On the minimax optimality and superiority of deep neural network learning over sparse parameter spaces

Satoshi Hayakawa [a], Taiji Suzuki [a,b,*]

[a] *Department of Mathematical Informatics, Graduate School of Information Science and Technology, The University of Tokyo, Japan*
[b] *Center for Advanced Intelligence Project, RIKEN, Japan*

A R T I C L E   I N F O

A B S T R A C T

Deep learning has been applied to various tasks in the field of machine learning and has shown superiority to other common procedures such as kernel methods. To provide a better theoretical understanding of the reasons for its success, we discuss the performance of deep learning and other methods on a nonparametric regression problem with a Gaussian noise. Whereas existing theoretical studies of deep learning have been based mainly on mathematical theories of well-known function classes such as Hölder and Besov classes, we focus on function classes with discontinuity and sparsity, which are those naturally assumed in practice. To highlight the effectiveness of deep learning, we compare deep learning with a class of linear estimators representative of a class of shallow estimators. It is shown that the minimax risk of a linear estimator on the convex hull of a target function class does not differ from that of the original target function class. This results in the suboptimality of linear methods over a simple but non-convex function class, on which deep learning can attain nearly the minimax-optimal rate. In addition to this extreme case, we consider function classes with sparse wavelet coefficients. On these function classes, deep learning also attains the minimax rate up to log factors of the sample size, and linear methods are still suboptimal if the assumed sparsity is strong. We also point out that the parameter sharing of deep neural networks can remarkably reduce the complexity of the model in our setting.

1. Introduction

Deep learning has been successfully applied to a number of machine learning problems, including image analysis and speech recognition (Goodfellow, Bengio, Courville, & Bengio, 2016; Schmidhuber, 2015). However, the rapid expansion of its applications has preceded a thorough theoretical understanding, and thus the theoretical properties of neural networks and their learning have not yet been fully understood. This paper aims to summarize recent developments in theoretical analyses of deep learning and to provide new approximation and estimation error bounds that theoretically confirm the superiority of deep learning to other representative methods.

In this section, we present an overview of the paper; here, we prioritize understandability over strict mathematical rigor. Some formal definitions and restrictions, such as for measurability and integrability, are presented in later sections.

1.1. Nonparametric regression

Throughout this paper, our intent is to demonstrate the superiority of the deep learning approach to other methods. To do so, we consider a simple nonparametric regression problem and compare the performance of various approaches in that setting. The nonparametric regression problem we analyze is formulated as follows:

We observe n i.i.d. input–output pairs $(X_i, Y_i) \in [0, 1]^d \times \mathbb{R}$ generated by the model

$$Y_i = f^\circ(X_i) + \xi_i, \quad i = 1, \ldots, n,$$

where ξ_i is an i.i.d. noise independent of inputs. The object is to estimate f° from the observed data.

This problem setting has been commonly used in statistical learning theory and is not limited to deep learning (Tsybakov, 2008; Yang & Barron, 1999; Zhang, Wong, & Zheng, 2002). In this paper, we assume the noise follows a Gaussian distribution.

In this scenario, a neural network (architecture) is treated as a set of functions $\mathcal{F} \subset \{f : [0, 1]^d \to \mathbb{R}\}$. Other estimation methods such as kernel ridge regression and wavelet

* Corresponding author.
E-mail addresses: satoshi_hayakawa@mist.i.u-tokyo.ac.jp (S. Hayakawa), taiji@mist.i.u-tokyo.ac.jp (T. Suzuki).

threshold estimators are also regarded as such mappings (Bishop, 2006; Donoho & Johnstone, 1994). In this paper, we evaluate the performance of estimators by the expected mean squared error $\mathrm{E}\left[\|\widehat{f} - f^\circ\|_{L^2}^2\right]$ (we call this quantity the "estimation error" for simplicity) dependent on n following convention (Schmidt-Hieber, 2017; Suzuki, 2019; Wang, Paterlini, Gao, & Yang, 2014), where the expectation is taken with respect to the training data. Usually, the $L^2(\mathrm{P}_X)$ norm (where P_X is the distribution of X_i) has been used in existing studies instead of the Lebesgue L^2 norm, but in evaluation the upper- (and lower-) boundedness of the density is typically assumed, so for simplicity, we treat X_i as uniformly distributed.

If we fix the set of true functions $\mathcal{F}^\circ$ (called a hypothesis space),

$$\sup_{f^\circ \in \mathcal{F}^\circ} \mathrm{E}\left[\|\widehat{f} - f^\circ\|_{L^2}^2\right]$$

is the worst-case performance of the estimator $(X_i, Y_i)_{i=1}^n \mapsto \widehat{f}$. We are interested in its asymptotic convergence rate with respect to n, the sample size. The minimax rate is determined by the convergence rate of

$$\inf_{(X_i, Y_i) \mapsto \widehat{f}} \sup_{f^\circ \in \mathcal{F}^\circ} \mathrm{E}\left[\|\widehat{f} - f^\circ\|_{L^2}^2\right],$$

where inf is taken over all possible estimators. We compare this to the convergence rate of fixed (with respect to n) sequences of estimators determined by some learning procedure such as deep learning to evaluate how efficient the estimation method is.

As a competitor of deep learning, a class of "linear estimators" is considered. Here, we say an estimator is linear if it depends linearly on the outputs Y_i; it is expressed as

$$\widehat{f}(x) = \sum_{i=1}^n Y_i \varphi_i(x; X_1, \ldots, X_n).$$

This estimator class includes several practical estimators such as kernel ridge regression and the Nadaraya–Watson estimator. The minimax rate in the class of linear estimators can be slower under some settings; e.g.,

$$\inf_{(X_i, Y_i) \mapsto \widehat{f}} \sup_{f^\circ \in \mathcal{F}^\circ} \mathrm{E}\left[\|\widehat{f} - f^\circ\|_{L^2}^2\right] \leq n^{-\gamma} \inf_{\widehat{f} : \text{linear}} \sup_{f^\circ \in \mathcal{F}^\circ} \mathrm{E}\left[\|\widehat{f} - f^\circ\|_{L^2}^2\right]$$

holds for some $\gamma > 0$ (for most cases, we consider only the polynomial order). Such situations were reported earlier by several authors (Donoho & Johnstone, 1998; Korostelev & Tsybakov, 1993; Zhang et al., 2002). In terms of deep learning analysis, a comparison of deep learning with linear methods has been performed by Imaizumi and Fukumizu (2019). The present paper also shows the suboptimality of linear methods (Table 1) for sparse function classes, which we define later.

Our main contribution here is that we find a quite simple and natural function class $\mathcal{I}_\Phi^0$ for which deep learning attains nearly the optimal rate, whereas linear methods are not able to converge faster than the suboptimal rate $O(n^{-1/2})$. In the next subsection, we explain how to treat and analyze deep learning in the context of statistical learning theory.

1.2. Related work on estimation of deep neural networks

Deep neural networks have a structure of alternating linear (or affine) transformations and nonlinear transformations; i.e., in one layer x is transformed to $\rho(Wx - v)$, where W is a matrix and v is a vector and ρ is a nonlinear function called an activation function. It is known that the repeated operation of this transformation gives a nice approximation of a wide class of nonlinear functions.

Traditionally, sigmoidal functions have been commonly used as activation functions:

$$\sigma : \mathbb{R} \to \mathbb{R}, \quad \text{with} \quad \lim_{t \to \infty} \sigma(t) = 1, \ \lim_{t \to -\infty} \sigma(t) = 0.$$

It is known that the set of functions realized by shallow networks with continuous sigmoidal activation is dense in any L^p space unless the number of parameters is not limited (Cybenko, 1989). However, a similar result has also been shown for non-sigmoidal activation cases (Sonoda & Murata, 2017). In particular, the Rectified Linear Unit (ReLU) activation function $\rho(x) = \max\{x, 0\}$ has shown practical performance (Glorot, Bordes, & Bengio, 2011) and is now widely used.

Basically, deep learning trains a network by minimizing the empirical risk with some regularization:

$$\text{minimize} \ \frac{1}{n} \sum_{i=1}^n (f(X_i) - Y_i)^2 + \lambda(f) \quad \text{subject to } f \in \mathcal{F},$$

where $\lambda(f)$ is the regularization term and $\mathcal{F}$ is the set of functions that are realizations of a specific neural network architecture. This optimization is usually carried out by Stochastic Gradient Descent (SGD) or a variant of it, and the output is not necessarily the global minimum (Goodfellow et al., 2016). In the present paper, however, we do not treat this optimization aspect, and we assume an ideal optimization.

The number of parameters in deep learning tends to be much larger than the sample size, and hence without any regularization, deep models can overfit the training sample. To overcome this issue, existing studies have utilized sparse regularization to obtain networks with a small number of nonzero parameters. This enables us to obtain a tight estimate of the error bounds using the result of approximation error analysis (see Section 2.2).

Yarotsky (2017) reported the effectiveness of ReLU activation in terms of approximation ability, and the result has been exploited in estimation theory for deep learning (Schmidt-Hieber, 2017; Suzuki, 2019). Their target function classes are Hölder space C^s and Besov space $B_{p,q}^s$, which are compatible with functional analysis or the theory of differential equations. Schmidt-Hieber (2017) also pointed out that deep learning is superior to linear estimators when the target function is of the form $f^\circ(x) = g(w^\top x)$, with g having some Hölder smoothness. In addition, Imaizumi and Fukumizu (2019) treated estimation theory for piecewise smooth functions using the approximation theory described in Petersen and Voigtlaender (2018). This paper investigates new target classes to demonstrate the effectiveness of deep learning with ReLU activation.

1.3. Contribution of this paper

The situations where deep learning shows superiority to linear estimators have been studied with particular theories:

- $f^\circ(x) = g(w^\top x)$ with g being smooth (Schmidt-Hieber, 2017),
- piece-wise smooth functions (Imaizumi & Fukumizu, 2019),
- Besov spaces in a certain range of parameters (Suzuki, 2019).

These situations can be understood mathematically through the non-convexity of models. To highlight this property, we introduce a sparse function class (and therefore non-convex) and investigate the generalization ability of deep learning and linear estimators over the space to see how sparsity (non-convexity) affects the estimation error. We indeed show that linear estimators have a critical disadvantage when the target class is non-convex (Section 3). In contrast, deep learning is shown to have the optimality over the sparse function classes we define (Section 5). These classes are natural in the sense we describe in the next paragraph.

The major difference between existing studies and this work is that we assume an explicit sparsity of target classes, which are defined parametrically. This kind of scenario seems to occur

Table 1
Estimated error bound: deep learning vs. linear methods.

Target class	$B^s_{p,q}([0,1])$ $(p < 2)$	$\mathcal{I}^0_\Phi$	$\mathcal{K}^p_\psi$ $(p < 1)$
Deep learning	$\widetilde{O}(n^{-\frac{2s}{2s+1}})$	$\widetilde{O}(n^{-1})$ (Section 5.3)	$\widetilde{O}(n^{-\frac{2\alpha}{2\alpha+1}})$ $(\alpha = \frac{1}{p} - \frac{1}{2})$ (Section 5.4)
Reference	Suzuki (2019)		This work
Linear methods	$\Omega(n^{-\frac{2\gamma}{2\gamma+1}})$ $(\gamma = s + \frac{1}{2} - \frac{1}{p})$		$\Omega(n^{-\frac{1}{2}})$ (Sections 3.2, 5.3, 5.4)
Reference	Donoho and Johnstone (1998) and Zhang et al. (2002)		This work

Note. More precisely, the actual target function classes are the unit balls of the classes as written. $\widetilde{O}$ means O up to poly-log factors, and for each class shown in this table, deep learning attains the minimax-optimal rate in the sense of $\widetilde{O}$. $B^s_{p,q}$ denotes the Besov space with parameters (s, p, q), and the parameters are additionally required to satisfy $s > 1/p$ and $p, q \geq 1$ or else $s = p = q = 1$.

Table 2
Correspondence between mathematical features and the real-world things.

| Feature	Real-world counterpart	Examples
Nonparametric regression	Real-world estimation problems	Image denoising, classification problem, style transfer
Basis function	Localized pattern	Pronunciation of vowels, Painting style, fractal structure in nature
Sparse combination of basis functions	Real data	Speech of one person, a painting, a noisy picture
Sparsity	Low dimensionality of actual information	Natural language, handwriting alphabets
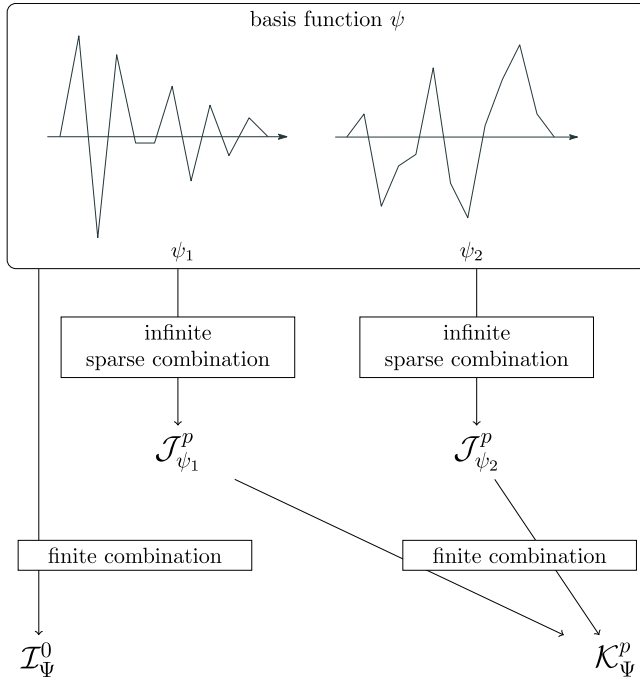
Fig. 1. Interpretation of function classes.

in practice; for example, speech data for a specific person are supposed to be a sparse linear combination of the person's pronunciation of each letter, and paintings by a specific painter may be regarded as combinations of patterns (see Table 2). Indeed, when we carry out the wavelet expansion to a natural image, its relatively few large coefficients can well reconstruct the original image, as is exploited in the field of compressive sensing (Candès & Wakin, 2008).

We reflect this property to define new function classes (Section 4). In particular, we introduce a function class $\mathcal{I}^0_\Phi$ and $\mathcal{K}^p_\psi$ (and $\mathcal{J}^p_\psi$ as well) with a parameter $p > 0$ controlling the "sparsity" of the function class. Fig. 1 gives an interpretation of the motivations and definitions of these function classes. More precisely, p controls the sparsity of coefficients of "infinite sparse combination" appearing in Fig. 1 through a criterion called "weak

ℓ^p norm" (see Definition 4.4 for details). Here, "combination" roughly means linear combination, but precisely includes affine transform in the input; i.e., the combination of $f_1, \ldots, f_m$ can be generally expressed as $f = \sum_{i=1}^m c_i f_i(A_i \cdot - b_i)$.

On the basis of our new function classes, we show that deep learning is superior to linear methods and actually attains nearly the minimax-optimal rate over each target class (Table 1). As an extreme case (corresponding to the sparsity level $p = 0$), we treat the class of piecewise constant functions, for which the convergence rate of the linear estimators is $\Omega(n^{-1/2})$, whereas deep learning attains the near-minimax rate $\widetilde{O}(n^{-1})$. This quite simply demonstrates the scenario described in Imaizumi and Fukumizu (2019). For $0 < p < 1$, we also show that deep learning attains the nearly minimax-optimal rate $\widetilde{O}(n^{-\frac{2\alpha}{2\alpha+1}})$ in estimating a function in $\mathcal{K}^p_\psi$, where $\alpha = 1/p - 1/2$. Here, we have $\frac{2\alpha}{2\alpha+1} > \frac{1}{2}$, and the difference between deep and linear becomes larger as p becomes smaller (i.e., as the sparsity becomes stronger). K^p_ψ has another parameter, β, which controls the rate of decay of coefficients of the function class. Surprisingly, we even find that the minimax rate of linear estimators can become arbitrarily slow under the same sparsity p (and the same order of covering entropy) if the value of β is varied. Although we do not yet have the upper bound for the convergence rate of deep learning over the range of parameter values producing this situation (see Theorem 4.15 and Remark 4.16), this indicates that the difference between deep learning and linear estimators could be arbitrary large. These differences essentially arise from the non-convexity of the model. That is, as the non-convexity of the model becomes stronger, the difference becomes larger.

In addition, we see that deep learning takes advantage of wavelet expansions with sparsity because a neural network can efficiently approximate functions of the form $\sum_i c_i f(A_i \cdot - b_i)$ if its subnetwork can approximate the "basis" function f precisely as is also mentioned in Bölcskei, Grohs, Kutyniok, and Petersen (2017). From this perspective, we see that parameter sharing, mentioned in Section 5.5, is also effective. It can also be said that this paper expands the approximation theory argued in Bölcskei et al. (2017) to estimation theory over sparse parameter spaces.

Thus, the contribution of this paper is summarized as follows:

- To deal with sparsity in machine learning, we define function classes $\mathcal{I}^0_\Phi$ and $\mathcal{K}^p_\psi$ with a parameter p controlling the sparsity. We also consider the nonparametric regression problem on these target classes and derive the minimax lower bounds of estimation error.

- We consider linear estimators, which are a competitor of deep learning, investigating them by evaluating their estimation error over sparse target classes. We show that linear estimators can only attain suboptimal rates on sparse and non-convex models and even become arbitrarily slow under the same sparsity with other parameters varying. This also gives a unified understanding of existing studies which describe situations where deep learning is superior to linear methods.
- To demonstrate the learning ability of the deep ReLU network on sparse spaces, we construct sparse neural networks that nearly attain minimax-optimal rates. It is also shown that parameter sharing in the construction of neural networks is effective on sparse target classes.

We give a brief overview of each section in the following.

In Section 2, we introduce general methods used in statistical learning theory, presenting our own proofs or arguments to the maximum extent possible. Section 2.2 presents an information-theoretic way to obtain a lower bound for the minimax rate. Also, the method for evaluating an estimation error by using an approximation error is given. Evaluations of linear minimax rates are given in Section 3. We prove that linear estimators cannot distinguish between a function class and its convex hull, and as a consequence linear minimax rates can be rather slower than ordinal minimax rates. Section 4 provides the definitions of our own target function classes. The ℓ^0 norm and the $w\ell^p$ quasi-norm of coefficients in linear combinations are introduced as indicators of sparsity. The minimax lower bounds for the defined classes are also given (which are revealed to be optimal up to log factors in the section that follows). In Section 5, we show that deep learning attains the nearly minimax rate for defined function classes. In addition, we propose that parameter sharing can be a means of reducing complexities in regularized networks. Finally, Section 6 provides a summary and presents future directions for this work.

1.4. Notation

We use the following notation throughout the paper.

- $\|\cdot\|_\infty$ and $\|\cdot\|_0$ are defined as

$$\|v\|_\infty := \max_{1 \le i \le m} |v_i|, \quad \|v\|_0 := |\{1 \le i \le m \mid v_i \ne 0\}|$$

for a vector $v = (v_1, \ldots, v_m)^\top \in \mathbb{R}^m$. They are defined similarly for real matrices.

- As a natural extension of $\|\cdot\|_0$, $\|a\|_{\ell^0}$ denotes the number of nonzero terms in the sequence $a = (a_i)_{i=1}^\infty$.
- For $p > 0$ and a real sequence $a = (a_i)_{i=1}^\infty$, the ℓ^p norm of a is defined as

$$\|a\|_{\ell^p} := \left(\sum_{i=1}^\infty |a_i|^p \right)^{1/p},$$

and ℓ^p denotes the set of all real sequences with a finite ℓ^p norm.

2. General theories in statistical estimation

2.1. General settings and notation

Let us consider the following regression model. We observe i.i.d. random variables (X_i, Y_i) generated by

$$Y_i = f^\circ(X_i) + \xi_i, \qquad i = 1, 2, \ldots, n. \tag{1}$$

Here, each ξ_i is an observation noise independent of other variables. In this paper, we use settings such that each X_i is

d-dimensional and uniformly distributed on $[0, 1]^d$, each Y_i is one-dimensional, and ξ_i's are i.i.d. centered Gaussian variables with variance σ^2 ($\sigma > 0$). For simplicity, we sometimes use the notation $X^n := (X_1, \ldots, X_n)$, $Y^n := (Y_1, \ldots, Y_n)$, and $Z^n := (X_i, Y_i)_{i=1}^n$.

Remark 2.1. In the following, we often write only $\widehat{f}$ to indicate an estimator where we should write $(X_i, Y_i)^n \mapsto \widehat{f}$ (this mapping is supposed to be measurable). For example, $\inf_{(X_i, Y_i)_{i=1}^n \mapsto \widehat{f} \in \mathcal{F}}$ is simply denoted by $\inf_{\widehat{f} \in \mathcal{F}}$. In addition, for the case $\mathcal{F} = L^2([0, 1]^d)$, we omit $\mathcal{F}$ and simply write $\inf_{\widehat{f}}$.

To evaluate the quality of estimators, we need to adopt some evaluation criteria. For a fixed f° and a function $f \in L^2([0, 1]^d)$, we have

$$\begin{aligned} E[(f(X) - Y)^2] &= E[(f(X) - f^\circ(X))^2] \\ &\quad - 2E[\xi(f(X) - f^\circ(X))] + E[\xi^2] \\ &= E[(f(X) - f^\circ(X))^2] + \sigma^2 \\ &= \|f - f^\circ\|_{L^2}^2 + \sigma^2. \end{aligned}$$

This implies that the magnitude of the expected error $E[(f(X_i) - Y_i)^2]$ depends only on that of the L^2 distance $\|f - f^\circ\|_{L^2}^2$. This leads to the following definition for a performance criterion.

Definition 2.2. The L^2 *risk* for an estimator $\widehat{f}$ is defined as

$$R(\widehat{f}, f^\circ) := E\left[\|\widehat{f} - f^\circ\|_{L^2}^2 \right].$$

For a model $\mathcal{F}^\circ \subset L^2([0, 1]^d)$, the *minimax L^2 risk* over $\mathcal{F}^\circ$ is defined as

$$\inf_{\widehat{f}} \sup_{f^\circ \in \mathcal{F}^\circ} R(\widehat{f}, f^\circ) = \inf_{\widehat{f}} \sup_{f^\circ \in \mathcal{F}^\circ} E\left[\|\widehat{f} - f^\circ\|_{L^2}^2 \right],$$

where $\widehat{f}$ runs over all estimators (measurable functions).

We evaluate the quality of an estimator $\widehat{f}$ by this L^2 risk and compare it with the minimax-optimal risk.

Remark 2.3. We omit n from the notation because it is treated as a constant when we consider a single regression problem. However, as n goes to ∞, the minimax risk converges to 0, and in this paper we are interested in the convergence rate of the minimax risk.

2.2. Relationships between complexity and minimax risk

In this section, we introduce a key procedure to evaluate the minimax risk. To do so, we define complexity measures called ε-*entropy* (ε may be taken place by any positive real number), which formally represent complexities of (totally bounded) metric spaces. This kind of complexity of $\mathcal{F}^\circ$ profoundly affects the convergence rate of the minimax risk (Yang & Barron, 1999).

Definition 2.4 (*van der Vaart & Wellner, 1996; Yang & Barron, 1999*)**.** For a metric space (S, d) and $\varepsilon > 0$,

- a finite subset T is called ε-*packing* if $d(x, y) > \varepsilon$ holds for any $x, y \in T$ with $x \ne y$, and the logarithm of the maximum cardinality of an ε-packing subset is called the *packing ε-entropy* and is denoted by $M_{(S,d)}(\varepsilon)$;
- a finite set $U \subset \overline{S}$ is called ε-*covering* if for any $x \in S$ there exists $y \in U$ such that $d(x, y) \le \varepsilon$, and the logarithm of the minimum cardinality of an ε-covering set is called the *covering ε-entropy* and is denoted by $V_{(S,d)}(\varepsilon)$.

Here, $\overline{S}$ is the completion of S with respect to the metric d.

The concept of ε-entropy is useful to obtain a lower bound of the minimax risk of some function class $\mathcal{F}^\circ$. Let $\mathcal{F}^\circ \subset L^2([0,1]^d)$ be the class of true functions, equipped with the L^2 metric. For simplicity, let $V(\varepsilon) = V_{(\mathcal{F}^\circ, \|\cdot\|_{L^2})}(\varepsilon)$ and $M(\varepsilon) = M_{(\mathcal{F}^\circ, \|\cdot\|_{L^2})}(\varepsilon)$. Then, the following theorem holds.

Theorem 2.5 (*Yang & Barron, 1999, Theorem 1*). *In the Gaussian regression model, suppose there exist $\delta, \varepsilon > 0$ such that*

$$V(\varepsilon) \leq \frac{n\varepsilon^2}{2\sigma^2}, \quad M(\delta) \geq \frac{2n\varepsilon^2}{\sigma^2} + 2\log 2.$$

Then we have

$$\inf_{\widehat{f}} \sup_{f \in \mathcal{F}^\circ} P_f\left(\|\widehat{f} - f\|_{L^2} \geq \frac{\delta}{2} \right) \geq \frac{1}{2}, \quad \inf_{\widehat{f}} \sup_{f \in \mathcal{F}^\circ} E_f\left[\|\widehat{f} - f\|_{L^2}^2 \right] \geq \frac{\delta^2}{8},$$

where P_f is the probability law with $f^\circ = f$, and E_f is the expectation determined by P_f.

The above theorem states the relationship between the complexity of *the function class* and the *lower* bound of the minimax risk. On the other hand, the following theorem relates the complexity of *the estimator* and the *upper* bound of the generalization error (and the minimax risk, at the same time). The following is also useful for evaluating the convergence rate of the empirical risk minimizer of some explicit model, such as neural networks.

Theorem 2.6 (*Schmidt-Hieber, 2017, Lemma 4*). *In the Gaussian regression model (1), let $\widehat{f}$ be the empirical risk minimizer, taking values in $\mathcal{F} \subset L^2([0,1]^d)$. Suppose every element $f \in \mathcal{F}$ satisfies $\|f\|_{L^\infty} \leq F$ for some fixed $F > 0$. Then, for an arbitrary $\delta > 0$, if $V_{(\mathcal{F}, \|\cdot\|_{L^\infty})}(\delta) \geq 1$, then*

$$R(\widehat{f}, f^\circ) \leq 4 \inf_{f \in \mathcal{F}} \|f - f^\circ\|_{L^2}^2 + C\left(\frac{(F^2 + \sigma^2)V_{(\mathcal{F}, \|\cdot\|_{L^\infty})}(\delta)}{n} + (F+\sigma)\delta \right)$$

holds, where $C > 0$ is an absolute constant.

This sort of evaluation has been obtained earlier in Giné and Koltchinskii (2006), Györfi, Kohler, Krzyzak, and Walk (2006), Koltchinskii (2006) and the proof for this is essentially the same as earlier ones. For completeness, however, we give a proof for this assertion in the appendix (Appendix B.1).[1]

3. Suboptimality of linear estimators

We consider linear estimators as a competitor to deep learning, and in this section, we characterize their suboptimality by the convexity of the target model. Linear estimators, represented by kernel methods, are classically applied to regression problems. Indeed, some linear estimators have minimax optimality over smooth function classes such as Hölder classes and Besov classes with some constraint on their parameters (with fixed design: Donoho & Johnstone, 1998; Tsybakov, 2008). However, as has been pointed out in the literature (Imaizumi & Fukumizu, 2019; Korostelev & Tsybakov, 1993), linear estimators can attain only suboptimal rates with function classes having discontinuity. We here show that the suboptimality of linear estimators arises even with a quite simple target class. Our first contribution is to point out that the concept of the convex hull gives the same explanation to such suboptimality for several target classes, and based on that argument, we then show that linear estimators perform suboptimally even on a quite simple target class.

3.1. Linear estimators and its minimax risk

Definition 3.1. The estimation scheme $(X_i, Y_i)_{i=1}^n \mapsto \widehat{f}$ is called *linear* if $\widehat{f}$ has the form

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

where we suppose $E\left[\|\varphi_i(\cdot; X^n)\|_{L^2}^2 \right] < \infty$. Also, we call an estimator $\widehat{f}$ *affine* if $\widehat{f}$ has the form

$$\widehat{f}(x) = \widehat{f}_L(x) + \varphi(x; X^n),$$

where φ has the same condition as φ_i, and $\widehat{f}_L$ is a linear estimator.

Remark 3.2. The condition $E\left[\|\varphi_i(\cdot; X^n)\|_{L^2}^2 \right] < \infty$ may be replaced by a weaker version. This actually assures that

- $\varphi_i(\cdot; X^n) \in L^2([0,1]^d)$ holds almost surely;
- $E\left[\varphi_i(x; X^n)^2 \right] < \infty$ holds almost everywhere.

The latter condition is only needed in the justification of (29).

A linear estimator is of course an affine estimator as well. Linear or affine estimators are classically used often; they include linear (ridge) regression, the Nadaraya–Watson estimator, and kernel ridge regression (Bishop, 2006; Friedman, Hastie, & Tibshirani, 2001; Tsybakov, 2008). For example, the estimator given by kernel ridge regression can be explicitly written as

$$\widehat{f}(x) := (k(x, X_1), \ldots, k(x, X_n))(K + \lambda I_n)^{-1}(Y_1, \ldots, Y_n)^\top,$$

where λ is a positive constant, $k : [0,1]^d \times [0,1]^d \to \mathbb{R}$ is a positive semi-definite kernel, and the matrix $K \in \mathbb{R}^{n \times n}$ is defined as $K := (k(X_i, X_j))_{i,j}$. We can see that the difference in performance between deep learning and linear estimators becomes large in a non-convex model, which can be explained by the following theorem. (This theorem can also be seen as a generalization of Cai and Low (2004, Theorem 5).)

Let $\mathrm{conv}(\mathcal{F}^\circ)$ denote the convex hull of $\mathcal{F}^\circ$; i.e.,

$$\mathrm{conv}(\mathcal{F}^\circ) := \left\{ \sum_{i=1}^k t_i f_i \; \middle| \; t_1, \ldots, t_k \geq 0, \right.$$
$$\left. \sum_{i=1}^k t_i = 1, \; f_1, \ldots, f_k \in \mathcal{F}^\circ, \; k \geq 1 \right\}.$$

Notice that the $\mathrm{conv}(\mathcal{F}^\circ)$ is larger than the original set $\mathcal{F}^\circ$. Let $\overline{\mathrm{conv}}(\mathcal{F}^\circ)$ be the closure of $\mathrm{conv}(\mathcal{F}^\circ)$ with respect to the L^2 metric (caller closed convex hull). The following assertion holds.

Theorem 3.3. *For affine methods, the minimax risk over $\mathcal{F}^\circ$ coincides with the minimax risk over $\overline{\mathrm{conv}}(\mathcal{F}^\circ)$; i.e., the following equality holds:*

$$\inf_{\widehat{f}: \text{affine}} \sup_{f^\circ \in \mathcal{F}^\circ} R(\widehat{f}, f^\circ) = \inf_{\widehat{f}: \text{affine}} \sup_{f^\circ \in \overline{\mathrm{conv}}(\mathcal{F}^\circ)} R(\widehat{f}, f^\circ).$$

Proof. We only prove the assertion

$$\inf_{\widehat{f}: \text{affine}} \sup_{f^\circ \in \mathcal{F}^\circ} R(\widehat{f}, f^\circ) = \inf_{\widehat{f}: \text{affine}} \sup_{f^\circ \in \mathrm{conv}(\mathcal{F}^\circ)} R(\widehat{f}, f^\circ). \tag{2}$$

For the case of $\overline{\mathrm{conv}}(\mathcal{F}^\circ)$, see the appendix (Appendix B.2).

Fix an estimator $\widehat{f}$ and let

$$\widehat{f}(x) = \varphi(x; X^n) + \sum_{i=1}^n Y_i \varphi_i(x; X^n). \tag{3}$$

---

[1] We noticed some technical flaws in an earlier version of the proof of Schmidt-Hieber (2017), so we include the proof in the appendix for completeness.

For $f^\circ, g^\circ \in \mathcal{F}^\circ$ and $t \in (0, 1)$, let $h^\circ := tf^\circ + (1 - t)g^\circ$. Then

$$
\begin{aligned}
R(\widehat{f}, h^\circ) &= \mathrm{E}\left[\int_{[0,1]^d} \left(\widehat{f}(x) - h^\circ(x)\right)^2 dx\right] \\
&= \int_{[0,1]^d} \mathrm{E}\left[\left(\widehat{f}(x) - h^\circ(x)\right)^2\right] dx
\end{aligned}
\tag{4}
$$

holds by Fubini's theorem (the integrated value is nonnegative). By the convexity of the square, we have

$$
\begin{aligned}
&\left(\widehat{f}(x) - h^\circ(x)\right)^2 \\
&= \left(\varphi(x; X^n) + \sum_{i=1}^n Y_i \varphi_i(x; X^n) - h^\circ(x)\right)^2 \\
&= \left(\varphi(x; X^n) + \sum_{i=1}^n \xi_i \varphi_i(x; X^n) + \sum_{i=1}^n h^\circ(X_i)\varphi_i(x; X^n) - h^\circ(x)\right)^2 \\
&\leq t\left(\varphi(x; X^n) + \sum_{i=1}^n \xi_i \varphi_i(x; X^n) + \sum_{i=1}^n f^\circ(X_i)\varphi_i(x; X^n) - f^\circ(x)\right)^2 \\
&\quad + (1 - t)\left(\varphi(x; X^n) + \sum_{i=1}^n \xi_i \varphi_i(x; X^n) \right. \\
&\qquad\qquad \left. + \sum_{i=1}^n g^\circ(X_i)\varphi_i(x; X^n) - g^\circ(x)\right)^2 \\
&= t\left(\widehat{f}(x)\bigg|_{Y_i = f^\circ(X_i) + \xi_i} - f^\circ(x)\right)^2 \\
&\quad + (1 - t)\left(\widehat{f}(x)\bigg|_{Y_i = g^\circ(X_i) + \xi_i} - g^\circ(x)\right)^2.
\end{aligned}
\tag{5}
$$

Here, notice that $\widehat{f}$ is dependent on whether we choose $f^\circ, g^\circ,$ or h°. Therefore, we integrate this inequality to obtain

$$
R(\widehat{f}, h^\circ) \leq tR(\widehat{f}, f^\circ) + (1 - t)R(\widehat{f}, g^\circ).
$$

This means that $R(\widehat{f}, \cdot)$ is a convex functional, and so LHS $\geq$ RHS holds in (2). Since it is clear that LHS $\leq$ RHS, the equality of (2) holds. □

Remark 3.4. Indeed, Donoho, Liu, and MacGibbon (1990) and Donoho and Johnstone (1998) pointed out that the convex hull in the above assertion can be replaced by the *quadratic hull*, which is generally larger than a convex hull in a similar setting (fixed design). However, their propositions require the assumption of fixed design and *orthosymmetricity* with some wavelet expansion. Hence, we have explicitly noted Theorem 3.3 under milder conditions.

By this theorem, we see that linear estimators hardly achieve the minimax rate in a non-convex model. This also explains the difference between deep learning and linear methods argued in Schmidt-Hieber (2017), Imaizumi and Fukumizu (2019) and Suzuki (2019) in a unified manner. In the following section, we demonstrate a simple example where linear estimators are suboptimal.

3.2. Functions of bounded total variation

Let us consider a specific function class as a simple but instructive example, a class whose convex hull becomes larger in terms of the covering entropy. In addition, the convex hull is dense in BV(C) (defined below), over which linear estimators can only attain a suboptimal rate.
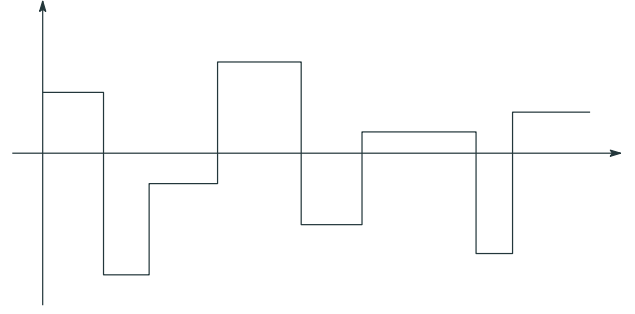

Fig. 2. An example of piece-wise constant functions.

Definition 3.5. For $k \geq 1$ and $C > 0$, define

$$
J_k(C) := \left\{ a_0 + \sum_{i=1}^k a_i \mathbf{1}_{[t_i, 1]} \,\middle|\, t_i \in (0, 1], \ |a_0| \leq C, \ \sum_{i=1}^k |a_i| \leq C \right\}
$$

as functions from $[0, 1]$ to $\mathbb{R}$ with jumps occurring at most k times.

We can also understand $J_k(C)$ as the set of piece-wise constant functions. (See Fig. 2.)

We next introduce a function class well-known in the field of real analysis, which is indeed related to J_k's (Lemma 3.8).

Definition 3.6. For any real numbers $a < b$ and a function $f : [a, b] \to \mathbb{R}$, define the *total variation* of f on $[a, b]$ as

$$
\mathrm{TV}^f([a, b]) := \sup_{M \geq 1, \ a = t_0 < \cdots < t_M = b} \sum_{i=0}^{M-1} |f(t_{i+1}) - f(t_i)|.
$$

Also, for $C > 0$, define the set of functions with bounded total variation as

$$
\mathrm{BV}(C) := \left\{ f : [0, 1] \to \mathbb{R} \,\middle|\, |f(0)| \leq C, \ \mathrm{TV}^f([0, 1]) \leq C \right\}.
$$

Remark 3.7. The condition $|f(0)| \leq C$ is needed to bound the size of the set, and it may be replaced by other similar bounding conditions such as $\sup_{t \in [0,1]} |f(t)| \leq C$ or $\int_0^1 |f(t)|\, dt \leq C$ (e.g., Donoho, 1993). These conditions are equivalent up to constant multiplications of C (i.e., $\mathrm{BV}(C) \subset \mathrm{BV}'(\alpha C) \subset \mathrm{BV}(\beta C)$ holds for some $\alpha, \beta > 0$, where BV' is a set defined with another constraint). Hence, we adopt $|f(0)| \leq C$ for simplicity of arguments.

Then, we have the following assertion. The proof is given in the appendix (Appendix B.3)

Lemma 3.8. $\overline{\mathrm{conv}}(J_k(C)) \supseteq \mathrm{BV}(C)$ *holds for each $k \geq 1$ and $C > 0$.*

From the above, it follows that linear estimators cannot distinguish $J_k(C)$ and $\mathrm{BV}(C)$ in terms of minimax convergence rates (Theorem 3.3).

Since it is known that the unit ball of $B^1_{1,1}([0, 1])$ is included in $\mathrm{BV}(C)$ for some $C > 0$ (Peetre, 1976), the following theorem can be seen as a special case of Theorem 1 in Zhang et al. (2002) (see also Table 1).

Theorem 3.9. *There exists a constant $c > 0$ dependent only on C such that*

$$
\inf_{\widehat{f}: \text{linear}} \sup_{f^\circ \in \mathrm{BV}(C)} R(\widehat{f}, f^\circ) \geq cn^{-1/2}
$$

holds.

The following corollary is one of the main results in this paper.

Corollary 3.10. *For $k = 1, 2, \ldots$, there exists a constant $c > 0$ dependent only on C such that*

$$\inf_{\widehat{f}:\text{linear}} \sup_{f^\circ \in J_k(C)} R(\widehat{f}, f^\circ) \geq cn^{-1/2}$$

holds.

Proof. The result is clear from Theorem 3.3, Lemma 3.8, and Theorem 3.9. We can of course take the same c as in Theorem 3.9. □

Remark 3.11. On the one hand, the minimax-optimal rate of the unit ball of $B_{1,1}^1([0, 1])$ is $\widetilde{\Theta}(n^{-2/3})$ (Table 1), whereas the counterpart of $J_k(C)$ is $\widetilde{\Theta}(n^{-1})$ as is attained by deep learning (proved later; see Corollary 5.8). On the other hand, the fact that the unit ball of $B_{1,1}^1([0, 1])$ is included in BV(C) implies that the linear minimax rate of BV(C) is not faster than that of $B_{1,1}^1([0, 1])$'s unit ball. Since BV(C) and $J_k(C)$ have the same linear minimax rate, $J_k(C)$ is a quite extreme example, even in comparison with $B_{1,1}^1([0, 1])$.

4. Sparse target function classes

As $J_k(C)$'s given in Section 3.2 are too simple, we consider sparse target function classes which are generalizations of $J_k(C)$'s and are also related to wavelets. We investigate the performance of deep learning and other methods over these sparse classes. The minimax lower bound for each class is also given by applying the arguments in Section 2.2. Sparsity well characterizes the spaces whose convex hulls are much larger than the original spaces, a property that is essential for the proofs that were given in Section 3.

4.1. The ℓ^0-bounded affine class

The definition of the following class is inspired by the concept of "affine class" treated in Bölcskei et al. (2017).

Definition 4.1. Given a set $\Phi \subset L^2([0, 1]^d)$ with $\|\varphi\|_{L^2} = 1$ for each $\varphi \in \Phi$ along with constants $n_s \in \mathbb{Z}_{>0}$ and $C > 0$, we define an ℓ^0-bounded affine class $\mathcal{I}_\Phi^0$ as

$$\mathcal{I}_\Phi^0(n_s, C) := \left\{ \sum_{i=1}^{n_s} c_i \varphi_i(A_i \cdot -b_i) \,\middle|\, |\det A_i|^{-1}, \ \|A_i\|_\infty, \ \|b_i\|_\infty, \right.$$

$$\left. |c_i| \leq C, \ \varphi_i \in \Phi, \ i = 1, \ldots, n_s \right\}.$$

If we adopt a jump-type function $\varphi \in \Phi$ such as $\varphi = 1_{[0,1/2)}$, we can see $\mathcal{I}_\Phi^0(n_s, C)$ as a generalization of $J_k(C)$'s defined in the previous section. The condition for c_i is also regarded as $\|c\|_{\ell^0} \leq n_s$, where the ℓ^0 norm is used as the most extreme measurement of sparsity (Raskutti, Wainwright, & Yu, 2011; Wang et al., 2014).

Let us derive a minimax lower bound for this class. Although the proof for this assertion can easily be given by applying the argument appearing in Tsybakov (2008), we provide it in the appendix (Appendix B.4).

Theorem 4.2. *There exists a constant $C_0 > 0$ depending only on σ^2 such that*

$$\inf_{\widehat{f}} \sup_{f^\circ \in \mathcal{I}_\varphi^0} R(\widehat{f}, f^\circ) \geq \frac{C_0}{n}$$

holds for each $n \geq 1$.

Remark 4.3. If the set Φ is simple enough; e.g., a finite set of piece-wise constants or piece-wise polynomials, then this rate can be almost attained by nonlinear estimators. Indeed, in such cases deep learning attains the rate $O(n^{-1}(\log n)^3)$ (Theorem 5.6). However, with a single noncontinuous $\varphi \in \Phi$, linear estimators become suboptimal; the linear minimax rate is lower-bounded by $\Omega(n^{-1/2})$ (Corollary 3.10).

4.2. The $w\ell^p$-bounded function classes

The function class we treat in the previous section seems too simple to approximate the real-world data. Therefore, we are going to consider larger function classes with "sparsity". We introduce concepts for measuring the sparsity of function classes in order to present a simple treatment of several sparse spaces. These concepts were introduced and discussed previously in Donoho (1993), Donoho (1996), and Yang and Barron (1999).

Definition 4.4. For a sequence $a = (a_i)_{i=1}^\infty \in \ell^2$, let each $|a|_{(i)}$ denote the ith largest absolute value of terms in a. For $0 < p < 2$, the *weak ℓ^p norm* of a is defined as

$$\|a\|_{w\ell^p} := \sup_{i \geq 1} i^{1/p} |a|_{(i)}. \tag{6}$$

Here, notice that $\| \cdot \|_{w\ell^p}$ is not a norm, as $(|a|_{(i)})_{i=1}^\infty$ is a permutation of $(|a_i|)_{i=1}^\infty$. However, we call it a "weak ℓ^p norm" following the notation used in Donoho (1993) and Donoho (1996).

Definition 4.5. Given an orthonormal set $\varphi = (\varphi_i)_{i=1}^\infty \subset L^2([0, 1]^d)$ and constants $C_1, C_2, \beta > 0$ and $0 < p < 2$, we define a *sparse ℓ^p-approximated set* $\mathcal{I}_\varphi^p$ as

$$\mathcal{I}_\varphi^p(C_1, C_2, \beta) := \left\{ \sum_{i=1}^\infty a_i \varphi_i \,\middle|\, \|a\|_{w\ell^p} \leq C_1, \ \sum_{i=m+1}^\infty a_i^2 \leq C_2 m^{-\beta}, \right.$$

$$\left. m = 1, 2, \ldots \right\}.$$

The constraint $\sum_{i=m+1}^\infty a_i^2 \leq Cm^{-\beta}$ is called β-*minimally tail compactness* of $(a_i)_{i=1}^\infty$, which is required to make the set compact in the L^2 metric.

Remark 4.6. To represent sparsity, the ℓ^p norm of coefficients is also used (see, e.g., Raskutti et al., 2011; Wang et al., 2014). Note here that $\|a\|_{\ell^p} \leq C$ implies $\|a\|_{w\ell^p} \leq C$. Indeed, $\|a\|_{\ell^p} \leq C$ means that for each i,

$$i^{1/p}|a|_{(i)} \leq \left(\sum_{j=1}^i |a|_{(j)}^p \right)^{1/p} \leq \left(\sum_{j=1}^\infty |a_j|^p \right)^{1/p} \leq C.$$

Thus, a weak ℓ^p ball contains an ordinary ℓ^p ball. In addition, consider the case in which $d = 1$ and φ is an orthonormal basis generated by a wavelet in $C^r([0, 1])$ with $r \in \mathbb{Z}_{>0}$ satisfying $r > \alpha := 1/p - 1/2$. Then, the Besov norm $\| \cdot \|_{B_{p,p}^\alpha}$ of a function is equivalent to the ℓ^p-norm $\| \cdot \|_{\ell^p}$ of wavelet coefficients (Donoho & Johnstone, 1998, Theorem 2). In this case, $\mathcal{I}_\varphi^p$ may be just a slight expansion of existing space, but our main interest is the case in which φ has a discontinuity (e.g., when φ is defined by the Haar wavelet), which makes things different. Furthermore, notice that Besov spaces with such parameters are omitted in Table 1 (see the note; the upper bounds are given in Suzuki (2019) for a wider range of parameters, but the range for the given lower bounds for linear estimators is limited).

Hereinafter, we fix p, C_1, C_2, and β and often write $\mathcal{I}_\varphi^p(C_1, C_2, \beta)$ as $\mathcal{I}_\varphi^p$ if there is no confusion; therefore, constants appearing in

the following may depend on these values. In the following arguments, we first derive a minimax lower bound for $\mathcal{I}_\varphi^p$, and then we introduce a broader function class that is well approximated by neural networks.

To use Theorem 2.5, we exploit the following lemma. As stated in Section 2.2, the covering entropy of the function class is important. The proof of the lemma is given in the appendix (Appendix B.5).

Lemma 4.7. *Let $\alpha := 1/p - 1/2$, and suppose β satisfies $\beta \leq 2\alpha$. Then there exists a constant $C_{\text{low}}, C_{\text{up}} > 0$ such that*

$$C_{\text{low}}\varepsilon^{-1/\alpha} \leq V_{(\mathcal{I}_\varphi^p, \|\cdot\|_{L^2})}(\varepsilon) \leq C_{\text{up}}\varepsilon^{-1/\alpha}(1 + \log(1/\varepsilon))$$

holds for each $\varepsilon > 0$.

Next, we derive a nearly tight minimax lower bound for $\mathcal{I}_\varphi^p$. In this case, "nearly" means "up to log factors".

Theorem 4.8. *There exists a constant $C = C(p, C_1, C_2) > 0$ such that*

$$\inf_{\widehat{f}} \sup_{f^\circ \in \mathcal{I}_\varphi^p} R(\widehat{f}, f^\circ) \geq Cn^{-\frac{2\alpha}{2\alpha+1}}(\log n)^{-\frac{4\alpha^2}{2\alpha+1}}$$

holds for each $n \geq 2$.

Proof. In this proof, we write the ε-entropies of $\mathcal{I}_\varphi^p$ simply as $V(\varepsilon)$ and $M(\varepsilon)$.

First, let $\varepsilon_n := c\left(\dfrac{\log n}{n}\right)^{\frac{\alpha}{2\alpha+1}}$ for some constant $c > 0$. Then by Lemma 4.7, we have

$$V(\varepsilon_n) \leq C_{\text{up}}c^{-1/\alpha}\left(\frac{\log n}{n}\right)^{-\frac{1}{2\alpha+1}}\left(1 + \frac{\alpha}{2\alpha+1}(\log n - \log\log n)\right)$$

$$\leq c_{\text{up}}c^{-1/\alpha}n^{\frac{1}{2\alpha+1}}(\log n)^{\frac{2\alpha}{2\alpha+1}},$$

where $c_{\text{up}} > 0$ is some constant independent of ε_n, and we have used $n \geq 2$. Thus we have

$$\frac{V(\varepsilon_n)}{n\varepsilon_n^2} \leq c_{\text{up}}c^{-2-1/\alpha} \leq \frac{1}{2\sigma^2} \tag{7}$$

for a sufficiently large c.

Second, notice that $M(\varepsilon) \geq V(\varepsilon)$ holds. Indeed, given a maximal ε-packing of $\mathcal{I}_\varphi^p$, the maximality implies that the set also satisfies the condition for being an ε-covering. Now, let $\delta_n := C'n^{-\frac{\alpha}{2\alpha+1}}(\log n)^{-\frac{2\alpha^2}{2\alpha+1}}$ for some constant $C' > 0$. Then we have, by Lemma 4.7,

$$M(\delta_n) \geq V(\delta_n) \geq C_{\text{low}}C'^{-1/\alpha}n^{\frac{1}{2\alpha+1}}(\log n)^{\frac{2\alpha}{2\alpha+1}}$$

$$\geq C'^{-1/\alpha}c_{\text{low}}\left(\frac{2n\varepsilon_n^2}{\sigma^2} + 2\log 2\right) \tag{8}$$

for some constant $c_{\text{low}} > 0$ independent of C', where we have used $n \geq 2$.

By (7), (8), and Theorem 2.5, for a sufficiently small C', we have

$$\inf_{\widehat{f}} \sup_{f \in \mathcal{I}_\varphi^p} \mathrm{E}\left[\|f - \widehat{f}\|_{L^2}^2\right] \geq \frac{1}{8}C'^2 n^{-\frac{2\alpha}{2\alpha+1}}(\log n)^{-\frac{4\alpha^2}{2\alpha+1}},$$

and the proof is complete. $\square$

Remark 4.9. This minimax lower bound is nearly tight, especially for the wavelet case treated in the following section. Indeed, deep learning (if necessary, with parameter sharing) achieves the rate $O(n^{\frac{2\alpha}{2\alpha+1}}(\log n)^3)$, for a broad range of wavelets (Theorem 5.10, 5.14).

4.3. Sparsity conditions for wavelet coefficients

In this subsection, we apply the argument in the previous subsection to orthogonal wavelets. They already have a broad application area from engineering and physics to pure and applied mathematics, especially in signal processing, numerical analysis, and so on (Daubechies, 1992). Because wavelets are the mathematical model of localized patterns, they are suitable for our motivation in setting function classes (see Section 1.3 and Table 2). For simplicity, we only consider the 1-dimensional case in this section, and the treatment of multi-dimensional cases is deferred to the appendix (Appendix A.2).

Definition 4.10. Let $\psi : [0, 1] \to \mathbb{R}$ be a function with $\|\psi\|_{L^2} = 1$. For such a function, we define, for integers k, ℓ,

$$\psi_{k,\ell}(x) := 2^{k/2}\psi(2^k x - \ell), \quad k \geq 0, \ 0 \leq \ell < 2^k,$$

where ψ is treated as 0 outside $[0, 1]$. Also, ψ is called an *orthogonal wavelet* if ψ satisfies

$$\int_0^1 \psi_{k,\ell}(x)\psi_{k',\ell'}(x)\,dx = 0$$

for all $(k, \ell) \neq (k', \ell')$.

Definition 4.11. Given a 1-dimensional orthonormal wavelet $\psi(x)$ and constants $C_1, C_2, \beta > 0$ and $0 < p < 2$, define

$$\mathcal{J}_\psi^p(C_1, C_2, \beta) := \left\{ \sum_{k\geq 0, \, 0\leq \ell < 2^k} a_{k,\ell}\psi_{k,\ell} \,\middle|\, \|a\|_{w\ell^p} \leq C_1, \right.$$

$$\left. \sum_{k\geq m} a_{k,\ell}^2 \leq C_2 2^{-\beta m}, \ m = 0, 1, \ldots \right\}.$$

For the definition of multi-dimensional cases, see Definition A.4.

Remark 4.12. If we consider the lexical order on (k, ℓ)'s, then $\mathcal{J}_\psi^p(C_1, C_2, \beta)$ is revealed to be β-minimally tail compact (see Definition 4.5). Thus, $\mathcal{I}_\varphi^p \subset \mathcal{J}_\psi^p$ holds with some modification of constants and they have the same degree of sparsity.

In the following, we introduce the class $\mathcal{K}_\psi^p$ as an expansion of $\mathcal{J}_\psi^p$. Though the following definition is treating the d-dimensional case, considering only the case $d = 1$ is sufficient for readers to understand the essential properties. For multi-dimensional cases, see also Appendix A.2.

Definition 4.13. Let $\Psi \subset L^2([0, 1]^d)$ consist of orthonormal wavelets. Then, for an integer $n_s > 0$ and constants $C_1, C_2, C_3, \beta > 0$ and $0 < p < 2$, define

$$\mathcal{K}_\psi^p(n_s, C_1, C_2, C_3, \beta)$$

$$:= \left\{ \sum_{j=1}^{n_s} f_j(A_j \cdot -b_j) \,\middle|\, \begin{array}{l} A_j \in \mathbb{R}^{d\times d}, \ b_j \in \mathbb{R}^d, \ |\det A_j|^{-1}, \|A_j\|_\infty, \|b_j\|_\infty \leq C_3, \\ f_j \in \mathcal{J}_{\psi_j}^p(C_1, C_2, \beta), \ \psi_j \in \Psi, \ j = 1, \ldots, n_s \end{array} \right\}.$$

Remark 4.14. By Remark 4.12 (and Remark A.5) the bound given in Theorem 4.8 is also the minimax lower bound for $\mathcal{K}_\psi^p$. Moreover, $J_k(C)$, introduced in Section 3, is included in K_ψ^p, with $k \leq n_s$ and a specific Ψ such as one containing the Haar wavelet. Thus, the bounds given in the previous sections are still applicable to this function class. Indeed, as described in Fig. 1, we define $\mathcal{K}_\psi^p$ as an expansion of $\mathcal{I}_\psi^0$ as well.

4.4. Suboptimality of linear estimators on $\mathcal{J}_\psi^p$ and $\mathcal{K}_\psi^p$

For a wavelet ψ with compact support, linear estimator shows suboptimality on $\mathcal{J}_\psi^p$ and $\mathcal{K}_\psi^p$, as is shown in Theorem 4.15. This can be regarded as a stronger version of Corollary 3.10 and shows the non-efficiency of linear estimators in sparse classes. The proof of this theorem is similar to that of Theorem 1 in Zhang et al. (2002) and is given in the appendix (Appendix B.6).

Theorem 4.15. *Let $d = 1$ and ψ be a bounded and compactly supported wavelet. For any constants $C_1, C_2, \beta > 0$ and $0 < p < 2$, there exists a constant C dependent only on C_1, C_2, and β such that*

$$\inf_{\widehat{f}: \text{linear}} \sup_{f \in \mathcal{J}_\psi^p(C_1, C_2, \beta)} R(\widehat{f}, f^\circ) \geq Cn^{-\frac{\beta}{1+\beta}}$$

holds for each $n \geq 1$. This bound also holds on $\mathcal{K}_\psi^p$.

Remark 4.16. From this result, we see that the minimax-optimal rate for linear estimators can be *arbitrarily slow* even with the same sparsity p, i.e., with a bounded covering entropy (by Lemma 4.7). The nearly optimal rates attained by deep learning given in Section 5 are unfortunately limited to the case $\beta > 1$ (because of the assumption of boundedness), but this still serves as evidence for the non-effectiveness of linear methods in estimating sparse classes.

5. Learning ability of deep ReLU neural networks

5.1. Mathematical formulation of deep ReLU neural networks

For mathematical treatments of neural networks, we have referenced some recent papers on approximation theory and estimation theory (Bölcskei et al., 2017; Keiper, Kutyniok, & Petersen, 2017; Schmidt-Hieber, 2017; Suzuki, 2019; Yarotsky, 2017). In the following, we define neural networks mathematically and evaluate their covering entropies.

Definition 5.1. Let $\rho : \mathbb{R} \to \mathbb{R}$. For $L, S, D \in \mathbb{Z}_{>0}$ and $B \geq 1$ (with $D \geq d$), define $\mathcal{N}(L, S, D, B)$ as the set of all functions $f : \mathbb{R}^d \to \mathbb{R}$ of the form

$$f = W_{L+1} \circ \rho(W_L \cdot -v_L) \circ \cdots \circ \rho(W_1 \cdot -v_1),$$

satisfying

$$W_1 \in \mathbb{R}^{D \times d}, \ W_2, \ldots, W_L \in \mathbb{R}^{D \times D}, \ W_{L+1} \in \mathbb{R}^{1 \times D},$$

$$v_1 \in \mathbb{R}^d, \ v_2, \ldots, v_L \in \mathbb{R}^D$$

and

$$\|v_i\|_\infty, \|W_i\|_\infty \leq B, \ \sum_{i=1}^{L+1} \|W_i\|_0 + \sum_{i=1}^{L} \|v_i\|_0 \leq S,$$

where ρ is operated elementwise, and L, S, and D denote the number of hidden layers, the sparsity, and the dimensionality of the layers, respectively. Also, for $F > 0$, we consider a function class

$$\mathcal{N}_F = \mathcal{N}_F(L, S, D, B) := \{\text{sgn}(f)\min\{|f|, F\} \mid f \in \mathcal{N}(L, S, D, B)\}.$$

5.2. Complexity of neural network models

Hereinafter, we use the ReLU activation function $\rho(x) = \max\{x, 0\}$. Notice that $\mathcal{N}_F$ can be realized easily using ReLU activation after an element of $\mathcal{N}$ is computed. To evaluate the generalization error of deep learning, we need bounds for the complexity of neural network models. The following lemma gives the required bound as a function of parameters L, S, D, B.

Lemma 5.2 (*Schmidt-Hieber, 2017; Suzuki, 2019*)**. *For any $0 < \delta < 1$, the δ-covering entropy with respect to $\|\cdot\|_{L^\infty}$ of $\mathcal{N}(L, S, D, B)$ (limiting the domain to $[0, 1]^d$) can be bounded as*

$$V_{(\mathcal{N}(L,S,D,B), \|\cdot\|_{L^\infty})}(\delta) \leq 2S(L+1) \log\left(\frac{B(L+1)(D+1)}{\delta}\right).$$

The proof for this lemma is given in the appendix (Appendix B.7). We next introduce a lemma which evaluates the approximation ability of a neural network given the approximation ability of subnetworks. This lemma is also stated in Bölcskei et al. (2017) in another form. Thanks to this lemma, we can only consider the approximation of the wavelet basis, where we can exploit many existing studies.

Lemma 5.3. *Let $\Phi \subset L^2([0, 1]^d)$ satisfy $\|\varphi\|_{L^2} = 1$ for each $\varphi \in \Phi$. Suppose for any $\varphi \in \Phi$ there exists a function $g \in \mathcal{N}(L, S, D, B)$ such that $\|g - \varphi\|_{L^2} \leq \varepsilon$. Then for any $f^\circ \in \mathcal{I}_\Phi^0(n_s, C)$, there exists a function*

$$f \in \mathcal{N}(L + 2, n_s(S + 2Dd + d^2 + d + 1), n_s D, \max\{B, C\})$$

such that $\|f - f^\circ\|_{L^2} \leq C^{3/2} n_s \varepsilon$ holds.

Proof. The approximation of $f^\circ(x) = \sum_{i=1}^{n_s} c_i \varphi_i(A_i x - b_i) \in \mathcal{I}_\Phi^0(n_s, C)$ can be constructed as shown in Fig. 3 (we use the ReLU activation function, and so we compute $\max\{\widetilde{\varphi}_i, 0\}$ and $\max\{-\widetilde{\varphi}_i, 0\}$ and combine them afterward), where each $\widetilde{\varphi}_i$ approximates φ_i with an L^2-error of at least ε. In this construction,

$$\|\widetilde{f} - f^\circ\|_{L^2} \leq \sum_{i=1}^{n_s} |c_i| \|\varphi_i(A_i \cdot -b_i) - \widetilde{\varphi}_i(A_i \cdot -b_i)\|_{L^2}$$

$$\leq \sum_{i=1}^{n_s} C|\det A_i|^{-1/2} \|\widetilde{\varphi}_i - \varphi_i\|_{L^2} \leq C^{3/2} n_s \varepsilon$$

holds. □

Remark 5.4. These lemmas evaluate the covering entropy and approximation ability of neural networks. Given these two informations, we can exploit Theorem 2.6 to estimate the generalization error of deep learning by a fixed size of neural network. However, we should know how large the parameters L, S, D, B get as ε goes to zero (or the sample size n goes to infinity). As one can easily see, a shallow ReLU network can approximate piecewise constant functions easily (more precisely, see e.g., Fig. 4 and the proof of Corollary 5.8). Moreover, deep ReLU networks can approximate polynomials very efficiently (Yarotsky, 2017). Combining these, for example, we can state that

there exists an architecture $\mathcal{N}(L, S, D, B)$ with $L, S, D = O(\log(1/\varepsilon))$ and $B = O(1/\varepsilon)$ that can approximate piece-wise polynomials with $O(\varepsilon)$ L^2-error.

Rigorously speaking, we of course have to restrict the function class (the number of non-smooth points, the degree of polynomials, the magnitude of coefficients, etc.). However, as the main concern of this paper is statistical viewpoints, we do not get deeper. The important thing is that the above size of network architecture is reasonable to approximate "cheap" functions.

Motivated by the above remark, we define the function class AP, which can be approximated by "light" networks, as follows.

Definition 5.5. For $C_1, C_2 > 0$, define $AP(C_1, C_2)$ as the set of all functions $\varphi \in L^2([0, 1]^d)$ satisfying that, for each $0 < \varepsilon < 1/2$, there exist parameters $L_\varepsilon, S_\varepsilon, D_\varepsilon, B_\varepsilon > 0$ such that

- $L_\varepsilon, S_\varepsilon, D_\varepsilon \leq C_1 \log(1/\varepsilon)$ and $B_\varepsilon \leq C_2/\varepsilon$ hold;
- there exists a $\widetilde{\varphi} \in \mathcal{N}(L_\varepsilon, S_\varepsilon, D_\varepsilon, B_\varepsilon)$ such that $\|\widetilde{\varphi} - \varphi\|_{L^2} \leq \varepsilon$.
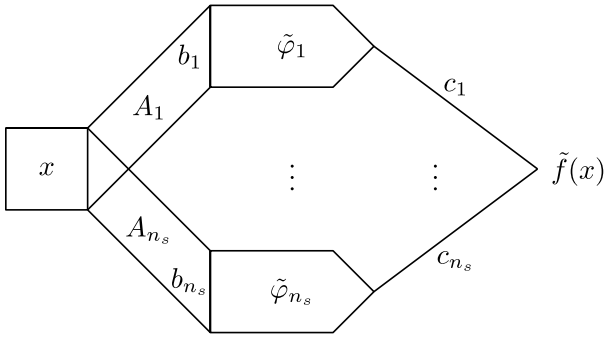
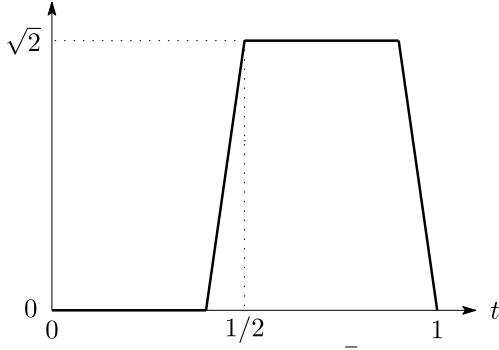**Fig. 3.** Construction of a larger neural network.

Fig. 4. An approximation of $\sqrt{2} \cdot 1_{[1/2,1]}$.

5.3. Generalization ability for an extreme case ($\mathcal{I}_\Phi^0$)

The learning ability of neural networks over $\mathcal{I}_\Phi^0$ is shown in the following. This is the most extreme case in terms of the difference between the performance of deep learning and linear methods.

Theorem 5.6. *Let $\Phi \subset L^2([0,1]^d)$ satisfy $\|\varphi\|_{L^2} = 1$ for each $\varphi \in L^2([0,1]^d)$ and $\sup_{\varphi \in \Phi} \|\varphi\|_{L^\infty} < \infty$. Suppose also $\Phi \subset \text{AP}(C_1, C_2)$ holds for some constants $C_1, C_2 > 0$. Then, for each $n_s, C > 0$, there exist constants $F, C_3 > 0$ dependent only on n_s, C (independent of n) such that the empirical risk minimizer $\widehat{f}$ over $\mathcal{N}_F^{(n)}$ satisfies*

$$\sup_{f^\circ \in \mathcal{I}_\Phi^0(n_s, C)} R(\widehat{f}, f^\circ) \leq C_3 \frac{(\log n)^3}{n}$$

for $n \geq 2$, where $\mathcal{N}_F^{(n)}$ denotes

$$\mathcal{N}_F \left(L_{1/n} + 2, n_s(S_{1/n} + 2D_{1/n}d + d^2 + d + 1), n_s D_{1/n}, \right.$$
$$\left. \max\{B_{1/n}, C\} \right).$$

Remark 5.7. This theorem implies that the empirical risk minimizer over $\mathcal{N}(L, S, D, B)$ with size

$$L, S, D = O(\log n), \quad B = O(n),$$

almost attains the minimax-optimal rate. Therefore, from the viewpoint of real-world application, a reasonable size of network can behave optimally.

Proof of Theorem 5.6. Let $M := \sup_{\varphi \in \Phi} \|\varphi\|_{L^\infty}$. If we define $F := n_s CM$, each $f^\circ \in \mathcal{I}_\Phi^0(n_s, C)$ satisfies $\|f\|_{L^\infty} \leq F$. Indeed, f° has some expression $f^\circ = \sum_{i=1}^{n_s} c_i \varphi_i(A_i \cdot -b_i)$, and so we have

$$\|f^\circ\|_{L^\infty} \leq \sum_{i=1}^{n_s} |c_i| \|\varphi_i\|_{L^\infty} \leq n_s CM = F.$$

Hence, for $f \in \mathcal{N}(L, S, D, B)$ and $f^\circ \in \mathcal{I}_\Phi^0(n_s, C)$, $\widetilde{f} := \text{sgn}(f) \min\{|f|, F\}$ satisfies

$$\|\widetilde{f} - f^\circ\|_{L^2}^2 = \int_{[0,1]^d} (\widetilde{f}(x) - f^\circ(x))^2 \, dx$$
$$\leq \int_{[0,1]^d} (f(x) - f^\circ(x))^2 \, dx \qquad (\because f^\circ(x) \in [-F, F])$$
$$= \|f - f^\circ\|_{L^2}^2.$$

Also, notice that $\mathcal{N}_F$'s covering entropy is not greater than that of $\mathcal{N}$, and so we have, by Lemma 5.2 and the assumption of the assertion,

$$V_{(\mathcal{N}_F^{(n)}, \|\cdot\|_{L^\infty})} \left(\frac{1}{n}\right) \leq C_0 (\log n)^3$$

for some constant $C_0 > 0$. Then, by Theorem 2.6 and Lemma 5.3,

$$\sup_{f^\circ \in \mathcal{I}_\Phi^0(n_s, C)} R(\widehat{f}, f^\circ)$$
$$\leq \frac{4C^{3/2} n_s}{n} + C' \left(C_0 (F^2 + \sigma^2) \frac{(\log n)^3}{n} + \frac{F + \sigma}{n} \right)$$
$$\leq C_3 \frac{(\log n)^3}{n}$$

holds for some $C_3 > 0$. $\square$

Corollary 5.8. *Let $d = 1$. For $J_k(C)$ in Definition 3.5, there exist a constant $F > 0$ and a sequence of neural networks $(\mathcal{N}^{(n)})_{n=2}^\infty$ such that the empirical risk minimizer $\widehat{f}$ satisfies*

$$\sup_{f^\circ \in J_k(C)} R(\widehat{f}, f^\circ) \leq C_3 \frac{(\log n)^3}{n}$$

for some constant $C_3 > 0$ independent of n and each $n \geq 2$.

Proof. By Theorem 5.6, it suffices to show that $\varphi = \sqrt{2} \cdot 1_{[1/2,1]}$ can be approximated within ε-error in L^2 by a neural network satisfying the condition of Theorem 5.6. This can be actually realized by a shallow network, as

$$\frac{4\sqrt{2}}{\varepsilon} \rho \left(t - \frac{1 - \varepsilon/2}{2} \right) - \frac{4\sqrt{2}}{\varepsilon} \rho \left(t - \frac{1}{2} \right)$$
$$- \frac{4\sqrt{2}}{\varepsilon} \rho \left(t - \left(1 - \frac{\varepsilon}{2}\right) \right) + \frac{4\sqrt{2}}{\varepsilon} \rho(t - 1)$$

(Fig. 4) satisfies the desired condition. $\square$

Remark 5.9. By Corollary 3.10, Theorem 4.2, and Corollary 5.8, $J_k(C)$ demonstrates an extreme situation, wherein neural network learning attains the optimal rate up to log factors whereas linear methods are suboptimal.

This result can easily be expanded to the case of $d = 2$ (if we properly define J_k for higher dimensions). In addition, we can treat a set broader than $J_k(C)$ as $\mathcal{I}_\Phi^0$ because smooth functions such as polynomials can be well approximated by $O(\log(1/\varepsilon))$ weights as has been mentioned in Remark 5.4.

5.4. Generalization ability for the wavelet case ($\mathcal{J}_\psi^p, \mathcal{K}_\psi^p$)

Let us consider the case in which the target function class is $\mathcal{K}_\psi^p$ in Definition 4.13. Note that, by Lemma 5.3, we only have to consider approximating functions in $\mathcal{J}_\psi^p$ in Definition 4.11 for fixed n_s. We defer the proof of the following main result to the appendix (Appendix B.8).

Theorem 5.10. *Let $\psi \in L^2([0,1]^d)$ be an orthonormal wavelet. Suppose there exist constants $C_1', C_2' > 0$ such that $\psi \in \mathrm{AP}(C_1', C_2')$ holds. Then, for each $C_1, C_2, \beta > 0$ and $0 < p < 2$, there exists a constant $C > 0$ dependent only on constants $C_1', C_2', p, C_1, C_2, \beta$ (independent of n) such that the empirical risk minimizer $\widehat{f}$ over $\mathcal{N}_F^{(n)}$ (with some network architecture) satisfies*

$$\sup_{\substack{f^\circ \in \mathcal{J}_\psi^p(C_1, C_2, \beta) \\ \|f^\circ\|_{L^\infty} \le F}} R(\widehat{f}, f^\circ) \le CF^2 n^{-\frac{2\alpha}{2\alpha+1}} (\log n)^3$$

for each $F \ge \max\{1, \sigma\}$ and $n \ge 2$, where $\alpha := 1/p - 1/2$. Moreover, for a set of wavelets $\Psi \subset \mathrm{AP}(C_1', C_2')$, the same evaluation is valid for $\mathcal{K}_\Psi^p$;

$$\sup_{\substack{f^\circ \in \mathcal{K}_\Psi^p(n_s, C_1, C_2, C_3, \beta) \\ \|f^\circ\|_{L^\infty} \le F}} R(\widehat{f}, f^\circ) \le C'F^2 n^{-\frac{2\alpha}{2\alpha+1}} (\log n)^3$$

holds for some $C' > 0$, where $\widehat{f}$ is the empirical risk minimizer of some larger network.

Remark 5.11. The actual network size of $\mathcal{N}(L, S, D, B)$ here is

$$L = O(\log n), \quad S, D = O(n^{\frac{1}{2\alpha+1}} \log n), \quad B = O\left(n^{\max\left\{1, \frac{4\alpha}{\beta(2\alpha+1)}\right\}}\right)$$

(see (37) in the Appendix). Though this network is larger than that of Remark 5.7, this is still acceptable size, especially with large α; i.e., in the case the parameter space is very sparse.

Concretely, ψ constructed by using the Haar wavelet satisfies the desired condition. Also, in the case of $d = 1$ and $\beta > 1$, we can remove the restriction by the constant F, because $\sup\{\|f\|_\infty \mid f \in J_\psi^p\} < \infty$ holds. Note also that this result is nearly minimax optimal; i.e., $\widehat{f}$ attains the minimax lower bound derived in Theorem 4.8 up to log factors and the constraint of boundedness. As p gets smaller, $\mathcal{K}_\psi^p$ gets sparser and more reasonable size of neural network can achieve the minimax optimality. Moreover, as is stated in the next remark, linear estimators cannot exploit the sparsity of $\mathcal{K}_\psi^p$ in typical settings (e.g., when Ψ contains non-continuous functions). From the viewpoint of approximation ability, this situation can be interpreted as follows; neural networks are allowed to approximate just the sparse set $\mathcal{K}_\psi^p$, whereas linear estimators must be able to approximate all the functions contained in $\mathrm{conv}(\mathcal{K}_\psi^p)$. This difference yields the difference of complexity; linear estimators require far more parameters than neural networks. Finally, this situation results in the great difference of generalization ability between deep and linear.

Remark 5.12. If Ψ contains the Haar wavelet and C_3 is sufficiently large, $\mathcal{K}_\psi^p$ includes J_k in Definition 3.5 (also, notice that J_k is bounded in the L^∞-norm sense). To be more precise, $\{f \in [0,1]^d \to \mathbb{R} \mid f(x_1, \ldots, x_d) = g(x_1), \ g \in J_k(C)\}$ is included in K_ψ^p for some $C > 0$. The proof of Theorem 3.9 can easily be modified for this case, and we have

$$\inf_{\widehat{f}:\text{linear}} \sup_{f^\circ \in \mathcal{K}_\psi^p, \|f^\circ\|_\infty \le F} R(\widehat{f}, f^\circ) \ge cn^{-1/2}$$

for some $c > 0$. If $\alpha > 1/2$ (equivalent to $p < 1$) holds, then the neural network learning is superior to linear methods.

5.5. Parameter sharing technique to restrict the covering entropy

The assumption of the ability for φ to be approximated by $\widetilde{\varphi}$ imposed in Theorems 5.6 and 5.10 is quite strong, and thus we cannot treat a broad range of wavelets. In the proof of Theorem 5.10, however, we do not exploit the full degree of freedom depicted in Fig. 3 because subnetworks share the same approximator $\widetilde{\psi}$. In this subsection, we consider neural networks with parameter sharing.

Definition 5.13. Let N be a positive integer. For a given neural network architecture $\mathcal{N}(L, S, D, B)$, denote the *N-sharing* of $\mathcal{N}(L, S, D, B)$ by $\mathcal{N}^N(L, S, D, B)$, defined as

$$\left\{ \sum_{i=1}^{N} c_i f(A_i \cdot - b_i) \,\middle|\, A_i \in \mathbb{R}^{d \times d}, \ b_i \in \mathbb{R}^d, \ c_i \in \mathbb{R}, \ \|A_i\|_\infty, \|b_i\|_\infty, \right.$$
$$\left. |c_i| \le B, \ i = 1, \ldots, d, \ f \in \mathcal{N}(L, S, D, B) \right\}.$$

Theorem 5.14. *Given a positive integer N and $\mathcal{N}(L, S, D, B)$ with $L \ge 2$, the δ-covering entropy with respect to $\|\cdot\|_{L^\infty}$ of $\mathcal{N}^N(L, S, D, B)$ (limiting the domain to $[0,1]^d$) can be bounded as*

$$V_{(\mathcal{N}^N(L,S,D,B), \|\cdot\|_{L^\infty})}(\delta)$$
$$\le \left(N(d+1)^2 + 2S(L+1)\right)(L+3)\log\left(\frac{NB(L+1)(D+1)}{\delta}\right)$$

for any $0 < \delta < 1$.

The proof is straightforward but a bit technical; thus, we defer it to the appendix (Appendix B.9).

Remark 5.15. If we use neural networks with parameter sharing, we can use non-trivial wavelets with some smoothness; i.e., in Theorem 5.10, the assumption for ψ can be weakened to the following:

For each $0 < \varepsilon < 1/2$, there exist $L_\varepsilon, S_\varepsilon, D_\varepsilon, B_\varepsilon > 0$ satisfying

- $L_\varepsilon \le C_1' \log(1/\varepsilon)$, $D_\varepsilon, B_\varepsilon \le C_2' \varepsilon^{-\gamma}$ and $S_\varepsilon \le C_3' \varepsilon^{-\frac{1}{\alpha+1}}$ hold for some constants $C_1', C_2', C_3', \gamma > 0$;
- there exists $\widetilde{\psi} \in \mathcal{N}(L_\varepsilon, S_\varepsilon, D_\varepsilon, B_\varepsilon)$ such that $\|\widetilde{\psi} - \psi\|_{L^2} \le \varepsilon$.

This class of ψ is actually broadened as there exist compactly supported wavelets with high regularity (a large Hölder exponent) (Daubechies, 1992), and such functions can be approximated well by networks with a small number of parameters. Indeed, Yarotsky (2017) proved that

A unit ball of Sobolev space $W^{n,\infty}([0,1]^d)$ can be approximated, with L^∞-error at most ε, by a neural network with $O(\log(1/\varepsilon))$ depth, and $O(\varepsilon^{-d/n})$ complexity (corresponds to S, D).

Though we cannot use this in the original form because L^∞ should be replaced by L^2 and the upper bound of B should be given, the refinement is not so difficult (e.g., Schmidt-Hieber, 2017).

6. Summary and discussion

6.1. Summary

In this paper, we have shown that deep learning outperforms other commonly used methods such as linear estimators even in a simple case. To evaluate the learning ability of estimators, we employed a Gaussian regression problem with a sparse target function space. In such a problem setting, neural network learning attains nearly the minimax-optimal rate of convergence with respect to the sample size, whereas a linear estimator can only achieve a suboptimal rate. The main novelty is that the target function spaces were selected to have natural sparsity, instead of following the well-known settings developed by the existing mathematical analyses. We have also shown that parameter sharing is quite effective for widening function classes where (near) minimax optimality holds.

6.2. Discussion and future work

There are two main limitations in this work that remain to be addressed in future investigations.

First, $\mathcal{I}_{\Phi}^0(n_s, C)$ is the most extreme case in the sense that deep learning outperforms linear estimators. This class is very simple, and we have additionally defined $w\ell^p$-bounded classes $\mathcal{K}_{\psi}^p$ for $0 < p < 2$, with the assumption of orthonormal wavelets. However, we should remove the orthogonality if we follow the philosophy of defining $\mathcal{I}_{\Phi}^0(n_s, C)$. For example, using the definition

$$\mathcal{L}_{\Phi}^p(C_1, C_2, C_3, \beta)$$
$$:= \left\{ \sum_{i=1}^{\infty} c_i \varphi(A_i \cdot - b_i) \, \middle| \, \begin{array}{l} |c_i| \leq C_1 i^{-1/p}, \ \|A_i\|_{\infty}, \|b_i\|_{\infty} \leq C_2 i^{\beta}, \\ |\det A_i|^{-1} \leq C_3, \ i = 1, 2, \ldots, \ \varphi \in \Phi \end{array} \right\}$$

would be one possible way. Of course, for some range of (p, β), we can show that deep learning attains a rate faster than do linear estimators. However, we could not have shown that the convergence rate satisfies minimax optimality, even up to log factors. This difficulty arises from the fact that we have fully exploited the orthogonality in the proof of deep learning's minimax optimality over $\mathcal{K}_{\psi}^p$. It is possible that we can find both a better minimax lower bound for $\mathcal{L}_{\Phi}^p$ and a better approximation bound by neural networks.

Second, parameter sharing, mentioned in Section 5.5, is used mainly in the context of convolutional neural networks (CNNs), and this implies the superiority of CNNs in solving a regression problem. However, some of the arguments in this paper are not directly applicable to the analysis of CNNs. Although CNNs have achieved notable success in pattern recognition, theories of CNNs with respect to regression problems have not yet been well argued in the literature.

In addition to these issues, a theoretical analysis of stochastic optimization as used in deep learning is needed, which is not treated in this paper.

Acknowledgments

Appendix A. Mathematical supplements

A.1. Generalities related to the minimax risk

In the Gaussian regression model, for two true functions $f, g \in \mathcal{F}^{\circ}$, it is well known that the Kullback–Leibler (KL) divergence between two distributions of (X, Y) generated by f and g is easy to calculate. Let $d_{\mathrm{KL}}(f, g)$ be the square root of the KL divergence. The following lemma is essential when one treats a regression problem as a parameter estimation problem (Schmidt-Hieber, 2017; Suzuki, 2019; Yang & Barron, 1999); its proof is given in Appendix B.10.

Lemma A.1. It holds that $d_{\mathrm{KL}}(f, g)^2 = \dfrac{1}{2\sigma^2} \|f - g\|_{L^2}^2$.

Therefore, the square root of the KL divergence is a metric equivalent to the L^2 metric in regression problems with a Gaussian noise.

We also introduce a lemma that is useful for deriving a lower bound of the metric entropy.

Lemma A.2 (Donoho, 1993, Lemma 4). Let $\mathcal{C}_k \subset \ell^2$ be a k-dimensional hypercube of side $2\delta > 0$ defined as

$$\mathcal{C}_k := \{a \in \ell^2 \mid |a_1|, \ldots, |a_k| \leq \delta, \ |a_{k+1}| = |a_{k+2}| = \cdots = 0\}.$$

Then there exists a constant $A > 0$ such that

$$V_{(\mathcal{C}_k, \|\cdot\|_{\ell^2})} \left(\frac{\delta \sqrt{k}}{2} \right) \geq Ak, \quad k = 1, 2, \ldots.$$

A.2. Treatment of multi-dimensional wavelets

In this section, we give a multi-dimensional extension of arguments in Section 4.3. Here, we adopt the multiplication of elementwise wavelets as a multi-dimensional wavelet.

Lemma A.3. For orthogonal wavelets $\psi^{(1)}, \ldots, \psi^{(d)}$,

$$\psi(x) := \prod_{i=1}^{d} \psi^{(i)}(x_i), \quad x = (x_1, \ldots, x_d) \in \mathbb{R}^d,$$

is a d-dimensional orthogonal wavelet; i.e.,

$$\left\{ \prod_{i=1}^{d} \psi_{k_i, \ell_i}^{(i)} \, \middle| \, k_i \geq 0, \ 0 \leq \ell_i < 2^{k_i}, \ i = 1, \ldots, d \right\}$$

is an orthonormal subset of $L^2([0, 1]^d)$.

Proof. The normality is clear by Fubini's theorem. Also, for distinct wavelets ψ, ψ' in the set, there exists i such that $(k_i, \ell_i) \neq (k_i', \ell_i')$. Since we have $\psi\psi' \in L^1([0, 1]^d)$ by the AM–GM inequality, Fubini's theorem leads to the conclusion. □

Definition A.4. Given an orthonormal wavelet

$$\psi(x) = \psi^{(1)}(x_1) \cdots \psi^{(d)}(x_d)$$

and constants $C_1, C_2, \beta > 0$ and $0 < p < 2$, define

$$\mathcal{J}_{\psi}^p(C_1, C_2, \beta) := \left\{ \sum_{(k, \ell) \in T_0} a_{k, \ell} \psi_{k, \ell} \, \middle| \, \|a\|_{w\ell^p} \leq C_1, \right.$$
$$\left. \sum_{(k, \ell) \in T_m} a_{k, \ell}^2 \leq C_2 2^{-\beta m}, \ m = 0, 1, \ldots \right\},$$

where the sets T_m ($m = 0, 1, \ldots$) are defined as

$$T_m := \left\{ (k, \ell) \in \mathbb{Z}^d \times \mathbb{Z}^d \, \middle| \, \begin{array}{l} k = (k_1, \ldots, k_d), \ \ell = (\ell_1, \ldots, \ell_d), \\ k_i \geq 0, \ 0 \leq \ell_i < 2^{k_i}, \ i = 1, \ldots, d, \\ \max_{1 \leq i \leq d} k_i \geq m \end{array} \right\},$$

and $\psi_{k, \ell}$ denotes $\prod_{i=1}^{d} \psi_{k_i, \ell_i}^{(i)}$ for each

$$(k, \ell) = ((k_1, \ldots, k_d), (\ell_1, \ldots, \ell_d)) \in T_0.$$

Remark A.5. If we define a partial order on S_0 by $(k, \ell) \preceq (k', \ell') \Leftrightarrow \max_i k_i \leq \max_i k_i'$ and then sort it, $\mathcal{J}_{\psi}^p(C_1, C_2, \beta)$ is revealed to be β/d-minimally tail compact. Thus, $\mathcal{I}_{\varphi}^p \subset \mathcal{J}_{\psi}^p$ holds with some modification of constants.

Appendix B. Proofs

B.1. Proof of Theorem 2.6

(Mainly following the original proof.) First, we evaluate the value of

$$D := \left| \mathbb{E} \left[\frac{1}{n} \sum_{i=1}^{n} (\widehat{f}(X_i) - f^{\circ}(X_i))^2 \right] - R(\widehat{f}, f^{\circ}) \right|.$$

Let $X'_1, \ldots, X'_n$ be i.i.d. random variables generated to be independent of $(X_i, Y_i)_{i=1}^n$. Then we have

$$R(\widehat{f}, f^\circ) = \frac{1}{n} \sum_{i=1}^n \mathrm{E}\left[(\widehat{f}(X'_i) - f^\circ(X'_i))^2\right],$$

and so we obtain

$$D = \left| \mathrm{E}\left[\frac{1}{n} \sum_{i=1}^n \left((\widehat{f}(X_i) - f^\circ(X_i))^2 - (\widehat{f}(X'_i) - f^\circ(X'_i))^2\right) \right] \right|$$

$$\leq \frac{1}{n} \mathrm{E}\left[\left| \sum_{i=1}^n \left((\widehat{f}(X_i) - f^\circ(X_i))^2 - (\widehat{f}(X'_i) - f^\circ(X'_i))^2\right) \right| \right].$$

Here, let $G_\delta = \{f_1, \ldots, f_N\}$ be a δ-covering of $\mathcal{F}$ with the minimum cardinality in the L^∞ metric. Notice that $\log N \geq 1$. If we define $g_j(x, x') := (f_j(x) - f^\circ(x))^2 - (f_j(x') - f^\circ(x'))^2$ and a random variable J taking values in $\{1, \ldots, N\}$ such that $\|\widehat{f} - f_J\|_{L^\infty} \leq \delta$, we have

$$D \leq \frac{1}{n} \mathrm{E}\left[\left| \sum_{i=1}^n g_J(X_i, X'_i) \right| \right] + 8F\delta. \tag{9}$$

In the above evaluation, we have used the inequality

$$\left| (\widehat{f}(x) - f^\circ(x))^2 - (f_J(x) - f^\circ(x))^2 \right|$$
$$= \left| \widehat{f}(x) - f_J(x) \right| \left| \widehat{f}(x) + f_J(x) - 2f^\circ(x) \right| \leq 4F\delta.$$

Define constants $r_j := \max\{A, \|f_j - f^\circ\|_{L^2}\}$ $(j = 1, \ldots, N)$ and a random variable

$$T := \max_{1 \leq j \leq N} \left| \sum_{i=1}^n \frac{g_j(X_i, X'_i)}{r_j} \right|,$$

where $A > 0$ is a deterministic quantity fixed afterward. Then, because of (9), we have

$$D \leq \frac{1}{n} \mathrm{E}[r_J T] + 8F\delta \leq \frac{1}{n} \sqrt{\mathrm{E}[r_J^2]\mathrm{E}[T^2]} + 8F\delta$$
$$\leq \frac{1}{2} \mathrm{E}[r_J^2] + \frac{1}{2n^2} \mathrm{E}[T^2] + 8F\delta \tag{10}$$

by the Cauchy–Schwarz inequality and the AM–GM inequality. Here, by the definition of J, $\mathrm{E}[r_J^2]$ can be evaluated as follows:

$$\mathrm{E}[r_J^2] \leq A^2 + \mathrm{E}\left[\|f_J - f^\circ\|_{L^2}^2\right] \leq A^2 + \mathrm{E}\left[\|\widehat{f} - f^\circ\|_{L^2}^2\right] + 4F\delta$$
$$= R(\widehat{f}, f^\circ) + A^2 + 4F\delta. \tag{11}$$

Because of the independence of the defined random variables,

$$\mathrm{E}\left[\left(\sum_{i=1}^n \frac{g_j(X_i, X'_i)}{r_j} \right)^2 \right]$$
$$= \sum_{i=1}^n \mathrm{E}\left[\left(\frac{g_j(X_i, X'_i)}{r_j} \right)^2 \right]$$
$$= \sum_{i=1}^n \left(\mathrm{E}\left[\frac{(f_j(X_i) - f^\circ(X_i))^4}{r_j^2} \right] + \mathrm{E}\left[\frac{(f_j(X'_i) - f^\circ(X'_i))^4}{r_j^2} \right] \right)$$
$$\leq 2F^2 n$$

holds, where we have used the fact that each $g_j(X_i, X'_i)$ is centered. Then, using Bernstein's inequality, we have, in terms of $r := \min_{1 \leq j \leq N} r_j$,

$$P(T^2 \geq t) = P(T \geq \sqrt{t}) \leq 2N \exp\left(-\frac{t}{2F^2\left(2n + \frac{\sqrt{t}}{3r}\right)} \right), \quad t \geq 0.$$

Let us evaluate $\mathrm{E}[T^2]$. For arbitrary $t_0 > 0$, it holds that

$$\mathrm{E}[T^2] = \int_0^\infty P(T^2 \geq t)\,\mathrm{d}t$$
$$\leq t_0 + \int_{t_0}^\infty P(T^2 \geq t)\,\mathrm{d}t$$
$$\leq t_0 + 2N \int_{t_0}^\infty \exp\left(-\frac{t}{8F^2 n} \right) \mathrm{d}t$$
$$+ 2N \int_{t_0}^\infty \exp\left(-\frac{3r\sqrt{t}}{4F^2} \right) \mathrm{d}t.$$

We compute the values of these two integrals in terms of t_0:

$$\int_{t_0}^\infty \exp\left(-\frac{t}{8F^2 n} \right) \mathrm{d}t = \left[-8F^2 n \exp\left(-\frac{t}{8F^2 n} \right) \right]_{t_0}^\infty$$
$$= 8F^2 n \exp\left(-\frac{t_0}{8F^2 n} \right),$$

$$\int_{t_0}^\infty \exp\left(-\frac{3r\sqrt{t}}{4F^2} \right) \mathrm{d}t = \int_{t_0}^\infty \exp(-a\sqrt{t})\,\mathrm{d}t \qquad (a := 3r/4F^2)$$
$$= \left[-\frac{2(a\sqrt{t} + 1)}{a^2} \exp(-a\sqrt{t}) \right]_{t_0}^\infty$$
$$= \frac{8F^2\sqrt{t_0}}{3r} \exp\left(-\frac{3r\sqrt{t_0}}{4F^2} \right)$$
$$+ \frac{32F^2}{9r^2} \exp\left(-\frac{3r\sqrt{t_0}}{4F^2} \right).$$

Now we determine $A = \sqrt{t_0}/6n$. Since we have $r \geq A = \sqrt{t_0}/6n$,

$$\mathrm{E}[T^2] \leq t_0 + 2N \left(8F^2 n + 16F^2 n + \frac{128F^2 n^2}{t_0} \right) \exp\left(-\frac{t_0}{8F^2 n} \right)$$
$$\leq t_0 + 16NF^2 n \left(3 + \frac{16n}{t_0} \right) \exp\left(-\frac{t_0}{8F^2 n} \right)$$

holds. Letting $t_0 = 8F^2 n \log N$, the above evaluation can be rewritten as

$$\mathrm{E}[T^2] \leq 8F^2 n \left(\log N + 6 + \frac{2}{F^2 \log N} \right). \tag{12}$$

Finally, we combine (10), (11), (12), and $A^2 = \dfrac{2F^2 \log N}{9n}$ to obtain

$$D \leq \left(\frac{1}{2}R(\widehat{f}, f^\circ) + \frac{1}{2}A^2 + 2F\delta \right)$$
$$+ \frac{4F^2}{n} \left(\log N + 6 + \frac{2}{F^2 \log N} \right) + 8F\delta$$
$$\leq \frac{1}{2}R(\widehat{f}, f^\circ) + \frac{F^2}{n} \left(\frac{37}{9} \log N + 32 \right) + 10F\delta,$$

where we have used the fact that $\log N \geq 1$. Thus, we obtain the evaluation

$$R(\widehat{f}, f^\circ) \leq 2\mathrm{E}\left[\frac{1}{n} \sum_{i=1}^n \widehat{f}(X_i) - f^\circ(X_i))^2 \right] + \frac{2F^2}{n} \left(\frac{37}{9} \log N + 32 \right)$$
$$+ 20F\delta. \tag{13}$$

Next, we evaluate the quantity

$$\widehat{R} := \mathrm{E}\left[\frac{1}{n} \sum_{i=1}^n \widehat{f}(X_i) - f^\circ(X_i))^2 \right]. \tag{14}$$

Since $\widehat{f}$ is an empirical risk minimizer, for arbitrary $f \in \mathcal{F}$,

$$\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(\widehat{f}(X_i) - Y_i)^2\right] \leq \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(f(X_i) - Y_i)^2\right]$$

holds. As $Y_i = f^\circ(X_i) + \xi_i$, we have

$$\mathrm{E}\left[(f(X_i) - Y_i)^2\right] - \mathrm{E}\left[(\widehat{f}(X_i) - Y_i)^2\right]$$
$$= \mathrm{E}\left[(f(X_i) - f^\circ(X_i))^2\right] - 2\mathrm{E}\left[\xi_i f(X_i)\right] - \mathrm{E}\left[(\widehat{f}(X_i) - f^\circ(X_i))^2\right]$$
$$\quad + 2\mathrm{E}\left[\xi_i \widehat{f}(X_i)\right]$$
$$= \left(\|f - f^\circ\|_{L^2}^2 + 2\mathrm{E}\left[\xi_i \widehat{f}(X_i)\right]\right) - \mathrm{E}\left[(\widehat{f}(X_i) - f^\circ(X_i))^2\right].$$

Here we have used the fact that

$$\mathrm{E}[\xi_i f(X_i)] = \mathrm{E}[\xi_i]\mathrm{E}[f(X_i)] = 0$$

holds because of the independence between ξ_i and X_i and the fact that both ξ_i and $f(X_i)$ have a finite L^1 norm. Thus we have

$$\widehat{R} \leq \|f - f^\circ\|_{L^2}^2 + \mathrm{E}\left[\frac{2}{n}\sum_{i=1}^{n}\xi_i \widehat{f}(X_i)\right]. \tag{15}$$

Let us evaluate the second term on the right-hand side.

$$\left|\mathrm{E}\left[\frac{2}{n}\sum_{i=1}^{n}\xi_i \widehat{f}(X_i)\right]\right| = \left|\mathrm{E}\left[\frac{2}{n}\sum_{i=1}^{n}\xi_i(\widehat{f}(X_i) - f^\circ(X_i))\right]\right|$$
$$\leq \frac{2\delta}{n}\mathrm{E}\left[\sum_{i=1}^{n}|\xi_i|\right]$$
$$\quad + \left|\mathrm{E}\left[\frac{2}{n}\sum_{i=1}^{n}\xi_i(f_J(X_i) - f^\circ(X_i))\right]\right|. \tag{16}$$

Here, the first term is upper-bounded by applying the Cauchy–Schwarz inequality:

$$\frac{2\delta}{n}\mathrm{E}\left[\sum_{i=1}^{n}|\xi_i|\right] \leq \frac{2\delta}{n}\mathrm{E}\left[n^{1/2}\left(\sum_{i=1}^{n}\xi_i^2\right)^{1/2}\right]$$
$$\leq \frac{2\delta}{\sqrt{n}}\mathrm{E}\left[\sum_{i=1}^{n}\xi_i^2\right]^{1/2} = 2\sigma\delta. \tag{17}$$

Let ε_j $(j = 1, \ldots, N)$ be random variables defined as

$$\varepsilon_j := \frac{\sum_{i=1}^{n}\xi_i(f_j(X_i) - f^\circ(X_i))}{\left(\sum_{i=1}^{n}(f_j(X_i) - f^\circ(X_i))^2\right)^{1/2}},$$

where $\varepsilon_j := 0$ if the denominator equals 0. Notice that each ε_j follows a centered Gaussian distribution with variance σ^2 (conditional on $X_1, \ldots, X_n$). Now we have, using the Cauchy–Schwarz inequality and the AM–GM inequality,

$$\left|\mathrm{E}\left[\frac{2}{n}\sum_{i=1}^{n}\xi_i(f_J(X_i) - f^\circ(X_i))\right]\right|$$
$$= \frac{2}{n}\left|\mathrm{E}\left[\left(\sum_{i=1}^{n}(f_J(X_i) - f^\circ(X_i))^2\right)^{1/2}\varepsilon_J\right]\right|$$
$$\leq \frac{2}{\sqrt{n}}\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(f_J(X_i) - f^\circ(X_i))^2\right]^{1/2}\mathrm{E}\left[\max_{1\leq j\leq N}\varepsilon_j^2\right]^{1/2}$$
$$\leq \frac{2}{\sqrt{n}}\sqrt{\widehat{R} + 4F\delta}\,\mathrm{E}\left[\max_{1\leq j\leq N}\varepsilon_j^2\right]^{1/2}$$
$$\leq \frac{1}{2}(\widehat{R} + 4F\delta) + \frac{2}{n}\mathrm{E}\left[\max_{1\leq j\leq N}\varepsilon_j^2\right]. \tag{18}$$

By a similar argument as given in the proof of Lafferty, Liu, and Wasserman (2008, Theorem 7.47), for any $0 < t < 1/2\sigma^2$,

$$\exp\left(t\mathrm{E}\left[\max_{1\leq j\leq N}\varepsilon_j^2\right]\right) \leq \mathrm{E}\left[\max_{1\leq j\leq N}\exp\left(t\varepsilon_j^2\right)\right]$$

(by Jensen's inequality)

$$\leq N\mathrm{E}\left[\exp\left(t\varepsilon_1^2\right)\right]$$
$$= \frac{N}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{\infty}e^{tx^2}e^{-\frac{x^2}{2\sigma^2}}\,dx = \frac{N}{\sqrt{1 - 2\sigma^2 t}}$$

holds. Therefore we have, by determining $t = 1/4\sigma^2$,

$$\mathrm{E}\left[\max_{1\leq j\leq N}\varepsilon_j^2\right] \leq 4\sigma^2\log(\sqrt{2}N) \leq 4\sigma^2(\log N + 1). \tag{19}$$

Now we combine (15)–(19) to obtain

$$\widehat{R} \leq \|f - f^\circ\|_{L^2}^2 + 2\sigma\delta + \frac{1}{2}(\widehat{R} + 4F\delta) + \frac{8\sigma^2}{n}(\log N + 1),$$

and so

$$\widehat{R} \leq 2\|f - f^\circ\|_{L^2}^2 + 4(\sigma + F)\delta + \frac{16\sigma^2}{n}(\log N + 1) \tag{20}$$

holds.

Finally, since f is an arbitrary element of $\mathcal{F}$, we combine (13), (14), and (20) to have

$$R(\widehat{f}, f^\circ) \leq 4\inf_{f\in\mathcal{F}}\|f - f^\circ\|_{L^2}^2$$
$$+ \frac{1}{n}\left(\left(\frac{37}{9}F^2 + 32\sigma^2\right)\log N + 32(F^2 + \sigma^2)\right)$$
$$+ (18F + 8\sigma)\delta,$$

and this leads to the conclusion.

B.2. Proof of Theorem 3.3

We first prove the following lemma.

Lemma B.1. *For any affine estimator $\widehat{f}$ and a sequence $f_1^\circ, f_2^\circ, \ldots \in L^2([0,1]^d)$ convergent to $f_\infty^\circ \in L^2([0,1]^d)$ almost everywhere,*

$$R(\widehat{f}, f_\infty^\circ) \leq \sup_{m\geq 1}R(\widehat{f}, f_m^\circ)$$

holds.

Proof. The conclusion follows immediately from Fatou's lemma and Eqs. (4) and (5). $\square$

Remark B.2. In the proof of Lemma B.1, we have not used the linearity of $\widehat{f}$. Indeed, if $\widehat{f}(X) - f_m^\circ(X)$ is convergent to $\widehat{f}(X) - f_\infty^\circ(X)$ with probability 1, then we have the same conclusion (where X is a uniformly distributed random variable independent of other observed random variables). Hence, Lemma B.1 is applicable to a broader class of estimators, such as estimators continuous with respect to observed data in some metric.

The assertion of Theorem 3.3 is now clear from Eq. (2), Lemma B.1, and the fact that a sequence convergent to a function in L^2 has a subsequence that is convergent to the same function almost everywhere.

B.3. Proof of Lemma 3.8

The following lemma is a well-known property of functions of bounded total variation.

Lemma B.3 (*Stein & Shakarchi, 2005*). *For each function $f : [0,1] \to \mathbb{R}$ with $TV^f([0,1]) < \infty$, there exist increasing $f^+, f^- :*

$[0, 1] \to \mathbb{R}$ *such that*

$$f = f^+ - f^- + f(0), \quad f^+(0) = f^-(0) = 0,$$
$$f^+(t) + f^-(t) = TV^f([0, t]), \quad 0 \le t \le 1,$$

holds.

By using this lemma, Lemma 3.8 can be proven as follows.

Since $J_1(C) \subset J_k(C)$ holds for each k, it suffices to show the assertion for $k = 1$. By the definition of convex hull, we have

$$\text{conv}(J_1(C)) = \left\{ a_0 + \sum_{i=1}^{k} a_i 1_{[t_i, 1]} \,\middle|\, t_i \in (0, 1], \ |a_0| \le C, \right.$$
$$\left. \sum_{i=1}^{k} |a_i| \le C, \ k \ge 1 \right\} = \bigcup_{k=1}^{\infty} J_k(C).$$

It is obvious that $J_k(C) \subset BV(C)$ for each k. Thus, we have only to show that for each $f \in BV(C)$ and $\varepsilon > 0$, there exist some $k \ge 1$ and $f_k \in J_k(C)$ such that $\|f_k - f\|_{L^2} \le \varepsilon$ holds.

Let $f \in BV(C)$, and take f^+ and f^- satisfying the condition of Lemma B.3. Then f can be written as $f = a_0 + f^+ - f^-$, where $a_0 := f(0)$ is a constant. Let $g : [0, 1] \to \mathbb{R}$ be an increasing function satisfying $g(0) = 0$, and define

$$g_k := \sum_{i=1}^{k} \left(g\left(\frac{i}{k}\right) - g\left(\frac{i-1}{k}\right) \right) 1_{[i/k, 1]}.$$

Then we have $g_k \in J_k(g(1))$ and $g_k(t) = g(i/k)$ for $t \in [i/k, (i+1)/k)$, and so

$$\int_0^1 (g_k(t) - g(t))^2 \, dt \le \sum_{i=0}^{k-1} \frac{1}{k} \left(g\left(\frac{i+1}{k}\right) - g\left(\frac{i}{k}\right) \right)^2$$
$$\le \frac{g(1)}{k} \sum_{i=0}^{k-1} \left(g\left(\frac{i+1}{k}\right) - g\left(\frac{i}{k}\right) \right)$$
$$= \frac{g(1)^2}{k}$$

holds because g is increasing. Take f_k^+, f_k^- similarly, and $f_k := a_0 + f_k^+ - f_k^-$ satisfies

$$f_k \in J_{2k}(C), \quad \|f_k - f\|_{L^2}^2 \le \frac{2C^2}{k},$$

and the proof is complete. We can of course take f_k directly only from f, but we have chosen an easier argument.

B.4. Proof of Theorem 4.2

Since a smaller set makes the minimax risk smaller, it suffices to consider

$$\mathcal{I} := \{ c\varphi \mid |c| \le C \} \subset \mathcal{I}_\Phi^0.$$

For simplicity, assume $C \ge 1$ (otherwise, take the functions and constants appearing below to be smaller at that rate). Define for each $n = 1, 2, \ldots$

$$f_n^+ := \frac{1}{2\sqrt{n}} \varphi, \quad f_n^- := -\frac{1}{2\sqrt{n}} \varphi.$$

Then, by applying the argument appearing in Yang and Barron (1999), we have

$$\inf_{\widehat{f}} \sup_{f \in \mathcal{I}} P_f \left(\|f - \widehat{f}\|_{L^2} \ge \frac{1}{2\sqrt{n}} \right)$$
$$\ge \inf_{\widehat{f}} \sup_{f \in \{f_n^+, f_n^-\}} P_f \left(\|f - \widehat{f}\|_{L^2} \ge \frac{1}{2\sqrt{n}} \right)$$
$$\ge \inf_{\widehat{f}} \sup_{f \in \{f_n^+, f_n^-\}} P_f(\widetilde{f} \ne f), \tag{21}$$

where $\widetilde{f}$ is the closer of the two (f_n^+, f_n^-) to $\widehat{f}$. Following the argument in Tsybakov (2008, Proposition 2.1), we have for any $t > 0$

$$P_{f_n^+} \left(\widetilde{f} \ne f_n^+ \right) = E_{f_n^-} \left[1_{\{\widetilde{f} = f_n^-\}}(Z^n) \frac{dP_{f_n^+}}{dP_{f_n^-}}(Z^n) \right]$$
$$\ge t P_{f_n^-} \left(\widetilde{f} = f_n^-, \ \frac{dP_{f_n^+}}{dP_{f_n^-}}(Z^n) \ge t \right)$$
$$\ge t \left(P_{f_n^-}(\widetilde{f} = f_n^-) - P_{f_n^-} \left(\frac{dP_{f_n^+}}{dP_{f_n^-}}(Z^n) < t \right) \right). \tag{22}$$

Here, Z^n denotes the i.i.d. sequence $(X_i, Y_i)_{i=1}^n$, and $dP_{f_n^+}/dP_{f_n^-}$ represents the Radon–Nikodym derivative. Then we have, by (21) and (22),

$$\inf_{\widehat{f}} \sup_{f \in \mathcal{I}} P_f \left(\|f - \widehat{f}\|_{L^2} \ge \frac{1}{2\sqrt{n}} \right)$$
$$\ge \frac{1}{1+t} \left(t P_{f_n^-} \left(\widetilde{f} \ne f_n^- \right) + P_{f_n^+} \left(\widetilde{f} \ne f_n^+ \right) \right)$$
$$\ge \frac{t}{1+t} \left(1 - P_{f_n^-} \left(\frac{dP_{f_n^+}}{dP_{f_n^-}}(Z^n) < t \right) \right)$$
$$= \frac{t}{1+t} P_{f_n^-} \left(\frac{dP_{f_n^+}}{dP_{f_n^-}}(Z^n) \ge t \right). \tag{23}$$

When $f^\circ = f_n^-$ holds, the Radon–Nikodym derivative appearing in (23) can be explicitly written as

$$\frac{dP_{f_n^+}}{dP_{f_n^-}}(Z^n)$$
$$= \exp \left(\frac{1}{2\sigma^2} \sum_{i=1}^n (Y_i - f_n^-(X_i))^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (Y_i - f_n^+(X_i))^2 \right)$$
$$= \exp \left(\frac{1}{2\sigma^2} \sum_{i=1}^n \left(\xi_i^2 - (\xi_i + f_n^-(X_i) - f_n^+(X_i))^2 \right) \right)$$
$$= \exp \left(\frac{2}{2\sigma^2 \sqrt{n}} \sum_{i=1}^n \xi_i \varphi(X_i) - \frac{1}{2\sigma^2 n} \sum_{i=1}^n \varphi(X_i)^2 \right). \tag{24}$$

Let us consider the right-hand side of (24). First, $\sum_{i=1}^n \xi_i \varphi_1(X_i)$ is a sum of independent symmetric random variables, and so the sum itself is also symmetric (X is *symmetric* if $-X$ has the same distribution as X), and thus we have $P\left(\sum_{i=1}^n \xi_i \varphi_1(X_i) \ge 0 \right) \ge 1/2$. Second, for any $s > 0$, by Markov's inequality,

$$P \left(\frac{1}{2\sigma^2 n} \sum_{i=1}^n \varphi(X_i)^2 \ge s \right) \le \frac{1}{s} E \left[\frac{1}{2\sigma^2 n} \sum_{i=1}^n \varphi(X_i)^2 \right] = \frac{1}{2\sigma^2 s}$$

holds. We determine $s = 2/\sigma^2$ to obtain the evaluation

$$P \left(-\frac{1}{2\sigma^2 n} \sum_{i=1}^n \varphi_1(X_i)^2 \ge -\frac{2}{\sigma^2} \right) \ge \frac{3}{4}.$$

Finally, we have

$$P \left(\frac{2}{2\sigma^2 \sqrt{n}} \sum_{i=1}^n \xi_i \varphi_1(X_i) - \frac{1}{2\sigma^2 n} \sum_{i=1}^n \varphi(X_i)^2 \ge -\frac{2}{\sigma^2} \right) \ge \frac{1}{4}. \tag{25}$$

By (23)–(25) and letting $t = e^{-2/\sigma^2}$, we have

$$\inf_{\widehat{f}} \sup_{f \in \mathcal{I}} P_f \left(\|f - \widehat{f}\|_{L^2} \ge \frac{1}{2\sqrt{n}} \right) \ge \frac{e^{-2/\sigma^2}}{1 + e^{-2/\sigma^2}} \cdot \frac{1}{4} \ge \frac{1}{8} e^{-2/\sigma^2},$$

and so we finally obtain the evaluation

$$\inf_{\widehat{f}} \sup_{f \in \mathcal{I}} \mathrm{E}_f \left[\|f - \widehat{f}\|_{L^2}^2 \right] \geq \frac{e^{-2/\sigma^2}}{32n},$$

and this is the desired result.

B.5. Proof of Lemma 4.7

First, we prove the lower-bound part, partially using the proofs in Donoho (1996). When we consider the covering entropy, we have only to consider the coefficients, as $\varphi = (\varphi_i)_{i=1}^\infty$ is an orthonormal set. Thus, define $\mathcal{A} \subset \ell^2$ as

$$\mathcal{A} := \left\{ a \in \ell^2 \ \middle| \ \|a\|_{w\ell^p} \leq C_1, \ \sum_{i=m+1}^\infty a_i^2 \leq C_2 m^{-\beta}, \ m = 1, 2, \ldots \right\}. \tag{26}$$

Then it suffices to evaluate $V(\varepsilon) := V_{(\mathcal{A}, \|\cdot\|_{\ell^2})}(\varepsilon)$.

For each $k = 1, 2, \ldots$, let $a^{(k)} \in \ell^2$ be defined as

$$a^{(k)} := (\underbrace{C_1 k^{-1/p}, \ldots, C_1 k^{-1/p}}_{k \text{ times}}, 0, 0, \ldots).$$

Then $\|a^{(k)}\|_{w\ell^p} = C_1$ holds. Let us consider the second condition. For $1 \leq m \leq k$,

$$m^\beta \sum_{i=m+1}^\infty (a_i^{(k)})^2 = m^\beta (k-m)(C_1 k^{-1/p})^2 \tag{27}$$

holds. Since $x^\beta(k-x)$ is maximized over $x \in [0, k]$ at $x = \frac{\beta}{1+\beta}k$, the left-hand side of (27) is bounded independent of m as

$$\sup_{m \geq 1} m^\beta \sum_{i=m+1}^\infty (a_i^{(k)})^2 \leq C_1^2 \frac{\beta^\beta}{(1+\beta)^{1+\beta}} k^{1+\beta-2/p} \leq C_1^2 \frac{\beta^\beta}{(1+\beta)^{1+\beta}},$$

where we have used the assumption $\beta \leq 2\alpha$ for the latter inequality. If we define a constant

$$C := \min \left\{ 1, \ \frac{C_2^{1/2}(1+\beta)^{(1+\beta)/2}}{C_1 \beta^{\beta/2}} \right\},$$

then each $Ca^{(k)}$ is an element of $\mathcal{A}$. For each k, let us consider a hyperrectangle defined as

$$\mathcal{A}_k := \{a \in \ell^2 \mid |a_i| \leq a_i^{(k)}, \ i = 1, 2, \ldots\}.$$

Obviously, each $\mathcal{A}_k$ is a subset of $\mathcal{A}$ (this actually is based on the fact that φ is an unconditional basis of $\mathcal{I}_\varphi^p$), and so we have $V(\varepsilon) \geq V_{(\mathcal{A}_k, \|\cdot\|_{\ell^2})}(\varepsilon)$. For each pair of distinct vertices of $\mathcal{A}_k$ (which has 2^k vertices), the ℓ^2 distance between the two is at least $CC_1 k^{-1/p}$, and so, by setting $\delta = k^{-1/p}$ in Lemma A.2, we have, for A appearing in the lemma,

$$V \left(\frac{CC_1}{2} k^{1/2-1/p} \right) \geq V_{(\mathcal{A}_k, \|\cdot\|_{\ell^2})} \left(\frac{CC_1}{2} k^{1/2-1/p} \right) \geq Ak$$

for each k. If we write $C' = CC_1/2$, then we have $V(C'k^{-\alpha}) \geq Ak$. Therefore, for $\varepsilon \in [C'2^{-(j+1)\alpha}, C'2^{-j\alpha}]$,

$$V(\varepsilon) \geq V(C'2^{-j\alpha}) \geq 2^j A = 2^{j+1} \frac{A}{2} \geq \frac{A}{2} \left(\frac{\varepsilon}{C'} \right)^{-1/\alpha} = c\varepsilon^{-1/\alpha}$$

holds, where $c := AC'^{1/\alpha}/2$. This evaluation holds only for $0 < \varepsilon \leq C'$; however, we have $V(\varepsilon) \geq 1$ for $\varepsilon > C'$, and hence $V(\varepsilon) \geq C'^{1/\alpha}\varepsilon^{-1/\alpha}$ holds. Thus, we have reached the desired result.

Then, we deal with the upper-bound part. By the same logic as in the proof of lower-bound, it suffices to evaluate the metric

entropy of $\mathcal{A}$ in (26). Let $V(\varepsilon) := V_{(\mathcal{A}, \|\cdot\|_{\ell^2})}(\varepsilon)$ similarly. For an arbitrary element $a \in \mathcal{A}$, let i_j be the index of the term of a having the jth largest absolute value; i.e.,

$$|a_{i_1}| \geq |a_{i_2}| \geq \cdots,$$

which is a permutation of $(|a_i|)_{i=1}^\infty$. By (6) and the definition of $\mathcal{A}$,

$$\sum_{j=k+1}^\infty a_{i_j}^2 \leq \sum_{j=k+1}^\infty C_1 j^{-2/p} \leq \int_k^\infty C_1 x^{-2/p}\,dx = \frac{C_1}{2\alpha} k^{-2\alpha}$$

holds for each k. Also, by the second condition of $\mathcal{A}$,

$$\sum_{i \geq k^{2\alpha/\beta}+1} a_i^2 \leq C_2 k^{-\beta \cdot \frac{2\alpha}{\beta}} = C_2 k^{-2\alpha}$$

holds. Thus, if we define $\widetilde{a} := (a_1, \ldots, a_{\lceil k^{2\alpha/\beta} \rceil}, 0, 0, \ldots)$ and $\widetilde{i}_1, \widetilde{i}_2, \ldots$ similarly,

$$b := (b_i)_{i=1}^\infty, \quad b_i := \begin{cases} \widetilde{a}_i \ (= a_i) & \text{if } i \in \{\widetilde{i}_1, \ldots, \widetilde{i}_k\} \\ 0 & \text{otherwise,} \end{cases}$$

satisfies $\|a - b\|_{\ell^2} \leq \sqrt{C_1/2\alpha + C_2} \cdot k^{-\alpha}$. Then its quantization

$$\widetilde{b} := \left(\mathrm{sgn}(b_i) \frac{\lfloor k^{1/2+\alpha}|b_i| \rfloor}{k^{1/2+\alpha}} \right)_{i=1}^\infty$$

satisfies $\|b - \widetilde{b}\|_{\ell^2} \leq k^{-\alpha}$ as b has at most k nonzero terms. Since $|b_i| \leq C_1$, the number of values possibly taken by $\widetilde{b}_i$ is at most $2C_1 k^{1/2+\alpha}+1$. Hence, the logarithm of the number of such $\widetilde{b}$ values can be upper-bounded by

$$\log \left(\binom{\lceil k^{2\alpha/\beta} \rceil}{k} (2C_1 k^{1/2+\alpha} + 1)^k \right)$$

$$\leq \frac{2\alpha}{\beta} k \log k + \left(\frac{1}{2} + \alpha \right) k \log k + k \log(2C_1 + 1)$$

$$\leq C_0 k(\log k + 1), \tag{28}$$

where $C_0 > 0$ is a constant. Since $\|a - \widetilde{b}\|_{\ell^2} \leq (1 + \sqrt{C_1/2\alpha + C_2})k^{-\alpha}$ holds, if we take $k \sim \varepsilon^{-1/\alpha}$ in the same way as used in the proof of lower-bound, then we reach the conclusion.

Remark B.4. In the case in which $\beta \geq 2\alpha$ holds, we can obtain a more accurate bound by using Stirling's approximation. However, we can see that such a case no longer requires the concept of weak ℓ^p norms, or else its conditions are too strong. Therefore, we have not treated this case.

B.6. Proof of Theorem 4.15

This proof is a refinement of the proof of Theorem 1 in Zhang et al. (2002). First, we prove the following lemma.

Lemma B.5. *Let $\gamma \in (0, 1)$, and let m be a positive integer such that $m \leq n^\gamma \leq 2m$. For $k = 0, \ldots, m - 1$, let A_k be the number of $X_1, \ldots, X_n$ contained in $[k/m, (k+1)/m)$. Then there exists a constant $c = c(\gamma) > 0$ such that*

$$\mathrm{P} \left(\max_{0 \leq k \leq m-1} A_k \geq \frac{cn}{m} \right) \leq 2^{-n^{1-\gamma}}$$

holds for any m, n satisfying the condition.

Proof. For a fixed k, A_k can be written as $A_k = \sum_{j=1}^n \eta_j$, where $(\eta_j)_{j=1}^n$ is an i.i.d. sequence with $\mathrm{P}(\eta_j = 0) = 1/m$ and $\mathrm{P}(\eta_j = 1) = 1 - 1/m$. Then, by Chernoff's inequality, we have for $t > 0$

$$\mathrm{P} \left(A_k \geq \frac{cn}{m} \right) \leq e^{-cnt/m} \mathrm{E} \left[e^{A_k t} \right] = e^{-cnt/m} \mathrm{E} \left[e^{\eta_1 t} \right]^n$$

$$= e^{-cnt/m} \left(1 - \frac{1}{m} + \frac{1}{m} e^t \right)^n.$$

Setting $t = \log 2$ and assuming $c > \log 2$, we obtain

$$P\left(A_k \geq \frac{cn}{m}\right) \leq 2^{-cn/m}\left(1 + \frac{1}{m}\right)^n \leq (2^{-c}e)^{n/m} \leq (2^{-c}e)^{2n^{1-\gamma}},$$

where we have used the fact that $(1+1/x)^x$ is increasing on $x > 0$. Then we finally have

$$P\left(\max_{0 \leq k \leq m-1} A_k \geq \frac{cn}{m}\right) \leq m(2^{-c}e)^{2n^{1-\gamma}} \leq n^\gamma (2^{-c}e)^{2n^{1-\gamma}}$$

$$= 2^{-n^{1-\gamma}} \cdot n^\gamma \left(2^{-(2c-1)}e^2\right)^{n^{1-\gamma}}.$$

Considering the logarithm of the last term, it is sufficient to take c as large enough to satisfy

$$(2c - 1)\log 2 \geq 2 + \gamma \max_{n \geq 1} \frac{\log n}{n^{1-\gamma}},$$

and we obtain the conclusion. □

We now prove Theorem 4.15. Let $\gamma = \frac{1}{1+\beta}$, and let

$$R^* := \inf_{\widehat{f}:\text{linear}} \sup_{f^\circ \in \mathcal{J}_\psi^p(C_1, C_2, \beta)} E\left[\|\widehat{f} - f^\circ\|_{L^2}^2\right].$$

Fix a linear estimator $\widehat{f}(x) = \sum_{i=1}^n Y_i \varphi_i(x; X^n)$. For any $f^\circ \in BV(C)$, we have by Fubini's theorem

$$R^* \geq E\left[\|\widehat{f} - f^\circ\|_{L^2}^2\right]$$

$$= E\left[\left\|\sum_{i=1}^n f^\circ(X_i)\varphi_i(\cdot; X^n) - f^\circ\right\|_{L^2}^2\right] + \sigma^2 \sum_{i=1}^n E\left[\|\varphi_i(\cdot; X^n)\|_{L^2}^2\right]. \tag{29}$$

Take a sufficiently large n, and let m be a power of 2 in $[\frac{1}{2}n^\gamma, n^\gamma]$. Notice that it holds that $m \leq n^\gamma \leq 2m$. Then there exists an integer $0 \leq k < m$ such that

$$\int_{k/m}^{(k+1)/m} E\left[\sum_{i=1}^n \varphi_i(x; X^n)^2\right] dx \leq \frac{R^*}{\sigma^2 m}. \tag{30}$$

Let $f^\circ = Fm^{-\beta/2} \cdot m^{1/2} \psi(m \cdot - k)$, where $F := \min\{C_1, C_2^{1/2}\}$. Then we have $f^\circ \in \mathcal{J}_\psi^p(C_1, C_2, \beta)$. Let A denote an event assured to have a probability of at least $1 - 2^{-n^{1-\gamma}}$ in Lemma B.5, and we obtain

$$\int_{k/m}^{(k+1)/m} E\left[\left(\sum_{i=1}^n f^\circ(X_i)\varphi_i(x; X^n)\right)^2, A\right] dx$$

$$\leq \int_{k/m}^{(k+1)/m} E\left[\left(\sum_{i=1}^n f^\circ(X_i)^2\right)\left(\sum_{i=1}^n \varphi_i(x; X^n)^2\right), A\right] dx$$

$$\leq M \cdot \frac{cn}{m}(Fm^{(1-\beta)/2}\|\psi\|_\infty)^2$$

$$\times \int_{k/m}^{(k+1)/m} E\left[\sum_{i=1}^n \varphi_i(x; X^n)^2, A\right] dx$$

$$\leq \frac{McF^2\|\psi\|_\infty^2}{\sigma^2} \cdot \frac{R^*n}{m^{1+\beta}}, \tag{31}$$

where M is the number of sections $[\ell, \ell+1)$ such that ℓ is an integer and $[\ell, \ell+1) \cap \text{supp}(\psi)$ is not empty. The last inequality has been derived by (30). By (29), (31), and the triangle inequality,

$$\sqrt{R^*} \geq \left(\int_{k/m}^{(k+1)/m} E\left[\left(\sum_{i=1}^n f^\circ(X_i)\varphi_i(x; X^n) - f^\circ(x)\right)^2, A\right] dx\right)^{1/2}$$

$$\geq \left(P(A)\|f^\circ\|_{L^2}^2\right)^{1/2} - \left(E\left[\left(\sum_{i=1}^n f^\circ(X_i)\varphi_i(x; X^n)\right)^2, A\right] dx\right)^{1/2}$$

$$\geq (1 - 2^{-n^{1-\gamma}})^{1/2}Fm^{-\beta/2}$$

$$- \left(\frac{McF^2\|\psi\|_\infty^2}{\sigma^2}\right)^{1/2} m^{-(1+\beta)/2}n^{1/2}\sqrt{R^*}.$$

Since n is sufficiently large, we can assume $1 - 2^{-n^{1-\gamma}} \geq 1/2$. Define a constant G by

$$G := \left(\frac{McF^2\|\psi\|_\infty^2}{\sigma^2}\right)^{1/2}.$$

We have $m^{1+\beta} = m^{1/\gamma} \geq (\frac{1}{2}n^\gamma)^{1/\gamma} = 2^{-1/\gamma}$ by assumption, and so we obtain

$$R^* \geq \frac{(F^2/2)m^{-\beta}}{(1 + Gm^{-(1+\beta)/2}n^{1/2})^2} \geq \frac{F^2}{2(1 + 2^{1/\gamma}G)^2}m^{-\beta}$$

$$\geq \frac{F^2}{2(1 + 2^{1/\gamma}G)^2}n^{-\frac{\beta}{1+\beta}}$$

as desired.

B.7. Proof of Lemma 5.2

(Mainly following Suzuki, 2019.) For $f \in \mathcal{N}(L, S, D, B)$ expressed as

$$f = W_{L+1} \circ \rho(W_L \cdot - v_L) \circ \cdots \circ \rho(W_1 \cdot - v_1),$$

let us define

$$\mathcal{A}_k(f) := \rho(W_{k-1} \cdot - v_{k-1}) \circ \cdots \circ \rho(W_1 \cdot - v_1),$$
$$\mathcal{B}_k(f) := W_{L+1} \circ \rho(W_L \cdot - v_L) \circ \cdots \circ \rho(W_k \cdot - v_k)$$

for $k = 1, \ldots, L+1$, where $\mathcal{A}_1(f)$ denotes the identity map, and $\mathcal{B}_{L+1}(f) = W_{L+1}$. Then $f = \mathcal{B}_{k+1}(f) \circ \rho(W_k \cdot - v_k) \circ \mathcal{A}_k(f)$ holds for $k = 1, \ldots, L$. Here, notice that for each $x \in [0, 1]^d$ and $1 \leq k \leq L+1$,

$$\|\mathcal{A}_k(f)(x)\|_\infty \leq D\|W_{k-1}\|_\infty\|\mathcal{A}_{k-1}(f)(x)\|_\infty + \|v_{k-1}\|_\infty$$
$$\leq DB\|\mathcal{A}_{k-1}(x)\|_\infty + B$$
$$\leq B + DB^2 + D^2B^3 + \cdots + D^{k-2}B^{k-1} + D^{k-1}B^{k-1}$$
$$\leq B^{k-1}(D+1)^{k-1} \tag{32}$$

holds, where we have used the assumption $B \geq 1$ at the last inequality. Also, the Lipschitz continuity of $\mathcal{B}_k(f)$ can be derived as

$$\|\mathcal{B}_k(f)(x) - \mathcal{B}_k(f)(x')\|_\infty \leq (BD)^{L-k+2}\|x - x'\|_\infty. \tag{33}$$

Let $\varepsilon > 0$. Suppose $f, g \in \mathcal{N}(L, S, D, B)$ satisfy

$$f = W_{L+1} \circ \rho(W_L \cdot - v_L) \circ \cdots \circ \rho(W_1 \cdot - v_1),$$
$$g = W'_{L+1} \circ \rho(W'_L \cdot - v'_L) \circ \cdots \circ \rho(W'_1 \cdot - v'_1)$$

and $\|W_i - W'_i\|_\infty \leq \varepsilon$, $\|v_i - v'_i\|_\infty \leq \varepsilon$ for each i. Then we have by (32) and (33)

$$|f(x) - g(x)| \leq \sum_{k=1}^L |\mathcal{B}_{k+1}(f) \circ \rho(W_k \cdot - v_k) \circ \mathcal{A}_k(g)(x)$$
$$- \mathcal{B}_{k+1}(f) \circ \rho(W'_k \cdot - v'_k) \circ \mathcal{A}_k(g)(x)|$$
$$+ |(W_{L+1} - W'_{L+1})\mathcal{A}_{L+1}(g)(x)|$$
$$\leq \sum_{k=1}^L (BD)^{L-k+1}\left(\varepsilon D(B(D+1))^{k-1} + \varepsilon\right)$$
$$+ \varepsilon DB^L(D+1)^L$$
$$\leq \varepsilon(L+1)B^L(D+1)^{L+1}.$$

Therefore, for a fixed sparsity pattern, and letting $\varepsilon = ((L+1)B^L(D+1)^{L+1})^{-1}\delta$, the δ-covering number is bounded by

$$\left(\frac{2B}{\varepsilon}\right)^S = \delta^{-S}\left(2(L+1)B^{L+1}(D+1)^{L+1}\right)^S.$$

The number of such patterns is bounded by $\binom{(D+1)^L}{S} \leq (D+1)^{LS}$, and so the δ-covering entropy is bounded by

$$\log\left((d+1)^{LS}\delta^{-S}\left(2(L+1)B^{L+1}(D+1)^{L+1}\right)^S\right) \tag{34}$$

$$= S\log\left(2\delta^{-1}(L+1)(D+1)^{2L+1}B^{L+1}\right)$$

$$\leq 2S(L+1)\log\left(\frac{B(L+1)(D+1)}{\delta}\right), \tag{35}$$

as desired.

B.8. Proof of Theorem 5.10

Let N be an integer in $[n^{\frac{1}{2\alpha+1}}, 2n^{\frac{1}{2\alpha+1}}]$. Also, suppose that we have an integer m in $[\frac{2\alpha}{\beta(2\alpha+1)}\log_2 n, \frac{4\alpha}{\beta(2\alpha+1)}\log_2 n]$ (for sufficiently large n). Fix the target function

$$f^\circ = \sum_{(k,\ell)\in T_0} a_{k,\ell}\psi_{k,\ell}.$$

Then, let $(k^1,\ell^1),\dots,(k^N,\ell^N) \in T_0 \setminus T_m$ be the N (absolutely) largest coefficients $a_{k^1,\ell^1},\dots,a_{k^N,\ell^N}$. Let

$$T := \{(k^1,\ell^1),\dots,(k^N,\ell^N)\}.$$

Then we have

$$\left\|\sum_{(k,\ell)\in T} a_{k,\ell}\psi_{k,\ell} - f^\circ\right\|_{L^2}^2 = \sum_{(k,\ell)\in T_0\setminus T} a_{k,\ell}^2$$

$$\leq \sum_{(k,\ell)\in T_m} a_{k,\ell}^2 + \sum_{i=N+1}^{\infty} C_1(i^{-1/p})^2$$

$$\leq C_2 n^{-\frac{2\alpha}{2\alpha+1}} + C_1 \int_N^\infty x^{-2/p}\,dx$$

$$= C_2 n^{-\frac{2\alpha}{2\alpha+1}} + C_1\frac{1-p}{p}N^{-2\alpha} \leq \left(C_2 + C_1\frac{1-p}{p}\right)n^{-\frac{2\alpha}{2\alpha+1}}. \tag{36}$$

Next, we approximate $\sum_{(k,\ell)\in T} a_{k,\ell}\psi_{k,\ell}$ by some neural network. Now, $(k,\ell)\in T$ implies that

$$(k,\ell) = ((k_1,\dots,k_d),(\ell_1,\dots,\ell_d))$$

satisfies $\max_{1\leq i\leq d} k_i < m$. Let $\widetilde{\psi} \in \mathcal{N}(L_\varepsilon, S_\varepsilon, D_\varepsilon, B_\varepsilon)$ satisfy $\|\widetilde{\psi} - \psi\|_{L^2} \leq \varepsilon$. Since we have

$$\psi_{k,\ell}(x_1,\dots,x_d) = 2^{\frac{k_1+\cdots+k_d}{2}}\psi\left(2^{k_1}t_1 - \ell_1,\dots,2^{k_d}t_d - \ell_d\right),$$

we can construct the approximator

$$\widetilde{f} \in \mathcal{N}(L_\varepsilon + 2, N(S_\varepsilon + 2D_\varepsilon d + d^2 + d + 1), ND_\varepsilon, \max\{B_\varepsilon, 2^m\}) \tag{37}$$

in a manner similar to that shown in Fig. 3 such that

$$\left\|\widetilde{f} - \sum_{(k,\ell)} a_{k,\ell}\psi_{k,\ell}\right\|_{L^2} \tag{38}$$

$$\leq \sum_{(k,\ell)\in T} |a_{k,\ell}|\|\widetilde{\psi}_{k,\ell} - \psi_{k,\ell}\|_{L^2} \leq C_1 N\varepsilon \leq 2C_1\varepsilon n^{\frac{1}{2\alpha+1}}$$

holds. If we determine $\varepsilon = 1/n$ and define $\mathcal{N}_F$ by using the set defined in (37), we have, by Lemma 5.2 and the assumption,

$$V_{(\mathcal{N}_F,\|\cdot\|_{L^\infty})}\left(\frac{1}{n}\right)$$

$$\leq 2N\left(S_{1/n} + 2D_{1/n}d + d^2 + d + 1\right)(L_{1/n} + 3)$$

$$\quad\times \log\left(\max\{B_{1/n}, 2^m\}(L_{1/n} + 3)(ND_{1/n} + 1)n\right)$$

$$\leq 4n^{\frac{1}{2\alpha+1}}\left(C_1'(1 + 2d)\log n + d^2 + d + 1\right)(C_1'\log n + 3)$$

$$\quad\cdot\left(\log\log(C_2'n) + \frac{2\alpha}{\beta(2\alpha+1)}\log n + \log(C_1'\log n + 3)\right.$$

$$\quad\left. + \log(C_1'N\log n + 1) + \log n\right)$$

$$\leq C'n^{\frac{1}{2\alpha+1}}(\log n)^3 \tag{39}$$

for some constant $C' > 0$.

Combining (36), (37), (39), and Theorem 2.6, we obtain an evaluation

$$R(\widehat{f}, f^\circ) \leq 4\left(\left(C_2 + C_1\frac{1-p}{p}\right)^{1/2}n^{-\frac{\alpha}{2\alpha+1}} + 2C_1 n^{-\frac{2\alpha}{2\alpha+1}}\right)^2$$

$$\quad + C''\left(C'(F^2 + \sigma^2)n^{-\frac{2\alpha}{2\alpha+1}}(\log n)^3 + \frac{F+\sigma}{n}\right)$$

$$\leq CF^2 n^{-\frac{2\alpha}{2\alpha+1}}(\log n)^3$$

for some $C > 0$, where $\widehat{f}$ denotes the empirical risk minimizer over $\mathcal{N}_F$.

B.9. Proof of Theorem 5.14

We consider functions expressed as $f^N = \sum_{i=1}^N c_i f(A_i \cdot -b_i) \in \mathcal{N}^N(L, S, D, B)$, where $f \in \mathcal{N}(L, S, D, B)$. Then it suffices to consider f defined over $[-B(d+1), B(d+1)]^d \subset [-B(D+1), B(D+1)]^d$. This changes evaluation (32) to $\|\mathcal{A}_k(f)(x)\|_\infty \leq B^k(D+1)^k$, and so $B^{L+1}(D+1)^{L+1}$ in (34) is replaced by $B^{L+2}(D+1)^{L+2}$ in the evaluation of the covering entropy with the domain limited to $[-B(D+1), B(D+1)]^d$. However, the upper bound (35) is still valid if $L \geq 2$, and so we have a set $\mathcal{N}_\varepsilon$ for each $0 < \varepsilon < 1$ satisfying

- $\mathcal{N}_\varepsilon$ is an ε-covering of $\mathcal{N}(L, S, D, B)$ with respect to $\|\cdot\|_{L^\infty}$, where the domain is limited to $[-B(D+1), B(D+1)]^d$;
- $\log|\mathcal{N}_\varepsilon| \leq 2S(L+1)\log\left(\frac{B(L+1)(D+1)}{\varepsilon}\right)$.

For $x \in [0,1]^d$, we have $|f(A_i x - b_i)| \leq B^{L+2}(D+1)^{L+2}$ by an evaluation similar to the one in (32). Also, we have the Lipschitz continuity $|f(x) - f(x')| \leq B^{L+1}D^{L+1}\|x - x'\|_\infty$, similar to (33).

Let us consider two functions in $\mathcal{N}^N(L, S, D, B)$ expressed as

$$f^N = \sum_{i=1}^N c_i f(A_i \cdot -b_i),$$

$$g^N = \sum_{i=1}^N c_i' g(A_i' \cdot -b_i'), \quad f, g \in \mathcal{N}(L, S, D, B),$$

such that $\|f - g\|_{L^\infty([-B(D+1),B(D+1)]^d)} \leq \varepsilon$ holds and $\|A_i - A_i'\|_\infty, \|b_i - b_i'\|_\infty, |c_i - c_i'| \leq \varepsilon$ holds for each i. Then we have for each

$x \in [0, 1]^d$

$|f^N(x) - g^N(x)|$

$$\leq \sum_{i=1}^{N} |c_i f(A_i x - b_i) - c_i' g(A_i' x - b_i')|$$

$$\leq \sum_{i=1}^{N} |c_i - c_i'| |f(A_i x - b_i)| + \sum_{i=1}^{N} |c_i'| |f(A_i x - b_i) - g(A_i x - b_i)|$$

$$+ \sum_{i=1}^{N} |c_i'| |g(A_i x - b_i) - g(A_i' x - b_i')|$$

$$\leq N\varepsilon B^{L+2}(D+1)^{L+2} + NB\varepsilon + NB \cdot B^{L+1}D^{L+1}(\varepsilon d + \varepsilon)$$

$$\leq 3N\varepsilon B^{L+2}(D+1)^{L+2}.$$

If we determine $\varepsilon = (3NB^{L+2}(D+1)^{L+2})^{-1}\delta$ for $0 < \delta < 1$, the δ-covering entropy of $\mathcal{N}^N(L, S, D, B)$ is now bounded by

$$\log\left(\left(\frac{2B}{\varepsilon}\right)^{N(d^2+d+1)} |\mathcal{N}_\varepsilon|\right)$$

$$= N(d^2 + d + 1)\log\left(\frac{2B}{\varepsilon}\right) + \log|\mathcal{N}_\varepsilon|$$

$$\leq N(d+1)^2 \log\left(\frac{6NB^{L+3}(D+1)^{L+2}}{\delta}\right)$$

$$+ 2S(L+1)\log\left(\frac{3NB^{L+3}(L+1)(D+1)^{L+3}}{\delta}\right)$$

$$\leq \left(N(d+1)^2 + 2S(L+1)\right)(L+3)\log\left(\frac{NB(L+1)(D+1)}{\delta}\right),$$

as desired.

B.10. Proof of Lemma A.1

The probability law P_f generated by f is regarded as being on $\mathbb{R}^d \times \mathbb{R}$. Hence, its density at $z = (x, y)$ is

$$p_f(z) = p_X(x)p_{Y|X}(y \mid x) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(y-f(x))^2}{2\sigma^2}\right).$$

As p_g can be calculated in the same way, we have

$$d_{KL}(f, g)^2 = \int_{\mathbb{R}^d \times \mathbb{R}} p_f(z) \log \frac{p_f(z)}{p_g(z)}\, dz$$

$$= \int_{\mathbb{R}^d \times \mathbb{R}} p_f(z) \frac{1}{2\sigma^2}\left((y - g(x))^2 - (y - f(x))^2\right) dz.$$

This coincides with the expectation of $\frac{1}{2\sigma^2}\left((Y - g(X))^2 - (Y - f(X))^2\right)$ with $Y = f(X) + \xi$. The term in parentheses is calculated as

$$E\left[(Y - g(X))^2 - (Y - f(X))^2\right]$$

$$= E\left[(f(X) - g(X) + \xi)^2 - \xi^2\right]$$

$$= E\left[(f(X) - g(X))^2 - 2\xi(f(X) - g(X))\right]$$

$$= E\left[(f(X) - g(X))^2\right] - 2E[\xi] \cdot E[f(X) - g(X)] = \|f - g\|_{L^2}^2,$$

where we have used the facts that each X follows the uniform distribution over $[0, 1]^d$ and that each ξ is independent of X. Thus, we obtain the desired result.

References

Bishop, C. M. (2006). *Pattern recognition and machine learning (information science and statistics)*. Springer.

Bölcskei, H., Grohs, P., Kutyniok, G., & Petersen, P. (2017). Optimal approximation with sparsely connected deep neural networks. arXiv preprint arXiv:1705.01714.

Cai, T. T., & Low, M. G. (2004). Minimax estimation of linear functionals over nonconvex parameter spaces. *The Annals of Statistics*, 32, 552–576.

Candès, E. J., & Wakin, M. B. (2008). An introduction to compressive sampling [a sensing/sampling paradigm that goes against the common knowledge in data acquisition]. *IEEE Signal Processing Magazine*, 25, 21–30.

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

Daubechies, I. (1992). *Ten lectures on wavelets*. Society for Industrial and Applied Mathematics.

Donoho, D. L. (1993). Unconditional bases are optimal bases for data compression and for statistical estimation. *Applied and Computational Harmonic Analysis*, 1, 100–115.

Donoho, D. L. (1996). Unconditional bases and bit-level compression. *Applied and Computational Harmonic Analysis*, 3, 388–392.

Donoho, D. L., & Johnstone, J. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81, 425–455.

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

Donoho, D. L., Liu, R. C., & MacGibbon, B. (1990). Minimax risk over hyperrectangles. *and implications. The Annals of Statistics*, 141, 6–1437.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning*. Springer.

Giné, E., & Koltchinskii, V. (2006). Concentration inequalities and asymptotic results for ratio type empirical processes. *The Annals of Probability*, 34, 1143–1216.

Glorot, X., Bordes, A., & Bengio, Y. (2011). Deep sparse rectifier neural networks. In *Proceedings of the 14th International conference on artificial intelligence and statistics* (pp. 315–323).

Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). *Deep learning*. MIT Press Cambridge.

Györfi, L., Kohler, M., Krzyzak, A., & Walk, H. (2006). *A distribution-free theory of nonparametric regression*. Springer.

Imaizumi, M., & Fukumizu, K. (2019). Deep neural networks learn non-smooth functions effectively. In *Proceedings of the 22nd international conference on artificial intelligence and statistics*.

Keiper, S., Kutyniok, G., & Petersen, P. (2017). DGD approximation theory workshop.

Koltchinskii, V. (2006). Local rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34, 2593–2656.

Korostelev, A. P., & Tsybakov, A. B. (1993). *Minimax theory of image reconstruction*. Springer.

Lafferty, J., Liu, H., & Wasserman, L. (2008). Concentration on measure. http://www.stat.cmu.edu/~larry/=sml/Concentration.pdf.

Peetre, J. (1976). *New thoughts on besov spaces*. Mathematics Department, Duke University.

Petersen, P., & Voigtlaender, F. (2018). Optimal approximation of piecewise smooth functions using deep ReLU neural networks. *Neural Networks*, 108, 296–330.

Raskutti, G., Wainwright, M. J., & Yu, B. (2011). Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Transactions on Information Theory*, 57, 6976–6994.

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

Schmidt-Hieber, J. (2017). Nonparametric regression using deep neural networks with ReLU activation function. *The Annals of Statistics* (in press) arXiv:1708.06633.

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

Stein, E. M., & Shakarchi, R. (2005). Real analysis: Measure theory, integration, and Hilbert spaces. In *Princeton Lectures in Analysis, Book 3*. Princeton University Press.

Suzuki, T. (2019). Adaptivity of deep ReLU network for learning in Besov and mixed smooth Besov spaces: optimal rate and curse of dimensionality. In *ICLR 2019*.

Tsybakov, A. B. (2008). *Introduction to nonparametric estimation*. Springer.

van der Vaart, A. W., & Wellner, J. A. (1996). *Weak convergence and empirical processes*. Springer.

Wang, Z., Paterlini, S., Gao, F., & Yang, Y. (2014). Adaptive minimax regression estimation over sparse ℓ_q-hulls. *Journal of Machine Learning Research (JMLR)*, 15, 1675–1711.

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

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

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