 and 3 indeed achieve the minimax optimal rate up to a poly-$\log(n)$ factor.

6 Suboptimality of linear estimators

In this section, we give the minimax optimal rate in the class of *linear estimators*. The linear estimator is a class of estimators that can be written as

$$\widehat{f}(x) = \sum_{i=1}^n y_i \varphi_i(x, X^n),$$

where $X^n = (x_1, \ldots, x_n)$ and $\varphi_i(x, X^n)$ $(i = 1, \ldots, n)$ are (measurable) functions that only depend on x and X^n. This is linearly dependent on $Y^n = (y_1, \ldots, y_n)$. We notice that the kernel

ridge regression is included in this class because it can be written as $\widehat{f}(x) = k_{x;X^n}(k_{X^n;X^n} + \quad \mathrm{I})^{-1}Y^n$, which linearly depends on Y^n. This class includes other important estimators, such as the Nadaraya–Watson estimator, the k-nearest neighbor estimator, and the sieve estimator. We compare deep learning with the linear estimators in terms of minimax risk. For this purpose, we define the minimax risk of the class of linear estimators:

$$R^{(\mathrm{lin})}(F\ ) := \inf_{\widehat{f}:\ \text{linear}} \sup_{f^o 2 F} \mathsf{E}_{D_n}[\|f^o \quad \widehat{f}\|^2_{L^2(P_X)}];$$

where $\widehat{f}$ runs over all *linear estimators*. We can see that linear estimators suffer from the sub-optimal rate because of the following two points: (i) they do not have adaptivity, and (ii) they significantly suffer from the curse of dimensionality.

Theorem 5. *(i) Suppose that the input distribution P_X is the uniform distribution on $\quad = [0;1]^d$ and assume that $\widetilde{\ } > 1/p$ and $1 \quad p;q \quad 1$. Then, the minimax optimal rate of the linear estimators is lower bounded as*

$$R^{(\mathrm{lin})}(U(B_{p;q})) \ \& \ n^{-\frac{2\widetilde{\ } \quad v}{2\widetilde{\ } + 1 \quad v}}; \tag{6}$$

where $v = 2(1/p \quad 1/2)_+$.

(ii) In addition to the above conditions, we assume that the true function is included in the affine composition model with $d^* \quad d$, $\_ = {}_1 = \quad = {}_{d^*}$ and $0 < p \quad 2$. Let $a_d = 1 + \quad$ with arbitrary small $\quad > 0$ when $d^* < d-2$, and let $a_d = 0$ when $d^* \quad d-2$. Then, the minimax rate of the linear estimators on the affine composition model is lower bounded by

$$R^{(\mathrm{lin})}(H_a\ ) \ \& \ n^{-\frac{2(\_ \quad d^*/p + d/2 + a_d)}{2(\_ \quad d^*/p + d/2 + a_d) + d}}. \tag{7}$$

The proof is provided in Appendix F. (i) The lower bound (7) reveals the suboptimality of linear estimators in terms of input dimensionality. Actually, if we consider a particular case where $d^* = 1$, $p = 1$ and $d \quad d^*$, then the obtained minimax rate of linear estimators and the estimation error rate of deep learning can be summarized as

$$\text{linear}: \ n^{-\frac{2\_ + d}{2\_ + 2d}}; \qquad \text{deep}: \ n^{-\frac{2\_}{2\_ + 1}};$$

by Theorem 2 when $\_ > 1$ (which can be checked by noticing $d^* = \_ = \widetilde{\ } = 1$ in this situation). We can see that the dependence on the dimensionality of linear estimators is significantly worse than that of deep leaning. This indicates poor adaptivity of linear estimators to the intrinsic dimensionality of data. Actually, as d becomes large, the rate for the linear estimator approaches to $1/n$ but that for the deep learning is not affected by d and still faster than $1/n$. To show the theorem, we used the "convex-hull argument" developed by Hayakawa & Suzuki (2019); Donoho & Johnstone (1998). We combined this technique with the so-called *Irie-Miyake's integral representation* (Irie & Miyake, 1988; Hornik et al., 1990). Note that this difference appears because there is an affine transformation in the first layer of the affine composition model. Deep learning is flexible against such a coordinate transform so that it can find directions to which the target function is smooth. In contrast, kernel methods do not have such adaptivity because there is no feature extraction layer. (ii) The lower bound (6) states that when $p < 2$ (that is, $v > 0$), the minimax rate of the linear estimators is outperformed by that of deep learning (Theorem 2). This is due to the "adaptivity" of deep leaning. When p is small, the smoothness of the target function is less homogeneous, and it requires an adaptive approximation scheme to achieve the best estimation error. Linear estimators do not have adaptivity and thus fail to achieve the minimax optimal rate. Our bound (6) extends the result by Zhang et al. (2002) to a multivariate anisotropic Besov space while Zhang et al. (2002) investigated the univariate space ($d = 1$).

7 Conclusion

We investigated the approximation error and estimation error of deep learning in the anisotropic Besov spaces. It was proved that the convergence rate is determined by the average of the anisotropic smoothness, which results in milder dependence on the input dimensionality. If the smoothness is

highly anisotropic, deep learning can avoid overfitting. We also compared the error rate of deep learning with that of linear estimators and showed that deep learning has better dependence on the input dimensionality. Moreover, it was shown that deep learning can achieve the adaptive rate and outperform non-adaptive approximation methods and linear estimators if the homogeneity p of smoothness is small. These analyses strongly support the practical success of deep learning from a theoretical perspective.

Limitations of this work Our work does not cover the optimization aspect of deep learning. It is assumed that the regularized least squares (5) can be executed. It would be nice to combine our study with recent developments of non-convex optimization techniques (Vempala & Wibisono, 2019; Suzuki & Akiyama, 2021).

Potential negative societal impact Since this is purely theoretical result, it is not expected that there is a direct negative societal impact. However, revealing detailed properties of the deep learning could promote an opportunity to pervert deep learning.

Acknowledgment

TS was partially supported by JSPS KAKENHI (18H03201), Japan Digital Design and JST CREST. AN was partially supported by JSPS Kakenhi (19K20337) and JST-PRESTO.

References

M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.

A. Bhattacharya, D. Pati, and D. B. Dunson. Adaptive dimension reduction with a gaussian process prior. *arXiv preprint arXiv:1111.1044*, 1445, 2011.

A. Bhattacharya, D. Pati, and D. Dunson. Anisotropic function estimation using multi-bandwidth gaussian processes. *Annals of statistics*, 42(1):352, 2014.

P. J. Bickel and B. Li. Local polynomial regression on unknown manifolds. In *Complex datasets and inverse problems*, pp. 177–186. Institute of Mathematical Statistics, 2007.

M. Chen, H. Jiang, W. Liao, and T. Zhao. Nonparametric Regression on Low-Dimensional Manifolds using Deep ReLU Networks. *arXiv e-prints*, art. arXiv:1908.01842, Aug 2019.

M. Chen, H. Jiang, W. Liao, and T. Zhao. Efficient approximation of deep relu networks for functions on low dimensional manifolds. In *Advances in Neural Information Processing Systems*, pp. 8172–8182, 2019.

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

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv e-prints*, art. arXiv:1810.04805, Oct 2018.

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

R. A. DeVore and V. A. Popov. Interpolation of Besov spaces. *Transactions of the American Mathematical Society*, 305(1):397–414, 1988.

R. A. DeVore, G. Kyriazis, D. Leviatan, and V. M. Tikhomirov. Wavelet compression and nonlinearn-widths. *Advances in Computational Mathematics*, 1(2):197–214, 1993.

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

D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard. Density estimation by wavelet thresholding. *The Annals of Statistics*, 24(2):508–539, 1996.

D. Dũng. B-spline quasi-interpolant representations and sampling recovery of functions with mixed smoothness. *Journal of Complexity*, 27(6):541–567, 2011a.

D. Dũng. Optimal adaptive sampling recovery. *Advances in Computational Mathematics*, 34(1): 1–41, 2011b.

S. Gaiffas and G. Lecue. Hyper-sparse optimal aggregation. *Journal of Machine Learning Research*, 12(Jun):1813–1833, 2011.

E. Giné and R. Nickl. *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2015.

X. Glorot, A. Bordes, and Y. Bengio. Deep sparse rectifier neural networks. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 315–323, 2011.

H. Hang and I. Steinwart. Optimal learning with anisotropic gaussian svms. *arXiv preprint arXiv:1810.02321*, 2018.

S. Hayakawa and T. Suzuki. On the minimax optimality and superiority of deep neural network learning over sparse parameter spaces. *arXiv preprint arXiv:1905.09195*, 2019.

M. Hoffman and O. Lepski. Random rates in anisotropic regression (with a discussion and a rejoinder by the authors). *The Annals of Statistics*, 30(2):325–396, 04 2002.

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

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

I. Ibragimov and R. Khas'minskii. More on the estimation of distribution densities. *Journal of Soviet Mathematics*, 25(3):1155–1165, 1984.

B. Irie and S. Miyake. Capabilities of three-layered perceptrons. In *IEEE 1988 International Conference on Neural Networks*, pp. 641–648, 1988.

G. Kerkyacharian and D. Picard. Density estimation in Besov spaces. *Statistics & Probability Letters*, 13:15–24, 1992.

G. Kerkyacharian, O. Lepski, and D. Picard. Nonlinear estimation in anisotropic multi-index denoising. *Probability Theory and Related Fields*, 121(2):137–170, Oct 2001.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105, 2012.

C. Leisner. Nonlinear wavelet approximation in anisotropic besov spaces. *Indiana University mathematics journal*, pp. 437–455, 2003.

S. Mallat. *A Wavelet Tour of Signal Processing*. Academic Press, 1999.

V. Myronyuk. Trigonometric approximations and kolmogorov widths of anisotropic besov classes of periodic functions of several variables. *Ukrainian Mathematical Journal*, 66(8), 2015.

V. V. Myronyuk. Kolmogorov widths of the anisotropic besov classes of periodic functions of many variables. *Ukrainian Mathematical Journal*, 68(5):718–727, Oct 2016.

V. V. Myronyuk. Widths of the anisotropic besov classes of periodic functions of several variables. *Ukrainian Mathematical Journal*, 68(8):1238–1251, Jan 2017. ISSN 1573-9376. doi: 10.1007/s11253-017-1290-1. URL https://doi.org/10.1007/s11253-017-1290-1.

V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning*, pp. 807–814, 2010.

R. Nakada and M. Imaizumi. Adaptive approximation and generalization of deep neural network with intrinsic dimensionality. *Journal of Machine Learning Research*, 21(174):1–38, 2020. URL http://jmlr.org/papers/v21/20-002.html.

S. M. Nikol'skii. *Approximation of functions of several variables and imbedding theorems*, volume 205. Springer-Verlag Berlin Heidelberg, 1975.

M. Nyssbaum. Optimal filtration of a function of many variables in white gaussian noise. *Problems of Information Transmission*, 19:23–29, 1983.

M. Nyssbaum. Nonparametric estimation of a regression function that is smooth in a domain in R^k. *Theory of Probability & Its Applications*, 31(1):108–115, 1987.

A. Radford, L. Metz, and S. Chintala. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. *arXiv e-prints*, art. arXiv:1511.06434, Nov 2015.

G. Raskutti, M. J. Wainwright, and B. Yu. Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *The Journal of Machine Learning Research*, 13(1):389–427, 2012.

J. Schmidt-Hieber. Nonparametric regression using deep neural networks with ReLU activation function. *ArXiv preprint arXiv:1708.06633(v3)*, 2018.

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

J. Schmidt-Hieber. Nonparametric regression using deep neural networks with relu activation function. *The Annals of Statistics*, 48(4):1875–1897, 2020.

P. Y. Simard, D. Steinkraus, and J. C. Platt. Best practices for convolutional neural networks applied to visual document analysis. In *Proceedings of the Seventh International Conference on Document Analysis and Recognition-Volume 2*, pp. 958. IEEE Computer Society, 2003.

S. Sonoda and N. Murata. Neural network with unbounded activation functions is universal approximator. *Applied and Computational Harmonic Analysis*, 43(2):233–268, 2017.

T. Suzuki. Fast generalization error bound of deep learning from a kernel perspective. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pp. 1397–1406. PMLR, 2018.

T. 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 (ICLR2019)*, 2019. URL https://openreview.net/forum?id=H1ebTsActm.

T. Suzuki and S. Akiyama. Benefit of deep learning with non-convex noisy gradient descent: Provable excess risk bound and superiority to kernel methods. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=2m0g1wEafh.

V. Temlyakov. *Approximation of Periodic Functions*. Nova Science Publishers, 1993.

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

H. Triebel. Entropy numbers in function spaces with mixed integrability. *Revista matemática complutense*, 24(1):169–188, 2011.

A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer New York, 2008.

A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, New York, 1996.

S. Vempala and A. Wibisono. Rapid convergence of the unadjusted langevin algorithm: Isoperimetry suffices. In *Advances in Neural Information Processing Systems*, pp. 8094–8106, 2019.

J. Vybiral. Function spaces with dominating mixed smoothness. *Dissertationes Math. (Rozprawy Mat.)*, 436:3–73, 2006.

Y. Yang and A. Barron. Information-theoretic determination of minimax rates of convergence. *The Annals of Statistics*, 27(5):1564–1599, 1999.

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

Y. Yang and S. T. Tokdar. Minimax-optimal nonparametric regression in high dimensions. *The Annals of Statistics*, 43(2):652–674, 2015.

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

S. Zhang, M.-Y. Wong, and Z. Zheng. Wavelet threshold estimation of a regression function with random design. *Journal of Multivariate Analysis*, 80(2):256–284, 2002.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [No]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] Each theorem explicitly describes assumptions on which the theorem is established.
   (b) Did you include complete proofs of all theoretical results? [Yes] See Appendix in the supplementary material. All proofs are included there.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

——Appendix——

A Approximation error of Affine composition model

Theorem 6 (Affine composition model)**.** *Assume that the distribution of* $x = Ax + b \in \mathbb{R}^{d'}$ *has a bounded density function on* $[0;1]^{d'}$ *when* x *obeys the uniform distribution on* , *and each element in* A *and* b *is bounded by a constant* C*. Assume that* $0 < p; q; r$ 1 *and* $\in \mathbb{R}^{d'}_{++}$ *satisfy* $\tilde{} > (1/p \quad 1/r)_+$*. Then, it holds that*

$$R_r(\ (L_1(d'); W_1(d'); S_1(d'); (d'C + 1)B_1(d')); H_a\ )\ .\ N\ ^{\tilde{}}; \tag{8}$$

where $L_1(\ );\ W_1(\ );\ S_1(\ ),\ B_1(\ )$ *are defined in Eq.* (3)*.*

The assumption $\tilde{} > (1/p \quad 1/r)_+$ ensures the L_r-integrability of the target function, and the inequality (without equality) admits a near-optimal wavelet approximation of the target function in terms of L_r-norm. From this theorem, the approximation error is almost identical to that for $B_{p;q}(\ )$ (Proposition 2).

B Proofs of approximation error bounds

To show the approximation accuracy, a key step is to show that the ReLU neural network can approximate the *cardinal B-spline* with high accuracy. Let $N(x) = 1\ (x \in [0;1]);\ 0$ (otherwise), then the *cardinal B-spline of order* m is defined by taking $m + 1$-times convolution of N:

$$N_m(x) = (\underbrace{N \quad N \quad\quad\quad N}_{m\ +\ 1\ \text{times}})(x);$$

where $f \quad g(x) := \int f(x \quad t)g(t)\mathrm{d}t$. It is known that N_m is a piece-wise polynomial of order m. For $k \in \mathbb{Z}^d_+$ and $j = (j_1; : : : ; j_d) \in \mathbb{Z}^d_+$, let

$$M^d_{k;j}(x) = \prod_{i=1}^{d} N_m(2^{bk\ {}^0_i c}x_i \quad j_i);$$

where $\in \mathbb{R}^d_{++}$ is a given smoothness parameter (we omit the dependency on from the notation which would be obvious from the context). Here, k controls the spatial "resolution" and j specifies the location on which the basis is put. Basically, we approximate a function f in an anisotropic Besov space via super-position of $M^m_{k;j}(x)$, which is closely related to wavelet analysis (Mallat, 1999). The following is a key lemma that was proven by Suzuki (2019).

Lemma 1 (Approximation of cardinal B-spline basis by the ReLU activation)**.** *There exists a constant* $c_{(d;m)}$ *depending only on* d *and* m *such that, for all* > 0*, there exists a neural network* $M \in (L_0; W_0; S_0; B_0)$ *with* $L_0 := 3 + 2\left\lceil \log_2\left(\frac{3^{d\_m}}{c_{(d;m)}}\right) + 5\right\rceil d\log_2(d\_m)e$*,* $W_0 := 6dm(m+2) + 2d$*,* $S_0 := L_0W_0^2$ *and* $B_0 := 2(m+1)^m$ *that satisfies*

$$kM^d_{0;0} \quad M\,k_{L^1(\mathbb{R}^d)} \quad ;$$

and $M(x) = 0$ *for all* $x \notin [0; m + 1]^d$*.*

Let

$$kkk_{\_^=} := \sum_{j=1}^{d} bk\_^= {}_j c$$

for a $k \in \mathbb{Z}$. For order $m \in \mathbb{N}$ of the cardinal B-spline bases, let

$$J_i(k) = f \quad m; \quad m + 1; : : : ; 2^{bk\ {}^0_i c} \quad 1; 2^{bk\ {}^0_i c}g$$

and

$$J(k) := J_1(k) \quad J_2(k) \qquad J_d(k):$$

and the quasi-norm of the coefficient $(\ _{k;j})_{k;j}$ for $k\ 2\ \mathbb{Z}_+$ and $j\ 2\ J(k)$ be

$$k(\ _{k;j})_{k;j}k_{b_{p;q}}:=\left\{\sum_{k=0}^{1}\left[2^{k[\_\ (\sum_{i=1}^{d}bk\ \ {}^{0}_{i}c=k)=p]}\Big(\sum_{j\ 2\ J(k)}j\ \ _{k;j}j^{p}\Big)^{1=p}\right]^{q}\right\}^{1=q}:$$

For $p\ =\ 1$ or $q\ =\ 1$, the definition should be appropriately modified as usual.

Lemma 2. *Assume the condition* $\tilde{}\ >\ (1=p\quad 1=r)_+$ *in Proposition 2 and* $0\ <\ \bar{}\ <\ \min(m;m\quad 1+1=p)$ *where* $m\ 2\ \mathbb{N}$ *is the order of the cardinal B-spline bases. Then,* $f\ 2\ B_{p;q}$ *admits the following decomposition:*

$$f\ =\sum_{k=0}^{1}\sum_{j\ 2\ J(k)}\ \ _{k;j}M_{k;j}^{d}(x) \tag{9}$$

with convergence in the sense of L^p*, and the coefficient* $(\ _{k;j})$ *yields the following norm equivalence*

$$kfk_{B_{p;q}}\ '\ k(\ _{k;j})_{k;j}k_{b_{p;q}}: \tag{10}$$

For an integer $K\ 2\ \mathbb{N}$*, let* $N\ =\ d2^{kKk_{\_}=}\ e$*, then for any* $f\ 2\ B_{p;q}(\ \ )$*, there exists* f_N *that satisfies*

$$kf\quad f_Nk_{L^{r}(\ \ )}\ .\ N^{-\tilde{}}kfk_{B_{p;q}};$$

and has the following form:

$$f_N(x)\ =\sum_{k=0}^{K}\sum_{j\ 2\ J(k)}\ \ _{k;j}M_{k;j}^{d}(x)\ +\sum_{k=K+1}^{K}\sum_{i=1}^{n_k}\ \ _{k;j_i}M_{k;j_i}^{d}(x); \tag{11}$$

where $K\ \ =\ dK(1\ +\ 1=\ )e$*,* $n_k\ =\ d2^{kKk_{\_}=}\ \ (kkk_{\_}=\ \ kKk_{\_}=\ \ )e\ (k\ =\ K\ +\ 1;:::;K\ )$ *for* $\ \ =\ (1=p\quad 1=r)_+$ *and* $\ \ =\ (\tilde{}\ \ \ \ )=(2\ )$*, and* $(j_i)_{i=1}^{n_k}\ \ J(k)$*.*

Proof of Lemma 2. Leisner (2003) showed that there exists a bounded linear operator P_k that can be expressed as

$$P_k(f)(x)\ =\sum_{j\ 2\ J(k)}a_{k;j}M_{k;j}^{d}(x) \tag{12}$$

where $\ _{k;j}$ is constructed in a certain way, and for every $f\ 2\ L^p([0;1]^d)$ with $0\ <\ p\ \ \ 1$, it holds

$$kf\quad P_k(f)k_{L^p}\ \ \ Cw_{r;p}(f;(2^{-\ {}^{k}_{1}};:::;2^{-\ {}^{k}_{d}}));$$

(See Theorem 3.2.4 of Leisner (2003) and DeVore & Popov (1988)). Let

$$p_k(f)\ :=\ P_k(f)\quad P_{k\ \ 1}(f);\ \ P_{\ \ 1}(f)\ =\ 0:$$

Then, Leisner (2003) showed that when $0\ <\ p;q\ \ \ 1$ and $0\ <\ \bar{}\ <\ \min(m;m\quad 1\ +\ 1=p)$, f belongs to $B_{p;q}$ if and only if f can be decomposed into

$$f\ =\sum_{k=0}^{1}p_k(f);$$

with the convergence condition

$$k(p_k(f))_{k=0}^{1}k_{b_q(L^p)}\ :=\ \Big[\sum_{k=0}^{1}(2^{-\_k}kp_kk_{L^p})^q\Big]^{1=q}\ <\ 1:$$

In particular, it is shown that

$$kfk_{B_{p;q}^{s}}\ '\ k(p_k(f))_{k=0}^{1}k_{b_p^s(L^p)}: \tag{13}$$

Here, each p_k can be expressed as $p_k(x) = \sum_{j \in J(k)} \alpha_{k;j} M^d_{k;j}(x)$ for a coefficient $(\alpha_{k;j})_{k;j}$ which could be different from $(a_{k;j})_{k;j}$ appearing in Eq. (12), and thus $f \in B_{p;q}$ can be decomposed into

$$f = \sum_{k=0}^{1} \sum_{j \in J(k)} \alpha_{k;j} M^d_{k;j}(x)$$

with convergence in the sense of L^p. Moreover, it is shown that $\|p_k\|_{L^p} \lesssim (2^{-kd} \sum_{j \in J(k)} |\alpha_{k;j}|^p)^{1/p}$ and thus

$$\|f\|_{B_{p;q}} \lesssim \|(\alpha_{k;j})_{k;j}\|_{b_{p;q}}.$$

This yields the first assertion.

Next, we move to the second assertion. If $p \geq r$, the assertion can be shown in the same manner as Theorem 3.1 of Dũng (2011a). More precisely, we can show the assertion in a similar line to the following proof for $p < r$ by setting $K = K'$. Thus, we show the assertion only for $p < r$. In this regime, we need to use an adaptive approximation method. In the following, we assume $p < r$. For a given K, by appropriately choosing K' later, we set

$$R_K(f)(x) = \sum_{0 \leq k \leq K} p_k + \sum_{k \in \mathbb{Z}_+ : K < k \leq K'} G_k(p_k);$$

where $G_k(p_k)$ is given as

$$G_k(p_k) = \sum_{1 \leq i \leq n_k} \alpha_{k;j_i} M^d_{k;j_i}(x)$$

where $(\alpha_{k;j_i})_{i=1}^{\#J(k)}$ is the sorted coefficients in decreasing order of absolute value: $|\alpha_{k;j_1}| \geq |\alpha_{k;j_2}| \geq \cdots \geq |\alpha_{k;j_{\#J(k)}}|$. Then, it holds that

$$\|p_k - G_k(p_k)\|_r \leq \|p_k\|_p 2^{-k(k-K)\tau} n_k^{-\tau};$$

where $\tau := (1/p - 1/r)$ (see the proof of Theorem 3.1 of Dũng (2011b) and Lemma 5.3 of Dũng (2011a)). Moreover, we also have

$$\|p_k\|_r \leq \|p_k\|_p 2^{k(k-K)\tau}$$

for $k \in \mathbb{Z}_+$ with $k > K$.

Here, we define N as

$$N = \lceil 2^{K-K'} e \rceil.$$

Let $\tilde{\alpha} = \tilde{\alpha}/(2\alpha)$,

$$K' = \lceil K(1 + 1/\alpha) \rceil e;$$

and

$$n_k = \left\lceil 2^{(K-K')\tau} \left(\frac{k-K}{k-K'}\right)^{\tilde{\alpha}\tau} \right\rceil$$

for $k \in \mathbb{Z}_+$ with $K + 1 \leq k \leq K'$.

Then, by Lemma 5.3 of Dũng (2011a), we have

$$\begin{aligned} \|f - R_K(f)\|_r^r &\lesssim \sum_{K < k \leq K'} \|p_k - G_k(p_k)\|_r^r + \sum_{K' < k} \|p_k\|_r^r \\ &\lesssim \sum_{K < k \leq K'} [\|p_k\|_p 2^{-k(k-K)\tau} n_k^{-\tau}]^r + \sum_{K' < k} [2^{-k(k-K)\tau} \|p_k\|_p]^r. \end{aligned} \tag{14}$$

(a) Suppose that $q \leq r$ and $r < 1$. Then,

$$\begin{aligned} \|f - R_K(f)\|_r^q &= \|f - R_K(f)\|_r^{r \frac{q}{r}} \\ &\lesssim \left\{ \sum_{K < k \leq k_1 \leq K'} [2^{-k(k-K)\tau} n_k^{-\tau} \|p_k\|_p]^r + \sum_{K' < k} [2^{-k(k-K)\tau} \|p_k\|_p]^r \right\}^{\frac{q}{r}} \quad (* \text{ Eq. (14)}) \end{aligned}$$

$$
\begin{aligned}
&\cdot \sum_{K<k\le K} [2^{\widetilde{kkk}_{-=}} n_k \, kp_k k_{L^p}]^q + \sum_{K<k} [2^{\widetilde{kkk}_{-=}} kp_k k_{L^p}]^q \\
&N^{q} 2^{(\widetilde{\ })kKk_{-=}q} \sum_{K<k\le K} [\underbrace{2^{(\widetilde{\ })(kkk_{-=} - kKk_{-=})}}_{1} 2^{\widetilde{kkk}_{-=}} kp_k k_{L^p}]^q \\
&+ 2^{q(\widetilde{\ })kK\,k_{-=}} \sum_{K<k} [2^{\widetilde{kkk}_{-=}} kp_k k_{L^p}]^q \\
&\overset{(i)}{\cdot} (N^{\ } 2^{(\widetilde{\ })kKk_{-=}} + 2^{(\widetilde{\ })K})^q kfk^q_{MB^s_{p;q}} \quad (* \text{ Eq. (13)}) \\
&\overset{(ii)}{\cdot} (N^{\ \widetilde{\ }})^q kfk^q_{MB_{p;q}}
\end{aligned}
$$

where we used $2^{\widetilde{kkk}_{-=}} \prime 2^{-k}$ in (i), and $N \prime 2^{kKk_{-=}}$ and $= (\widetilde{\ })/(2^{\ })$ in (ii).

(b) Suppose that $q > r$ and $r < 1$. Then, letting $= q/r\ (> 1)$ and $^{0} = 1/(1 - 1/\ ) = q/(q - r)$ (note that $\frac{1}{\ } + \frac{1}{^{0}} = 1$), we have

$$
\begin{aligned}
&kf - R_K(f)k^r_{L^r} \cdot \sum_{K<k\le K} [2^{\widetilde{kkk}_{-=}} n_k \, kp_k k_{L^p}]^r + \sum_{K<k} [2^{\widetilde{kkk}_{-=}} kp_k k_{L^p}]^r \quad (* \text{ Eq. (14)}) \\
&2^{\widetilde{kKk}_{-=}r} \sum_{K<k\le K} [2^{(\widetilde{\ })(kkk_{-=} - kKk_{-=})} 2^{\widetilde{kkk}_{-=}} kp_k k_{L^p}]^r \\
&+ \sum_{K<k} [2^{\widetilde{kkk}_{-=}} kp_k k_{L^p}]^r (2^{(\widetilde{\ })kkk_{-=}})^r \\
&(2^{\widetilde{kKk}_{-=}} + 2^{(\widetilde{\ })kK\,k_{-=}})^r \Big\{ \sum_{K<k\le K} [2^{(\widetilde{\ })(kkk_{-=} - kKk_{-=})} 2^{\widetilde{kkk}_{-=}} kp_k k_{L^p}]^r \\
&+ \sum_{K<k} [2^{\widetilde{kkk}_{-=}} kp_k k_{L^p}]^r 2^{(\widetilde{\ })(kkk_{-=} - kK\,k_{-=})r} \Big\} \\
&(2^{\widetilde{kKk}_{-=}r} + 2^{(\widetilde{\ })kK\,k_{-=}})^r \left\{ \sum_{K<k\le K} [2^{\widetilde{kkk}_{-=}} kp_k k_{L^p}]^r + \sum_{K<k} [2^{\widetilde{kkk}_{-=}} kp_k k_{L^p}]^r \right\}^{1/\ } \\
&\left\{ \sum_{K<k\le K} [2^{(\widetilde{\ })(kkk_{-=} - kKk_{-=})}]^{r/^{0}} + \sum_{K<k} [2^{(s\ )(kkk_{-=} - K\ )}]^{r/^{0}} \right\}^{1/\ ^{0}} \\
&\cdot (2^{\widetilde{kKk}_{-=}} + 2^{(\widetilde{\ })kK\,k_{-=}})^r kfk^r_{B_{p;q}} \quad (* \text{ Eq. (13) and } 2^{\widetilde{kkk}_{-=}} \prime 2^{-k}) \\
&\cdot (N^{\ \widetilde{\ }})^r kfk^r_{B_{p;q}};
\end{aligned}
$$

(c) Suppose that $r = 1$. Then, similarly to the analysis in (b), we can evaluate

$$
\begin{aligned}
&kf - R_K(f)k_{L^r} \\
&\cdot 2^{\widetilde{kKk}_{-=}} \sum_{K<k\le K} [2^{(\widetilde{\ })(kkk_{-=} - kKk_{-=})} 2^{\widetilde{kkk}_{-=}} kp_k k_{L^p}] \\
&\quad + \sum_{K<k} [2^{\widetilde{kkk}_{-=}} kp_k k_{L^p}](2^{(\widetilde{\ })kkk_{-=}}) \\
&\cdot (2^{\widetilde{kKk}_{-=}} + 2^{(\widetilde{\ })kK\,k_{-=}}) kfk_{B_{p;q}} \\
&\cdot N^{\ \widetilde{\ }} kfk_{B_{p;q}};
\end{aligned}
$$

This concludes the proof. □

Proof of Proposition 2. We adopt the proof line employed by Suzuki (2019). Basically, we combine Lemma 1 and Lemma 2. We substitute the approximated cardinal B-spline basis M into the

decomposition of f_N (11). Let the set of indexes $(k; j) \in Z \quad Z$ that consists f_N given in Eq. (11) be E_N, i.e., $f_N = \sum_{(k;j)\in E_N} \quad _{k;j} M^d_{k;j}$. Accordingly, we set $f := \sum_{(k;j)\in E_N} \quad _{k;j} M^d_{k;j}$. Note that for each x, the number of $(k; j) \in E_N$ that satisfy $M_{k;j}(x) \neq 0$ is bounded by $(m + 1)^d(1 + K \ )$, and $\max_{(k;j)\in E_N} j \ _{k;j}j \ . \ 2^{K \ =(\tilde{} \ 1=p)_+}$ by the norm equivalence Eq. (10). For each $x \in R^d$, it holds that

$$\begin{aligned} jf_N(x) \quad f(x)j &\quad \sum_{(k;j)\in E_N} j \ _{k;j}jjM^d_{k;j}(x) \quad M^d_{k;j}(x)j \\ &\quad \sum_{(k;j)\in E_N} j \ _{k;j}j1fM^d_{k;j}(x) \neq 0g \\ &\ . \ (m + 1)^d(1 + K \ )2^{K \ (\_=\tilde{})(\tilde{} \ 1=p)_+} kfk_{B^s_{p;q}} \\ &\ . \ \log(N)N^{(1+ \quad ^1)(\tilde{} \ 1=p)_+} kfk_{B^s_{p;q}}; \end{aligned}$$

where we used the definition of K in the last inequality. This evaluation yields that, for each $f \in U(B_{p;q}(\ ))$, it holds that

$$kf \quad fk_{L^r} \ . \ kf \quad f_Nk_{L^r} + kf_N \quad fk_{L^r} \ . \ \log(N)N^{(1+ \quad ^1)(1=p \ \tilde{})_+} kfk_{B^s_{p;q}} \ + N \ \tilde{};$$

By taking $$ to satisfy $\log(N)N^{(1+ \quad ^1)(1=p \ \tilde{})_+} \quad N \ \tilde{}$, we obtain the approximation error bound.

As we have seen above $\max_{(k;j)\in E_N} j \ _{k;j}j \ . \ 2^{K \ =(\tilde{} \ 1=p)_+} \quad N^{(1+ \quad ^1)(1=p \ \tilde{})_+}$. The max of the absolute values of parameters used in $M^d_{k;j}$ can be bounded by 2^K (see Suzuki (2019)) which is bounded by $N^{d(1+ \quad ^1)(1=p \ \tilde{})_+}$. Then, we obtain the assertion.

□

B.1 Proof of Theorem 6 and Theorem 1

Proof of Theorem 6. This proof is almost obvious from Proposition 2. We know that, from Proposition 2, for $g \in U(B_{p;q}([0; 1]^d))$, there exists $f \in \quad (L_1(d); W_1(d); S_1(d); B_1(d))$ such that

$$kf \quad gk_r \ . \ N \ \tilde{};$$

Because the density of the distribution of $Ax + b$ is bounded above when x obeys the uniform distribution on $$, this also yields

$$kf \quad (A \ +b) \quad g \quad (A \ +b)k_r \ . \ N \ \tilde{};$$

(note that the Lebesgue measure on $= [0; 1]^d$ corresponds to the uniform distribution on $$). If f can be written as

$$f(x) = (W^{(L_1)} \ ( ) + b^{(L)}) \qquad (W^{(1)}x + b^{(1)});$$

then we have

$$f \ (A \ +b) = (W^{(L_1)} \ ( )+b^{(L)}) \qquad (W^{(1)}A \ +b^{(1)}+W^{(1)}b) \in \quad (L_1(d); W_1(d); S_1(d); (dC+1)B_1(d)):$$

□

Proof of Theorem 1.

$$H_{\text{deep}} := fh_H \qquad h_1(x) \ j \ h_` : [0; 1]^{m_`} \ ! \ [0; 1]^{m_{`+1}}; \ h_{`;k} \in U(B^{(`)}_{p;q}([0; 1]^{m_`})) \ (8k \in [m_{`+1}])g:$$

Since $\tilde{}(`) \quad > \quad 1=p$, we can show that for each $h_{`;k}$, there exists $f_{`;k} \quad \in \quad (L_1(m_`); W_1(m_`); S_1(m_`); B_1(m_`))$ such that

$$kf_{`;k} \quad h_{`;k}k_1 \ . \ N \ \tilde{};$$

Moreover, from the proof of Proposition 2, we can share all parameters other than the last layer among $f_{`;k}$ $(k = 1; : : : ; m_{`+1})$. If necessary, we may modify $f_{`;k}$ so that $0 \quad f_{`;k}(x) \quad 1 \ (8x \in$

$[0;1]^{m_\ell}$) by adding one additional clipping layer which can be realized by ReLU (actually, the clipping operator can be constructed by a linear combination of 2 nodes with ReLU activation as $f(x) = \max\{x; 0\} \quad \max\{x \quad 1; 0\} = \min\{\max\{x; 0\}; 1\}$ for $x \in \mathbb{R}$). The approximation error of the whole layer can be evaluated as

$$\begin{aligned}
&\|h_H \quad h_1 \quad f_H \quad f_1\|_1 \\
&\sum_{\ell=1}^{H} \|h_{\ell+1} \quad h_\ell \quad f_\ell \quad _1 \quad f_1 \quad h_H \quad h_{\ell+1} \quad f_\ell \quad f_\ell \quad _1 \quad f_1\|_1 \\
&\sum_{\ell=1}^{H} \|h_{\ell+1} \quad h_\ell \quad h_H \quad h_{\ell+1} \quad f_\ell\|_1 :
\end{aligned}$$

Proposition 1 tells that $h_{\ell';k} \in C(\_^{(\ell')} \quad 1=p) \wedge 1$; thus, $h_{\ell';k}$ is ℓ'-Hölder continuous where $\ell' := (\_^{(\ell')} \quad 1=p) \wedge 1$. Their composition $h_H \quad h_H \quad _1 \quad h_{\ell+1}$ is ℓ-Hölder continuous where $B_\ell = \prod_{\ell'=\ell+1}^{H} \ell'$. Therefore, we have

$$\|h_H \quad h_{\ell+1} \quad h_\ell \quad h_H \quad h_{\ell+1} \quad f_\ell\|_1 \quad . \quad \|h_\ell \quad f_\ell\|_1^{B_\ell};$$

where $\|\quad\|_1$ for a vector-valued function $g : \mathbb{R}^{d'} \to \mathbb{R}^{d''}$ is defined as $\sup_x \|g(x)\|$. Summing up this evaluation for $\ell = 1; \ldots; H$ concludes that

$$\|h_H \quad h_1 \quad f_H \quad f_1\|_1 \quad . \quad \sum_{\ell=1}^{H} N^{B_\ell \tilde{}(\ell)} \quad . \quad \max_{\ell \in H} N^{\tilde{} \quad (\ell)}:$$

Consequently, the whole network can be realized as an element of $(L; W; S; B)$ where

$$\begin{aligned}
L &= \sum_{\ell=1}^{H} (L_1(m_\ell) + 1); \quad W = \max(W_1(m_\ell) _ m_{\ell+1}); \\
S &= \sum_{\ell=1}^{H} (S_1(m_\ell) + 3m_{\ell+1}); \quad B = \max B_1(m_\ell):
\end{aligned}$$

□

C Proofs of estimation error bound (Theorem 2 and Theorem 3)

Proof of Theorem 2. We follow the proof strategy from Schmidt-Hieber (2018); Suzuki (2019) which uses Proposition 4. It suffices to the covering number of $\hat{F} = \{f \mid f \in (L; W; S; B)\}$ for $(L; W; S; B)$ given in Theorem 6 where f is the clipped version of a function f. Note that the covering number of $\hat{F}$ is not larger than that of $(L; W; S; B)$. Hence, it is sufficient to evaluate that of $(L; W; S; B)$. From Lemma 6, the covering number of this class is upper bounded by

$$\log N(\ ; \hat{F}; \|\quad\|_1) \quad . \quad N \log(N)[\log(N)^2 + \log(\quad ^1)]:$$

From Proposition 2, there exists $f \in (L; W; S; B)$ such that

$$\|f^o \quad R_K(f^o)\|_2 \quad . \quad N^{\tilde{}}:$$

Moreover, we notice that $\|f \quad f^o\|_{L^2(P_X)} \quad R\|f \quad f^o\|_2$: for any $f : [0;1]^d \to \mathbb{R}$ because the density p_X of P_X is bounded by R. Therefore, by applying Proposition 4 with $\quad = 1=n$, we have

$$\mathbb{E}_{D_n}[\|\hat{f} \quad f^o\|_{L^2(P_X)}] \quad . \quad N^{2\tilde{}} + \frac{N \log(N)(\log(N)^2 + \log(n))}{n} + \frac{1}{n}:$$

Here, we can minimize the right hand side by setting $N \quad n^{\frac{1}{2\tilde{}+1}}$ up to $\log(n)^3$-order, and then we obtain the estimation error of the least squares estimator as

$$n^{-\frac{2\tilde{}}{2\tilde{}+1}} \log(n)^3:$$

This yields the assertion. □

Proof of Theorem 3. The proof is almost identical to the proof of Theorem 2, except that we use Theorem 1 as an approximation error bound. □

D Embedding theorem

Lemma 3. *For* $0 < p^{(1)}; p^{(2)} \quad 1$, *let* $^{(1)}; \ ^{(2)} \ 2 \ \mathrm{R}^d_{+\,+}$ *such that they satisfy*

$$\begin{aligned} &\sim^{(1)} \quad \sim^{(2)} \quad \frac{1}{p^{(1)}} \quad \frac{1}{p^{(2)}}; \\ &{}^{(2)} = \quad {}^{(1)}; \\ &p^{(1)} < p^{(2)}; \end{aligned} \tag{15}$$

for $0 < \quad < 1$. *Then, it holds that*

$$B^{(1)}_{p^{(1)};q} \ , ! \ B^{(2)}_{p^{(2)};q}:$$

Proof. We show the assertion only for the situation where $p^{(1)} \not= 1$, $p^{(2)} \not= 1$, and $q \not= 1$. The proof for the setting in which $p^{(1)} = 1$, $p^{(2)} = 1$, or $q = 1$ is satisfied is almost identical. Recall the following norm equivalence shown in Lemma 2:

$$\|f\|_{B^s_{p;q}} \ ' \ \|(\ _{k;j})_{k;j}\|_{b_{p;q}} = \left\{ \sum_{k=0}^{1} \left[2^{k[\_ \ (\sum_{i=1}^d bk \ {}^{0}_{i}c=k)=p]} \Big(\sum_{j \, 2 J(k)} j \ _{k;j} j^p \Big)^{1=p} \right]^q \right\}^{1=q};$$

when $p; q < 1$. Since $\frac{p^{(1)}}{p^{(2)}} < 1$, it holds that

$$\Big(\sum_{j \, 2 J(k)} j \ _{k;j} j^{p^{(1)}} \Big)^{1=p^{(1)}} = \Big(\sum_{j \, 2 J(k)} j \ _{k;j} j^{p^{(2)} \frac{p^{(1)}}{p^{(2)}}} \Big)^{1=p^{(1)}}$$

$$\Big(\sum_{j \, 2 J(k)} j \ _{k;j} j^{p^{(2)}} \Big)^{\frac{p^{(1)}}{p^{(2)}} \frac{1}{p^{(1)}}} = \Big(\sum_{j \, 2 J(k)} j \ _{k;j} j^{p^{(2)}} \Big)^{\frac{1}{p^{(2)}}}:$$

Moreover, we have

$$\begin{aligned} &2^{k[\_^{(1)} \ (\sum_{i=1}^d bk \ {}^{0(1)}_{i}c=k)=p^{(1)}]} \\ &' \ 2^{k\_^{(1)} \ \sum_{i=1}^d \ {}^{0(1)}_{i}=p^{(1)}} = 2^{k \frac{(1)}{\sim^{(1)}} \left(\sim^{(1)} \ \frac{1}{p^{(1)}} \right)} \overset{(a)}{=} 2^{k \frac{(2)}{\sim^{(2)}} \left(\sim^{(1)} \ \frac{1}{p^{(1)}} + \frac{1}{p^{(2)}} \ \frac{1}{p^{(2)}} \right)} \\ &\overset{(b)}{} 2^{k \frac{(2)}{\sim^{(2)}} \left(\sim^{(2)} \ \frac{1}{p^{(2)}} \right)} = 2^{k\_^{(2)} \ \sum_{i=1}^d \ {}^{0(2)}_{i}=p^{(2)}} \ ' \ 2^{k[\_^{(2)} \ (\sum_{i=1}^d bk \ {}^{0(2)}_{i}c=k)=p^{(2)}]}; \end{aligned}$$

where we used the condition $^{(2)} = \ ^{(1)}$ in (a), and we used the condition from Eq. (15) in (b). These relations yield the following evaluation:

$$\begin{aligned} &\|f\|_{B^{(1)}_{p^{(1)};q}} \\ &' \ \|(\ _{k;j})_{k;j}\|_{b^{(1)}_{p^{(1)};q}} \\ &= \left\{ \sum_{k=0}^{1} \left[2^{k[\_^{(1)} \ (\sum_{i=1}^d bk \ {}^{0(1)}_{i}c=k)=p^{(1)}]} \Big(\sum_{j \, 2 J(k)} j \ _{k;j} j^{p^{(1)}} \Big)^{1=p^{(1)}} \right]^q \right\}^{1=q} \\ &\& \left\{ \sum_{k=0}^{1} \left[2^{k[\_^{(2)} \ (\sum_{i=1}^d bk \ {}^{0(2)}_{i}c=k)=p^{(2)}]} \Big(\sum_{j \, 2 J(k)} j \ _{k;j} j^{p^{(2)}} \Big)^{1=p^{(2)}} \right]^q \right\}^{1=q} \\ &' \ \|f\|_{B^{(2)}_{p^{(2)};q}}: \end{aligned}$$

This yields the assertion. □

By combining Lemma 3 with the relation $B_{1;1} \ , ! \ C \ -$ (Triebel, 2011), we immediately obtain the following corollary.

Corollary 1. *Suppose that* $\sim > p$, *then for* $= \frac{\sim p}{\sim -}$; *it holds that*

$$B_{p;q} \ , ! \ B_{1;q} \ , ! \ B_{1;1} \ , ! \ C \ -:$$

E Minimax optimality

In this section, we demonstrate the proof of Theorem 4. Before this, we prepare the basic notions. The -covering number $N(\ ;C;\hat{d})$ of a metric space C equipped with a metric $\hat{d}$ that is the minimal number of balls with radius measured by the metric $\hat{d}$ required to cover the set C (van der Vaart & Wellner, 1996). Similarly, the -packing number $M(\ ;C;\hat{d})$ is defined as the largest number of elements $f f_1;:::;f_M g$ C such that $\hat{d}(f_i;f_j)$ for all $i \neq j$.

Raskutti et al. (2012) showed the following inequality in their proof of Theorem 2(b) by utilizing the result by Yang & Barron (1999).

Lemma 4. *Let F be the model of the true function. For a given $_n > 0$ and $"_n > 0$, let Q be the $_n$-packing number $M(\ _n;F\ ;L^2(P_X))$ of F and N be the $"_n$ covering number of that. Suppose that they satisfy the following condition:*

$$\begin{aligned} &\frac{n}{2^{\ 2}}"_n^2 \quad \log(N); \\ &8\log(N) \quad \log(Q);\ 4\log(2) \quad \log(Q): \end{aligned} \tag{16}$$

Then, the minimax learning rate is lower bounded as

$$\inf_{\hat{f}}\sup_{f\ 2F} \mathsf{E}_{D_n}[k\hat{f} \quad f\ k^2_{L^2(P_X)}] \quad \frac{2}{4}:$$

This concludes the assertion.

Now, we are ready to show Theorem 4.

Proof of Theorem 4. Proposition 10 of Triebel (2011) showed that the -covering number of the unit ball of anisotropic Besov spaces $B_{p;q}(\ )$ can be evaluated as

$$\log N(\ ;U(B_{p;q}(\ ));k\ \ k_r)\ '\ \ ^{1=\tilde{\ }};$$

for $0 < p;q \quad 1, 1 \quad r < 1$, and $2\ \mathsf{R}^d_{+\ +}$ that satisfy

$$\tilde{\ } > \max\left\{\frac{1}{p} \quad \frac{1}{r};\frac{1}{p} \quad 1;0\right\}:$$

Affine composition model:

Apparently, $U(B_{p;q}(\ ))$ is included in H_{a}. Hence, noting that P_X is the uniform distribution and $k\ \ k_2 = k\ \ k_{L^2(P_X)}$, the covering number of H_{a} can be lower bounded by

$$\log N(H_{\mathsf{a}}\ ;k\ \ k_{L^2(P_X)}) \ \& \quad ^{1=\tilde{\ }}:$$

From this evaluation, Lemma 4 yields that there exists $C_1 > 0$ independent of n such that

$$\inf_{\hat{f}}\sup_{f\ 2H_{\mathsf{aff}}} \mathsf{E}_{D_n}[k\hat{f} \quad f\ k^2_{L^2(P_X)}] \quad C_1 n^{\ \frac{2\tilde{\ }}{2\tilde{\ }+1}}:$$

To see this, we may just set $_n\ '\ \ _n\ '\ n^{\ \frac{2\tilde{\ }}{2\tilde{\ }+1}}$ in Eq. (16) of Lemma 4.

Deep composition model:

Next, we show the minimax rate for the deep composition model. Basically, we follow the same strategy developed by Schmidt-Hieber (2018), but we need to modify some technical details because we are dealing with anisotropic Besov spaces while Schmidt-Hieber (2018) analyzed isotropic Hölder space. Let $` := \min_{\cdot\ 2[H]}\tilde{\ }^{(`)}$, and $s^{(`)} := (\ ^{(`)} \quad 1=p + \ ) \wedge 1$ where > 0 can be arbitrary small for $q < 1$ and $= 0$ for $q = 1$. Without loss of generality, we may assume that $^{(`)}_1 \quad ^{(`)}_2 \quad ^{(`)}_d$ for $` \ 2\ [H]$. Let us consider a sub-model H^0_{deep} of H_{deep} defined as

$$H^0_{\mathsf{deep}} := fg_H \qquad g_1\ j$$

$$g_\ell(x) = x \quad (\ell = 1, \dots, \ell - 1);$$
$$g_\ell(x) = (g_{\ell,1}(x), 0, \dots, 0)^\top \text{ where } g_{\ell,1} \in U(B_{p,q}^{(\alpha)}(R));$$
$$g_\ell(x) = (x_1^{s(\alpha)}, 0, \dots, 0)^\top \quad (\ell = \ell + 1, \dots, H) g.$$

For $\ell = \ell + 1, \dots, H$, through a cumbersome calculation, we can verify that $x_1^{s(\alpha)} \in B_{\bar{p},q}^{(\alpha)}([0,1])$ for $x \in [0,1]$, which ensures $g_{\ell,j}(x) \in B_{p,q}^{(\alpha)}([0,1]^d)$ for $j = 1, \dots, d$. To lower bound the covering number, we concretely construct a subset the cardinality of which can be easily estimated. For that purpose, we use the expansion $f = \sum_{k=0}^{\infty} \sum_{j \in J(k)} \omega_{k,j} M_{k,j}^d(x)$ and the norm equivalence $\|f\|_{B_{p,q}} \asymp \|(\omega_{k,j})_{k,j}\|_{b_{p,q}}$ given in Lemma 2. For a while, we let $\alpha := \alpha(\ell)$ and $B := \prod_{q=\ell+1}^{H} s(\alpha)$. We define $k \in \mathbb{N}$ so that k satisfies $2^{k\alpha} \asymp n^{\frac{1}{1+2B\alpha}}$. For this choice of k, take a subset $\tilde{J}(k) \subset J(k)$ such that $|\tilde{J}(k)| \asymp |J(k)|$ and for each $j, j' \in \tilde{J}(k)$ with $j \neq j'$, the supports of $M_{k,j}^d$ and $M_{k,j'}^d$ are disjoint. Using this index set $\tilde{J}(k)$, we consider a set of functions that is given by

$$\tilde{H}_\ell := \left\{ f = \sum_{j \in \tilde{J}(k)} \omega_{k,j} M_{k,j}^d(x) \mid \omega_{k,j} \in \{0, 2^{-k\alpha}\} \right\}.$$

We can check that $|\tilde{H}_\ell| = |\tilde{J}(k)| \asymp 2^{k \sum_{j=1}^{d} \alpha_j} = 2^{k/\tilde{\alpha}}$ and $\|f\|_{B_{p,q}} \lesssim 1$ for all $f \in \tilde{H}_\ell$ from the norm equivalence (10). For any $g_w = \sum_{j \in \tilde{J}(k)} w_j 2^{-k\alpha} M_{k,j}^d(x) \in \tilde{H}_\ell$ $(w \in \{0,1\}^{|\tilde{J}_k|})$, we can see that

$$f_w(x) = g_H \circ g_{\ell+1} \circ g_w \circ g_{\ell-1} \circ \cdots \circ g_1(x)$$
$$= \sum_{j \in \tilde{J}(k)} w 2^{-Bk\alpha} M_{k,j}^{dB}(x).$$

If $w \neq w'$, then we can see that

$$\|f_w - f_{w'}\|_{L^2(P_X)}^2 \gtrsim \mathrm{Ham}(w, w') 2^{-2BK\alpha} \asymp 2^{-k/\tilde{\alpha}}$$
$$\gtrsim \mathrm{Ham}(w, w') 2^{-k(2B\tilde{\alpha}+1)/\tilde{\alpha}};$$

where Ham is the Hamming distance because $\|M_{k,j}^d\|_{L^2(P_X)}^2 \asymp 2^{-k/\tilde{\alpha}}$.

Then, by the Varshamov–Gilbert bound (see Lemma 2.9 of Tsybakov (2008), for example), there exists a subset $W_k \subset \{0,1\}^{|\tilde{J}(k)|}$ such that $|W_k| \geq 2^{|\tilde{J}(k)|/8}$ and $\mathrm{Ham}(w, w') \geq |\tilde{J}(k)|/8$ for all $w, w' \in W_k$ with $w \neq w'$. This yields

$$\|f_w - f_{w'}\|_{L^2(P_X)}^2 \gtrsim 2^{-k/\tilde{\alpha}} 2^{-k(2B\tilde{\alpha}+1)/\tilde{\alpha}} = 2^{-2Bk\alpha} \asymp n^{-\frac{2B\tilde{\alpha}}{2B\tilde{\alpha}+1}};$$

where the definition of k is used. This implies that there exists a subset $H_{\mathrm{deep}}'' \subset H_{\mathrm{deep}}'(R) \subset H_{\mathrm{deep}}$ such that

$$\log(N(\varepsilon_n, H_{\mathrm{deep}}'', \|\cdot\|_{L^2(P_X)})) \gtrsim n^{\frac{1}{1+2B\alpha}}$$

for $\varepsilon_n \lesssim n^{-\frac{B\tilde{\alpha}}{2B\tilde{\alpha}+1}}$. Then, by Lemma 4, we obtain that the minimax optima rate on H_{deep} is lower bounded as

$$\inf_{\hat{f}} \sup_{f \in F} \mathsf{E}_{D_n}[\|\hat{f} - f\|_{L^2(P_X)}^2] \gtrsim n^{-\frac{B\tilde{\alpha}}{2B\tilde{\alpha}+1}}.$$

□

F Minimax optimal rate of linear estimators

Define the convex hull of a function class F as

$$\mathrm{conv}(F) := \left\{ f(x) = \sum_{j=1}^{M} \lambda_j f_j(x) \mid M = 1, 2, \dots, f_j \in F, \lambda_j \geq 0, \sum_{j=1}^{M} \lambda_j = 1 \right\}.$$

Let $\overline{\mathrm{conv}}(\ )$ is the closure of $\overline{\mathrm{conv}}(\ )$ with respect to $L_2(P_X)$-norm.

Proposition 3 (Hayakawa & Suzuki (2019)). *The minimax optimal rate of linear estimators on a target function class F is the same as that on the convex hull of F :*

$$\inf_{\widehat{f}:\text{ linear}} \sup_{f^o \in F} \mathbb{E}_{D_n}[\|f^o - \widehat{f}\|^2_{L^2(P_X)}] = \inf_{\widehat{f}:\text{ linear}} \sup_{f^o \in \overline{\mathrm{conv}}(F)} \mathbb{E}_{D_n}[\|f^o - \widehat{f}\|^2_{L^2(P_X)}].$$

See Hayakawa & Suzuki (2019) for the proof of this proposition.

Proof of Theorem 5. We basically follow the strategy developed by Zhang et al. (2002). Let be the uniform measure on . They essentially showed the following statement in their Theorem 1. Suppose that the space has even partition A such that $|A| = 2^K$ for an integer $K \in \mathbb{N}$, each A has equivalent measure $(A) = 2^{-K}$ for all $A \in A$, and A is indeed a partition of , i.e., $\bigcup_{A \in A} A =$, $A \setminus A' = \emptyset$ for $A, A' \in$ and $A \neq A'$. Then, if K is chosen as $n^{\alpha_1} \le 2^K \le n^{\alpha_2}$ for constants $\alpha_1, \alpha_2 > 0$ that are independent of n, then there exists an event E such that, for a constant $C' > 0$,

$$|\{x_i \mid x_i \in A\ (i \in \{1, \dots, n\})\}| \le C' n / 2^K \quad (\forall A \in A),$$
$$P(E) \ge 1 + o(1).$$

We call this property of A "Condition A."

Here, we consider a set F of functions on for which there exists > 0 that satisfies the following conditions:

1. There exists $F > 0$ such that, for any $A \in A$, there exists $g \in F$ that satisfies $g(x) \ge \frac{1}{2} F$ for all $x \in A$,
2. There exists K' and $C'' > 0$ such that $\frac{1}{n}\sum_{i=1}^n g(x_i)^2 \le C'' 2^{K'}$ for any $g \in F$ on the event E.

We call this condition of the function class F "Condition B."

Let the minimax optimal rate of linear estimators on the function class F be

$$R = \inf_{\widehat{f}:\text{linear}} \sup_{f^o \in F} \mathbb{E}_{D^n}[\|\widehat{f} - f^o\|^2_{L^2(P_X)}].$$

Then, under Conditions A and B, there exists a constant F_1 such that at least one of the following inequalities holds:

$$\frac{F^2}{4F_1 C''} \frac{2^{K'}}{n} \le R, \tag{17a}$$
$$\frac{F^3}{32} 2^{-K} \le R, \tag{17b}$$

for sufficiently large n.

(i) Proof of Eq. (6).

For given $k \in \mathbb{N}$ (which will be fixed later), let $= 2^{k[\_ (\sum_{i=1}^d b_k {}^{\prime}_i c = k) = p]}$. Then, from the wavelet expansion of anisotropic Besov space (9),

$$f_w = \sum_{j \in J(k)} w_j M^d_{k;j}(x) \in CU(B_{p;q}( )),$$

where $C > 0$ is a constant and $w = (w_j)_{j \in J(k)}$ is a one-hot vector, i.e., $w_j = 1$ for some $j \in J(k)$ and $w_{j'} = 0$ for all $j' \in J(k)$ with $j' \neq j$. This expansion ensures that, for $K = \sum_{i=1}^d b_k {}^{\prime}_i c$, there exists a partition A of that satisfies Condition A, and for any $A \in A$, there exists w such that $f_w(x) \&$ for all $x \in A$ and

$$\frac{1}{n}\sum_{i=1}^n f_w(x_i)^2 \le \frac{1}{n} |\{i \mid x_i \in A\ (i = 1, \dots, n)\}| \cdot 2^{-K};$$

on the event E, which ensures that $F = \{f_w \mid w$ is a one-hot vector$\}$ satisfies Condition B. Hence, by choosing $k \in \mathbb{N}$ so that $2^k \asymp n^{\frac{1}{2(\tilde{\alpha}+\frac{1}{2}-\frac{1}{p})+1}}$ (recall that $K = \sum_{i=1}^{d} b_k \binom{k}{i} c$ by definition), and setting $K = K'$, then Eq. (17) gives

$$R \gtrsim n^{-\frac{2\tilde{\alpha}}{2\tilde{\alpha}+1}};$$

for $\nu = 2(1/p - 1/2)$. This yields the assertion because $F \subset CU(B_{p,q}(\cdot))$ for a constant C.

(ii) Proof of Eq. (7).

Let $\alpha := \bar{\alpha} = \alpha_1 = \cdots = \alpha_d = \underline{\alpha}$. For m such that $m < \min\{m, m - 1 + 1/p\}$, let $\phi_\alpha(x) = \prod_{j=1}^{d} N_m(x_j - (m+1)/2)$ $(x \in \mathbb{R}^d)$.

(ii-a) *Setting of* d *=* d*=2:*

Let $V_{d,d} := \{U \in \mathbb{R}^{d \times d} \mid UU^\top = I_d\}$ be the Stiefel manifold and let $\nu_{d,d}$ be the invariant measure on the Stiefel manifold (i.e., the uniform distribution). Then, let $\phi_\alpha : \mathbb{R}^d \to \mathbb{R}$ be

$$\phi_\alpha(x) = \int \phi_\alpha(Ux) \mathrm{d}\nu_{d,d}(U) \quad (x \in \mathbb{R}^d).$$

We can see that ϕ_α is spherically symmetric and there exists $F, C > 0$ such that

$$\phi_\alpha(x) \geq F \quad (\forall x \in \mathbb{R}^d \text{ s.t. } \|x\| \leq 1),$$

and

$$\phi_\alpha(x) \begin{cases} \geq C\|x\|^{-\alpha} & (\|x\| \geq 1), \\ \leq 1 & (\|x\| \leq 1). \end{cases}$$

The last inequality can be checked by the fact that for a sufficiently large $R > 0$, the measure of the set $\nu_R(\{x \mid \|x\| = R, \phi_\alpha(x) > 0\}) \geq 1 - R^{d-\alpha-1}/R^{d-1} = 1 - R^{-\alpha}$ (here, ν_R is the uniform probability measure on the sphere $S_{d-1}(R) = \{x \in \mathbb{R}^d \mid \|x\| = R\}$) and $\|\phi_\alpha\|_{L_1} \geq 1$.

By the construction of ϕ_α and the wavelet expansion of anisotropic Besov space (9) with the norm equivalence (10), we have that there exists a constant $c > 0$ such that, for any $k \in \mathbb{N}$ and $b = \lceil \frac{1}{2} \cdot 2^k \cdot (\frac{m+1}{2} - \lfloor \frac{m+1}{2} \rfloor) \rceil (1, \ldots, 1)^\top \in \mathbb{R}^d$, it holds that

$$c \cdot 2^{k(\alpha - d/p)} \phi_\alpha(2^k(\cdot - b)) \in U(B_{p,q}([0,1]^d));$$

where $\cdot = 2^{-k(\alpha - d/p)}$. Here, let $0 < c < 1$ be a constant such that $cU(x - b') + b \in [0,1]^d$ for any $x, b' \in [0,1]^d$ and any $U \in V_{d,d}$. Then, we have that, for any $b' \in [0,1]^d$,

$$c \cdot 2^{k(\alpha - d/p)} \phi_\alpha(2^k cU(\cdot - b')) = c \cdot 2^{k(\alpha - d/p)} \phi_\alpha(2^k(\cdot - b)) \quad (cU(\cdot - b') + b) \in H_a,$$

for any $U \in V_{d,d}$. By the convex hull argument (Proposition 3), this yields that

$$R^{\mathrm{lin}}(H_a) = R^{\mathrm{lin}}(\overline{\mathrm{conv}}(H_a)) \geq R^{\mathrm{lin}}(\{c \cdot 2^{k(\alpha - d/p)} \phi_\alpha(2^k c(\cdot - b')) \mid b' \in \cdot\}).$$

Hence, it suffices to lower bound the far right-hand side of this inequality. We consider a partition A of $\cdot$, where $A \in \mathcal{A}$ has the form $A = [2^{-k} j_1, 2^{-k}(j_1+1)) \times \cdots \times [2^{-k} j_d, 2^{-k}(j_d+1))$ for $0 \leq j_i \leq 2^k - 1$ $(i = 1, \ldots, d)$. Let $\hat{J}(k) = \{(j_1, \ldots, j_d) \mid 0 \leq j_i \leq 2^k - 1\}$ and $A_j = [2^{-k} j_1, 2^{-k}(j_1+1)) \times \cdots \times [2^{-k} j_d, 2^{-k}(j_d+1)) \in \mathcal{A}$ for $j \in \hat{J}(k)$. Let $\psi_{A_j} = c \cdot 2^{k(\alpha - d/p)} \phi_\alpha(2^k c(\cdot - b_{A_j}))$, where $b_{A_j} = (2^{-k}(j_1 + 1/2), \ldots, 2^{-k}(j_d + 1/2))^\top$ for $j \in \hat{J}(k)$. We can see that $|\mathcal{A}| = 2^{dk}$. Hence, $\mathcal{A}$ satisfies Condition A with $K = dk$ if 2^k is in polynomial order with respect to n.

Moreover, there exists $F > 0$ such that $\psi_A(x) \geq F$ for all $x \in \mathcal{A}$. Next, we evaluate $\frac{1}{n}\sum_{i=1}^{n} \psi_A(x_i)^2$. On the event E, there exists C' such that $\{\{i \in [n] \mid x_i \in A'\} \geq C' n / 2^K = C' n \cdot \cdot(A')\}$ for all $A' \in \mathcal{A}$. Here, let

$$\psi_A(x) := \begin{cases} cCk2^k c(x - b_A)\|^{-\alpha} & (k2^k c(x - b_A)\| \geq 1), \\ c & (\text{otherwise}), \end{cases}$$

then $\psi_A(x) \le \psi_A(x)$. Thus, we can upper bound $\frac{1}{n}\sum_{i=1}^n \psi_A(x_i)^2$ as

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n \psi_A(x_i)^2 &\le \frac{1}{n}\sum_{i=1}^n \psi_A(x_i)^2 = \frac{1}{n}\sum_{A' \in \mathcal{A}}\sum_{x_i \in A'} \psi_A(x_i)^2 \le \frac{1}{n}\sum_{A' \in \mathcal{A}} C' \frac{n}{2^k} \max_{x \in A'} \psi_A(x)^2 \\
&= C' \sum_{A' \in \mathcal{A}} (A') \max_{x \in A'} \psi_A(x)^2 = C' \sum_{A' \in \mathcal{A}} (A') \min_{x \in A'} \psi_A(x)^2 \frac{\max_{x \in A'} \psi_A(x)^2}{\min_{x \in A'} \psi_A(x)^2} \\
&\le C' \sum_{A' \in \mathcal{A}} (A') \min_{x \in A'} \psi_A(x)^2 \frac{\max_{x \in A'} \psi_A(x)^2}{\min_{x \in A'} \psi_A(x)^2} \\
&\le C' \sum_{A' \in \mathcal{A}} (A') \min_{x \in A'} \psi_A(x)^2 \max_{x : k2^k c(x - b_A)k^{\infty} \ge 1} \frac{k2^k c(x - b_A)k^{\infty}}{(k2^k c(x - b_A)k + ck1k)^{\infty}} \\
&\le C' \sum_{A' \in \mathcal{A}} (A') \min_{x \in A'} \psi_A(x)^2 (1 + c\sqrt{d})^{2\sigma} \\
&\le C'(1 + c\sqrt{d})^{2\sigma} \int \psi_A(x)^2 dx.
\end{aligned}$$

The quantity $\int \psi_A(x)^2 dx$ on the right-hand side can be evaluated as

$$\begin{aligned}
\int \psi_A(x)^2 dx &\le \int_{x : kx - b_A k \le 2\sqrt{d}} \psi_A(x)^2 dx \\
&\le \int_{x : kx - b_A k \le c \cdot 1 2^{-k}} \psi_A(x)^2 dx + \int_{x : c \cdot 1 2^{-k} < kx - b_A k \le 2\sqrt{d}} \psi_A(x)^2 dx \\
&\le 2^{-kd} + Cc^{2\sigma} 2^{2k\sigma} \int_{c \cdot 1 2^{-k} \le r \le 2\sqrt{d}} r^{2\sigma} r^{d-1} dr \\
&\le 2^{-kd} + 2^{-2k\sigma} \max\{2^{k(2\sigma - d)}; 1\} \\
&\le \max\{2^{-kd}; 2^{-2k\sigma}\}.
\end{aligned}$$

Therefore, we have that, for a constant C'', on the event E, we have that

$$\frac{1}{n}\sum_{i=1}^n \psi_A(x_i)^2 \le C''(2^{-kd} \vee 2^{-2k\sigma}).$$

Let $\mathcal{F} = \{\psi_A \mid A \in \mathcal{A}\}$, then $\mathcal{F}$ satisfies Condition B. When $\sigma \le d/2$, by choosing k so that $c2^k \asymp n^{\frac{1}{2(\sigma + d - \sigma/p)}}$ and $K = K' = dk$, then Eq. (17) yields

$$R^{\mathrm{lin}}(\mathcal{F}) \asymp n^{-\frac{2(\sigma - \sigma/p + d/2)}{2(\sigma - \sigma/p + d/2) + d}}.$$

This concludes the proof.

(ii-b) *Setting of* $\sigma < d/2$*:*

Let $\mathcal{A}$ be the partition of $\mathcal{X}$ as defined in the proof for $\sigma \ge d/2$, i.e., $|\mathcal{A}| = 2^{dk}$ and each $A \in \mathcal{A}$ can be written as $A = [2^{-k} j_1; 2^{-k}(j_1 + 1)) \times \cdots \times [2^{-k} j_d; 2^{-k}(j_d + 1))$ for $0 \le j_i \le 2^k - 1$ $(i = 1, \ldots, d)$. Pick up $A \in \mathcal{A}$ and let $j \in \hat{J}(k)$ be the index such that $A = A_j$. For a while, we fix A and let $b = b_{A_j} = (2^{-k}(j_1 + 1/2), \ldots, 2^{-k}(j_d + 1/2))^\top$ accordingly. For $\theta = (w, b) \in \mathbb{R}^d \times \mathbb{R}$, let $\Lambda_A : \mathbb{R}^d \to \mathbb{R}^\sigma$ be

$$\Lambda_A(x) := 2^k [x_1 - b_1, \ldots, x_{\sigma-1} - b_{\sigma-1}, w^\top (x_{\sigma:d} - b_{\sigma:d}) + b],$$

and consider

$$\Psi(x) := \Psi_\sigma(\Lambda_A(x)).$$

We take its convex hull with respect to θ. We note that

$$\Psi(x) = \left(\prod_{j=1}^{\sigma-1} N_m(2^k(x_j - b_j)) \, (m+1)/2 \right) N_m\left(2^k [w^\top (x_{\sigma:d} - b_{\sigma:d}) + b] \, (m+1)/2\right).$$

To analyze its convex hull, it suffices to consider the convex hull of the last term $N_m\left(2^k[w^\top(x_{d:d} \quad b_{d:d}) + b] \quad (m+1)/2\right)$. Hence, we set $(\,) := N_m(\quad (m+1)/2)$ and consider a set of functions

$$F_{C;}^{(\,)} := \{x \in \mathrm{R}^{d \quad d+1} \mapsto a \ (\ (w^\top x + b))) \mid |a| \quad 2C;\ \|kwk\| \quad 1;\ |jbj| \quad 2 \ (a; b \in \mathrm{R};\ w \in \mathrm{R}^{d \quad d+1})\}$$

for $C > 0, \quad > 0$. We also define the Fourier transform of $\quad$ as $\hat{}(!) := (2\ )^{\ 1} \int e^{\ i!\top x} \ (x)\mathrm{d}x$ ($! \in \mathrm{R}$). Then, by Lemma 5, we have that, for $h = 2^{\ k}$ and $\quad = h^{\ 1}$,

$$\inf_{g \in \overline{\mathrm{conv}}(F_{C;}^{(\,)})} \sup_{x \in [0;1]^d} \left| g(x) \quad \exp\left(\quad \frac{\|x\| \quad ck^2}{2h^2} \right) \right|$$
$$\frac{4}{|j2 \quad \hat{}(1)j} \left[C_{d \quad d+1} R^{2(d \quad d \quad 1)} \exp(\quad R^2/2) + \exp(\quad R) \right];$$

where $C = \frac{}{|j\hat{}(1)j|} = \quad (h^{\ 1} \quad)$ and $R = h \quad (2\sqrt{d} + 1)$. This indicates that, for a fixed $A \in \mathcal{A}$, the convex hull of the set $\{a \quad j \quad = (w; b) \in \mathrm{R}^{d \quad d+1} \quad \mathrm{R};\ \|kwk\| \quad 1;\ |jbj| \quad 2;\ |aj| \quad \}$ where $\quad = 2^{\ k(\quad d/p)}$ contains $'_A$ which satisfies

$$\left\| '_A \quad (2C)^{\ 1} \left(\prod_{j=1}^{d \quad 1} N_m(2^k(x_j \quad b_j) \quad (m+1)/2) \right) \exp\left(\quad \frac{\|kx_{d:d} \quad b_{d:d}k^2}{2h^2} \right) \right\|_1$$
$$= O\left(\quad 2^{\ k(1+\ )}(h^{\ (2(d \quad d \quad 1))} \exp(\quad h^{\ 2}/2) + \exp(\quad h^{\ \ })) \right):$$

We can see that on the event E, it holds that

$$\frac{1}{n}\sum_{i=1}^{n} '^2_A(x_i) \ . \ (A)(\ 2^{\ k(1+\ )})^2 \ . \ 2^{\ kd} 2^{\ 2k(\quad d/p+1)} 2^{\ 2k} = 2^{\ 2k(\quad d/p+1+d/2) \ 2k};$$

for all $A \in \mathcal{A}$. Let $F = \{'_A \mid A \in \mathcal{A}\}$, then F satisfies Condition B. Note that, by the definition of $F_{C;}^{(\,)}$ is holds that $'_A \in \overline{\mathrm{conv}}(H_{a\ })$ for all $A \in \mathcal{A}$. Thus

$$R^{\mathrm{lin}}(H_{a\ }) = R^{\mathrm{lin}}(\overline{\mathrm{conv}}(H_{a\ })) \quad R^{\mathrm{lin}}(F\ ):$$

Therefore, by choosing k such that $2^k \ ' \ n^{\frac{1}{2(\quad d/p+1+d/2)+d+2}}$, and setting $K = K^0 = dk$, then Eq. (17) gives

$$R^{\mathrm{lin}}(F\ ) \ \& \ n^{\ \frac{2(\quad d/p+1+\quad +d/2)}{2(\quad d/p+1+\quad +d/2)+d}}:$$

□

Lemma 5 (Suzuki & Akiyama (2021)). *Let $h > 0$ and $R := h \ /(2\sqrt{d}+1)$. Then, for $C = \frac{}{|j\hat{}(1)j|}$, the Gaussian RBF kernel can be approximated by*

$$\inf_{g \in \overline{\mathrm{conv}}(F_{C;}^{(\,)})} \sup_{x \in [0;1]^d} \left| g(x) \quad \exp\left(\quad \frac{\|x\| \quad ck^2}{2h^2} \right) \right|$$
$$\frac{4}{|j2 \quad \hat{}(1)j} \left[C_d R^{2(d \quad 2)} \exp(\quad R^2/2) + \exp(\quad R) \right]$$

for any $c \in [0;1]^d$, where C_d is a constant depending only on d. In particular, the right hand side is $O(\exp(\quad n\ ))$ if $R = n$.

G Auxiliary lemmas

The following proposition which were shown in Schmidt-Hieber (2018); Hayakawa & Suzuki (2019); Suzuki (2018) is convenient to show the estimation error rate.

Proposition 4 (Schmidt-Hieber (2018); Hayakawa & Suzuki (2019)). *Let F be a set of functions. Let $\widehat{f}$ be the least-squares estimator in F:*

$$\widehat{f} = \operatorname*{argmin}_{f \in F} \sum_{i=1}^{n} (y_i \quad f(x_i))^2:$$

Assume that $kf^{o}k_1 \quad F$ and all $f \in F$ satisfies $kfk_1 \quad F$ for some $F \quad 1$. If > 0 satisfies $N(; F; k \quad k_1) \quad 3$, then it holds that

$$\mathrm{E}_{D_n}[k\widehat{f} \quad f^{o}k^2_{L^2(P_X)}] \quad C\left[\inf_{f \in F} kf \quad f^{o}k^2_{L^2(P_X)} + (F^2 + \ ^2)\frac{\log N(; F; k \quad k_1)}{n} + \ (F + \ )\right];$$

where C is a universal constant.

The following lemma provides the covering number of the deep neural network model.

Lemma 6 (Covering number evaluation). *The covering number of $(L; W; S; B)$ can be bounded by*

$$\log N(\ ; \ (L; W; S; B); k \quad k_1) \quad S\log(\ ^1L(B \_ 1)^{L \ 1}(W + 1)^{2L})$$
$$2SL\log((B \_ 1)(W + 1)) + S\log(\ ^1L):$$

Proof of Lemma 6. Given a network $f \in (L; W; S; B)$ expressed as

$$f(x) = (W^{(L)} \ ( ) + b^{(L)}) \quad (W^{(1)}x + b^{(1)});$$

let

$$A_k(f)(x) = \quad (W^{(k \ 1)} \ ( ) + b^{(k \ 1)}) \quad (W^{(1)}x + b^{(1)});$$

and

$$B_k(f)(x) = (W^{(L)} \ ( ) + b^{(L)}) \quad (W^{(k)} \ (x) + b^{(k)});$$

for $k = 2; : : : ; L$. Corresponding to the last and first layers, we define $B_{L+1}(f)(x) = x$ and $A_1(f)(x) = x$ respectively. Then, it is easy to see that $f(x) = B_{k+1}(f) \ (W^{(k)} + b^{(k)}) \ A_k(f)(x)$. Now, suppose that a pair of different two networks $f; g \in (L; W; S; B)$ given by

$$f(x) = (W^{(L)} \ ( ) + b^{(L)}) \quad (W^{(1)}x + b^{(1)}); \ g(x) = (W^{(L)^0} \ ( ) + b^{(L)^0}) \quad (W^{(1)^0}x + b^{(1)^0});$$

has parameters with distance $: kW^{(`)} \quad W^{(`)^0}k_1$ and $kb^{(`)} \quad b^{(`)^0}k_1$. Now, not that $kA_k(f)k_1 \quad \max_j kW^{(k \ 1)}_{j;:}k_1 kA_{k \ 1}(f)k_1 + kb^{(k \ 1)}k_1 \quad WBkA_{k \ 1}(f)k_1 + B \quad (B \_ 1)(W + 1)kA_{k \ 1}(f)k_1 \quad (B \_ 1)^{k \ 1}(W + 1)^{k \ 1}$, and similarly, the Lipshitz continuity of $B_k(f)$ with respect to $k \ k_1$-norm is bounded as $(BW)^{L \ k+1}$: Then, it holds that

$$\begin{aligned}
& jf(x) \quad g(x)j \\
&= \left|\sum_{k=1}^{L} B_{k+1}(g) \quad (W^{(k)} \ +b^{(k)}) \quad A_k(f)(x) \quad B_{k+1}(g) \quad (W^{(k)^0} \ +b^{(k)^0}) \quad A_k(f)(x)\right| \\
& \sum_{k=1}^{L} (BW)^{L \ k} k(W^{(k)} \ +b^{(k)}) \quad A_k(f)(x) \quad (W^{(k)^0} \ +b^{(k)^0}) \quad A_k(f)(x)k_1 \\
& \sum_{k=1}^{L} (BW)^{L \ k} \ [W(B _ 1)^{k \ 1}(W + 1)^{k \ 1} + 1] \\
& \sum_{k=1}^{L} (BW)^{L \ k} \ (B _ 1)^{k \ 1}(W + 1)^{k} \quad L(B _ 1)^{L \ 1}(W + 1)^{L}:
\end{aligned}$$

Thus, for a fixed sparsity pattern (the locations of non-zero parameters), the covering number is bounded by $(\ =[L(B \_ 1)^{L \ 1}(W + 1)^{L}])^{S}$. There are the number of configurations of the sparsity pattern is bounded by $\binom{(W+1)^{L}}{S} \quad (W + 1)^{LS}$. Thus, the covering number of the whole space is bounded as

$$(W + 1)^{LS}\{ \ =[L(B \_ 1)^{L \ 1}(W + 1)^{L}]\}^{S} = [\ ^1L(B \_ 1)^{L \ 1}(W + 1)^{2L}]^{S};$$

which yields the assertion.

□