DEEP NEURAL NETWORKS LEARN NON-SMOOTH FUNCTIONS EFFECTIVELY

MASAAKI IMAIZUMI AND KENJI FUKUMIZU

The Institute of Statistical Mathematics

Abstract. We theoretically discuss why deep neural networks (DNNs) performs better than other models in some cases by investigating statistical properties of DNNs for non-smooth functions. While DNNs have empirically shown higher performance than other standard methods, understanding its mechanism is still a challenging problem. From an aspect of the statistical theory, it is known many standard methods attain optimal convergence rates, and thus it has been difficult to find theoretical advantages of DNNs. This paper fills this gap by considering learning of a certain class of non-smooth functions, which was not covered by the previous theory. We derive convergence rates of estimators by DNNs with a ReLU activation, and show that the estimators by DNNs are almost optimal to estimate the non-smooth functions, while some of the popular models do not attain the optimal rate. In addition, our theoretical result provides guidelines for selecting an appropriate number of layers and edges of DNNs. We provide numerical experiments to support the theoretical results.

1. Introduction

Deep neural networks (DNNs) have shown outstanding performance on various tasks of data analysis (Schmidhuber, 2015; LeCun *et al.*, 2015). Enjoying their flexible modeling by a multi-layer structure and many elaborate computational and optimization techniques, DNNs empirically achieve higher accuracy than many other machine learning methods such as kernel methods (Hinton *et al.*, 2006; Le *et al.*, 2011; Kingma and Ba, 2014). Hence, DNNs are employed in many successful applications, such as image analysis (He *et al.*, 2016), medical data analysis (Fakoor *et al.*, 2013), natural language processing (Collobert and Weston, 2008), and others.

Despite such outstanding performance of DNNs, little is yet known why DNNs outperform the other methods. Without sufficient understanding, practical use of DNNs could be inefficient or unreliable. To reveal the mechanism, numerous studies have investigated theoretical properties of neural networks from various aspects. with approximation theory, the expressive power of neural networks have been analyzed(Cybenko, 1989; Barron, 1993; Bengio and Delalleau, 2011; Montufar *et al.*, 2014; Yarotsky, 2017; Petersen and Voigtlaender, 2017), statistics and learning theories have elucidated generalization errors (Barron, 1994; Neyshabur *et al.*, 2015; Schmidt-Hieber, 2017; Zhang *et al.*, 2017; Suzuki, 2018), and optimization theory has discussed the landscape of the objective function and dynamics of learning(Baldi and

E-mail address: imaizumi@ism.ac.jp.

Date: February 14, 2018.

Hornik, 1989; Fukumizu and Amari, 2000; Dauphin *et al.*, 2014; Kawaguchi, 2016; Soudry and Carmon, 2016).

One limitation in the existing statistical analysis of DNNs is a *smoothness assumption* for data generating processes. It makes one of the reasons for difficulties, when we try to reveal the advantage of DNNs. In the statistical theory, it is assumed that data are generated from smooth (i.e. differentiable) functions, namely, data $\{(Y_i, X_i)\}$ are given

$$Y_i = f(X_i) + \xi_i,\ \xi_i \sim \mathcal{N}(0, \sigma^2),$$

where f is a β-times differentiable function with D-dimensional input. With this setting, however, not only DNNs but also other popular methods (kernel methods, Gaussian processes, series methods, and so on) achieve generalization errors bounded as

$$O\left(n^{-2\beta/(2\beta+D)}\right),$$

which is known to be optimal in the minimax sense (Stone, 1982; Tsybakov, 2009; Giné and Nickl, 2015). Hence, as long as we employ the smoothness assumption, it is not possible to show a theoretical evidence for the empirical advantage of DNNs.

This paper considers learning of *non-smooth* functions for the data generating processes to break the difficulty. We prove that DNNs certainly have a theoretical advantage under the non-smooth setting. Specifically, we discuss a nonparametric regression problem with a class of *piecewise smooth functions* which are non-smooth on boundaries of pieces in their domains. Then, we derive convergence rates of least square and Bayes estimators by DNNs with a ReLU activation as

$$O\left(\max\left\{n^{-2\beta/(2\beta+D)}, n^{-\alpha/(\alpha+D-1)}\right\}\right),$$

up to log factors (Theorems 1, 2, and Corollary 1). Here, α and β denote a degree of smoothness of piecewise smooth functions, and D is the dimensionality of inputs. We prove also that the convergence rate by DNNs is optimal in the minimax sense (Theorem 3). In addition, we show that some of other popular methods, such as kernel methods and orthogonal series methods with some specified bases, cannot estimate the piecewise smooth functions with the optimal convergence rate (Proposition 1 and 2). Notably, in contrast to these models, our result shows that DNNs with a ReLU achieve the optimal rate in estimating non-smooth functions, although the DNN realizes smooth functions. We provide some numerical results supporting our results.

Contributions of this paper are as follows:

- We derive the convergence rates of the estimators by DNNs for the class of piecewise smooth functions. Our convergence results are more general than existing studies, since the class is regarded as a generalization of smooth functions.
- We prove that DNNs theoretically outperform other standard methods for data from non-smooth generating processes, as a consequence the proved convergence rates.
- We provide a practical guideline on the structure of DNNs; namely, we show a necessary number of layers and parameters of DNNs to achieve the optimal convergence rate. It is shown in Table 1.

All of the proofs are deferred to the supplementary material.

Element	Number
# of layers	$\leqslant c(1+\max\{\beta/D, \alpha/2(D-1)\})$
# of parameters	$c' n^{\max\{D/(2\beta+D),(2D-2)/(2\alpha+2D-2)\}}$

Table 1. Architecture for DNNs which are necessary to achieve the optimal convergence rate. $c, c' > 0$ are some constants.

1.1. **Notation.** We use notations $I := [0,1]$ and $\mathbb{N}$ for natural numbers. The j-th element of vector b is denoted by b_j, and $\|\cdot\|_q := (\sum_j b_j^q)^{1/q}$ is the q-norm ($q \in [0,\infty]$). $\mathrm{vec}(\cdot)$ is a vectorization operator for matrices. For $z \in \mathbb{N}$, $[z] := \{1, 2, \ldots, z\}$ is the set of positive integers no more than z. For a measure P on I and a function $f : I \to \mathbb{R}$, $\|f\|_{L^2(P)} := (\int_I |f(x)|^2 dP(x))^{1/2}$ denotes the $L^2(P)$ norm. $\otimes$ denotes a tensor product, and $\bigotimes_{j\in[J]} x_j := x_1 \otimes \cdots \otimes x_J$ for a sequence $\{x_j\}_{j\in[J]}$.

2. Regression with Non-Smooth Functions

We formulate a regression problem when a function for generating data is non-smooth. Firstly, we summarize a brief outline of the regression problem, and secondly, we introduce a class of non-smooth functions.

2.1. **Regression Problem.** In this paper, we use the D-dimensional cube I^D ($D \geqslant 2$) for the domain of data. Suppose we have a set of observations $(X_i, Y_i) \in I^D \times \mathbb{R}$ for $i \in [n]$ which is independently and identically distributed with the data generating process

$$Y_i = f^*(X_i) + \xi_i, \tag{1}$$

where $f^* : I^D \to \mathbb{R}$ is an unknown true function and ξ_i is Gaussian noise with mean 0 and variance $\sigma^2 > 0$ for $i \in [n]$. We assume that the marginal distribution of X on I^D has a positive and bounded density function $P_X(x)$.

The goal of the regression problem is to estimate f^* from the set of observations $\mathcal{D}_n := \{(X_i, Y_i)\}_{i\in[n]}$. With an estimator $\widehat{f}$, its performance is measured by the $L^2(P_X)$ norm:

$$\|\widehat{f} - f^*\|^2_{L^2(P_X)} = \mathbb{E}_X\left[(\widehat{f}(X) - f^*(X))^2\right].$$

There are various methods to estimate f^* and their statistical properties are extensively investigated (For summary, see Wasserman (2006) and Tsybakov (2009)).

A classification problem can be also analyzed through the regression framework. For instance, consider a Q-classes classification problem with covariates X_i and labels $Z_i \in [Q]$ for $i \in [n]$. To describe the classification problem, we consider a Q-dimensional vector-valued function $f^*(x) = (f_1^*(x), ..., f_Q^*(x))$ and a generative model for Z_i as

$$Z_i = \operatorname*{argmax}_{q\in[Q]} f_q^*(X_i).$$

Here, estimating f^* can solve the classification problem. (For summary, see Steinwart and Christmann (2008)).

2.2. **Piecewise Smooth Functions.** To describe non-smoothness of functions, we introduce a notion of *piecewise smooth functions* which have a support divided into several pieces and smooth only within each of the pieces. On boundaries of the pieces, piecewise smooth functions are non-smooth, i.e. non-differentiable and even discontinuous. Figure 1 shows an example of piecewise smooth functions.

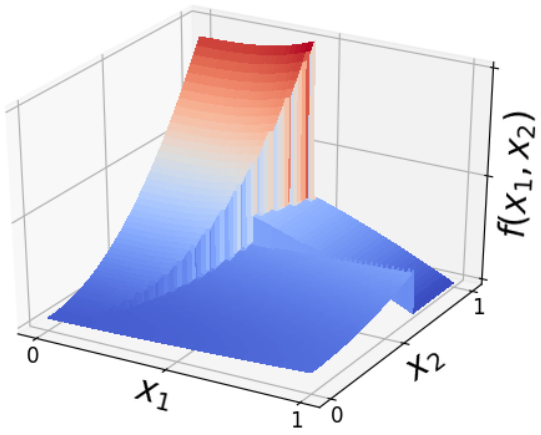

FIGURE 1. An example of piecewise smooth functions with a 2-dimensional input. The support $[0,1]^2$ is divided into three pieces and the function $f(x_1, x_2)$ is non-smooth (also discontinuous) on boundaries of the pieces.

As preparation, we introduce notions of (i) smooth functions and (ii) pieces in supports. Afterwards, we combine them and provide the notion of (iii) piecewise smooth functions.

(i). Smooth Functions

We introduce *the Hölder space* to describe smooth functions. With a parameter $\beta > 0$, the Hölder norm for $f : I^D \to \mathbb{R}$ is defined as

$$\|f\|_{H^\beta} := \max_{|a| \leqslant \lfloor \beta \rfloor} \sup_{x \in I^D} |\partial^a f(x)| + \max_{|a| = \lfloor \beta \rfloor} \sup_{x, x' \in I^D, x \neq x'} \frac{|\partial^a f(x) - \partial^a f(x')|}{|x - x'|^{\beta - \lfloor \beta \rfloor}},$$

where a denotes a multi-index of differentiation and ∂^a denotes a partial derivative. Then, the Hölder space H^β on I^D is defined as

$$H^\beta := \{f : I^D \to \mathbb{R} \mid \|f\|_{H^\beta} < \infty\}.$$

Intuitively, H^β contains functions such that they are $\lfloor \beta \rfloor$-times differentiable and the $\lfloor \beta \rfloor$-th derivatives are $\beta - \lfloor \beta \rfloor$-Hölder continuous.

The Hölder space is popularly used for representing smooth functions, and many statistical methods can effectively estimate functions in the Hölder space. (For summary, see Giné and Nickl (2015).)

(ii). Pieces in Supports

To describe pieces in supports, we introduce an extended notion of a *boundary fragment class* which is developed by Dudley (1974) and Mammen *et al.* (1999).

Preliminarily, we consider a sphere $\mathcal{S}^{D-1} := \{x \in \mathbb{R}^D : \|x\|_2 = 1\}$ in $\mathbb{R}^D$ and its center is the origin. With $J \in \mathbb{N}$, let $\{V_j\}_{j=1}^J$ be sets in $\mathcal{S}^{D-1}$ such as $\bigcup_{j \in [J]} \mathrm{cl}(V_j) = \mathcal{S}^{D-1}$ and $V_j \cap V_{j'} = \varnothing, \forall j \neq j', j, j' \in [J]$.

We provide a notion of boundaries of a piece in $\mathbb{R}^D$ using $\{V_j\}_{j\in[J]}$. Let $\bar{\mathcal{S}}^{D-1} := \{x \in [-1,1]^D : \|x\|_2 < 1\}$ be an open ball in $\mathbb{R}^D$, and $F_j : \bar{\mathcal{S}}^{D-1} \to V_j$ be a C^∞ surjective function for $j \in [J]$. With a parameter $\alpha \geqslant 1$, let $\mathcal{G}_{\alpha,J}$ be the set of boundaries, defined by

$$\mathcal{G}_{\alpha,J} := \left\{ (g_1, ..., g_D) \mid \text{injective}, g_d : S^{D-1} \to I, g_d \circ F_j \in H^\alpha(\bar{\mathcal{S}}^{D-1}), j \in [J], d \in [D], \right\},$$

where $H^\alpha(\bar{\mathcal{S}}^{D-1})$ denotes the Hölder space of smooth functions on $\bar{\mathcal{S}}^{D-1}$. Intuitively, boundaries $g = (g_1, ..., g_D)$ is $\lfloor\alpha\rfloor$-times differentiable expect at frontier points of V_j.

Given $g \in \mathcal{G}_{\alpha,J}$ as the boundary of a piece, we define $\text{Int}(g)$ as the interior of $g \in \mathcal{G}_{\alpha,J}$ (detailed definition is provided by Dudley (1974)). At last, we define $\mathcal{R}_{\alpha,J}$ as a set of pieces in I^D such as

$$\mathcal{R}_{\alpha,J} := \{\text{Int}(g) : g \in \mathcal{G}_{\alpha,J}\}.$$

Figure 2 shows a brief example.

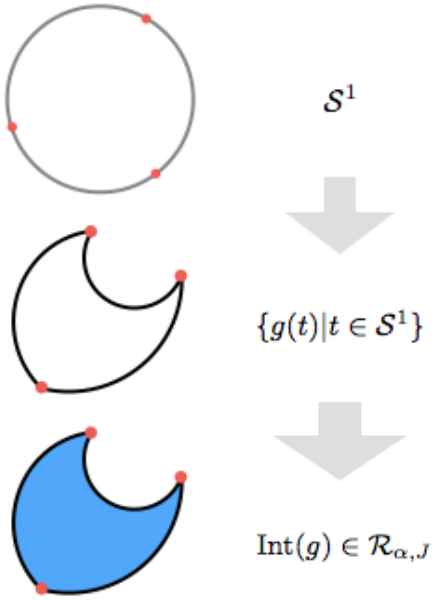

FIGURE 2. An example of pieces with $D = 2$ and $J = 3$. The top figure is a circle, and the middle figure is a boundary is obtained by reshaping the circle and it is smooth except the frontier points of V_j (the red dots). The bottom figure is the piece as $\text{Int}(g)$. The interior is shown as the blue area.

We mention that $\mathcal{R}_{\alpha,J}$ can describe a wide range of pieces (Dudley, 1974): $\mathcal{R}_{\alpha,J}$ with $\alpha = 2$ is dense in a set of all convex sets in I^D.

(iii). Piecewise Smooth Functions

Using H^β and $\mathcal{R}_{\alpha,J}$, we define piecewise smooth functions. Let $M \in \mathbb{N}$ be a number of pieces of the support I^D. With a piece $R \subset I^D$, let $\mathbf{1}_R : I^D \to \{0, 1\}$ be the indicator function

such that

$$\mathbf{1}_R(x) = \begin{cases} 1, & \text{if } x \in R, \\ 0, & \text{if } x \notin R. \end{cases}$$

We define a set of piecewise smooth functions as

$$\mathcal{F}_{M,J,\alpha,\beta} = \left\{ \sum_{m=1}^{M} f_m \otimes \mathbf{1}_{R_m} : f_m \in H^\beta, R_m \in \mathcal{R}_{\alpha,J} \right\}.$$

Since $f_m(x)$ realizes only when $x \in R_m$, the notion of $\mathcal{F}_{M,J,\alpha,\beta}$ can express a combination of smooth functions on each piece R_m. Hence, functions in $\mathcal{F}_{M,J,\alpha,\beta}$ are non-smooth (and even discontinuous) on boundaries of R_m. Obviously, $H^\beta \subset \mathcal{F}_{M,J,\alpha,\beta}$ with $M = 1$ and $R_1 = I^D$, hence the notion of piecewise smooth functions can describe a wider class of functions.

3. Analysis for Estimation by DNNs

In this section, we provide estimators for the regression problem by DNNs and derive their theoretical properties. Firstly, we define a statistical model by DNNs. Afterwards, we investigate two estimators by DNNs; a least square estimator and a Bayes-estimator.

3.1. Models by Deep Neural Networks.

Let $L \in \mathbb{N}$ be the number of layers in DNNs. For $\ell \in [L+1]$, let $D_\ell \in \mathbb{N}$ be the dimensionality of variables in the ℓ-th layer. For brevity, we set $D_{L+1} = 1$, i.e., the output is one-dimensional. We define $A_\ell \in \mathbb{R}^{D_{\ell+1} \times D_\ell}$ and $b_\ell \in \mathbb{R}^{D_\ell}$ be matrix and vector parameters to give the transform of ℓ-th layer. The *architecture* Θ of DNN is a set of L pairs of (A_ℓ, b_ℓ):

$$\Theta := ((A_1, b_1), ..., (A_L, b_L)).$$

We define $|\Theta| := L$ be a number of layers in Θ, $\|\Theta\|_0 := \sum_{\ell \in [L]} \|\operatorname{vec}(A_\ell)\|_0 + \|b_\ell\|_0$ as a number of non-zero elements in Θ, and $\|\Theta\|_\infty := \max\{\max_{\ell \in [L]} \|\operatorname{vec}(A_\ell)\|_\infty, \max_{\ell \in [L]} \|b_\ell\|_\infty\}$ be the largest absolute value of the parameters in Θ.

For an activation function $\eta : \mathbb{R}^{D'} \to \mathbb{R}^{D'}$ for each $D' \in \mathbb{N}$, this paper considers the ReLU activation $\eta(x) = (\max\{x_d, 0\})_{d \in [D']}$.

The model of neural networks with architecture Θ and activation η is the function $G_\eta[\Theta] : \mathbb{R}^{D_1} \to \mathbb{R}$, which is defined inductively as

$$G_\eta[\Theta](x) = x^{(L+1)},$$

with

$$\begin{aligned} x^{(1)} &:= x, \\ x^{(\ell+1)} &:= \eta(A_\ell x^{(\ell)} + b_\ell), \text{ for } \ell \in [L], \end{aligned}$$

where $L = |\Theta|$ is the number of layers. The set of model functions by DNNs is thus given by

$$\mathcal{F}_{NN,\eta}(S, B, L') := \left\{ G_\eta[\Theta] : I^D \to \mathbb{R} \mid \|\Theta\|_0 \leqslant S, \|\Theta\|_\infty \leqslant B, |\Theta| \leqslant L' \right\},$$

with $S \in \mathbb{N}$, $B > 0$, and $L' \in \mathbb{N}$. Here, S bounds the number of non-zero parameters of DNNs by Θ, namely, the number of edges of an architecture in the networks. This also describes sparseness of DNNs. B is a bound for scales of parameters.

3.2. **Least Square Estimator.** Using the model of DNNs, we define a least square estimator by empirical risk minimization. Using the observations $\mathcal{D}_n$, we consider the minimization problem with respect to parameters of DNNs as

$$\widehat{f}^L \in \underset{f \in \mathcal{F}_{NN,\eta}(S,B,L)}{\operatorname{argmin}} \frac{1}{n} \sum_{i \in [n]} (Y_i - f(X_i))^2, \tag{2}$$

and use $\widehat{f}^L$ for an estimator of f^*.

Note that the problem (2) has at least one minimizer since the parameter set Θ is compact and η is continuous. If necessary, we can add a regularization term for the problem (2), because it is not difficult to extend our results to an estimator with regularization. Furthermore, we can apply the early stopping techniques, since they play a role as the regularization (LeCun *et al.*, 2015). However, for simplicity, we confine our arguments of this paper in the least square.

We investigate theoretical aspects of convergence properties of $\widehat{f}^L$ with a ReLU activation.

Theorem 1. *Suppose $f^* \in \mathcal{F}_{M,J,\alpha,\beta}$. Then, there exist constants $c_1, c_1', C_L > 0, s \in \mathbb{N}\backslash\{1\}$, and (S, B, L) satisfying*

(i) $S = c_1' \max\{n^{D/(2\beta+D)}, n^{(2D-2)/(2\alpha+2D-2)}\}$,
(ii) $B \geqslant c_1 n^s$,
(iii) $L \leqslant c_1(1 + \max\{\beta/D, \alpha/2(D-1)\})$,

such that $\widehat{f}^L \in \mathcal{F}_{NN,\eta}(S, B, L)$ provides

$$\|\widehat{f}^L - f^*\|^2_{L^2(P_X)} \leqslant C_L \max\{n^{-2\beta/(2\beta+D)}, n^{-\alpha/(\alpha+D-1)}\} \log n, \tag{3}$$

with probability at least $1 - c_1 n^{-2}$.

Proof of Theorem 1 is a combination of a set estimation (Dudley, 1974; Mammen and Tsybakov, 1995), an approximation theory of DNNs (Yarotsky, 2017; Petersen and Voigtlaender, 2017), and an applications of the empirical process techniques (Koltchinskii, 2006; Giné and Nickl, 2015; Suzuki, 2018).

The convergence rate in Theorem 1 is simply interpreted as follows. The first term $n^{-2\beta/(2\beta+D)}$ describes an effect of estimating $f_m \in H^\beta$ for $m \in [M]$. The rate corresponds to the minimax optimal rate for estimating smooth functions in H^β (For a summary, see Tsybakov (2009)). The second term $n^{-\alpha/(\alpha+D-1)}$ reveals an effect from estimation of $\mathbf{1}_{R_m}$ for $m \in [M]$ through estimating the boundaries of $R_m \in \mathcal{R}_{\alpha,J}$. The same convergence rate appears in a problem for estimating sets with smooth boundaries (Mammen and Tsybakov, 1995).

We remark that a larger number of layers decreases B. Considering the result by Bartlett (1998), which shows that large values of parameters make the performance of DNNs worse, the above theoretical result suggests that a deep structure can avoid the performance loss caused by large parameters.

We also mention that our theoretical result is independent of the non-convex optimization problem. Suppose an optimization method fails to obtain the minimizer, i.e. we obtain a solution $\breve{f} \in \mathcal{F}_{NN,\eta}(S, B, L)$ such that $\Delta = n^{-1} \sum_{i \in [n]} (Y_i - \breve{f}(X_i))^2 - (Y_i - \widehat{f}(X_i))^2$ with an

error $\Delta > 0$. Then, an error of $\breve{f}$ is evaluated as

$$\mathbb{E}_{f^*}\left[\|\breve{f} - f^*\|^2_{L^2(P_X)}\right] \leqslant C_L \max\{n^{-2\beta/(2\beta+D)}, n^{-\alpha/(\alpha+D-1)}\}\log n + \Delta,$$

since we can evaluate the estimation error and the optimization error independently. Here, $\mathbb{E}_{f^*}[\cdot]$ denotes an expectation with respect to the true distribution of (X, Y). Thus, combining the results on the magnitude of Δ (e.g. Kawaguchi (2016)), we can evaluate the error in the cases of non-convex optimization.

3.3. **Bayes Estimator.** We define a Bayes estimator for DNNs which can avoid the non-convexity problem in optimization. Fix architecture Θ and $\mathcal{F}_{NN,\eta}(S, B, L)$ with given S, B and L. Then, a prior distribution for $\mathcal{F}_{NN,\eta}(S, B, L)$ is defined through providing distributions for the parameters contained in Θ. Let $\Pi_\ell^{(A)}$ and $\Pi_\ell^{(b)}$ be distributions of A_ℓ and b_ℓ as

$$A_\ell \sim \Pi_\ell^{(A)},\ b_\ell \sim \Pi_\ell^{(b)}$$

for $\ell \in [L]$. We set $\Pi_\ell^{(A)}$ and $\Pi_\ell^{(b)}$ such that each of the S parameters of Θ is uniformly distributed on $[-B, B]$, and the other parameters degenerate at 0. Using these distributions, we define a prior distribution Π_Θ on Θ by

$$\Pi_\Theta := \bigotimes_{\ell\in[L]} \Pi_\ell^{(A)} \otimes \Pi_\ell^{(b)}.$$

Then, a prior distribution for $f \in \mathcal{F}_{NN,\eta}(S, B, L)$ is defined by

$$\Pi_f(f) := \Pi_\Theta(\Theta : G_\eta[\Theta] = f).$$

We consider the posterior distribution for f. Since the noise ξ_i in (1) is Gaussian with its variance σ^2, the posterior distribution is given by

$$d\Pi_f(f|\mathcal{D}_n) = \frac{\exp(-\sum_{i\in[n]}(Y_i - f(X_i))^2/\sigma^2)d\Pi_f(f)}{\int \exp(-\sum_{i\in[n]}(Y_i - f'(X_i))^2/\sigma^2)d\Pi_f(f')}.$$

Note that we do not discuss computational issues of the Bayesian approach since the main focus is a theoretical aspect. To solve the computational problems, see Hernández-Lobato and Adams (2015) and others.

We provide theoretical analysis on the rate of contraction for the posterior distribution. Same as the least square estimator cases, we consider a ReLU activation function.

Theorem 2. *Suppose $f^* \in \mathcal{F}_{M,J,\alpha,\beta}$. Then, there exist constants $c_2, c_2', C_B > 0, s \in \mathbb{N}\backslash\{1\}$, architecture $\Theta : \|\Theta\|_0 \leqslant S, \|\Theta\|_\infty \leqslant B, |\Theta| \leqslant L$ satisfying following conditions:*

(i) $S = c_2' \max\{n^{D/(2\beta+D)}, n^{(2D-2)/(2\alpha+2D-2)}\}$,
(ii) $B \geqslant c_2 n^s$,
(iii) $L \leqslant c_2(1 + \max\{\beta/D, \alpha/2(D-1)\})$,

and a prior distribution Π_f, such that the posterior distribution $\Pi_f(\cdot|\mathcal{D}_n)$ provides

$$\begin{aligned}&\mathbb{E}_{f^*}\Big[\Pi_f\Big(f : \|f - f^*\|^2_{L^2(P_X)} \geqslant rC_B \times \max\{n^{-2\beta/(2\beta+D)}, n^{-\alpha/(\alpha+D-1)}\}\log n|\mathcal{D}_n\Big)\Big]\\ &\leqslant \exp\left(-r^2 c_2 \max\{n^{D/(2\beta+D)}, n^{(D-1)/(\alpha+D-1)}\}\right), \qquad (4)\end{aligned}$$

for all $r > 0$.

To provide proof of Theorem 2, we additionally apply studies for statistical analysis for Bayesian nonparametrics (van der Vaart and van Zanten, 2008, 2011).

Based on the result, we define a Bayes estimator as

$$\widehat{f}^B := \int f d\Pi_f(f|\mathcal{D}_n),$$

by the Bochner integral in $L^\infty(I^D)$. Then, we obtain the convergence rate of $\widehat{f}^B$ by the following corollary.

Corollary 1. *With the same setting in Theorem 2, consider $\widehat{f}^B$. Then, we have*

$$\mathbb{E}_{f^*}\left[\|\widehat{f}^B - f^*\|^2_{L^2(P_X)}\right] \leqslant C_B \max\{n^{-2\beta/(2\beta+D)}, n^{-\alpha/(\alpha+D-1)}\}\log n.$$

This result states that the Bayes estimator can achieve the same convergence rate as the least square estimator shown in Theorem 1. Since the Bayes estimator does not use optimization, we can avoid the non-convex optimization problem, while the computation of the posterior and mean are not straightforward.

4. Discussion: Why DNNs work better?

We discuss why DNNs work better than some other popular methods. Firstly, we show that the convergence rates by DNNs in Theorem 1 and 2 are optimal for estimating a function in the piecewise smooth function class. Secondly, we provide additional shreds of evidence that other methods are not suitable for the piecewise smooth functions. At last, we add some discussions.

4.1. Optimality of the DNN Estimators.

We will show optimality of the convergence rates by the DNN estimators in Theorem 1 and Corollary 1. To this end, we employ a theory of minimax optimal rate which is known in the field of mathematical statistics (Giné and Nickl, 2015). The theory derives a lower bound of a convergence rate with arbitrary estimators, thus we can obtain a theoretical limitation of convergence rates.

The result of the minimax optimal rate for the class of piecewise smooth functions $\mathcal{F}_{M,J,\alpha,\beta}$ is shown in the following theorem.

Theorem 3. *Consider $\bar{f}$ is an arbitrary estimator for $f^* \in \mathcal{F}_{M,J,\alpha,\beta}$. Then, there exists a constant $C_{mm} > 0$ such that*

$$\inf_{\bar{f}} \sup_{f^* \in \mathcal{F}_{M,J,\alpha,\beta}} \mathbb{E}_{f^*}\left[\|\bar{f} - f^*\|^2_{L^2(P_X)}\right] \geqslant C_{mm} \max\left\{n^{-2\beta/(2\beta+D)}, n^{-\alpha/(\alpha+D-1)}\right\}.$$

Proof of Theorem 3 employs techniques in the minimax theory developed by Yang and Barron (1999) and Raskutti *et al.* (2012).

We show that the convergence rates by the estimators with DNNs are optimal in the minimax sense, since the rates in Theorems 1 and 2 correspond to the lower bound of Theorem 3 up to a log factor. In other words, for estimating $f^* \in \mathcal{F}_{M,J,\alpha,\beta}$, no other methods could achieve a better convergence rate than the estimators by DNNs.

4.2. **Inefficiency of Other Methods.** We consider kernel methods and orthogonal series methods as representatives of other standard methods, then show that these methods are not optimal for estimating piecewise smooth functions.

Kernel methods are popular to estimate functions in the field of machine learning (Rasmussen and Williams, 2006; Steinwart and Christmann, 2008). Also, it is well known that theoretical aspects of kernel methods are equivalent to that of the Gaussian process regression (van der Vaart and van Zanten, 2008). An estimator by the kernel method is defined as

$$\widehat{f}^K(x) := \underset{f \in \mathcal{H}_K}{\operatorname{argmin}} \frac{1}{n} \sum_{i \in [n]} (Y_i - f(X_i))^2 + \mu \|f\|^2_{\mathcal{H}_K},$$

where $K : I^D \times I^D \to \mathbb{R}$ is a kernel function, $\mathcal{H}_K$ is a reproducing kernel Hilbert space given by K with its norm $\|\cdot\|_{\mathcal{H}_K}$, and $\mu > 0$ is a regularization coefficient as a hyper-parameter. Here, we consider two standard kernel functions such as the Gaussian kernel and the polynomial kernel. In the Gaussian kernel case, it is known that $\widehat{f}^K(x)$ are optimal when $f^* \in H^\beta$ (Steinwart and Christmann, 2008). We provide a theoretical result about $\widehat{f}^K(x)$ for estimating non-smooth functions.

Proposition 1. *Fix $D \in \mathbb{N}\backslash\{1\}, M, J \in \mathbb{N}, \alpha > 0$ and $\beta > 0$ arbitrary. Let $\widehat{f}^K(x)$ be the kernel estimator with the Gaussian kernel or the polynomial kernel. Then, there exists $f^* \in \mathcal{F}_{M,J,\alpha,\beta}$ and a constant $C_K > 0$ such that*

$$\mathbb{E}_{f^*}\left[\|\widehat{f}^K - f^*\|^2_{L^2(P_X)}\right] \to C_K,$$

as $n \to \infty$.

Since the kernel functions are not appropriate to express smooth structure of f^*, a set of functions by the kernel functions do not contain some $f^* \in \mathcal{F}_{M,J,\alpha,\beta}$. Although the Gaussian kernel is universal kernel, i.e. the RKHS by the Gaussian kernel is dense in a class of continuous functions, some $f^* \in \mathcal{F}_{M,J,\alpha,\beta}$ has a discontinuous structure, hence kernel methods with the kernel functions cannot estimate $f^* \in \mathcal{F}_{M,J,\alpha,\beta}$ consistently. Similar properties hold for other smooth kernel functions.

Orthogonal series methods, which is known as Fourier methods, estimate functions using an orthonormal basis. It is one of the most fundamental methods for nonparametric regression (For an introduction, see Section 1.7 in Tsybakov (2009)). Let $\phi_j(x)$ for $j \in \mathbb{N}$ be an orthonormal basis function in $L^2(P_X)$. An estimator for f^* by the orthogonal series method is defined as

$$\widehat{f}^S(x) := \sum_{j \in [J]} \widehat{\gamma}_j \phi_j(x),$$

where $J \in \mathbb{N}$ is a hyper-parameter and $\widehat{\gamma}_j$ is a coefficient calculated as $\widehat{\gamma}_j := \frac{1}{n}\sum_{i \in [n]} Y_i \phi_j(X_i)$. When the true function is smooth, i.e. $f^* \in H^\beta$, $\widehat{f}^S$ is known to be optimal in the minimax sense (Tsybakov, 2009). About estimation for $f^* \in \mathcal{F}_{M,J,\alpha,\beta}$, we can obtain the following proposition.

Proposition 2. *Fix $D \in \mathbb{N}\backslash\{1\}, M, J \in \mathbb{N}, \alpha > 2$ and $\beta > 1$ arbitrary. Let $\widehat{f}^S$ be the estimator by the orthogonal series method. Suppose $\phi_j, j \in \mathbb{N}$ are the trigonometric basis or the Fourier*

basis. Then, with sufficient large n, there exist $f^* \in \mathcal{F}_{M,J,\alpha,\beta}$, P_X, *a constant* $C_F > 0$, *and a parameter*

$$-\kappa > \max\{-2\beta/(2\beta + D), -\alpha/(\alpha + D - 1)\},$$

such that

$$\mathbb{E}_{f^*}\left[\|\widehat{f}^F - f^*\|^2_{L^2(P_X)}\right] > C_F n^{-\kappa}.$$

Proposition 2 shows that $\widehat{f}^S$ can estimate $f^* \in \mathcal{F}_{M,J,\alpha,\beta}$ consistently since the orthogonal basis in $L^2(P_X)$ can reveal all square integrable functions. Its convergence rate is, however, strictly worse than the optimal rate. Intuitively, the method requires many basis functions to express the non-smooth structure of $f^* \in \mathcal{F}_{M,J,\alpha,\beta}$, and a large number of bases increases variance of the estimator, hence they lose efficiency.

4.3. **Interpretation on Our Result.** According to the results, we can see that the estimators by DNNs have the theoretical advantage than the others for estimating $f^* \in \mathcal{F}_{M,J,\alpha,\beta}$, since the estimators by DNNs achieve the optimal convergence rate and the others do not.

We provide an intuition on why DNNs are optimal and the others are not. The most notable fact is that DNNs can realize non-smooth functions with a small number of parameters, due to activation functions and multi-layer structures. A combination of two ReLU functions can approximate step functions, and a composition of the step functions in a combination of other parts of the network can easily express smooth functions restricted to pieces. In contrast, even though the other methods have the universal approximation property, they require a larger number of parameters to represent non-smooth structures. By the statistical theory, a larger number of parameters increases variance of estimators and worsens the performance, hence the other methods lose the optimality.

About the inefficiency of the other methods, we do not claim that every statistical method except DNNs misses the optimality for estimating piecewise smooth functions. Our argument is the advantage of DNNs against the commonly used methods, such as the orthogonal series methods and the kernel methods. There may exist some other models which can achieve the optimality as DNNs, and this is an interesting future work.

An estimation using non-smooth kernels or basis functions is also an interesting direction. While some studies have investigated properties in such situations(van Eeden, 1985; Wu and Chu, 1993a,b; Wolpert *et al.*, 2011; Imaizumi *et al.*, 2018), these works focus on different settings such as density estimation or univariate data analysis, hence their setting does not fit problems discussed here.

5. Experiments

We carry out simple experiments to support our theoretical results.

5.1. **Non-smooth Realization by DNNs.** We show how the estimators by DNNs can estimate non-smooth functions. To this end, we consider the following data generating process with a piecewise linear function. Let $D = 2$, ξ be an independent Gaussian variable with a scale $\sigma = 0.5$, and X be a uniform random variable on I^2. Then, we generate n pairs of (X, Y) from (1) with a true function f^* as piecewise smooth function such that

$$f^*(x) = \mathbf{1}_{R_1}(x)(0.2 + x_1^2 + 0.1x_2) + \mathbf{1}_{R_2}(x)(0.7 + 0.01|4x_1 + 10x_2 - 9|^{1.5}), \tag{5}$$

with a set $R_1 = \{(x_1, x_2) \in I^2 : x_2 \geqslant -0.6x_1 + 0.75\}$ and $R_1 = I^2 \backslash R_1$. A plot of f in figure 3 shows its non-smooth structure.

About the estimation by DNNs, we employ the least square estimator (2). For the architecture Θ of DNNs, we set $|\Theta| = 4$ and dimensionality of each of the layers as $D_1 = 2, D_\ell = 3$ for $\ell \in \{2, 3, 4\}$, and $D_5 = 1$. We use a ReLU activation. To mitigate an effect of the non-convex optimization problem, we employ 100 initial points which are generated from the Gaussian distribution with an adjusted mean. We employ Adam (Kingma and Ba, 2014) for optimization.

We generate data with a sample size $n = 100$ f and obtain the least square estimator $\widehat{f}^L$ for f^*. Then, we plot $\widehat{f}^L$ in Figure 4 which minimize an error from the 100 trials with different initial points. We can observe that $\widehat{f}^L$ succeeds in approximating the non-smooth structure of f^*.

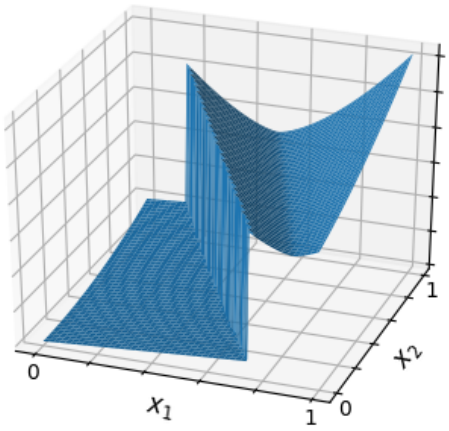

FIGURE 3. A plot for f^*.

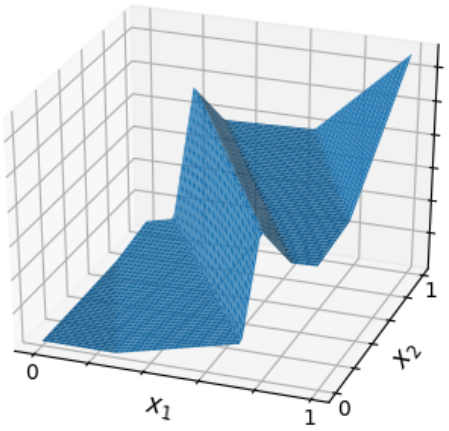

FIGURE 4. A plot for $\widehat{f}^L$.

5.2. **Comparison with the Other Methods.** We compare performances of the estimator by DNNs, the orthogonal series method, and the kernel methods. About the estimator by DNNs, we inherit the setting in Section 5.1. About the kernel methods, we employ estimators by the Gaussian kernel and the polynomial kernel. A bandwidth of the Gaussian kernel is selected from $\{0.01, 0.1, 0.2, ..., 2.0\}$ and a degree of the polynomial kernel is selected from [5]. Regularization coefficients of the estimators are selected from $\{0.01, 0.4, 0.8, ..., 2.0\}$. About the orthogonal series method, we employ the trigonometric basis which is a variation of the Fourier basis. All of the parameters are selected by a cross-validation.

We generate data from the process (1) with (5) with a sample size $n \in \{100, 200, ..., 1500\}$ and measure the expected loss of the methods. In figure 5, we report a mean and standard deviation of a logarithm of the loss by 100 replications. By the result, the estimator by DNNs always outperforms the other estimators. The other methods cannot estimate the non-smooth structure of f^*, although some of the other methods have the universal approximation property.

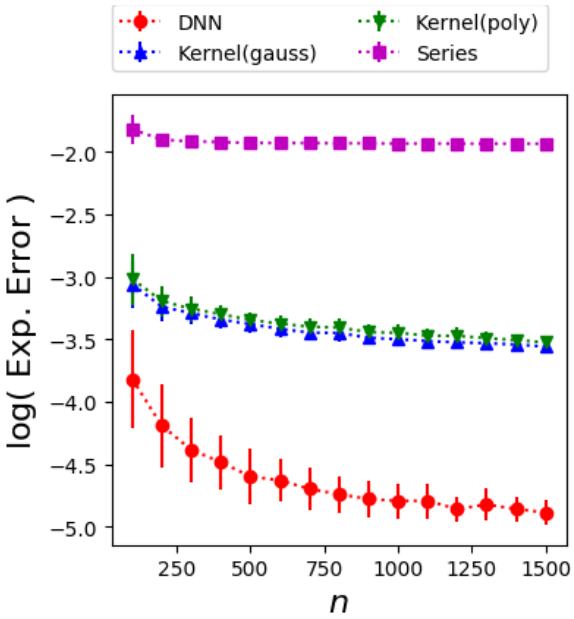

FIGURE 5. Comparison of a logarithm of the expected error by the methods. Markers are means and bars are standard deviations of 100 replications. Red circles denote a result by the estimator by DNNs, blue triangles are by the kernel estimator with the Gaussian kernel, green triangles are by the kernel estimator by the polynomial kernel, and purple squares are by the orthogonal series estimator.

6. Conclusion and Future Work

In this paper, we have derived theoretical results that explain why DNNs outperform other methods. To this goal, we considered a regression problem under the situation where the true function is piecewise smooth. We focused on the least square and Bayes estimators, and derived convergence rates of the estimators. Notably, we showed that the rates are optimal in the minimax sense. Furthermore, we proved that the commonly used orthogonal series methods and kernel methods are inefficient to estimate piecewise smooth functions, hence we show that the estimators by DNNs work better than the other methods for non-smooth functions. We also provided a guideline for selecting a number of layers and parameters of DNNs based on the theoretical results.

Investigating selection for architecture of DNNs has remained as a future work. While our results show the existence of an architecture of DNNs that achieves the optimal rate, we did not discuss how to learn the optimal architecture from data effectively. Practically and theoretically, this is obviously an important problem for analyzing a mechanism of DNNs.

References

Anthony, M. and Bartlett, P. L. (2009) *Neural network learning: Theoretical foundations*, cambridge university press.

Baldi, P. and Hornik, K. (1989) Neural networks and principal component analysis: Learning from examples without local minima, *Neural networks*, **2**, 53–58.

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

Barron, A. R. (1994) Approximation and estimation bounds for artificial neural networks, *Machine learning*, **14**, 115–133.

Bartlett, P. L. (1998) The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network, *IEEE transactions on Information Theory*, **44**, 525–536.

Bengio, Y. and Delalleau, O. (2011) On the expressive power of deep architectures, in *Algorithmic Learning Theory*, Springer, pp. 18–36.

Berlinet, A. and Thomas-Agnan, C. (2011) *Reproducing kernel Hilbert spaces in probability and statistics*, Springer Science & Business Media.

Collobert, R. and Weston, J. (2008) A unified architecture for natural language processing: Deep neural networks with multitask learning, in *Proceedings of the 25th international conference on Machine learning*, ACM, pp. 160–167.

Cover, T. M. and Thomas, J. A. (2012) *Elements of information theory*, John Wiley & Sons.

Cybenko, G. (1989) Approximation by superpositions of a sigmoidal function, *Mathematics of control, signals and systems*, **2**, 303–314.

Dauphin, Y. N., Pascanu, R., Gulcehre, C., Cho, K., Ganguli, S. and Bengio, Y. (2014) Identifying and attacking the saddle point problem in high-dimensional non-convex optimization, in *Advances in neural information processing systems*, pp. 2933–2941.

Dudley, R. M. (1974) Metric entropy of some classes of sets with differentiable boundaries, *Journal of Approximation Theory*, **10**, 227–236.

Fakoor, R., Ladhak, F., Nazi, A. and Huber, M. (2013) Using deep learning to enhance cancer diagnosis and classification, in *Proceedings of the International Conference on Machine Learning*.

Fukumizu, K. and Amari, S.-i. (2000) Local minima and plateaus in hierarchical structures of multilayer perceptrons, *Neural networks*, **13**, 317–327.

Giné, E. and Nickl, R. (2015) *Mathematical foundations of infinite-dimensional statistical models*, vol. 40, Cambridge University Press.

He, K., Zhang, X., Ren, S. and Sun, J. (2016) Deep residual learning for image recognition, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.

Hernández-Lobato, J. M. and Adams, R. (2015) Probabilistic backpropagation for scalable learning of bayesian neural networks, in *International Conference on Machine Learning*, pp. 1861–1869.

Hinton, G. E., Osindero, S. and Teh, Y.-W. (2006) A fast learning algorithm for deep belief nets, *Neural computation*, **18**, 1527–1554.

Imaizumi, M., Maehara, T. and Yoshida, Y. (2018) Statistically efficient estimation for non-smooth probability densities, in *Artificial Intelligence and Statistics*.

Kawaguchi, K. (2016) Deep learning without poor local minima, in *Advances in Neural Information Processing Systems*, pp. 586–594.

Kingma, D. P. and Ba, J. (2014) Adam: A method for stochastic optimization., *CoRR*, **abs/1412.6980**.

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

Le, Q. V., Ngiam, J., Coates, A., Lahiri, A., Prochnow, B. and Ng, A. Y. (2011) On optimization methods for deep learning, in *Proceedings of the 28th International Conference on International Conference on Machine Learning*, Omnipress, pp. 265–272.

LeCun, Y., Bengio, Y. and Hinton, G. (2015) Deep learning, *Nature*, **521**, 436–444.

Mammen, E. and Tsybakov, A. (1995) Asymptotical minimax recovery of sets with smooth boundaries, *The Annals of Statistics*, **23**, 502–524.

Mammen, E., Tsybakov, A. B. *et al.* (1999) Smooth discrimination analysis, *The Annals of Statistics*, **27**, 1808–1829.

Montufar, G. F., Pascanu, R., Cho, K. and Bengio, Y. (2014) On the number of linear regions of deep neural networks, in *Advances in neural information processing systems*, pp. 2924–2932.

Neyshabur, B., Tomioka, R. and Srebro, N. (2015) Norm-based capacity control in neural networks, in *Conference on Learning Theory*, pp. 1376–1401.

Petersen, P. and Voigtlaender, F. (2017) Optimal approximation of piecewise smooth functions using deep relu neural networks, *arXiv preprint arXiv:1709.05289*.

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

Rasmussen, C. E. and Williams, C. K. (2006) *Gaussian processes for machine learning*, vol. 1, MIT press Cambridge.

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, *arXiv preprint arXiv:1708.06633*.

Soudry, D. and Carmon, Y. (2016) No bad local minima: Data independent training error guarantees for multilayer neural networks, *arXiv preprint arXiv:1605.08361*.

Steinwart, I. and Christmann, A. (2008) *Support vector machines*, Springer Science & Business Media.

Stone, C. (1982) Optimal global rates of convergence for nonparametric regression, *The Annals of Statistics*, **10**, 1040–1053.

Suzuki, T. (2018) Fast generalization error bound of deep learning from a kernel perspective, in *Artificial Intelligence and Statistics*.

Tsybakov, A. B. (2009) Introduction to nonparametric estimation.

van der Vaart, A. and van Zanten, H. (2011) Information rates of nonparametric gaussian process methods, *Journal of Machine Learning Research*, **12**, 2095–2119.

van der Vaart, A. and van Zanten, J. (2008) Rates of contraction of posterior distributions based on gaussian process priors, *The Annals of Statistics*, **36**, 1435–1463.

van der Vaart, A. and Wellner, J. (1996) *Weak Convergence and Empirical Processes: With Applications to Statistics*, Springer Science & Business Media.

van Eeden, C. (1985) Mean integrated squared error of kernel estimators when the density and its derivative are not necessarily continuous, *Annals of the Institute of Statistical Mathematics*, **37**, 461–472.

Wasserman, L. A. (2006) *All of nonparametric statistics: with 52 illustrations*, Springer.

Wolpert, R. L., Clyde, M. A. and Tu, C. (2011) Stochastic expansions using continuous dictionaries: Lévy adaptive regression kernels, *The Annals of Statistics*, **39**, 1916–1962.

Wu, J. and Chu, C. (1993a) Kernel-type estimators of jump points and values of a regression function, *The Annals of Statistics*, **21**, 1545–1566.

Wu, J. and Chu, C. (1993b) Nonparametric function estimation and bandwidth selection for discontinuous regression functions, *Statistica Sinica*, pp. 557–576.

Yang, Y. and 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, C., Bengio, S., Hardt, M., Recht, B. and Vinyals, O. (2017) Understanding deep learning requires rethinking generalization, in *ICLR*.

Appendix A. Proof of Theorem 1

We provide additional notation. λ denotes the Lebesgue measure. For a function $f : I^D \to R$, $\|f\|_{L^\infty} = \sup_{x \in I^D}$ is a supremum norm. $\|f\|_{L^2} = \|f\|_{L^2(\lambda)}$ is a $L^2(\lambda)$-norm with the Lebesgue measure.

With the set of observations, let $\|\cdot\|_n$ and be an empirical norm such as

$$\|f\|_n = n^{-1} \sum_{i=1}^{n} f(X_i).$$

Also, we define the empirical norm of random variables such as

$$\|Y\|_n := \left(n^{-1} \sum_{i \in [n]} Y_i^2 \right)^{1/2} \text{ and } \|\xi\|_n := \left(n^{-1} \sum_{i \in [n]} \xi_i^2 \right)^{1/2}.$$

With a set $\mathcal{F}$ and a radius $\epsilon > 0$, we introduce a covering number as

$$\mathcal{N}(\epsilon, \mathcal{F}, \|\cdot\|) := \inf\Big\{ N \mid \{f_j\}_{j \in [N]}, \|f - f_j\| \leqslant \epsilon, \forall f \in \mathcal{F} \Big\},$$

with a norm $\|\cdot\|$.

By the definition of the least square estimator (2), we obtain the following basic inequality as

$$\|Y - \widehat{f}^L\|_n^2 \leqslant \|Y - f\|_n^2,$$

for all $f \in \mathcal{F}_{NN,\eta}(S, B, L)$. Since we have $Y_i = f^*(X_i) + \xi_i$, we obtain

$$\|f^* + \xi - \widehat{f}^L\|_n^2 \leqslant \|f^* + \xi - f\|_n^2.$$

By the simple calculation, it yields

$$\|f^* - \widehat{f}^L\|_n^2 \leqslant \|f^* - f\|_n^2 + \frac{2}{n} \sum_{i=1}^{n} \xi_i (\widehat{f}^L(X_i) - f(X_i)). \tag{6}$$

In the following, we will fix $f \in \mathcal{F}_{NN,\eta}(S, B, L)$ and evaluate each of the terms each of the terms of the RHS of (6) At the first subsection, we provide a result for approximating $f^* \in \mathcal{F}_{M,J,\alpha,\beta}$ by DNNs. At the second subsection, we evaluate a variance of $\widehat{f}^F$. At the last subsection, we combine the results and derive an overall convergence rate.

A.1. **Approximate piecewise functions by DNNs.** A purpose of this part is to bound the following value

$$\|f - f^*\|_{L^2(P_X)},$$

with a properly selected $f \in \mathcal{F}_{NN,\eta}(S, B, L)$. To this end, we consider an existing Θ with properly selected S, B and L. Our proof of this part is obtained by extending a technique by Yarotsky (2017) and Petersen and Voigtlaender (2017).

Fix $f^* \in \mathcal{F}_{M,J,\alpha,\beta}$ such that $f^* = \sum_{m \in [M]} f_m^* \mathbf{1}_{R_m^*}$ with f_m^* and R_m^* for $m \in [M]$. To approximate f^*, we consider neural networks $\Theta_{f,m}$ and $\Theta_{r,m}$ for $m \in [M]$, and their number of layers and non-zero parameters will be specified later. We also consider a network Θ_3

which approximates a multiplication and a summation, i.e. $G_\eta[\Theta_3](x_1,...,x_M,x'_1,...,x'_M) \approx \sum_{m\in[M]} x_m x'_m$.

We evaluate a distance between f^* and a combined neural network $G_\eta[\Theta_3](G_\eta[\Theta_1](\cdot), G_\eta[\Theta_2](\cdot))$ as

$$
\begin{aligned}
&\|f^* - G_\eta[\Theta_3](G_\eta[\Theta_{f,1}](\cdot),...,G_\eta[\Theta_{f,M}](\cdot), G_\eta[\Theta_{r,1}](\cdot),...,G_\eta[\Theta_{r,M}](\cdot))\|_{L^2} \\
&= \left\| \sum_{m\in[M]} f^*_m \mathbf{1}_{R^*_m} - G_\eta[\Theta_3](G_\eta[\Theta_{f,1}](\cdot),...,G_\eta[\Theta_{f,M}](\cdot), G_\eta[\Theta_{r,1}](\cdot),...,G_\eta[\Theta_{r,M}](\cdot)) \right\|_{L^2} \\
&\leqslant \left\| \sum_{m\in[M]} f^*_m \otimes \mathbf{1}_{R^*_m} - \sum_{m\in[M]} G_\eta[\Theta_{f,M}] \otimes G_\eta[\Theta_{r,M}] \right\|_{L^2} \\
&\quad + \Bigg\| \sum_{m\in[M]} G_\eta[\Theta_{f,m}] \otimes G_\eta[\Theta_{r,m}] \\
&\qquad\qquad - G_\eta[\Theta_3](G_\eta[\Theta_{f,1}](\cdot),...,G_\eta[\Theta_{f,M}](\cdot), G_\eta[\Theta_{r,1}](\cdot),...,G_\eta[\Theta_{r,M}](\cdot)) \Bigg\|_{L^2} \\
&\leqslant \sum_{m\in[M]} \left\| f^*_m \otimes \mathbf{1}_{R^*_m} - G_\eta[\Theta_{f,m}] \otimes G_\eta[\Theta_{r,m}] \right\|_{L^2} \\
&\quad + \Bigg\| \sum_{m\in[M]} G_\eta[\Theta_{f,m}] \otimes G_\eta[\Theta_{r,m}] \\
&\qquad\qquad - G_\eta[\Theta_3](G_\eta[\Theta_{f,1}](\cdot),...,G_\eta[\Theta_{f,M}](\cdot), G_\eta[\Theta_{r,1}](\cdot),...,G_\eta[\Theta_{r,M}](\cdot)) \Bigg\|_{L^2} \\
&\leqslant \sum_{m\in[M]} \|(f^*_m - G_\eta[\Theta_{f,m}]) \otimes G_\eta[\Theta_{r,m}]\|_{L^2} + \sum_{m\in[M]} \left\| f^*_m \otimes (\mathbf{1}_{R^*_m} - G_\eta[\Theta_{r,m}]) \right\|_{L^2} \\
&\quad + \Bigg\| \sum_{m\in[M]} G_\eta[\Theta_{f,m}] \otimes G_\eta[\Theta_{r,m}] \\
&\qquad\qquad - G_\eta[\Theta_3](G_\eta[\Theta_{f,1}](\cdot),...,G_\eta[\Theta_{f,M}](\cdot), G_\eta[\Theta_{r,1}](\cdot),...,G_\eta[\Theta_{r,M}](\cdot)) \Bigg\|_{L^2} \\
&=: \sum_{m\in[M]} B_{1,m} + \sum_{m\in[M]} B_{2,m} + B_3. \qquad (7)
\end{aligned}
$$

We will bound $B_{m,1}, B_{m,2}$ for $m \in [M]$ and B_3.

About the term $B_{1,m}$ for $m \in [M]$, we apply the Cauchy-Schwartz inequality and obtain

$$\left\|(f^*_m - G_\eta[\Theta_{f,m}]) \otimes \mathbf{1}_{R^*_m}\right\|_{L^2} \leqslant \|f^*_m - G_\eta[\Theta_{f,m}]\|_{L^2} \|G_\eta[\Theta_{r,m}]\|_{L^2}.$$

By Theorem 1 in Yarotsky (2017) and Theorem A.8 in Petersen and Voigtlaender (2017), we can assure that there exists a neural network $\Theta_{f,m}$ with $\|\Theta_{f,m}\|_0 \leqslant C'_f \epsilon^{D/\beta}, \|\Theta_{f,m}\|_{L^\infty} \leqslant \epsilon^{-2s}$ such that $\|f^*_m - G_\eta[\Theta_{f,m}]\|_{L^2} < \epsilon$. About $\|G_\eta[\Theta_{r,m}]\|_{L^2}$, we will employ a neural network

in Lemma 3.4 in Petersen and Voigtlaender (2017) and use the result that the $G_\eta[\Theta_{r,m}]$ is uniformly bounded by 1. Hence,

$$\|G_\eta[\Theta_{r,m}]\|_{L^2} \leqslant \left(\int_{[0,1]^D} 1 d\lambda\right)^{1/2} = 1.$$

Combining the results, we obtain

$$B_{1,m} < \epsilon.$$

For evaluating the term $B_{2,m}$ for $m \in [M]$, we consider decomposition of R_m. As the same discussion, we have

$$\left\|f_m^* \otimes (\mathbf{1}_{R_m^*} - G_\eta[\Theta_{r,m}])\right\|_{L^2} \leqslant \|f_m^*\|_{L^2} \left\|\mathbf{1}_{R_m^*} - G_\eta[\Theta_{r,m}]\right\|_{L^2}.$$

Since $f_m^* \in H^\beta$, there exists a constant $C_H > 0$ such that $\|f_m^*\|_{L^2} \leqslant C_H$. We divide R_m into J parts such as $R_{m,j}, j \in [J]$. Also, we describe boundaries of $R_{m,j}$ by $2^J + Q$ horizon functions with finite $Q \in \mathbb{N}$. Here, a horizon function $h : [0,1]^D \to \{0,1\}$ which is defined as

$$h = \Psi(x_1 + f'(x_2, ..., x_D), x_2, ..., x_D),$$

where Ψ is the Heaviside function, i.e. $\Phi(x_1, ..., x_D) := \mathbf{1}_{\{x_1 \geqslant 0\}}(x_1, ..., x_D)$, and $f' \in H^\alpha([0,1]^{D-1})$. We consider horizon functions $h_{m,j,k}$ for $k \in [2^D + Q]$ and represent $\mathbf{1}_{R_{m,j}}$ by $\prod_k h_{m,j,k}$. To approximate the product of horizon functions, we consider that $\Theta_{r,m}$ is constructed by a network for summation Θ_+, for multiplication $\Theta_\times$ and a network for horizon functions $\Theta_{m,j,k}$. The approximation error is written as

$$\begin{aligned}
&\|\mathbf{1}_{R_m} - G_\eta[\Theta_{r,m}]\|_{L^2} \\
&= \Bigg\| \sum_{j\in[J]} \mathbf{1}_{R_{m,j}} \\
&\qquad - G_\eta[\Theta_+](G_\eta[\Theta'_\times](G_\eta[\Theta_{m,1,1}](\cdot), ..., \\
&\qquad\qquad G_\eta[\Theta_{m,1,2^D+S}](\cdot)), ..., G_\eta[\Theta_\times](G_\eta[\Theta_{m,J,1}](\cdot), ..., G_\eta[\Theta_{m,J,2^D+S}](\cdot))) \Bigg\|_{L^2} \\
&= \left\| \sum_{j\in[J]} \mathbf{1}_{R_{m,j}} - \sum_{j\in[J]} G_\eta[\Theta'_\times](G_\eta[\Theta_{m,j,1}](\cdot), ..., G_\eta[\Theta_{m,j,2^D+S}](\cdot)) \right\|_{L^2} \\
&\leqslant \sum_{j\in[J]} \left\| \prod_{k\in[2^D+S]} h_{m,j,k} - \prod_{k\in[2^D+S]} G_\eta[\Theta_{m,j,k}](\cdot) \right\|_{L^2} \\
&\quad + \sum_{j\in[J]} \left\| \prod_{k\in[2^D+S]} G_\eta[\Theta_{m,j,k}](\cdot) - G_\eta[\Theta'_\times](G_\eta[\Theta_{m,j,1}](\cdot), ..., G_\eta[\Theta_{m,j,2^D+S}](\cdot)) \right\|_{L^2} \\
&=: \sum_{j\in[J]} B_{2,m,j} + \sum_{j\in[J]} B'_{2,m,j}.
\end{aligned}$$

To evaluate $B_{2,m,j}$, we apply Lemma 3.4 in Petersen and Voigtlaender (2017). By setting $\Theta_{m,j,k}$ as Θ_h, hence we obtain $0 \leqslant G_\eta[\Theta_{m,j,k}](x) \leqslant 1$. Thus, combining the fact $0 \leqslant h_{m,j,k}(\cdot) \leqslant 1$, we have

$$\begin{aligned}
&\prod_{k\in[2^D+Q]} h_{m,j,k}(\cdot) - \prod_{k\in[2^D+Q]} G_\eta[\Theta_{m,j,k}](\cdot) \\
&= \sum_{k\in[2^D+Q]} \{h_{m,j,k}(\cdot) - G_\eta[\Theta_{m,j,k}](\cdot)\} \prod_{k'\in[k-1]} h_{m,j,k}(\cdot) \prod_{k'\in[K]\setminus[k]} G_\eta[\Theta_{m,j,k}](\cdot) \\
&\leqslant \sum_{k\in[2^D+Q]} \{h_{m,j,k}(\cdot) - G_\eta[\Theta_{m,j,k}](\cdot)\}.
\end{aligned}$$

Then, we have

$$B_{2,m,j} \leqslant \sum_{k\in[2^D+Q]} \|h_{m,j,k} - G_\eta[\Theta_{m,j,k}]\|_{L^2} \leqslant (2^D+Q)\epsilon,$$

by applying Lemma 3.4 in Petersen and Voigtlaender (2017).

To evaluate $B'_{2,m,j}$, we apply Lemma 1 for multiple production.

Lemma 1. *Fix $\eta > 0$ arbitrary. Then, for each $\epsilon \in (0, 1/2)$, there exists a neural network $\Theta_{\times'}$ for a D'-dimensional input with at most $(1 + \log_2 D')/\eta$ layers such that $\|\Theta_{\times'}\|_0 \leqslant C_{\times'} D' \epsilon^{-\eta}$, $\|\Theta_{\times'}\|_\infty \leqslant \epsilon^{-2s}$ with some constants $C_{\times'} > 0$ and $s \in \mathbb{N}$, and it satisfies*

$$\left| \prod_{d\in[D']} x_d - G_\eta[\Theta_{\times'}](x_1, ..., x_{D'}) \right| \leqslant (D'-1)\epsilon.$$

Proof. We employ the neural network for multiplication $\Theta_\times$ as Proposition 3 in Yarotsky (2017) and Lemma A.2 in Petersen and Voigtlaender (2017) and consider a tree-shaped multiplication network. There are $D'-1$ multiplication networks and the tree has $1 + \log_2 D'$ depth. □

Using the result, we set $\Theta'_\times$ and bound $B'_{2,m,j}$ as

$$B'_{2,m,j} \leqslant (2^D + Q - 1)\epsilon.$$

Combining the results about $B_{2,m,j}$ and $B'_{2,m,j}$, we obtain

$$B_{2,m} \leqslant 2J(2^D + Q - 1/2)\epsilon.$$

About the term B_3, we consider Lemma 2.

Lemma 2. *Let $\eta > 0$ be arbitrary. Then, for each $\epsilon \in (0, 1/2)$, there exists a neural network Θ_3 for a $2D'$-dimensional input with at most $1 + L$ layers where $L > 1/\eta$ and $D' + C_\times D' \epsilon^{-\eta}$ non-zero parameters such that $\|\Theta_3\|_\infty \leqslant \epsilon^{-2s}$, and it satisfies*

$$\left| G_\eta[\Theta_3](x_1, ..., x_{2D'}) - \sum_{d\in[D']} x_d x_{D'+d} \right| \leqslant D'\epsilon.$$

Proof. Let us define the function by the neural network $G_\eta[\Theta_3]$ as

$$G_\eta[\Theta_3](x) = G_\eta[\Theta_+](G_\eta[\Theta_\times](x_1, x_{D'+1}), ..., G_\eta[\Theta_\times](x_{D'}, x_{2D'})),$$

where $\Theta_\times$ is defined in Proposition 3 in Yarotsky (2017) and Lemma A.2 in Petersen and Voigtlaender (2017). Here, Θ_+ is a summation network such that

$$\Theta_+ := (A, b) = \left(\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}, 0 \right).$$

Then, we evaluate the difference as

$$\begin{aligned}
&\left| G_\eta[\Theta_3](x_1, ..., x_{2D'}) - \sum_{d\in[D']} x_d x_{2d} \right| \\
&= \left| \sum_{d\in[D']} G_\eta[\Theta_\times](x_d, x_{D'+d}) - \sum_{d\in[D']} x_d x_{D'+d} \right| \\
&\leqslant \sum_{d\in[D']} |G_\eta[\Theta_\times](x_d, x_{D'+d}) - x_d x_{D'+d}| \\
&\leqslant D'\epsilon.
\end{aligned}$$

Here, the last inequality follows Proposition 3 in Yarotsky (2017).

□

Let Θ_3 be the neural network defined in the statement, then we obtain that

$$B_3 \leqslant 2M\epsilon.$$

We combine the result about $B_{1,m}, B_{2,m}$ and B_3, then define $f \in \mathcal{F}_{NN,\eta}(S, B, L)$ for approximating f^*. About Θ_1, it contains $\Theta_{f,m}$ for $m \in [M]$, thus we known $\|\Theta_1\|_0 \leqslant C'_f M\epsilon^{D/\beta}$. Here, we set $\epsilon = c_1 n^{-\beta/(2\beta+D)}$ with a constant c_1, thus we conclude $\|\Theta_1\|_0 \leqslant C'_f c_1 M n^{D/(2\beta+D)}$ and $\|\Theta_1\|_\infty \leqslant c^{-2s} n^{2s\beta/(2\beta+D)}$. About Θ_2, it contains $\Theta_{\Theta_{m,j,k}}$ for $m \in [M], J \in [J], k \in [2^D + Q]$, Θ_+ and $\Theta'_\times$. Hence, we know $\|\Theta_2\|_0 \leqslant MJ(2^D + Q)(\epsilon^{-2(D-1)/\alpha} + C'_\times \epsilon^{-\eta}) + MJ$. We set $\epsilon = c_2 n^{-\alpha/(2\alpha+2D-2)}$ with c_2 and $\eta = 2(D-1)/\alpha$, then we have $\|\Theta_2\|_0 \leqslant c_2 MJ(2^D + Q)(1 + 2C'_\times) n^{2(D-1)/(2\alpha+2D-2)} + MJ$ and $\|\Theta_2\|_\infty \leqslant c_2 n^{2s\alpha/(2\alpha+2D-2)}$. About Θ_3, we already define it in Lemma 2 such that Θ_3 has $1 + L$ layer with $L > 1/\eta$ and $\|\Theta_3\|_0 = 2M(1 + C_\times \epsilon^{-\eta})$. Then, we set $\eta = \max\{2(D-1)/\alpha, D/\beta\}$ and $\epsilon = c_3 \max\{n^{-\beta/(2\beta+D)}, n^{-\alpha/(2\alpha+2D-2)}\}$ with $c_3 > 0$.

We combine Θ_1, Θ_2 and Θ_3 as a unified neural network $\dot{\Theta}$. Here, we know that the number of layers is at most

$$\begin{aligned}
&C_{sl}(1 + \lceil \log_2(1 + \beta) \rceil)(1 + \beta/D) + C_h \log_2(2 + \alpha)(1 + \alpha/D) \\
&\quad + (1 + \log_2(2^D + Q))\alpha/D + (1 + \log_2 M) \max\{\beta/D, \alpha/2(D-1)\} \\
&\leqslant C_L(1 + \log_2(\max\{1 + \beta, 2 + \alpha, 2^D + Q, M\}))(1 + \max\{\beta/D, \alpha/2(D-1)\}).
\end{aligned}$$

Also, the number of non-zero parameters $\|\dot{\Theta}\|_0$ is at most

$$C'_f c_1 M n^{D/(2\beta+D)} + c_2 MJ(2^D + Q)(1 + 2C'_\times) n^{2(D-1)/(2\alpha+2D-2)} + MJ$$

$$
\begin{aligned}
&+ c_3 2M(1 + C_\times \max\{n^{D/(2\beta+D)}, n^{(2D-2)/(2\alpha+2D-2)}\}) \\
&\leqslant C_S M(1 + J(2^D + Q)\max\{n^{D/(2\beta+D)}, n^{(2D-2)/(2\alpha+2D-2)}\}), \qquad (8)
\end{aligned}
$$

with a constant $C_S > 0$. Then, there exists a function by the neural network $\dot{f} := G_\eta[\dot{\Theta}]$ which satisfies

$$
\begin{aligned}
&\|f^* - \dot{f}\|_{L^2} \\
&\leqslant Mn^{-\beta/(2\beta+D)} + 2JM(2^D + Q - 1/2)n^{-\alpha/(2\alpha+2D-2)} + 2M\max\{n^{-\beta/(2\beta+D)}, n^{-\alpha/(\alpha+2D-2)}\} \\
&\leqslant 2JM(2^D + Q - 1/2)\max\{n^{-\beta/(2\beta+D)}, n^{-\alpha/(\alpha+2D-2)}\}. \qquad (9)
\end{aligned}
$$

A.2. **Evaluate an entropy bound of the estimators by DNNs.** Here, we evaluate a variance term of $\|\widehat{f}^L - f^*\|_n$ in (6) through evaluating the term

$$
\left| \frac{2}{n} \sum_{i\in[n]} \xi_i (\widehat{f}^L(X_i) - f(X_i)) \right|.
$$

To bound the term, we employ the technique by the empirical process technique (Koltchinskii, 2006; Giné and Nickl, 2015; Suzuki, 2018).

We consider an expectation of the term. Let us define a subset $\widetilde{F}_{NN,\delta} \subset \mathcal{F}_{NN,\eta}(S, B, L)$ as $\widetilde{\mathcal{F}}_{NN,\delta} := \{f - \widehat{f}^L : \|f - \widehat{f}^L\|_n \leqslant \delta, f \in \mathcal{F}_{NN,\eta}(S, B, L)\}$. Here, we mention that $f \in \widetilde{F}_{NN,\delta}$ is bounded by providing the following lemma.

Lemma 3. *For any $f \in \widetilde{F}_{NN,\delta}$ with an activation function η satisfying Lipschitz continuity with a constant 1, we obtain*

$$
\|f\|_{L^\infty} \leqslant B_f,
$$

where $B_f > 0$ is a finite constant.

Proof. For each $\ell \in [L]$, consider a transformation

$$
f_\ell(x) := \eta(A_\ell x + b_\ell).
$$

When $\|x\|_\infty = B_x$ and $\|\operatorname{vec}(A_\ell)\|_\infty, \|b_\ell\|_\infty \leqslant B$, we obtain

$$
\|f_\ell\|_{L^\infty} \leqslant \|A_\ell x + b_\ell\|_\infty \leqslant D_\ell B_x B + B.
$$

Let $\bar{D} := \max_{\ell\in[L]} D_\ell$, when iteratively we have

$$
\|f\|_{L^\infty} \leqslant \sum_{\ell\in[L]\cup\{0\}} \prod_{\ell'\in[L]\setminus[\ell]} (\bar{D}B)^{\ell'} < \infty,
$$

by applying that $\|x\|_\infty \leqslant 1$ for an input. □

Due to Lemma 3, with given $\{X_i\}_{i\in[n]}$, we can apply the chaining (Theorem 2.3.6 in Giné and Nickl (2015)) and obtain

$$
2\mathbb{E}_\xi \left[\sup_{f'\in\widetilde{F}_{NN,\delta}} \left| \frac{1}{n} \sum_{i\in[n]} \xi_i f'(X_i) \right| \right] \leqslant 8\sqrt{2} \frac{\sigma}{n^{1/2}} \int_0^{\delta/2} \sqrt{\log 2\mathcal{N}(\epsilon', \mathcal{F}_{NN,\eta}(S, B, L), \|\cdot\|_n)} d\epsilon'.
$$

Here, to apply Theorem 2.3.6 in Giné and Nickl (2015), we set $n^{-1/2}\sum_{i\in[n]}\xi_i f(X_i)$ as the stochastic process and 0 as $X(t_0)$ in the theorem. Then, to bound the entropy term, we apply an inequality

$$\begin{aligned}
&\mathcal{N}(\epsilon, \widetilde{\mathcal{F}}_{NN,\delta}, \|\cdot\|_n) \\
&\leqslant \mathcal{N}(\epsilon, \mathcal{F}_{NN,\eta}(S,B,L), \|\cdot\|_{L^\infty}) \\
&\leqslant \log \mathcal{N}(\epsilon, \mathcal{F}_{NN,\eta}(S,B,L), \|\cdot\|_{L^\infty}) \\
&\leqslant (S+1)\log\left(\frac{2(L+1)N^2}{\epsilon}\right),
\end{aligned}$$

where the last inequality holds by Theorem 14.5 in Anthony and Bartlett (2009) and Lemma 8 in Schmidt-Hieber (2017). Then, we obtain

$$2\mathbb{E}_\xi\left[\sup_{f'\in\widetilde{F}_{NN,\delta}}\left|\frac{1}{n}\sum_{i\in[n]}\xi_i f'(X_i)\right|\right] \leqslant 4\sqrt{2}\frac{\sigma\sqrt{S+1}\delta}{n^{1/2}}\left(\log\frac{(L+1)N^2}{\delta}+1\right). \tag{10}$$

With the bound (10) for the expectation term, we apply the Gaussian concentration inequality (Theorem 2.5.8 in Giné and Nickl (2015)) by setting $n^{-1}\sum_{i\in[n]}\xi_i f'(X_i)$ as the stochastic process and $\delta^2 \geqslant \|f\|_n^2$ be B^2 and obtain

$$\begin{aligned}
&1-\exp(-nu^2/2\sigma^2\delta^2) \\
&\leqslant \Pr_\xi\left(4\sup_{f'\in\widetilde{F}_{NN,\delta}}\left|\frac{1}{n}\sum_{i\in[n]}\xi_i f'(X_i)\right| \leqslant 4\mathbb{E}_\xi\left[\sup_{f'\in\widetilde{F}_{NN,\delta}}\left|\frac{1}{n}\sum_{i=1}^n\xi_i f'(X_i)\right|\right]+u\right) \\
&\leqslant \Pr_\xi\left(4\sup_{f'\in\widetilde{F}_{NN,\delta}}\left|\frac{1}{n}\sum_{i\in[n]}\xi_i f'(X_i)\right| \leqslant 8\sqrt{2}\frac{\sigma\sqrt{S+1}\delta}{n^{1/2}}\left(\log\frac{(L+1)N^2}{\delta}+1\right)+u\right),
\end{aligned} \tag{11}$$

for any $u>0$. Let us introduce the following notation as

$$V_n := 8\sqrt{2}\frac{\sigma\sqrt{S+1}}{n^{1/2}}.$$

To evaluate the variance term, we reform the basic inequality (6) as

$$-\frac{2}{n}\sum_{i=1}^n\xi_i(\widehat{f}^L(X_i)-f(X_i))+\|f^*-\widehat{f}^L\|_n^2 \leqslant \|f^*-f\|_n^2,$$

and apply an inequality $\frac{1}{2}\|\widehat{f}^L-f\|_n^2 \leqslant \|f-f^*\|_n^2+\|f^*-\widehat{f}^L\|_n^2$, then we have

$$-\frac{2}{n}\sum_{i=1}^n\xi_i(\widehat{f}^L(X_i)-f(X_i))+\frac{1}{2}\|\widehat{f}^L-f\|_n^2-\|f-f^*\|_n^2 \leqslant \|f^*-f\|_n^2,$$

then we have

$$-\frac{2}{n}\sum_{i=1}^n\xi_i(\widehat{f}^L(X_i)-f(X_i))+\frac{1}{2}\|\widehat{f}^L-f\|_n^2 \leqslant 2\|f^*-f\|_n^2. \tag{12}$$

Let we consider a lower bound for $-\frac{2}{n}\sum_{i\in[n]}\xi_i(\widehat{f}^L(X_i)-f(X_i))$. To make the bound (11) be valid for all $f\in\mathcal{F}_{NN,\eta}(S,B,L)$, we let $\delta=\max\{\|\widehat{f}^L-f\|_n,V_n\}$. Then, we obtain the bound

$$\begin{aligned}&\left|\frac{2}{n}\sum_{i\in[n]}\xi_i(\widehat{f}^L(X_i)-f(X_i))\right|\\&\leqslant\max\{\|\widehat{f}^L-f\|_n,V_n\}\left\{V_n\left(\log\frac{(L+1)N^2}{V_n}+1\right)\right\}+u\\&\leqslant\frac{1}{4}\left(\max\{\|\widehat{f}^L-f\|_n,V_n\}\right)^2+2\left\{V_n\left(\log\frac{(L+1)N^2}{V_n}+1\right)\right\}^2+u,\end{aligned}$$

by using $xy\leqslant\frac{1}{4}x^2+2y^2$. Using this result to (12), we obtain

$$-\frac{1}{4}\left(\max\{\|\widehat{f}^L-f\|_n,V_n\}\right)^2-2\left\{V_n\left(\log\frac{(L+1)N^2}{V_n}+1\right)\right\}^2-u+\frac{1}{2}\|\widehat{f}^L-f\|_n^2\leqslant 2\|f^*-f\|_n^2.$$

If $\|\widehat{f}^L-f\|_n\geqslant V_n$ holds, we obtain

$$-\frac{1}{4}\|\widehat{f}^L-f\|_n^2-2\left\{V_n\left(\log\frac{(L+1)N^2}{V_n}+1\right)\right\}^2-u+\frac{1}{2}\|\widehat{f}^L-f\|_n^2\leqslant 2\|f^*-f\|_n^2.$$

Then, simple calculation yields

$$\|\widehat{f}^L-f\|_n^2\leqslant 4\left\{V_n\left(\log\frac{(L+1)N^2}{V_n}+1\right)\right\}^2+2u+4\|f^*-f\|_n^2. \tag{13}$$

If $\|\widehat{f}^L-f\|_n\leqslant V_n$, the same result holds.

We additionally apply an inequality $\frac{1}{2}\|\widehat{f}^L-f^*\|_{L^2}^2\leqslant\|f^*-f\|_n^2+\|\widehat{f}^L-f\|_n^2$ to (13), we obtain

$$\|\widehat{f}^L-f^*\|_n^2\leqslant 10\|f^*-f\|_n^2+8\left\{V_n\left(\log\frac{(L+1)N^2}{V_n}+1\right)\right\}^2+4u, \tag{14}$$

with probability at least $1-\exp(-nu^2/2\sigma^2\delta^2)$ for all $u>0$.

A.3. **Combine the results.** Combining the results in Section A.1 and A.2, we evaluate an expectation of the LHS of (14), i.e. $\|\widehat{f}^L-f^*\|_{L^2(P_X)}$. To this end, we substitute $\dot{f}$ in Section A.1 into f in (14) and obtain

$$\mathbb{E}_X\left[\|\dot{f}-f^*\|_n^2\right]=\int_{[0,1]^D}(\dot{f}-f^*)^2dP_X=\int_{[0,1]^D}(\dot{f}-f^*)^2d\lambda\frac{dP_X}{d\lambda}\leqslant\|\dot{f}-f^*\|_{L^2}^2\sup_{x\in[0,1]^D}p_X(x), \tag{15}$$

by the Hölder's inequality. Here, p_X is a density of P_X and $\sup_{x\in[0,1]^D}p_X(x)\leqslant B_P$ is finite by the setting.

Then, for all $n\in\mathbb{N}$ and $u>0$, we have

$$\|\widehat{f}^L-f^*\|_{L^2(P_X)}^2$$

$$\leqslant 10B_P\|\dot{f}-f^*\|_{L^2}^2+8\left\{V_n\left(\log\frac{(L+1)N^2}{V_n}+1\right)\right\}^2+4u$$
$$\leqslant 4J^2M^2(2^D+Q-1/2)^2\max\{n^{-2\beta/(2\beta+D)},n^{-\alpha/(\alpha+2D-2)}\}$$
$$+128\frac{\sigma^2(S+1)}{n}\left(\log\frac{(L+1)N^2}{V_n}+1\right)^2+\frac{4C_u}{n},$$

where $u=C_u/n$ with a constant $C_u>0$. Here, we know the number of non-zero parameters $S\leqslant C_SM(1+J(2^D+Q)\max\{n^{-D/(2\beta+D)},n^{-2D-2/(2\alpha+2D-2)}\})$. Then, we substitute it and obtain

$$\|\widehat{f}^L-f^*\|_{L^2(P_X)}^2$$
$$\leqslant\left\{4J^2M^2(2^D+Q-1/2)^2+128\sigma^2C_SM(1+J(2^D+Q)\left(\log\frac{(L+1)N^2}{V_n}+1\right)^2\right\}$$
$$\times\max\{n^{-2\beta/(2\beta+D)},n^{-\alpha/(\alpha+2D-2)}\}+\frac{128\sigma^2+4C_u}{n}.$$

□

Appendix B. Proof of Theorem 2

We follow a technique developed by van der Vaart and van Zanten (2011) and evaluate contraction of the posterior distribution. To this end, we consider the following two steps. At the first step, we consider a bound for the distribution with an empirical norm $\|\cdot\|_n$. Secondly, we derive a bound with an expectation with respect to the $L^2(P_X)$ norm.

In this section, we reuse $\dot{f}\in\mathcal{F}_{NN,\eta}(S,B,L)$ by the neural network $\dot{\Theta}$ which is defined in Section A.1. By employing $\dot{f}$, we can use the bounds for an approximation error $\|f^*-\dot{f}\|_{L^2}$, a number of layers in $\dot{\Theta}$, and a number of non-zero parameters $\|\dot{\Theta}\|_0$.

B.1. **Bound with an empirical norm. Step 1. Preparation**

To evaluate the convergence, we provide some notions for preparation.

We use addition notation for the dataset $Y_{1:n}:=(Y_1,...,Y_n)$ and $X_{1:n}:=(X_1,...,X_n)$ and a probability distribution of $Y_{1:n}$ given $X_{1:n}$ such as

$$P_{n,f}=\prod_{i\in[n]}\mathcal{N}(f(X_i),\sigma^2),$$

with some function f. Let $p_{n,f}$ be a density function of $P_{n,f}$.

Firstly, we provide an event which characterizes a distribution of a likelihood ratio. We apply Lemma 14 in van der Vaart and van Zanten (2011) we obtain that

$$P_{n,f^*}\left(\int\frac{p_{n,f}(Y_{1:n})}{p_{n,f^*}(Y_{1:n})}d\Pi_f(f)\geqslant\exp(-r^2)\Pi_f(f:\|f-f^*\|_n<r)\right)\geqslant 1-\exp(-nr^2/8),$$

for any f and $r>0$. To employ the entropy bound, we will update $\Pi_f(f:\|f-f^*\|_n<r)$ of this bound as $\Pi_f(f:\|f-\dot{f}\|_{L^\infty}<r)$. To this end, we apply Lemma 4 then it yields the following bound such for $\|f-f^*\|_n$ as

$$1-\exp(-nr^2/B_f^2)\leqslant\Pr_X\left(\|f-f^*\|_n\leqslant\|f-\dot{f}\|_{L^\infty}+B_p\|\dot{f}-f^*\|_{L^2}+r\right),$$

for any r and a parameter $B_f > 0$. Using the inequality (9) for $\|\dot{f} - f^*\|_{L^2}$, we define ϵ_n as

$$\epsilon_n \geqslant \|\dot{f} - f^*\|_{L^2},$$

and also substitute $r = B_p\epsilon_n$, then we have

$$1 - \exp(-nB_p^2\epsilon_n^2/B_f^2) \leqslant \Pr_X\left(\|f - f^*\|_n \leqslant \|f - \dot{f}\|_{L^\infty} + 2B_p\epsilon_n\right).$$

Then, we consider an event $\mathcal{E}_r$ as follows and obtain that

$$\begin{aligned} P_{n,f^*}(\mathcal{E}_r) &:= P_{n,f^*}\left(\int \frac{p_{n,f}(Y_{1:n})}{p_{n,f^*}(Y_{1:n})} d\Pi_f(f) \geqslant \exp(-r^2)\Pi_f(f : \|f - \dot{f}\|_{L^\infty} < B_p\epsilon_n)\right) \\ &\geqslant 1 - \exp(-n9B_p^2\epsilon_n^2/8) - \exp(-nB_p^2\epsilon_n^2/B_f^2), \end{aligned} \tag{16}$$

by substituting $r = 3B_p\epsilon_n$.

Secondly, we provide a test function $\phi : Y_{1:n} \mapsto z \in \mathbb{R}$ which can identify the distribution with f^* asymptotically. Let $\mathbb{E}_{n,f}[\cdot]$ be an expectation with respect to $P_{n,f}$. By Lemma 13 in van der Vaart and van Zanten (2011), there exists a test ϕ satisfying

$$\mathbb{E}_{n,f^*}[\phi_r] \leqslant 9\mathcal{N}(r/2, \mathcal{F}_{NN,\eta}(S, B, L), \|\cdot\|_n)\exp(-r^2/8),$$

and

$$\sup_{f \in \mathcal{F}_{NN,\eta}(S,B,L) : \|f - f^*\|_n \geqslant r} \mathbb{E}_{n,f}[1 - \phi_r] \leqslant \exp(-r^2/8),$$

for any $r > 0$ and $j \in \mathbb{N}$. By the entropy bound for $\mathcal{N}(r, \mathcal{F}_{NN,\eta}(S, B, L), \|\cdot\|_n) \leqslant \mathcal{N}(r, \mathcal{F}_{NN,\eta}(S, B, L), \|\cdot\|_{L^\infty})$, we have

$$\mathbb{E}_{n,f^*}[\phi_r] \leqslant r^{-1}18(L+1)N^2\exp(-r^2/8 + S + 1).$$

Step 2. Bound an error with fixed design.

To evaluate contraction of the posterior distribution, we decompose the expected posterior distribution as

$$\begin{aligned} &\mathbb{E}_{f^*}\left[\Pi_f\left(f : \|f - f^*\|_n \geqslant 4\epsilon r | \mathcal{D}_n\right)\right] \\ &\leqslant \mathbb{E}_{f^*}\left[\phi_r\right] + \mathbb{E}_{f^*}\left[\mathcal{E}_r^c\right] + \mathbb{E}_{f^*}\left[\Pi_f(f : \|f - f^*\|_n > 4\epsilon r | \mathcal{D}_n)(1 - \phi_r)\mathbf{1}_{\mathcal{E}_r}\right] \\ &=: A_n + B_n + C_n. \end{aligned}$$

Here, note that a support of Π_f is included in $\mathcal{F}_{NN,\eta}(S, B, L)$ due to the setting of Π.

About A_n, we use the bound about ϕ_r substitute $\sqrt{n}\epsilon r$ into r, then obtain

$$A_n \leqslant 18(\sqrt{n}\epsilon r)^{-1}(L+1)N^2\exp(-n\epsilon^2 r^2/8 + S + 1).$$

About B_n, by using the result of $\mathcal{E}_r$ as (16) and substitute $\sqrt{n}\epsilon r$ into r, then we have

$$B_n \leqslant \exp(-n9B_p^2\epsilon_n^2/8) + \exp(-nB_p^2\epsilon_n^2/B_f^2).$$

About C_n, we decompose the term as

$$C_n = \mathbb{E}_X\left[\mathbb{E}_{n,f^*}\left[\frac{\int_{\mathcal{F}_{NN,\eta}(S,B,L)} \mathbf{1}_{\{\|f - f^*\|_n > 4\epsilon r\}} p_{n,f}(Y_{1:n}) d\Pi_f(f)}{\int_{\mathcal{F}_{NN,\eta}(S,B,L)} p_{n,f}(Y_{1:n}) d\Pi_f(f)}(1 - \phi_r)\mathbf{1}_{\mathcal{E}_r}\right]\right]$$

$$
\begin{aligned}
&= \mathbb{E}_X \left[\mathbb{E}_{n,f*} \left[\frac{\int_{\mathcal{F}} \mathbf{1}_{\{\|f-f*\|_n > 4\epsilon r\}} \frac{p_{n,f}(Y_{1:n})}{p_{n,f*}(Y_{1:n})} d\Pi_f(f)}{\int_{\mathcal{F}} \frac{p_{n,f}(Y_{1:n})}{p_{n,f*}(Y_{1:n})} d\Pi_f(f)} (1-\phi_r) \mathbf{1}_{\mathcal{E}_r} \right] \right] \\
&\leqslant \mathbb{E}_X \left[\mathbb{E}_{n,f*} \left[\int_{f \in \mathcal{F}_{NN,\eta}(S,B,L): \|f-f*\|_n > \sqrt{2}\epsilon r} \frac{p_{n,f}(Y_{a:n})}{p_{n,f*}(Y_{1:n})} d\Pi_f(f) \right. \right. \\
&\qquad \left. \left. \times \exp(n\epsilon^2 r^2) \Pi_f(f : \|f - \dot{f}\|_{L^\infty} < B_p \epsilon_n)^{-1} (1-\phi_r) \mathbf{1}_{\mathcal{E}_r} \right] \right] \\
&= \mathbb{E}_X \left[\mathbb{E}_{n,f*} \left[\int_{f \in \mathcal{F}_{NN,\eta}(S,B,L): \|f-f*\|_n > \sqrt{2}\epsilon r} \frac{p_{n,f}(Y_{a:n})}{p_{n,f*}(Y_{1:n})} d\Pi_f(f) \right. \right. \\
&\qquad \left. \left. \times \exp(n\epsilon^2 r^2 - \log \Pi_f(f : \|f - \dot{f}\|_{L^\infty} < B_p \epsilon_n)) (1-\phi_r) \mathbf{1}_{\mathcal{E}_r} \right] \right]
\end{aligned}
$$

by the definition of $\mathcal{E}_r$. Here, we evaluate $-\log \Pi_f(f : \|f - \dot{f}\|_{L^\infty} < B_p \epsilon_n)$ as

$$
-\log \Pi_f(f : \|f - \dot{f}\|_{L^\infty} < B_p \epsilon_n) \leqslant -\log \Pi_\Theta(\Theta : \|\Theta - \dot{\Theta}\|_\infty < L_f B_p \epsilon_n) \leqslant S \log((B_f L_f \epsilon_n)^{-1}),
$$

where $\dot{\Theta}$ is the parameter which constitute $\dot{f}$ and L_f is a Lipschitz constant of $G_\eta[\cdot]$. Thus, the bound for C_n is rewritten as

$$
\begin{aligned}
C_n &\leqslant \mathbb{E}_X \left[\int_{f \in \mathcal{F}_{NN,\eta}(S,B,L): \|f-f*\|_n > \sqrt{2}\epsilon r} \frac{p_{n,f}(Y_{1:n})}{p_{n,f*}(Y_{1:n})} \mathbb{E}_{n,f} \left[(1-\phi_r) \mathbf{1}_{\mathcal{E}_r} \right] d\Pi_f(f) \right. \\
&\qquad \left. \times \exp(n\epsilon^2 r^2 + S \log((B_f L_f \epsilon_n)^{-1})) \right] \\
&\leqslant \exp\left(n\epsilon^2 r^2 + S \log((B_f L_f \epsilon_n)^{-1}) - \frac{r'^2}{8} \right),
\end{aligned}
$$

here, we introduce r' is a r for defining ϕ_r to identify r for $\mathcal{E}_r$. Here, we substitute $r' = 4\sqrt{n}\epsilon r$, then we have

$$
C_n \leqslant \exp\left(S \log((B_f L_f \epsilon_n)^{-1}) - 2n\epsilon^2 r^2 \right)
$$

Combining the results about A_n, B_n, C_n and D_n, we obtain

$$
\begin{aligned}
&\mathbb{E}_{f*}[\Pi_f(f : \|f - f^*\|_n \geqslant 4\epsilon r | \mathcal{D}_n)] \\
&\leqslant \exp(-n\epsilon^2 r^2/8 + S + 1 + \log 18 (\sqrt{n}\epsilon r)^{-1} (L+1) N^2) \\
&\quad + \exp(-n 9 B_p^2 \epsilon_n^2/8) + \exp(-n B_p^2 \epsilon_n^2 / B_f^2) + \exp\left(S \log((B_f L_f \epsilon_n)^{-1}) - 2n\epsilon^2 r^2 \right) \\
&\leqslant 2 \exp\left(-\max\{9B_p^2/8, B_p^2/B_f^2\} n \epsilon_n^2 \right) \\
&\quad + 2 \exp\left(2n\epsilon^2 r - 2 + C_S'' \max\{n^{-D/(2\beta+D)}, n^{-2D-2/(2\alpha+2D-2)}\}) \log n + 1 \right).
\end{aligned}
$$

by substituting the order or S as (8) as $S = C_S' \max\{n^{-D/(2\beta+D)}, n^{-2D-2/(2\alpha+2D-2)}\})$ where $C_S' = C_S M(1 + J(2^D + Q)$ and C_S'' is a constant as $C_S'' = C_S' \log \max\{-D/(2\beta+D), -2D -$

$2/(2\alpha+2D-2)\})/(B_f L_f)$. By substituting $r=1$ and

$$\epsilon=\epsilon_n\log n=2JM(2^D+Q-1/2)\max\{n^{-\beta/(2\beta+D)},n^{-\alpha/(\alpha+2D-2)}\}\log n,$$

then we obtain

$$\mathbb{E}_{f^*}\left[\Pi_f\left(f:\|f-f^*\|_n\geqslant C_\epsilon\max\{n^{-\beta/(2\beta+D)},n^{-\alpha/(\alpha+2D-2)}\}\log n|\mathcal{D}_n\right)\right]\to 0,$$

as $n\to\infty$ with a constant $C_\epsilon>0$.

□

B.2. **The bound with a $L^2(P_X)$ norm.** We evaluate an expectation of the posterior distribution with respect to the $\|\cdot\|_{L^2(P_X)}$ norm. The term is decomposed as

$$\begin{aligned}
&\mathbb{E}_{f^*}\left[\Pi_f(f:\|f-f^*\|_{L^2(P_X)}>r\epsilon|\mathcal{D}_n)\right]\\
&\leqslant\mathbb{E}_{f^*}\left[\mathbf{1}_{\mathcal{E}_r^c}\right]+\mathbb{E}_{f^*}\left[\mathbf{1}_{\mathcal{E}_r}\Pi_f(f:2\|f-f^*\|_n>r\epsilon|\mathcal{D}_n)\right]\\
&\quad+\mathbb{E}_{f^*}\left[\mathbf{1}_{\mathcal{E}_r}\Pi_f(f:2\|f-f^*\|_{L^2(P_X)}>r\epsilon>\|f-f^*\|_n|\mathcal{D}_n)\right]\\
&=:I_n+II_n+III_n.
\end{aligned}$$

for all $\epsilon>0$ and $r>0$. Since we already bound I_n and II_n in step 2, we will bound III_n.

To bound the empirical norm, we provide the following lemma.

Lemma 4. *Let a finite constant $B_f>0$ satisfy $B_f\geqslant\|\dot f-f^*\|_{L^\infty}$. Then, for any $r>0$ and $f\in\mathcal{F}_{NN,\eta}(S,B,L)$, we have*

$$1-\exp(-nr^2/B_f^2)\leqslant\Pr_X\left(\|f-f^*\|_n\leqslant\|f-\dot f\|_{L^\infty}+B_p\|\dot f-f^*\|_{L^2}+r\right).$$

Proof. We note that the finite B_f exists. We know that $\dot f\in\mathcal{F}_{NN,\eta}(S,B,L)$ is bounded by Lemma 3. Also, $f^*\in\mathcal{F}_{M,J,\alpha,\beta}$ is bounded since it is a finite sum of continuous functions with compact supports.

We evaluate $\|f-f^*\|_n$ as

$$\|f-f^*\|_n\leqslant\|f-\dot f\|_n+\|\dot f-f^*\|_n\leqslant\|f-\dot f\|_{L^\infty}+\|\dot f-f^*\|_n.$$

To bound the term $\|\dot f-f^*\|_n$, we apply the Hoeffding's inequality and obtain

$$1-\exp(-2nr^2/2B_f^2)\leqslant\Pr_X\left(\|\dot f-f^*\|_n\leqslant\|\dot f-f^*\|_{L^2(P_X)}+r\right).$$

Using the inequality (15), we have

$$\Pr_X\left(\|\dot f-f^*\|_n\leqslant\|\dot f-f^*\|_{L^2(P_X)}+r\right)\leqslant\Pr_X\left(\|f-f^*\|_n\leqslant B_p\|\dot f-f^*\|_{L^2}+r\right),$$

then obtain the desired result.

□

By Lemma 4, we know the bound

$$1-\exp(-2nr'^2/2B_f^2)\leqslant\Pr_X\left(\|f-f^*\|_n\leqslant\|f-f^*\|_{L^2(P_X)}+r'\right),$$

for all f such as $\|f\|_{L^\infty}\leqslant B$. WE set $r'=\|f-f^*\|_{L^2(P_X)}$, hence

$$1-\exp\left(-\frac{n\|f-f^*\|_{L^2(P_X)}^2}{B_f^2}\right)\leqslant\Pr_X\left(\|f-f^*\|_n\leqslant 2\|f-f^*\|_{L^2(P_X)}\right).$$

Using this result, we obtain

$$\begin{aligned}
III_n &\leqslant \mathbb{E}_X\left[\mathbb{E}_{n,f^*}\left[\int_{f\in\mathcal{F}_{NN,\eta}(S,B,L):\|f-f^*\|_{L^2(P_X)}>r\epsilon>2\|f-f^*\|_n} \frac{p_{n,f}(Y_{1:n})}{p_{n,f^*}(Y_{1:n})} d\Pi_f(f)\mathbf{1}_{\mathcal{E}_r}\right]\right] \\
&\qquad \times \exp\left(n\epsilon^2 r''^2 - \log\Pi_f(f:\|f-\dot{f}\|_{L^\infty}<B_p\epsilon_n)\right) \\
&\leqslant \int_{f\in\mathcal{F}_{NN,\eta}(S,B,L):\|f-f^*\|_{L^2(P_X)}>r\epsilon} \Pr_X\left(\|f-f^*\|_{L^2(P_X)}>2\|f-f^*\|_n\right) d\Pi_f(f) \\
&\qquad \times \exp\left(n\epsilon^2 r^2 + S\log((B_f L_f\epsilon_n)^{-1})\right) \\
&\leqslant \exp\left(n\epsilon^2 r''^2 + S\log((B_f L_f\epsilon_n)^{-1}) - \frac{nr^2\epsilon^2}{B_f^2}\right),
\end{aligned}$$

where r'' is a parameter for defining $\mathcal{E}_r$. We substitute $r'' = r/\sqrt{2}B$, then we have

$$III_n \leqslant \exp\left(S\log((B_f L_f\epsilon_n)^{-1}) - \frac{nr^2\epsilon^2}{2B_f^2}\right)$$

Following the same discussion in Section B.1, we combine the result and obtain

$$\begin{aligned}
&I_n + II_n + III_n \\
&\leqslant 3\exp\left(-\max\{9B_p^2/8, B_p^2/B_f^2\}n\epsilon_n^2\right) + \exp\left(S\log((B_f L_f\epsilon_n)^{-1}) - nr^2\epsilon^2/2B_f^2\right) \\
&\quad + 3\exp\left(2n\epsilon^2 r - 2 + C_S''\max\{n^{-D/(2\beta+D)}, n^{-2D-2/(2\alpha+2D-2)}\})\log n + 1\right),
\end{aligned}$$

and setting

$$\epsilon = \epsilon_n\log n = 2JM(2^D+Q-1/2)\max\{n^{-\beta/(2\beta+D)}, n^{-\alpha/(\alpha+2D-2)}\}\log n,$$

yields the same results.

□

Appendix C. Proof of Theorem 3

We discuss minimax optimality of the estimator and its convergence rate. We apply the techniques developed by Yang and Barron (1999) and utilized by Raskutti *et al.* (2012).

Let $\widetilde{\mathcal{F}}_{M,J,\alpha,\beta}(\delta)\subset\mathcal{F}_{M,J,\alpha,\beta}$ be a packing set of $\mathcal{F}_{M,J,\alpha,\beta}$ with respect to $\|\cdot\|_{L^2}$, namely, each pair of elements $f, f'\in\widetilde{F}_{M,J,\alpha,\beta}$ satisfies $\|f-f'\|_{L^2}\geqslant\delta$. Following the discussion by Yang and Barron (1999), the minimax estimation error is lower bounded as

$$\min_{\bar{f}}\max_{f^*\in\mathcal{F}_{M,J,\alpha,\beta}}\Pr_{f^*}\left(\|\bar{f}-f^*\|_{L^2(P_X)}\geqslant\frac{\delta_n}{2}\right)\geqslant\min_{\bar{f}}\max_{f^*\in\widetilde{\mathcal{F}}_{M,J,\alpha,\beta}(\delta)}\Pr_{f^*}\left(\|\bar{f}-f^*\|_{L^2(P_X)}\geqslant\frac{\delta_n}{2}\right).$$

Let $\widetilde{f}' := \operatorname{argmin}_{f'\in\widetilde{\mathcal{F}}_{M,J,\alpha,\beta}(\delta)}\|\widetilde{f}-\bar{f}\|$ be a projected estimator $\bar{f}$ onto $\widetilde{\mathcal{F}}_{M,J,\alpha,\beta}(\delta)$. Then, the value is lower bounded as

$$\begin{aligned}
&\min_{\bar{f}}\max_{f^*\in\widetilde{\mathcal{F}}_{M,J,\alpha,\beta}(\delta)}\Pr_{f^*}\left(\|\bar{f}-f^*\|_{L^2(P_X)}\geqslant\frac{\delta_n}{2}\right) \\
&\geqslant\min_{\bar{f}'}\max_{f\in\widetilde{\mathcal{F}}_{M,J,\alpha,\beta}(\delta)}\Pr_f(f\neq\bar{f}')
\end{aligned}$$

$$\geqslant \min_{\bar{f}'} \Pr_{\check{f}\sim U}(\bar{f}' \neq \check{f}),$$

where $\check{f}$ is uniformly generated from $\widetilde{\mathcal{F}}_{M,J,\alpha,\beta}(\delta)$ and Pr_U denotes a probability with respect to the uniform distribution.

We apply the Fano's inequality (summarized as Theorem 2.10.1 in Cover and Thomas (2012)), we obtain

$$\Pr_{\check{f}\sim U}(\bar{f}' \neq \check{f}) \geqslant 1 - \frac{I(F_U; D_n) + \log 2}{\log |\widetilde{\mathcal{F}}_{M,J,\alpha,\beta}(\delta)|},$$

where $I(F_U; Y_{1:n})$ is a mutual information between a uniform random variable F_U on $\widetilde{\mathcal{F}}_{M,J,\alpha,\beta}(\delta)$ and $Y_{1:n}$. The mutual information is evaluated as

$$\begin{aligned}
&I(F_U; Y_{1:n}) \\
&= \frac{1}{|\widetilde{\mathcal{F}}_{M,J,\alpha,\beta}(\delta)|} \sum_{f\in\widetilde{\mathcal{F}}_{M,J,\alpha,\beta}(\delta)} \int \log\left(\frac{p_{n,f}(Y_{1:n})}{E_{F_U}[p_{n,F_U}(Y_{1:n})]}\right) dP_{n,f}(Y_{1:n}) \\
&\leqslant \max_{f\in\widetilde{\mathcal{F}}_{M,J,\alpha,\beta}(\delta)} \int \log\left(\frac{p_{n,f}(Y_{1:n})}{E_{F_U}[p_{n,F_U}(Y_{1:n})]}\right) dP_{n,f}(Y_{1:n}) \\
&\leqslant \max_{f\in\widetilde{\mathcal{F}}_{M,J,\alpha,\beta}(\delta)} \max_{f'\in\widetilde{\mathcal{F}}_{M,J,\alpha,\beta}(\delta)} \int \log\left(\frac{p_{n,f}(Y_{1:n})}{|\widetilde{\mathcal{F}}_{M,J,\alpha,\beta}(\delta)|^{-1} p_{n,f'}(Y_{1:n})}\right) dP_{n,f}(Y_{1:n}) \\
&= \max_{f,f'\in\widetilde{\mathcal{F}}_{M,J,\alpha,\beta}(\delta)} \log|\widetilde{\mathcal{F}}_{M,J,\alpha,\beta}(\delta)| + \int \log\left(\frac{p_{n,f}(Y_{1:n})}{p_{n,f'}(Y_{1:n})}\right) dP_{n,f}(Y_{1:n}).
\end{aligned}$$

Here, we know that

$$\log|\widetilde{\mathcal{F}}_{M,J,\alpha,\beta}(\delta)| \leqslant \log \mathcal{N}(\delta, \mathcal{F}_{M,J,\alpha,\beta}, \|\cdot\|_{L^2}),$$

and

$$\int \log\left(\frac{p_{n,f}(Y_{1:n})}{p_{n,f'}(Y_{1:n})}\right) dP_{n,f}(Y_{1:n}) \leqslant \frac{n}{2}\mathbb{E}_X\left[\|f - f'\|_n^2\right] \leqslant \frac{n}{2}\delta^2,$$

since $f, f' \in \widetilde{\mathcal{F}}_{M,J,\alpha,\beta}(\delta)$.

We will provide a bound for $\log \mathcal{N}(\delta, \mathcal{F}_{M,J,\alpha,\beta}, \|\cdot\|_{L^2})$. Since $\mathcal{F}_{M,J,\alpha,\beta}$ is a sum of M functions in $\mathcal{F}_{1,J,\alpha,\beta}$, we have

$$\log \mathcal{N}(\delta, \mathcal{F}_{M,J,\alpha,\beta}, \|\cdot\|_{L^2}) \leqslant M \log \mathcal{N}(\delta, \mathcal{F}_{1,J,\alpha,\beta}, \|\cdot\|_{L^2}).$$

To bound $\log \mathcal{N}(\delta, \mathcal{F}_{1,J,\alpha,\beta}, \|\cdot\|_{L^2})$, we define $\mathcal{I}_{\alpha,J} := \{\mathbf{1}_R : I^D \to \{0,1\} | R \in \mathcal{R}_{\alpha,J}\}$. We know that $\mathcal{F}_{1,J,\alpha,\beta} = H^\beta(I^D) \otimes \mathcal{I}_{\alpha,J}$, hence we obtain

$$\log \mathcal{N}(\delta, \mathcal{F}_{1,J,\alpha,\beta}, \|\cdot\|_{L^2}) \leqslant \log \mathcal{N}(\delta, H^\beta(I^D), \|\cdot\|_{L^2}) + \log \mathcal{N}(\delta, \mathcal{I}_{\alpha,J}, \|\cdot\|_{L^2}).$$

By the entropy bound for smooth functions (e.g. Theorem 2.7.1 in van der Vaart and Wellner (1996)), we use the bound

$$\log \mathcal{N}(\delta, H^\beta(I^D), \|\cdot\|_{L^2}) \leqslant C_H \delta^{-D/\beta},$$

with a constant $C_H > 0$. Furthermore, about the covering number of $\mathcal{I}_{\alpha,J}$, we use the relation

$$\begin{aligned}\|\mathbf{1}_R - \mathbf{1}_{R'}\|_{L^2}^2 &= \int (\mathbf{1}_R(x) - \mathbf{1}_{R'}(x))^2 dx = \int (\mathbf{1}_R(x) - \mathbf{1}_{R'}(x)) dx \\ &= \int_{\mathbf{x}\in I^D} \mathbf{1}_R(\mathbf{x})(1 - \mathbf{1}_{R'}(\mathbf{x}))d\mathbf{x} =: d_1(R, R'),\end{aligned}$$

where $R, R' \in \mathcal{R}_{\alpha,J}$ and d_1 is a difference distance with a Lebesgue measure for sets by Dudley (1974). By Theorem 3.1 in Dudley (1974), we have

$$\log \mathcal{N}(\delta, \mathcal{R}_{\alpha,J}, d_1) =\leqslant C_\lambda \delta^{-(D-1)/\alpha},$$

with a constant $C_\lambda > 0$. Then, we bound the entropy of $\mathcal{I}_{\alpha,J}$ as

$$\log \mathcal{N}(\delta, \mathcal{I}_{\alpha,J}, \|\cdot\|_{L^2}) \leqslant \log \mathcal{N}(\delta^2, \mathcal{R}_{\alpha,J}, d_1) \leqslant C_\lambda \delta^{-2(D-1)/\alpha}.$$

Substituting the results yields

$$\log \mathcal{N}(\delta, \mathcal{F}_{M,J,\alpha,\beta}, \|\cdot\|_{L^2}) \leqslant MC_H \delta^{-D/\beta} + MC_\lambda \delta^{-2(D-1)/\alpha}.$$

Then, we provide a lower bound of $\Pr_{\check{f}\sim U}(\bar{f}' \neq \check{f})$ as

$$\Pr_{\check{f}\sim U}(\bar{f}' \neq \check{f}) \geqslant \frac{\frac{n}{2}\delta + \log 2}{M \max\{C_H \delta^{-D/\beta}, C_\lambda \delta^{-2(D-1)/\alpha}\}}.$$

By substituting $\delta_n = \max\{n^{-2\beta/(2\beta+D)}, n^{-\alpha/(\alpha+2D-2)}\}$, we finally obtain the statement of Theorem 3.

Appendix D. Proof of Propositions

D.1. **Proof of Proposition 1.** About the polynomial kernel, since the RKHS of the kernel is the Sobolev space, we can find $f^* \in \mathcal{F}_{M,J,\alpha,\beta}$ which is not differentiable. To see the properties of the RKHS, see Berlinet and Thomas-Agnan (2011). About the Gaussian kernel, we consider $f^* = \mathbf{1}_R \in \mathcal{F}_{M,J,\alpha,\beta}$, where $R \subset I^D$ is some open set. By Corollary 4.44 in Steinwart and Christmann (2008), f^* is not contained in the RKHS by the Gaussian kernel. Hence, we obtain the results.

□

D.2. **Proof of Proposition 2.** We will specify $f^* \in \mathcal{F}_{M,J,\alpha,\beta}$ and distribution of X, and derive a convergence rate of the estimator by the Fourier method.

For preparation, we consider $D = 1$ case. Let X be generated by a distribution which realize a specific case $X_i = i/n$. Also, we specify $f^* \in \mathcal{F}_{M,J,\alpha,\beta}$ as

$$f^*(x) = \mathbf{1}_{\{x_1 \geqslant 0.5\}},$$

with $x = (x_1, x_2) \in I^2$. We consider a decomposition of f^* by the trigonometric basis such as

$$\phi_j(x) = \begin{cases} 1 & \text{if } j = 0, \\ \sqrt{2}\cos(2\pi k x) & \text{if } j = 2k, \\ \sqrt{2}\sin(2\pi k x) & \text{if } j = 2k+1, \end{cases}$$

for $k \in \mathbb{N}$. Then, we obtain

$$f^* = \sum_{j \in \mathbb{N} \cup \{0\}} \theta_j^* \phi_j.$$

Here, θ_j^* is a true coefficient.

For the estimator, we review its definition as follows. The estimator is written as

$$\widehat{f}^F = \sum_{j \in [J] \cup \{0\}} \widehat{\theta}_j \phi_j,$$

where $\widehat{\theta}_{j_1, j_2}$ is a coefficient which is defined as

$$\widehat{\theta}_j = \frac{1}{n} \sum_{i \in [n]} Y_i \phi_j(X_i).$$

Also, $J \in \mathbb{N}$ are hyper-parameters. Since ϕ_j is an orthogonal basis in L^2 and the Parseval's identity, an expected loss by the estimator is decomposed as

$$\begin{aligned}
\mathbb{E}_{f^*}\left[\|\widehat{f}^F - f^*\|_{L^2(P_X)}^2\right] &= \mathbb{E}_{f^*}\left[\sum_{j \in \mathbb{N} \cup \{0\}} (\widehat{\theta}_j - \theta_j^*)^2\right] \\
&= \mathbb{E}_{f^*}\left[\sum_{j \in [J] \cup \{0\}} (\widehat{\theta}_j - \theta_j^*)^2 + \sum_{j > J} (\theta_j^*)^2\right] \\
&= \sum_{j \in [J] \cup \{0\}} \mathbb{E}_{f^*}\left[(\widehat{\theta}_j - \theta_j^*)^2\right] + \sum_{j > J} (\theta_j^*)^2.
\end{aligned}$$

Here, we apply Proposition 1.16 in Tsybakov (2009) and obtain

$$\begin{aligned}
\mathbb{E}_{f^*}\left[\|\widehat{f}^F - f^*\|_{L^2(P_X)}^2\right] &= \sum_{j \in [J] \cup \{0\}} \left(\frac{\sigma^2}{n} + \rho_j^2\right) + \sum_{j > J} (\theta_j^*)^2 \\
&\geqslant \sum_{j \in [J] \cup \{0\}} \frac{\sigma^2}{n} + \sum_{j > J} (\theta_j^*)^2 \\
&= \frac{\sigma^2 (J+1)}{n} + \sum_{j > J} (\theta_j^*)^2,
\end{aligned}$$

where $\rho_j := n^{-1} \sum_{i \in [n]} f(X_i)\phi_j(X_i) - \langle f, \phi_j \rangle$ is a residual.

Considering the Fourier transform of step functions, we obtain $\theta_j^* = \frac{1-(-1)^j}{2\pi j}$, hence

$$\sum_{j > J} (\theta_j^*)^2 = \frac{1}{4\pi^2} \Psi(J+1) = \frac{1}{4\pi^2} \sum_{k \in \mathbb{N} \cup \{0\}} \frac{1}{(J+1+k)^2} \geqslant \frac{1}{4\pi^2 (J+1)^2},$$

where Ψ is the digamma function.

Combining the results, we obtain

$$\mathbb{E}_{f^*}\left[\|\widehat{f}^F - f^*\|_{L^2(P_X)}^2\right] \geqslant \frac{\sigma^2 J + 1}{n} + \frac{1}{4\pi^2 (J+1)^2}.$$

We set $J = \lfloor c_J n^{1/3} - 1 \rfloor$ with a constant $c_J > 0$, then we finally obtain

$$\mathbb{E}_{f^*}\left[\|\widehat{f}^F - f^*\|^2_{L^2(P_X)}\right] \geqslant n^{-2/3}\left(\sigma^2 + \frac{1}{4\pi^2}\right).$$

Then, we obtain the lower bound for the $D = 1$ case.

For general $D \in \mathbb{N}$, we set a true function as

$$f^* = \bigotimes_{d\in[D]} \mathbf{1}_{\{\cdot \geqslant 0.5\}}.$$

Due to the tensor structure, we obtain the decomposed form

$$f^* = \sum_{j_1 \in \mathbb{N}\cup\{0\}} \cdots \sum_{j_D \in \mathbb{N}\cup\{0\}} \gamma_{j_1,\ldots,j_D} \bigotimes_{d\in[D]} \phi_{j_d},$$

where $\gamma_{j_1,\ldots,j_D}$ is a coefficient such as

$$\gamma_{j_1,\ldots,j_D} = \prod_{d\in[D]} \theta_{j_d},$$

using θ_{j_d} in the preceding part. Following the same discussion, we obtain the following lower bound as

$$\mathbb{E}_{f^*}\left[\|\widehat{f}^F - f^*\|^2_{L^2(P_X)}\right] \geqslant \frac{\sigma^2(J+1)^D}{n} + D\sum_{j>J}(\theta^*_j)^2.$$

Then, we set $J - 1 = \lfloor n^{1/(2+D)} \rfloor$, we obtain that the bound is written as

$$\mathbb{E}_{f^*}\left[\|\widehat{f}^F - f^*\|^2_{L^2(P_X)}\right] \geqslant n^{-2/(2+D)}\left(\sigma^2 + \frac{D}{2\pi^2}\right).$$

Then, we obtain the claim of the proposition for any $D \in N$.

□