Supplementary materials for "Deep Neural Networks Learn Non-Smooth Functions Effectively."

A Additional definitions

The Hölder Space

Let Ω be an open subset of $\mathbb{R}^D$ and $\beta > 0$ a constant. The Hölder space $H^\beta(\bar{\Omega})$, where $\bar{\Omega}$ is the closure of Ω, is the set of functions $f : \bar{\Omega} \to \mathbb{R}$ such that f is continuously differentiable on $\bar{\Omega}$ up to the order $\lfloor\beta\rfloor$, and the $\lfloor\beta\rfloor$-the derivatives of f are Hölder continuous with exponent $\beta - \lfloor\beta\rfloor$, namely,

$$\sup_{x,x' \in \bar{\Omega}, x \neq x'} \frac{|\partial^a f(x) - \partial^a f(x')|}{|x - x'|^{\beta - \lfloor\beta\rfloor}} < \infty$$

for any multi-index a with $|a| = \lfloor\beta\rfloor$, where ∂^a denotes a partial derivative.. The norm of the Hölder space is defined by

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

Basis Pieces defined by Continuous Embeddings

We redefine a piece as an intersection of J embeddings of D-dimensional balls. we first introduce an extended notion of a *boundary fragment class* which is developed by Dudley (1974) and Mammen et al. (1999).

Preliminarily, let $\mathbb{S}^{D-1} := \{x \in \mathbb{R}^D : \|x\|_2 = 1\}$ is the $D-1$ dimensional sphere, and let $(V_j, F_j)_{j=1}^\ell$ be its coordinate system as a C^∞-differentiable manifold such that $F_j : V_j \to \mathring{B}^{D-1} := \{x \in \mathbb{R}^{D-1} \mid \|x\| < 1\}$ is a diffeomorphism. A function $g : \mathbb{S}^{D-1} \to \mathbb{R}$ is said to be in the Hölder class $H^\alpha(\mathbb{S}^{D-1})$ with $\alpha > 0$ if $g \circ F_j^{-1}$ is in $H^\alpha(\mathring{B})$.

Let $B^D = \{x \in \mathbb{R}^D \mid \|x\| \leqslant 1\}$. A subset $R \subset I^D$ is called a *basic piece* if it satisfies two conditions: (i) there is a continuous embedding $g : B^D \to \mathbb{R}^D$ such that its restriction to the boundary $\mathbb{S}^{D-1}$ is in $H^\alpha(\mathbb{S}^{D-1})$ and $R = I^D \cap \mathrm{Image}(g)$, (ii) there is $1 \leqslant i \leqslant D$ and $h \in H^\alpha(I^{D-1})$ such that the indicator function of R is given by the graph

$$1_R = \Psi_d(x_1, \ldots, x_{i-1}, x_i + h(x_1, \ldots, \breve{x}_i, \ldots, x_D), x_{i+1}, ..., x_D),$$

where Ψ is the Heaviside function. The condition (i) tells that a basic piece belongs to the *boundary fragment class* which is developed by Dudley (1974) and Mammen et al. (1999), while (ii) means R is a set defined by a horizon function discussed in Petersen and Voigtlaender (2018).

B Proof of Theorem 1

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

Given a set of observations $\{X_1, \ldots, X_n\}$, let $\|\cdot\|_n$ be an empirical norm defined by

$$\|f\|_n^2 = n^{-1} \sum_{i=1}^n f(X_i)^2.$$

The empirical norm of a random variable is also defined by

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

The empirical norms are in fact seminorms, which do not satisfy the strong positivity.

Let $\mathcal{F}$ be a vector space with a norm $\|\cdot\|$. For $\epsilon > 0$, the covering number of $\mathcal{F}$ with $\|\cdot\|$ for radius δ is defined by

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

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

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

for all $f \in \Xi_{NN,\eta}(S, B, L)$. It follows from $Y_i = f^*(X_i) + \xi_i$ that

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

A simple calculation yields

$$\|f^* - \widehat{f}^L\|_n^2 \leqslant \|f^* - f\|_n^2 + \frac{2}{n}\sum_{i=1}^n \xi_i(\widehat{f}^L(X_i) - f(X_i)). \tag{7}$$

In the following, we will fix $f \in \Xi_{NN,\eta}(S, B, L)$ and evaluate each of the three terms in the RHS of (7). In the first subsection, we provide a result for approximating $f^* \in \mathcal{F}_{M,J,\alpha,\beta}$ by DNNs. In the second subsection, we evaluate the variance of $\widehat{f}^F$. In the last subsection, we combine the results and derive an overall rate.

B.1 Approximate piecewise functions by DNNs

The purpose of this part is to bound the following error

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

for properly selected $f \in \Xi_{NN,\eta}(S, B, L)$. To this end, we consider an existing Θ with properly selected S, B and L. Our proof is obtained by extending techniques by Yarotsky (2017) and Petersen and Voigtlaender (2018).

Fix $f^* \in \mathcal{F}_{M,J,\alpha,\beta}$ such that $f^* = \sum_{m\in[M]} f_m^* \mathbf{1}_{R_m^*}$ with $f_m^* \in H^\beta$ and $R_m^* \in \mathcal{R}_{\alpha,J}$ for $m \in [M]$. To approximate f^*, we introduce neural networks $\Theta_{f,m}$ and $\Theta_{r,m}$ for each $m \in [M]$, where the number of layers L and non-zero parameters S will be specified later.

For approximation, we introduce some specific architectures of DNNs as building blocks. The DNN $\Theta_+ := (A, b) = ((1, ..., 1)^\top, 0)$ works as *summation*: $G_\eta[\Theta_+](x_1, ..., x_D) = \sum_{d\in[D]} x_d$, and the DNN $\Theta_\times$ plays a role for multiplication: $G_\eta[\Theta_\times](x_1, ..., x_D) \approx \prod_{d\in[D]} x_d$. A network Θ_3 approximates the inner product, i.e., $G_\eta[\Theta_3](x_1, ..., x_M, x_1', ..., x_M') \approx \sum_{m\in[M]} x_m x_m'$. The existence and their approximation errors of $G_\eta[\Theta_\times]$ and $G_\eta[\Theta_3]$ will be shown in Lemma 1 and 2.

We construct a network given by $G_\eta[\Theta_3](G_\eta[\Theta_1](\cdot), G_\eta[\Theta_2](\cdot))$, where Θ_1 and Θ_2 consist of M-dimensional outputs $(\Theta_{f,1}, \ldots, \Theta_{f,M})$ and $(\Theta_{r,1}, \ldots, \Theta_{r,M})$, respectively. We evaluate the distance between f^* and the combined neural network :

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

$$
\begin{aligned}
&+ \left\| \sum_{m\in[M]} G_\eta[\Theta_{f,m}] \otimes G_\eta[\Theta_{r,m}] - G_\eta[\Theta_3](G_\eta[\Theta_{f,1}](\cdot), ..., G_\eta[\Theta_{f,M}](\cdot), G_\eta[\Theta_{r,1}](\cdot), ..., G_\eta[\Theta_{r,M}](\cdot)) \right\|_{L^2} \\
&\leqslant \sum_{m\in[M]} \|(f^*_m - G_\eta[\Theta_{f,m}]) \otimes G_\eta[\Theta_{r,m}]\|_{L^2} + \sum_{m\in[M]} \left\| f^*_m \otimes (\mathbf{1}_{R^*_m} - G_\eta[\Theta_{r,m}]) \right\|_{L^2} \\
&+ \left\| \sum_{m\in[M]} G_\eta[\Theta_{f,m}] \otimes G_\eta[\Theta_{r,m}] - G_\eta[\Theta_3](G_\eta[\Theta_{f,1}](\cdot), ..., G_\eta[\Theta_{f,M}](\cdot), G_\eta[\Theta_{r,1}](\cdot), ..., G_\eta[\Theta_{r,M}](\cdot)) \right\|_{L^2} \\
&=: \sum_{m\in[M]} B_{1,m} + \sum_{m\in[M]} B_{2,m} + B_3. \qquad (8)
\end{aligned}
$$

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

Bound of $B_{1,m}$. The Hölder inequality gives

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

Theorem 1 in Yarotsky (2017) and Theorem A.9 in Petersen and Voigtlaender (2018) guarantee that there exists a neural network $\Theta_{f,m}$ such that $|\Theta_{f,m}| \leqslant c'_1(1+log_2\lceil(1+\beta)\rceil\cdot(1+\beta/D)$, $\|\Theta_{f,m}\|_0 \leqslant C'_1\epsilon^{-D/\beta}0$, $\|\Theta_{f,m}\|_\infty \leqslant \epsilon^{-2s_1}$,and $\|f^*_m - G_\eta[\Theta_{f,m}]\|_{L^2} < \epsilon$, where $c_1, c'_1, s_1 > 0$ are constants depending only on f^*. The neural network $\Theta_{r,m}$ is given by Lemma 3.4 in Petersen and Voigtlaender (2018), for which $\|G_\eta[\Theta_{r,m}]\|_{L^\infty} \leqslant 1$. Combining these results, we obtain

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

Bound of $B_{2,m}$. We have

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

From $f^*_m \in H^\beta(I^D)$, there exists a constant $C_H > 0$ such that $\|f^*_m\|_{L^2} \leqslant C_H$.

Recall that each $R^*_m \in \mathcal{R}_{\alpha,J}$ takes the form $R^*_m = \cap_{j=1}^J R^j_m$ with $R^j_m \in \mathcal{R}_{\alpha,1}$ for some B, and thus $1_{R^j_m} \in \mathcal{HF}_{\alpha,D,B}$ defined in Petersen and Voigtlaender (2018). Then, from Lemma 3.4 in Petersen and Voigtlaender (2018), there are some constants c', c, and $s > 0$ depending on α, D, and B such that for any $\varepsilon > 0$ a neural network $\Theta_{m,j}$ can be found with

$$
\left\|1_{R^j_m} - G_\eta[\Theta_{m,j}]\right\|_{L^2} \leqslant \varepsilon,
$$

$|\Theta_{m,j}| \leqslant c'_2(1+\alpha/D)\log(2+\alpha)$, $\|\Theta_{m,j}\|_0 \leqslant c_2\varepsilon^{-2(D-1)/\alpha}$, and $\|\Theta_{m,j}\|_\infty \leqslant \varepsilon^{-2s_2}$. Note that c_2, c'_2, s_2 depend only on f^* for our purpose.

Define a neural network $\Theta_{r,m}$ by

$$
G[\Theta_{r,m}] := G[\Theta_{\times,J}](G_\eta[\Theta_{m,1}](\cdot), \ldots, G_\eta[\Theta_{m,J}](\cdot)),
$$

where $\Theta_{\times,J}$ is given in Lemma 1 below. It follows that

$$
\begin{aligned}
&\left\|\mathbf{1}_{R^*_m} - G_\eta[\Theta_{r,m}]\right\|_{L^2} \\
&\leqslant \left\|\mathbf{1}_{R^*_m} - \bigotimes_{j\in[J]} G_\eta[\Theta_{m,j}]\right\|_{L^2} + \left\|\bigotimes_{j\in[J]} G_\eta[\Theta_{m,j}] - G_\eta[\Theta_{r,m}]\right\|_{L^2} \qquad (9)
\end{aligned}
$$

The first term of the last line of (9) is bounded by

$$
\begin{aligned}
&\left\|\bigotimes_{j\in[J]} 1_{R_{m,j}} - \bigotimes_{j\in[J]} G_\eta[\Theta_{m,j}]\right\|_{L^2} \\
&= \sum_{j\in[J]} \left\|1_{R_{m,j}} - G_\eta[\Theta_{m,j}]\right\|_{L^2} \prod_{j'=1}^{j} \left\|1_{R_{m,j'}}\right\|_{L^2} \prod_{j''\in[J]\setminus[j]} \|G_\eta[\Theta_{m,j''}]\|_{L^2}
\end{aligned}
$$

$$\leqslant \sum_{j\in[J]} \left\| 1_{R_{m,j}} - G_\eta[\Theta_{m,j}] \right\|_{L^2},$$

where $\|G_\eta[\Theta_{r,m}]\|_\infty \leqslant 1$ is used in the last line. From Lemma 1, the second term of (9) is upper bounded by $(J-1)\varepsilon$. We finally obtain

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

Lemma 1. *Fix $\theta > 0$ arbitrary. There are absolute constants $C_\times > 0$ and $s_\times > 0$ such that for any $\epsilon \in (0, 1/2)$, $D' \in \mathbb{N}$ there exists a neural network $\Theta_{\times,D'}$ of D'-dimensional input with at most $(1 + \log_2 D')/\theta$ layers, $\|\Theta_{\times,D'}\|_0 \leqslant C_\times D'\epsilon^{-\theta}$, $\|\Theta'_2\|_\infty \leqslant \epsilon^{-2s}$, and*

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

Proof. We employ the neural network for multiplication $\Theta_{\times,D'}$ as Proposition 3 in Yarotsky (2017) and Lemma A.3 in Petersen and Voigtlaender (2018), and consider a tree-shaped multiplication network. There are $D'-1$ multiplication networks and the tree has $1 + \log_2 D'$ depth. □

Bound of B_3. Take Θ_3 as the neural network in Lemma 2. Then we obtain

$$B_3 \leqslant M\epsilon.$$

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

$$\left| G_\eta[\Theta_3](x_1, \ldots, x_{D'}, X_{D'+1}, \ldots, x_{2D'}) - \sum_{d\in[D']} x_d x_{D'+d} \right| \leqslant D'\epsilon.$$

Proof. Let Θ_3 be a neural network defined by

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

where $\Theta_\times$ is given by Lemma 1, and Θ_+ is the sammation network given by

$$\Theta_+ := (A, b) = ((1, \ldots, 1)\top, 0).$$

Then, we evaluate the difference as

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

where the last inequality uses Lemma 1. □

Combined bound We combine the results about $B_{1,m}, B_{2,m}$ and B_3, then define $\dot{f} \in \Xi_{NN,\eta}(S, B, L)$ for approximating f^*.

For Θ_1

$$|\Theta_1| \leqslant c_1'\big(1 + \lceil \log_2(1+\beta) \rceil \cdot (1 + \beta/D)\big), \quad |\Theta_1|_0 \leqslant M c_1 \varepsilon_1^{-D/\beta}, \quad \|\Theta_1\|_\infty \leqslant \varepsilon^{-2s_1}.$$

For Θ_2,

$$|\Theta_2| \leqslant c_2'\big(1 + \lceil \log_2(2+\alpha) \rceil \cdot (1 + \alpha/D)\big) + \frac{1 + \log_2 J}{\theta_2}, \quad |\Theta_2|_0 \leqslant MJ\big(c_2\varepsilon_2^{-(2D-2)/\alpha} + c_\times \varepsilon_2^{-\theta_2}\big), \quad \|\Theta_2\|_\infty \leqslant \varepsilon^{-2s_2}.$$

For Θ_3,

$$|\Theta_3| \leqslant 1 + 1/\theta_3, \quad |\Theta_3|_0 \leqslant M + c_\times M\varepsilon^{-\theta_3}, \quad \|\Theta_3\|_\infty \leqslant \varepsilon_3^{-2s_3}.$$

To balance the approximation error and estimation error, the latter of which will be discussed later, we choose the ε_i $(i = 1, 2, 3)$ and θ_i $(i = 2, 3)$ as follows:

$$\varepsilon_1 := a_1 n^{-\beta/(2\beta+D)}, \quad \varepsilon_2 := a_2 n^{-\alpha/(2\alpha+2D-2)}, \quad \varepsilon_3 := a_3 \max\{-\beta/(2\beta+D), n^{-\alpha/(2\alpha+2D-2)}\}, \tag{10}$$

$$\theta_2 := (2D-2)/\alpha, \qquad \theta_3 := \min\{(2D-2)/\alpha, D/\beta\}, \tag{11}$$

where a_1, a_2, a_3 are arbitrary positive constants.

The total network $\dot{\Theta}$ to give $\dot{f} := G_\eta[\Theta_3](G_\eta[\Theta_1], G_\eta[\Theta_2])$. With the above choice of ε_i and θ_i, the maximum numbers of layers, non-zero parameters, and maximum absolute value of parameters in are bounded by

$$\begin{aligned}
&|\Theta| \leqslant C_L(1 + \log_2\big(\max\{1+\beta, 2+\alpha, 1+\log_2 J\}\big))(1 + \max\{\beta/D, \alpha/(2D-2)\}),\\
&\|\Theta\|_0 \leqslant Mc_1\varepsilon_1^{-D/\beta} + MJ\big(c_2\varepsilon_2^{-(2D-2)/\alpha}\big) + M + c_\times M\varepsilon^{-\theta_3}\\
&\quad \leqslant C_S M\Big\{1 + J\max\{n^{D/(2\beta+D)}, n^{2(D-1)/(2\alpha+2D-2)}\}\Big\},\\
&\|\Theta\|_\infty \leqslant C_B \max\{n^{2s(2\beta+D)/\beta}, n^{2s(2\alpha+2D-2)/\alpha}\},
\end{aligned}$$

where $s > 0$ is a positive constant depending only on f^*. The approximation error is given by

$$\begin{aligned}
&\|f^* - \dot{f}\|_{L^2}\\
&\leqslant a_1' M n^{-\beta/(2\beta+D)} + C_H a_2' M(2J-1)n^{-\alpha/(2\alpha+2D-2)} + M\max\{n^{-\beta/(2\beta+D)}, n^{-\alpha/(2\alpha+2D-2)}\}\\
&\leqslant C_{apr}(2J+1)M\max\{n^{-\beta/(2\beta+D)}, n^{-\alpha/(2\alpha+2D-2)}\},
\end{aligned} \tag{12}$$

where $C_{apr} > 0$ is a constant.

B.2 Evaluate an entropy bound of the estimators by DNNs

Here, we evaluate a variance term of $\|\widehat{f}^L - f^*\|_n$ in (7) 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 $\widetilde{\mathcal{F}}_{NN,\delta} \subset \Xi_{NN,\eta}(S,B,L)$ by

$$\widetilde{\mathcal{F}}_{NN,\delta} := \{f - \widehat{f}^L : \|f - \widehat{f}^L\|_n \leqslant \delta, f \in \Xi_{NN,\eta}(S,B,L)\}.$$

Here, we mention that $f \in \widetilde{\mathcal{F}}_{NN,\delta}$ is bounded by providing the following lemma.

Lemma 3. *For any $f \in \Xi_{NN,\eta}(S,B,L)$ with an activation function η satisfying Lipschitz continuity with a constant 1, we obtain*

$$\|f\|_{L^\infty} \leqslant B_{\mathcal{F}},$$

where $B_{\mathcal{F}} > 0$ 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

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

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

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

the last inequality holds by Theorem 14.5 in Anthony and Bartlett (2009) and Lemma 12 in Schmidt-Hieber (2017), and the constant N is defined by

$$N := \prod_{\ell \in [L]} (N_\ell + 1),$$

where N_ℓ be a number of nodes in the ℓ-th layer, and we can obtain $N = O((S/L)^L)$. Then, we obtain

$$2\mathbb{E}_\xi \left[\sup_{f' \in \widetilde{F}_{NN,\delta}} \left| \frac{1}{n} \sum_{i \in [n]} \xi_i f'(X_i) \right| \right] \leqslant 4\sqrt{2} \frac{\sigma \sqrt{S+1} \delta}{n^{1/2}} \left(\log \frac{(L+1)N^2}{B\delta} + 1 \right). \tag{13}$$

With the bound (13) 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 (in Theorem 2.5.8, Giné and Nickl (2015)), and obtain

$$\begin{aligned} &1 - \exp(-nu^2/2\sigma^2\delta^2) \qquad (14) \\ &\leqslant \Pr_\xi \left(4 \sup_{f' \in \widetilde{F}_{NN,\delta}} \left| \frac{1}{n} \sum_{i \in [n]} \xi_i f'(X_i) \right| \leqslant 4\mathbb{E}_\xi \left[\sup_{f' \in \widetilde{F}_{NN,\delta}} \left| \frac{1}{n} \sum_{i=1}^n \xi_i f'(X_i) \right| \right] + u \right) \\ &\leqslant \Pr_\xi \left(4 \sup_{f' \in \widetilde{F}_{NN,\delta}} \left| \frac{1}{n} \sum_{i \in [n]} \xi_i f'(X_i) \right| \leqslant 8\sqrt{2} \frac{\sigma \sqrt{S+1} \delta}{n^{1/2}} \left(\log \frac{(L+1)N^2}{B\delta} + 1 \right) + u \right), \qquad (15) \end{aligned}$$

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

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

To evaluate the variance term, we reform the basic inequality (7) 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

$$-\frac{2}{n}\sum_{i=1}^{n}\xi_i(\widehat{f}^L(X_i)-f(X_i))+\frac{1}{2}\|\widehat{f}^L-f\|_n^2\leqslant 2\|f^*-f\|_n^2. \tag{16}$$

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 (15) be valid for all $f\in\Xi_{NN,\eta}(S,B,L)$, we let $\delta=\max\{\|\widehat{f}^L-f\|_n,V_n\}$. Then, we obtain the bound

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

by using $xy\leqslant\frac{1}{4}x^2+2y^2$. Using this result to (16), we obtain

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

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

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

Then, simple calculation yields

$$\|\widehat{f}^L-f\|_n^2\leqslant 4\left\{V_n\left(\log\frac{(L+1)N^2}{BV_n}+1\right)\right\}^2+2u+4\|f^*-f\|_n^2. \tag{17}$$

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 (17), we obtain

$$\|\widehat{f}^L-f^*\|_n^2\leqslant 10\|f^*-f\|_n^2+8\left\{V_n\left(\log\frac{(L+1)N^2}{BV_n}+1\right)\right\}^2+4u, \tag{18}$$

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

B.3 Combine the results

We combine the results in Sections B.1 and B.2 and evaluate $\|\widehat{f}^L-f^*\|_{L^2(P_X)}$. Wwe apply the inequality (I) in the proof of Lemma 10 of Schmidt-Hieber (2017), we obtain

$$\|\widehat{f}^L-f^*\|_{L^2(P_X)}^2\leqslant(1+\varepsilon)^2\left\{\mathbb{E}_X\left[\|\widehat{f}^L-f^*\|_n^2\right]+(1+\varepsilon)\frac{T^2}{n\varepsilon}(8\log\mathcal{N}(\delta,\Xi_{NN,\eta}(S,B,L),\|\cdot\|_\infty)+18)+19\delta T\right\},$$

for all $\varepsilon,\delta\in(0,1)$. Combining the basis inequality (7), the entropy bound (18), and Lemma 12 in Schmidt-Hieber (2017), we obtain

$$\begin{aligned}
\|\widehat{f}^L-f^*\|_{L^2(P_X)}^2\leqslant 3\Big\{&10\mathbb{E}_X\left[\|\dot{f}-f^*\|_n^2\right]+8\left\{V_n\left(\log\frac{(L+1)N^2}{BV_n}+1\right)\right\}^2+4u\\
&+\frac{3T^2}{n}\left(8(S+1)\log\left(\frac{2n(L+1)N^2}{B}\right)+18\right)+\frac{19T}{n}\Big\},
\end{aligned}$$

by setting $\varepsilon = 0.5$ and $\delta = 1/n$.

About the first term of the RHS, we have

$$\mathbb{E}_X\left[\|\dot{f} - f^*\|_n^2\right] = \int_{[0,1]^D} (\dot{f} - f^*)^2 dP_X = \int_{[0,1]^D} (\dot{f} - f^*)^2 d\lambda \frac{dP_X}{d\lambda} \leqslant \|\dot{f} - f^*\|_{L^2}^2 \sup_{x\in[0,1]^D} p_X(x) \tag{19}$$

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 assumption. Also, it follows from Bernstein's inequality that for any $u > 0$

$$\Pr\Big(\|\dot{f} - f^*\|_n^2 \leqslant \|\dot{f} - f^*\|_{L^2(P_X)}^2 + u\Big) \geqslant 1 - \exp\Big(-\frac{nu^2}{A^2 s_n + Au}\Big), \tag{20}$$

where A is a constant with $\|\dot{f}\|_\infty \leqslant A$ and $\|f^*\|_\infty \leqslant A$, and $s_n = \mathbb{E}|\dot{f}(X) - f^*(X)|^2$.

In (18), by the choice $f = \dot{f}$, we see that $\delta^2 \leqslant C\max\{n^{-2\beta/(2\beta+D)}, n^{-2\alpha/(2\alpha+2D-2)}\}$ with some constant $C > 0$, and thus $\exp(-nu^2/(2\sigma^2\delta^2))$ converges to zero for $u = C_u/n$ with a constant $C_n > 0$. Additionally, in (20), since $s_n \leqslant C\max\{n^{-2\beta/(2\beta+D)}, n^{-2\alpha/(2\alpha+2D-2)}\}$ with some constant $C > 0$, for $u = C_u/n$, we have $\exp(-nu^2/(A^2 s_n + Au))$ goes to zero. It follows then

$$\begin{aligned}
&\|\widehat{f}^L - f^*\|_{L^2(P_X)}^2 \\
&\leqslant 30B_P\|\dot{f} - f^*\|_{L^2}^2 + 24\left\{V_n\left(\log\frac{(L+1)N^2}{BV_n} + 1\right)\right\}^2 + 12u \\
&\quad + \frac{9T^2}{n}\left(8(S+1)\log\left(\frac{2n(L+1)N^2}{B}\right) + 18\right) + \frac{27T}{n} \\
&\leqslant C_a^2(2J+1)^2M^2\max\{n^{-2\beta/(2\beta+D)}, n^{-2\alpha/(2\alpha+2D-2)}\} \\
&\quad + \frac{S+1}{n}\left\{128\sigma^2\left(\log\frac{(L+1)N^2}{BV_n} + 1\right)^2 + 72T^2\log\left(\frac{2n(L+1)N^2}{B}\right)\right\} + \frac{12C_u + 27T + 182T^2}{n},
\end{aligned}$$

with probability converging to one, where $C_a, C_b > 0$ is a constant. Using the bound of the number of non-zero parameters $S \leqslant C_S M\Big\{1 + J\max\{n^{D/(2\beta+D)}, n^{2(D-1)/(2\alpha+2D-2)}\Big\}$, we obtain

$$\begin{aligned}
&\|\widehat{f}^L - f^*\|_{L^2(P_X)}^2 \\
&\leqslant \left\{C_a^2(2J+1)^2M^2 + 1024\sigma^2 C_S M(1+J)\left(\log\frac{(L+1)N^2}{BV_n} + 1\right)^2 + 72T^2\log\left(\frac{2n(L+1)N^2}{B}\right)\right\} \\
&\quad \times \max\{n^{-2\beta/(2\beta+D)}, n^{-\alpha/(\alpha+D-1)}\} + \frac{1024\sigma^2 + 12C_u + 27T + 182T^2}{n}.
\end{aligned}$$

Since V_n, V, L, B, N are polynomial to n, this completes the proof of Theorem 1.

□

C Proof of Theorem 2

We follow a technique developed by van der Vaart and van Zanten (2011) and evaluate contraction of the posterior distribution, and show that

$$\begin{aligned}
&\mathbb{E}_{f*}\Big[\Pi_f\Big(f : \|f - f^*\|_{L^2(P_X)}^2 \geqslant rC_B\max\{n^{-2\beta/(2\beta+D)}, n^{-\alpha/(\alpha+D-1)}\}(\log n)^2|\mathcal{D}_n\Big)\Big] \\
&\leqslant \exp\left(-r^2c_2\max\{n^{D/(2\beta+D)}, n^{(D-1)/(\alpha+D-1)}\}\right),
\end{aligned} \tag{21}$$

for all $r > 0$. By the contraction, we can immediately obtain the statement of Theorem 2. 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 $\dot{f} \in \Xi_{NN,\eta}(S, B, L)$ by the neural network $\dot{\Theta}$ which is defined in Section B.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 $\|\dot{\Theta}\|_0$.

C.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 (12) 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

$$\begin{aligned} P_{n,f*}(\mathcal{E}_r) &:= P_{n,f*}\left(\int \frac{p_{n,f}(Y_{1:n})}{p_{n,f*}(Y_{1:n})} d\Pi_f(f) \geqslant \exp(-r^2)\Pi_f(f : \|f - \dot{f}\|_{L^\infty} < B_p\epsilon_n)\right) \\ &\geqslant 1 - \exp(-n9B_p^2\epsilon_n^2/8) - \exp(-nB_p^2\epsilon_n^2/B_f^2), \end{aligned} \tag{22}$$

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

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

and

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

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

$$\mathbb{E}_{f*}\left[\Pi_f\left(f : \|f - f^*\|_n \geqslant 4\epsilon r|\mathcal{D}_n\right)\right]$$

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

Here, note that a support of Π_f is included in $\Xi_{NN,\eta}(S,B,L)$ 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 (22) 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

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

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

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

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

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

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

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

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

$$\begin{aligned}
&\mathbb{E}_{f*}[\Pi_f(f:\|f-f^*\|_n \geqslant 4\epsilon r|\mathcal{D}_n)] \\
&\leqslant \exp(-n\epsilon^2r^2/8 + S + 1 + \log 18(\sqrt{n}\epsilon r)^{-1}(L+1)N^2)
\end{aligned}$$

$$
\begin{aligned}
&+ \exp(-n9B_p^2\epsilon_n^2/8) + \exp(-nB_p^2\epsilon_n^2/B_f^2) + \exp\left(S\log((B_fL_f\epsilon_n)^{-1}) - 2n\epsilon^2r^2\right) \\
&\leqslant 2\exp\left(-\max\{9B_p^2/8, B_p^2/B_f^2\}n\epsilon_n^2\right) \\
&\quad + 2\exp\left(2n\epsilon^2r - 2 + C_S''\max\{n^{-D/(2\beta+D)}, n^{-2D-2/(2\alpha+2D-2)}\})\log n + 1\right).
\end{aligned}
$$

by substituting the order or S as (11) as $S = C_S'\max\{n^{-D/(2\beta+D)}, n^{-2D-2/(2\alpha+2D-2)}\})$ where $C_S' = C_SM(1 + J(2^D + Q)$ and C_S'' is a constant as $C_S'' = C_S'\log\max\{-D/(2\beta + D), -2D - 2/(2\alpha + 2D - 2)\})/(B_fL_f)$. 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 $C_\epsilon > 0$

C.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 $f \in \Xi_{NN,\eta}(S, B, L)$, 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 $\dot{f} \in \Xi_{NN,\eta}(S, B, L)$ 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 (19), 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

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

Using this result, we obtain

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

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 C.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. Then, we obtain the inequality (21), hence we show the result.

□

D Proof of Theorem 3

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

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

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

where $\breve{f}$ 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

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

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

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

and

$$\int \log\left(\frac{p_{n,f}(Y_{1:n})}{p_{n,f'}(Y_{1:n})}\right) dP_{n,f}(Y_{1:n}) \leqslant \frac{n}{2}\mathbb{E}_X\left[\|f - f'\|_n^2\right] \leqslant \frac{n}{8}\delta'^2,$$

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

We will provide a bound for $\log \mathcal{N}(\delta'/2, \mathcal{F}_{M,J,\alpha,\beta}, \|\cdot\|_{L^2})$. 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 $\mathcal{I}_{\alpha,J} := \{\boldsymbol{1}_R : I^D \to \{0,1\} | R \in \mathcal{R}_{\alpha,J}\}$. 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

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

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

$$\log \mathcal{N}(\delta', \mathcal{I}_{\alpha,J}, \|\cdot\|_{L^2}) = \log \mathcal{N}(\delta'^2, \mathcal{R}_{\alpha,J}, d_1). \tag{23}$$

To bound the term, we provide the following Lemma.

Lemma 5. *We obtain*

$$\log \mathcal{N}(\delta, \mathcal{R}_{\alpha,J}, d_1) \leqslant J\mathcal{N}(\delta/J, \mathcal{R}_{\alpha,1}, d_1).$$

Proof. Fix $\delta > 0$ arbitrary. Let $\widetilde{\mathcal{R}} \subset \mathcal{R}_{\alpha,1}$ be a centers of the δ-covering balls, and $|\widetilde{\mathcal{R}}| = \bar{R}$. Also define that $\widetilde{\mathcal{R}}^J := \{\cap_{j\in[J]} R_j \mid R_j \in \widetilde{\mathcal{R}}\}$. Obviously, we have $\widetilde{\mathcal{R}}^J \subset \mathcal{R}_{\alpha,J}$ and $|\widetilde{\mathcal{R}}^J| = \bar{R}^J$.

Consider $R \in \mathcal{R}_{\alpha,J}$. By its definition, there exist $\widetilde{R}_1, ..., \widetilde{R}_J \in \mathcal{R}_{\alpha,1}$ and satisfy $R = \cap_{j\in[J]} \widetilde{R}_j$. Since $\widetilde{\mathcal{R}}$ is a set of centers of the covering balls, there exist $\dot{R}_1, ..., \dot{R}_J \in \widetilde{\mathcal{R}}$ and $d_1(\dot{R}_j, \widetilde{R}_j) \leqslant \delta$ holds.

Here, we define $\dot{R} \in \widetilde{\mathcal{R}}^J$ as $\dot{R} = \cap_{j\in[J]} \dot{R}_j$. Now, we have

$$d_1(\dot{R}, R) \leqslant \sum_{j\in[J]} d_1(\dot{R}_j, \widetilde{R}_j) \leqslant J\delta.$$

Hence, for arbitrary $R \in \mathcal{R}_{\alpha,J}$, there exists $\dot{R}$ in $\widetilde{\mathcal{R}}^J$ and their distance is bounded by $J\delta$. Now, we can say that $\widetilde{\mathcal{R}}^J$ is a set of centers for covering balls for $\mathcal{R}_{\alpha,J}$ with radius $J\delta$. Since $|\widetilde{\mathcal{R}}^J| = \bar{R}^J$, the statement holds. □

Applying Lemma 5, we obtain

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

Substituting the results yields

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

We provide a lower bound for $\log|\widetilde{\mathcal{F}}_{M,J,\alpha,\beta}(\delta)|$. Let $D(\delta, \mathcal{F}_{M,J,\alpha,\beta}, \|\cdot\|_{L^2})$ be a notation for a packing number $|\widetilde{\mathcal{F}}_{M,J,\alpha,\beta}(\delta)|$. Now, we have

$$\begin{aligned}\log D(\delta, \mathcal{F}_{M,J,\alpha,\beta}, \|\cdot\|_{L^2}) &\geqslant \log D(\delta, \mathcal{F}_{1,J,\alpha,\beta}, \|\cdot\|_{L^2}) \\ &\geqslant \max\{\log D(\delta, H^\beta(I^D), \|\cdot\|_{L^2}), \log D(\delta, \mathcal{I}_{\alpha,J}, \|\cdot\|_{L^2})\}.\end{aligned}$$

Similar to (23),

$$\log D(\delta, \mathcal{I}_{\alpha,J}, \|\cdot\|_{L^2}) = \log D(\delta^2, \mathcal{R}_{\alpha,J}, d_1) \geqslant \log D(\delta^2, \mathcal{R}_{\alpha,1}, d_1).$$

About $\log D(\delta, H^\beta(I^D), \|\cdot\|_{L^2})$, we apply Lemma 3.5 in Dudley (1974) then

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

with some constant $c_{lh} > 0$. About $\log D(\delta^2, \mathcal{R}_{\alpha,1}, d_1)$, since the definition of $\mathcal{R}_{\alpha,1}$ follows the boundary fragmented class by restricting sets as a image of smooth embeddings, we apply Theorem 3.1 in Dudley (1974) and obtain

$$\log D(\delta^2, \mathcal{R}_{\alpha,1}, d_1) \geqslant \log \mathcal{N}(\delta^2, \mathcal{R}_{\alpha,1}, d_1) \geqslant c_{lr} \delta^{-2(D-1)/\alpha},$$

with some constant $c_{lr} > 0$.

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

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

By selecting δ and δ' as having an order $\max\{n^{-2\beta/(2\beta+D)}, n^{-\alpha/(\alpha+2D-2)}\}$ and satisfying

$$1 - \frac{C_H M \delta'^{-D/\beta} + C_\lambda M \delta'^{-2(D-1)/\alpha} + \frac{n}{2}\delta^2 + \log 2}{\max\{c_{lh}\delta^{-D/\beta}, c_{lr}\delta^{-2(D-1)/\alpha}\}} \geqslant \frac{1}{2}.$$

Then, we finally obtain the statement of Theorem 3.

E Proof of Corollary 4

Let us introduce a specific form of a piecewise smooth function $\bar{f} : \mathbb{R}^D \to \mathbb{R}$ as

$$\bar{f}(x) := f_0(x) \otimes \mathbf{1}_{x_1 \geqslant g_1(x_{-1})}(x) + f_1(x)\mathbf{1}_{x_1 \geqslant g_2(x_{-1})}(x),$$

where $f_0, f_1 \in H^\beta(I^D)$ such that $\sup_{x\in I^D} f_0(x) < \inf_{x\in I^D} f_1(x)$ and $g_1, g_2 \in H^\alpha(I^D)$. According to Corollary 6.4.2 in Korostelev and Tsybakov (2012), all linear estimators (4) do not attain the minimax optimal rate for estimating function with a form of $\bar{f}$. Then, combining the results in Theorem 1, 2 and 3, the statement holds.

F Specific Examples of Other Inefficient Methods

Orthogonal series 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

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

Proof. We will specify $f^* \in \mathcal{F}_{M,J,\alpha,\beta}$ and distribution of X, and derive an rate of convergence by 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

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

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 $D \in N_{\geqslant 2}$.

□

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 order is, however, strictly worse than the optimal order. 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.