J. Math. Anal. Appl. 505 (2022) 125561

Contents lists available at ScienceDirect

Journal of Mathematical Analysis and Applications

www.elsevier.com/locate/jmaa

Optimal nonparametric inference via deep neural network

Ruiqi Liu [a], Ben Boukai [b], Zuofeng Shang [c,*]

[a] *Department of Mathematics and Statistics, Texas Tech University, USA*
[b] *Department of Mathematical Sciences, Indiana University–Purdue University Indianapolis, USA*
[c] *Department of Mathematical Sciences, New Jersey Institute of Technology, USA*

ARTICLE INFO

Article history:
Received 29 April 2020
Available online 10 August 2021
Submitted by A. Jentzen

Keywords:
Deep neural network
Nonparametric inference
Tensor product B-splines
Optimal minimax risk bound
Asymptotic distribution
Nonparametric testing

ABSTRACT

Deep neural network is a state-of-art method in modern science and technology. Much statistical literature have been devoted to understanding its performance in nonparametric estimation, whereas the results are suboptimal due to a redundant logarithmic sacrifice. In this paper, we show that such log-factors are not necessary. We derive upper bounds for the L^2 minimax risk in nonparametric estimation. Sufficient conditions on network architectures are provided such that the upper bounds become optimal (without log-sacrifice). Our proof relies on an explicitly constructed network estimator based on tensor product B-splines. We also derive asymptotic distributions for the constructed network and a relating hypothesis testing procedure. The testing procedure is further proved as minimax optimal under suitable network architectures.

1. Introduction

With the remarkable development of modern technology, difficult learning problems can nowadays be tackled smartly via deep learning architectures. For instance, deep neural networks have led to impressive performance in fields such as computer vision, natural language processing, image/speech/audio recognition, social network filtering, machine translation, bioinformatics, drug design, medical image analysis, where they have demonstrated superior performance to human experts. The success of deep networks hinges on their rich expressiveness (see [3], [23], [22], [1], [28], [20] and [31,32]). Recently, deep networks have played an increasingly important role in statistics particularly in nonparametric curve fitting (see [15,11,16,18,24]). Applications of deep networks in other fields such as image processing or pattern recgnition include, to name a few, LeCun et al. [19], Deng et al. [4], Wan et al. [29], Gal and Ghahramani [9], etc.

* Corresponding author.
E-mail address: zshang@njit.edu (Z. Shang).

https://doi.org/10.1016/j.jmaa.2021.125561
0022-247X/

A fundamental problem in statistical applications of deep networks is how accurate they can estimate a nonparametric regression function. To describe the problem, let us consider i.i.d. observations $(Y_i, \mathbf{X}_i)$, $i = 1, 2, \ldots, n$ generated from the following nonparametric model:

$$Y_i = f_0(\mathbf{X}_i) + \epsilon_i, \tag{1.1}$$

where $\mathbf{X}_i \in [0,1]^d$ are i.i.d. d-dimensional predictors for a fixed $d \geq 1$, ϵ_i are i.i.d. random noise with $E(\epsilon_i) = 0$ and $Var(\epsilon_i) = \tau^2$, f_0 is an unknown function belonging to some function space $\mathcal{H}$. For any $L \in \mathbb{N}$ and $\mathbf{p} = (p_1, \ldots, p_L) \in \mathbb{N}^L$, let $\mathcal{F}(L, \mathbf{p})$ denote the collection of network functions from $\mathbb{R}^d$ to $\mathbb{R}$ consisting of L hidden layers with the lth layer including p_l neurons. The problem of interest is to find an order R_n that controls the L^2 minimax risk:

$$\inf_{\widehat{f} \in \mathcal{F}(L,\mathbf{p})} \sup_{f_0 \in \mathcal{H}} \mathbb{E}_{f_0} \left(\|\widehat{f} - f_0\|_{L^2}^2 \middle| \mathbb{X} \right) = O_P(R_n), \tag{1.2}$$

where $\mathbb{X} = \{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$, the infimum is taken over all estimators $\widehat{f} \in \mathcal{F}(L, \mathbf{p})$, and $\mathbb{E}_{f_0}$ represents the expectation taken over the conditional distribution of Y_i's given $\mathbf{X}_i$'s with $Y_i, \mathbf{X}_i$ generated from model (1.1), and O_P stands for stochastic boundedness, which will be formally defined at the end of Section 2. In other words, we are interested in the performance of the "best" network estimator in the "worst" scenario.

Existing results regarding (1.2) are sub-optimal. For instance, when $\mathcal{H}$ is a β-smooth Hölder class and $L, \mathbf{p}$ are properly selected, it has been argued that $R_n = n^{-\frac{2\beta}{2\beta+d}} (\log n)^s$ for some constant $s > 0$; see Kohler and Krzyżak [15], Hamers and Kohler [11], Kohler and Krzyżak [16], Kohler and Mehnert [18], Schmidt-Hieber [24], Suzuki [27], Farrell et al. [8], Liu et al. [21], Wang et al. [30]. Such results are mostly proved based on empirical processes techniques in which the logarithmic factors arise from the entropy bound of the neural network class. The aim of this paper is to fully remove the redundant logarithmic factors, i.e., under proper selections of $L, \mathbf{p}$ one actually has $R_n = n^{-\frac{2\beta}{2\beta+d}}$ in (1.2). This means that neural network estimators can exactly achieve minimax estimation rate. Our proof relies on an explicitly constructed neural network through tensor product B-splines which is proved minimax optimal. One technical contribution of this paper is to show that tensor product B-splines can be effectively expressed by deep networks. Compared with other basis structures such as local Taylor expansions, e.g., Yarotsky [31] and Schmidt-Hieber [24], the tensor product B-splines framework is convenient to our theoretical analysis due to its rich statistical literature; see Huang [12] and Huang [13].

Some interesting byproducts are worth mentioning. First, we will derive the pointwise asymptotic distribution of the constructed neural network estimator which will be useful to establish pointwise confidence interval. Second, the constructed neural network estimator will be further used as a test statistic which is proved optimal when $L, \mathbf{p}$ are properly selected. As far as we know, these are the first provably valid confidence interval and test statistic based on neural networks in nonparametric regression. Third, the rate R_n can be further improved when f_0 satisfies additional structures. Specifically, we will show that $R_n = n^{-\frac{2\beta}{2\beta+1}}$ if f_0 satisfies additive structure, i.e., f_0 is a sum of univariate β-Hölder functions. Such rate is minimax according to Stone [25].

This paper is organized as follows. Section 2 includes some preliminaries on deep networks and defines some notation. In Section 3, we derive upper bounds for the minimax risk and investigate their optimality. Section 4 provides the proof of the main result, which covers the construction of (optimal) network and relates results on network approximation of tensor product B-splines. As by products, we also provide limiting distribution and optimal testing results in Section 5. We further study the additive model using network approximation in Section 6. The Appendix contains the proofs of relevant lemmas and a table indexing some important symbols used in the proof.

2. Preliminaries and notation

In this section, we review some notion about deep networks and function spaces, as well as provide useful symbols or notation used throughout this paper. Throughout let σ denote the rectifier linear unit (ReLU) activation function, i.e., $\sigma(x) = (x)_+$ for $x \in \mathbb{R}$. For any real vectors $\mathbf{v} = (v_1, \ldots, v_r)^T$ and $\mathbf{y} = (y_1, \ldots, y_r)^T$, we define the shift activation function $\sigma_{\mathbf{v}}(\mathbf{y}) = (\sigma(y_1 - v_1), \ldots, \sigma(y_r - v_r))^T$. Let $\mathbf{p} = (p_1, \ldots, p_L) \in \mathbb{N}^L$, and we say $f \in \mathcal{F}(L, \mathbf{p})$ if

$$f(\mathbf{x}) = W_{L+1}\sigma_{\mathbf{v}_L} W_L \sigma_{\mathbf{v}_{L-1}} \ldots W_2 \sigma_{\mathbf{v}_1} W_1 \mathbf{x}, \ \mathbf{x} \in \mathbb{R}^d,$$

where $\mathbf{v}_l \in \mathbb{R}^{p_l}$ is a shift vector and $W_l \in \mathbb{R}^{p_l \times p_{l-1}}$ is a weight matrix, and $\mathbf{x}$ represents the argument of f. We adopt the representation $\mathbf{x} = (x_1, \ldots, x_d)^T$ with x_j being the jth component of $\mathbf{x}$ and the convention $p_0 = d$ and $p_{L+1} = 1$. For simplicity, we only consider fully connected networks and do not make any sparsity assumptions on the entries of $\mathbf{v}_l$ and W_l.

Next let us review the concept of Hölder space. Let $\Omega = [0,1]^d$ denote the domain of the functions. For f defined on Ω, we define the supnorm, L^2-norm and empirical norm of f by $\|f\|_{\sup} = \sup_{\mathbf{x}\in\Omega} |f(\mathbf{x})|$, $\|f\|_{L^2}^2 = \int_\Omega f(\mathbf{x})^2 Q(\mathbf{x}) d\mathbf{x}$ and $\|f\|_n^2 = \frac{1}{n}\sum_{i=1}^n f(\mathbf{X}_i)$, respectively. Here $Q(\cdot)$ is the probability density for the predictor $\mathbf{X}_i$'s. For any $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_d) \in \mathbb{N}^d$, we denote $|\boldsymbol{\alpha}| = \sum_{j=1}^d \alpha_j$ and $\partial^{\boldsymbol{\alpha}} f = \frac{\partial^{|\boldsymbol{\alpha}|} f}{\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d}}$ whenever the partial derivative exists. For any $\beta > 0$ and $F > 0$, let $\Lambda^\beta(F, \Omega)$ denote the ball of β-Hölder functions with radius F, i.e.,

$$\Lambda^\beta(F,\Omega) = \left\{ f : \Omega \to \mathbb{R} \middle| \sum_{\boldsymbol{\alpha}:|\boldsymbol{\alpha}|\le\lfloor\beta\rfloor} \|\partial^{\boldsymbol{\alpha}} f\|_{\sup} + \sum_{\boldsymbol{\alpha}:|\boldsymbol{\alpha}|=\lfloor\beta\rfloor} \sup_{\mathbf{x}_1\neq\mathbf{x}_2\in\Omega} \frac{|\partial^{\boldsymbol{\alpha}} f(\mathbf{x}_1) - \partial^{\boldsymbol{\alpha}} f(\mathbf{x}_2)|}{\|\mathbf{x}_1 - \mathbf{x}_2\|_2^{\beta - \lfloor\beta\rfloor}} \le F \right\},$$

in which $\lfloor\beta\rfloor$ is the largest integer strictly smaller than β.

At the end, we need some notation for vector, matrix and asymptotic analysis. For vector $v = (v_1, \ldots, v_p) \in \mathbb{R}^p$, let $\|v\|_\infty = \max_{1\le i\le p} |v_i|$ and $\|v\|_2 = \sqrt{\sum_{i=1}^p v_i^2}$ be its supnorm and Euclidean norm. Let $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ denote the smallest and largest eigenvalues of a squared matrix. For two positive sequences a_n and b_n, we define $a_n \asymp b_n$ if there exist positive constants C_1, C_2 such that $C_1 a_n \le b_n \le C_2 a_n$. We say a sequence of random variables $X_n = O_P(a_n)$ for some positive deterministic sequence a_n if for any $\varepsilon > 0$, there exists a constant $C_\varepsilon > 0$ such that $P(|X_n| \ge C_\varepsilon a_n) \le \varepsilon$ for all $n \ge 1$. Finally, we denote $X_n = o_P(a_n)$ if $\lim_{n\to\infty} P(|X_n| > \epsilon a_n) = 0$ for any $\epsilon > 0$.

3. Minimax neural network estimation

In this section, we derive an upper bound for the L^2 minimax risk in the problem (1.2). The risk bound will be proved optimal under suitable circumstances. To simplify the expressions, we only consider networks with architecture $(L, \mathbf{p}(T))$, where $\mathbf{p}(T) := (T, \ldots, T) \in \mathbb{N}^L$ for any $T \in \mathbb{N}$. In other words, we focus on networks with L hidden layers and each having T neurons. Our results hold under suitable conditions on L and T as well as the following assumption on the design and model error.

Assumption A1. The probability density $Q(\mathbf{x})$ of $\mathbf{X}$ is supported on Ω. There exists a constant $c > 0$ such that $c^{-1} \le Q(\mathbf{x}) \le c$ for any $\mathbf{x} \in \Omega$. The error terms ϵ_i's are independent of $\mathbf{X}_i$'s.

Theorem 1. *Let Assumption A1 be satisfied. Suppose that $L \to \infty$, $T \to \infty$ and $T \log T = o(n)$ as $n \to \infty$, then for any fixed constants $\beta, F > 0$, it follows that*

$$\inf_{\widehat{f}\in\mathcal{F}(L,\mathbf{p}(T))} \sup_{f_0\in\Lambda^\beta(F,\Omega)} \mathbb{E}_{f_0}\left(\|\widehat{f} - f_0\|_{L^2}^2 \middle| \mathbb{X} \right) = O_P\left(\frac{1}{T^{\frac{2\beta}{d}}} + \frac{T}{n} + \frac{T^2}{4^{\frac{L}{d+k}}} \right), \tag{3.1}$$

where k is the smallest integer satisfying $k \geq \max(\beta, 2)$. As a consequence, if $T \asymp n^{\frac{d}{2\beta+d}}$ and $n^{\frac{2\beta+2d}{2\beta+d}} = O(4^{\frac{L}{d+k}})$, then the following holds:

$$\inf_{\widehat{f} \in \mathcal{F}(L, \mathbf{p}(T))} \sup_{f_0 \in \Lambda^\beta(F,\Omega)} \mathbb{E}_{f_0}\left(\|\widehat{f} - f_0\|_{L^2}^2 \middle| \mathbb{X} \right) = O_P(n^{-\frac{2\beta}{2\beta+d}}).$$

The O_P in Theorem 1 represents stochastic boundedness as defined at the end of Section 2, which involves some fixed constant. The constant term in O_P relies on c (Assumption A1), β (smoothness of f_0), k (auxiliary integer related to β), F (radius of function space), and d (dimension of the design point), and is free of n, T, L. We ignore the constant term as the focus of this paper is to investigate the impact of T, L (network architecture) and n (sample size) on the minimax rate. Moreover, the choice $T \asymp n^{\frac{d}{2\beta+d}}$ would be satisfied if $T = c_0 n^{\frac{d}{2\beta+d}}$ for some fixed constant $c_0 > 0$.

Proof of Theorem 1 relies on an explicitly constructed network estimator based on tensor product B-splines of order k, where $k \geq \max(\beta, 2)$ is the constant specified in condition of Theorem 1. The minimax risk bound in (3.1) consists of three components $T^{-\frac{2\beta}{d}}, n^{-1}T, 4^{-\frac{L}{d+k}}T^2$ corresponding to the bias, variance and approximation error of the constructed network. The optimal risk bound is achieved through balancing the three terms. The approximation error of the constructed network decreases exponentially along with L. Networks constructed based on other methods such as local Taylor approximations ([31], [32] and [24] have similar approximation performance. However, their statistical properties are more challenging to deal with due to the unbalanced eigenvalues of the corresponding basis matrix. In contrast, the eigenvalues of the tensor product B-spline basis matrix are known to have balanced orders, e.g., see de Boor [2], which plays an important role in deriving the risk bound. Also notice that the risk bound will blow out when L is fixed, which partially explains the superior performance of deep networks compared with shallow ones; see Eldan and Shamir [7].

4. Construction of optimal networks

In this section, we prove Theorem 1 by explicitly constructing a network estimator $\widehat{f}_{\text{net}} \in \mathcal{F}(L, \mathbf{p}(T))$ and deriving its risk bound. The construction process starts from a pilot estimator $\widehat{f}_{\text{pilot}}$ obtained under tensor product B-splines with order $k \geq \max(\beta, 2)$. The tensor product B-spline basis functions are further approximated through explicitly constructed multi-layer networks, which will be aggregated to obtain the network estimator $\widehat{f}_{\text{net}}$. The key step is to show that the discrepancies between the tensor product B-spline basis functions and the corresponding network approximations are reasonably small such that $\widehat{f}_{\text{net}}$ will perform similarly as $\widehat{f}_{\text{pilot}}$, and thus, optimally.

Our construction is different from Yarotsky [31] and Schmidt-Hieber [24], where the basis functions are obtained through local Taylor approximation. We find that the eigenvalue performance of the local Taylor basis matrix is difficult to quantify so that the corresponding pilot estimator cannot be used effectively. Instead, the pilot estimator based on tensor product B-splines is more convenient to deal with. Other basis such as wavelets or smoothing splines may also work but this will be explored elsewhere.

4.1. A pilot estimator through tensor product B-splines

In this subsection, we review tensor product B-splines and construct the corresponding pilot estimator. For any integer $M \geq 2$, let $0 = t_0 < t_1 < \cdots < t_{M-1} < t_M = 1$ be knots that form a partition of the unit interval. The definition of univariate B-splines of order $k \geq 2$ depends on additional knots $t_{-k+1} < t_{-k+2} < \ldots < t_{-1} < 0$ and $1 < t_{M+1} < \ldots < t_{M+k-1}$. Given knots $t = (t_{-k+1}, \ldots, t_{M+k-1}) \in \mathbb{R}^{M+2k-1}$, the univariate B-spline basis functions of order k, denoted $B_{i,k}(x)$, $i = -k+1, -k+2, \ldots, M-1$, can be defined inductively by $B_{i,s}(x)$ for $s = 2, 3, \ldots, k$. For $s = 2$ and $-k+1 \leq i \leq M+k-3$, define

$$B_{i,2}(x) = \begin{cases} \frac{x-t_i}{t_{i+1}-t_i}, & \text{if } x \in [t_i, t_{i+1}] \\ \frac{t_{i+2}-x}{t_{i+2}-t_{i+1}}, & \text{if } x \in [t_{i+1}, t_{i+2}] \\ 0, & \text{elsewhere} \end{cases}.$$

Suppose that $B_{i,s}(x)$, $i = -k+1, \ldots, M+k-s-1$ have been defined. We recursively define

$$B_{i,s+1} = a_{i,s} B_{i,s,t} + b_{i,s} B_{i+1,s,t}, \text{ for } i = -k+1, -k+2, \ldots, M+k-s-2, \tag{4.1}$$

where

$$a_{i,s}(x) = \begin{cases} 0, & \text{if } x < t_i \\ \frac{x-t_i}{t_{i+s}-t_i}, & \text{if } t_i \le x \le t_{i+s} \\ 0, & \text{if } x > t_{i+s} \end{cases}, \quad b_{i,s}(x) = \begin{cases} 0, & \text{if } x < t_{t+1} \\ \frac{t_{i+s+1}-x}{t_{i+s+1}-t_{i+1}}, & \text{if } t_{i+1} \le x \le t_{i+s+1} \\ 0, & \text{if } x > t_{i+s+1} \end{cases}.$$

Proceeding with this construction, we can obtain $B_{i,k}(x)$.

To approximate a multivariate function, we adopt the tensor product B-splines. Let $\Gamma = \{-k+1, -k+2, \ldots, 0, 1, \ldots, M-1\}^d$ and $q = |\Gamma| = (M+k-1)^d$. For $\mathbf{i} = (i_1, i_2, \ldots, i_d) \in \Gamma$, we define $D_{\mathbf{i},k}(\mathbf{x}) = \prod_{j=1}^d B_{i_j,k}(x_j)$ and obtain the corresponding pilot estimator

$$\widehat{f}_{\text{pilot}}(\mathbf{x}) = \sum_{\mathbf{i} \in \Gamma} \widehat{b}_{\mathbf{i}} D_{\mathbf{i},k}(\mathbf{x}), \tag{4.2}$$

where $\{\widehat{b}_{\mathbf{i}}, \mathbf{i} \in \Gamma\}$ are the basis coefficients obtained by the following least square estimation:

$$\widehat{C} := [\widehat{b}_{\mathbf{i}}]_{\mathbf{i} \in \Gamma} = \underset{b_{\mathbf{i}}, \mathbf{i} \in \Gamma}{\arg\min} \sum_{i=1}^n \left(Y_i - \sum_{\mathbf{i} \in \Gamma} b_{\mathbf{i}} D_{\mathbf{i},k}(\mathbf{X}_i) \right)^2. \tag{4.3}$$

4.2. Network approximation of tensor product B-splines

In this subsection, we approximate $D_{\mathbf{i},k}$'s through multilayer neural networks. We first construct networks that approximate the univariate B-spline basis $B_{i,k}$'s, and then multiply these networks through a product network $\times_s$ introduced by Yarotsky [31] to approximate the tensor product B-spline basis. Here, the product network $\times_s(x_1, x_2, \ldots, x_s)$ is constructed to approximate the monomials $\prod_{j=1}^s x_j$. Unlike Yarotsky [31] and Schmidt-Hieber [24], our construction proceeds in an inductive manner due to the intrinsic induction structure of B-splines.

To proceed, let us introduce some notation. For $L, p_0, \ldots, p_{L+1} \in \mathbb{N}$, let us denote $\mathcal{NN}(L, (p_0, p_1, \ldots, p_L, p_{L+1}))$ as the class of p_0-input-p_{L+1}-output ReLU neural network functions of L hidden layers, with the jth layer consisting of p_j nodes, for $j = 1, \ldots, L$. In particular, with $p_0 = d$ and $p_{L+1} = 1$, $\mathcal{NN}(L, (p_0, p_1, \ldots, p_L, p_{L+1}))$ is equivalent to $\mathcal{F}(L, (p_1, \ldots, p_{L+1}))$. The following Propositions 1-3 quantify the approximation error of the product network $\times_s$.

Proposition 1. *For any integer $m \ge 1$, there exists $SQ \in \mathcal{NN}(2m, (1, 4, \ldots, 4, 1))$ such that*

$$|SQ(x) - x^2| \le 2^{-2m-2}, \quad \text{for all } x \in [0, 1].$$

Proof of Proposition 1. For $s \ge 1$, let g, g_s be functions taking values in $[0, 1]$ defined as

$$g(x) = \begin{cases} 2x, & \text{if } 0 \le x < 1/2 \\ 2(1-x), & \text{if } 1/2 \le x \ge\le 1 \end{cases}, \quad g_s = \underbrace{g \circ g \circ \cdots g}_{s \text{ times}}.$$

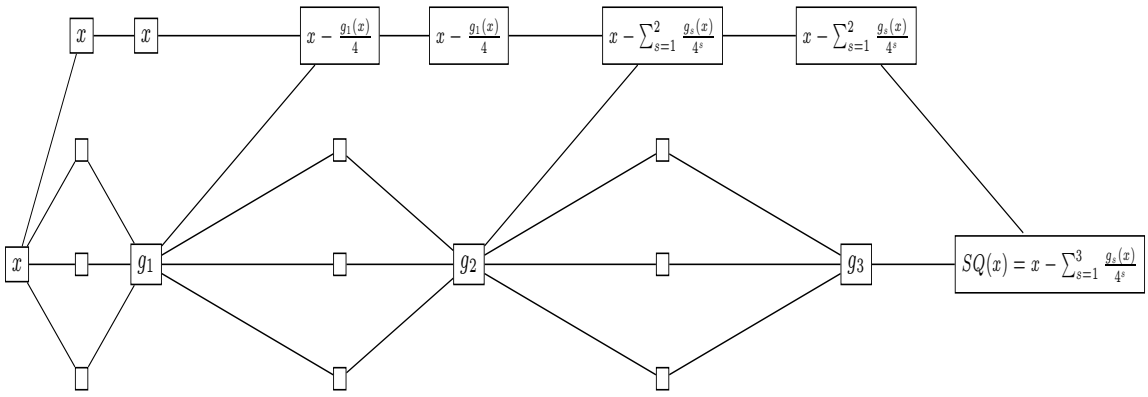

Fig. 1. Construction of SQ when $m = 3$. Clearly, SQ is a network of 6 hidden layers each consisting of at most 4 neurons. For general m, one just adds more layers to construct SQ while the number of neurons on each layer is still not exceeding 4.

It can be shown by induction that

$$g_s(x) = \begin{cases} 2^s \left(x - \frac{2k}{s^2}\right), & \text{if } x \in [\frac{2k}{2s}, \frac{2k+1}{2^s}] \\ 2^s \left(\frac{2k}{2^s} - x\right), & \text{if } x \in [\frac{2k-1}{2s}, \frac{2k}{2^s}] \end{cases}.$$

Let $h_m(x)$ be the linear interpolation of $h(x) = x^2$ at points $k2^{-m}$, for $k = 0, 1, \ldots, 2^m$. Namely,

$$h_m(x) = \frac{2k+1}{2^m} x - \frac{k(k+1)}{4^m}, \quad \text{if } x \in [k2^{-m}, (k+1)2^{-m}].$$

By direct examinations, we have

$$|h(x) - h_m(x)| \leq 2^{-2m-2}, \quad \text{for all } x \in [0, 1].$$

Moreover, by induction, it can be shown that

$$h_{m-1}(x) - h_m(x) = \frac{g_m(x)}{4^m}, \quad \text{for all } x \in [0, 1].$$

The above equation and the fact that $h_0(x) = x$ lead to

$$h_m(x) = x - \sum_{s=1}^{m} \frac{g_s(x)}{4^s}.$$

Since $g(x) = 2\sigma(x) - 4\sigma(x - \frac{1}{2}) + 2\sigma(x - 1)$, $g(x)$ is a neural network consisting of one hidden layer. Define $SQ = h_m$, then SQ is a single-input-single-output neural network of $2m$ hidden layers, and each layer contains 4 neurons, i.e., $SQ \in \mathcal{NN}(2m, (1, 4, \ldots, 4, 1))$; see Fig. 1 for the case when $m = 3$. □

Proposition 2. *For any integer $m \geq 1$, there exists $\times_2 \in \mathcal{NN}(2m + 2, (2, 12, \ldots, 12, 1))$ such that*

$$0 \leq \times_2(x, y) \leq 1, \quad \left|\times_2(x, y) - xy\right| \leq 4^{-m+1}, \quad \textit{for all } x, y \in [0, 1],$$

Proof of Proposition 2. The proof is a modification of Yarotsky [31] to incorporate normalization. Observe that

$$xy = 2\left(\frac{x+y}{2}\right)^2 - \frac{1}{2}x^2 - \frac{1}{2}y^2.$$

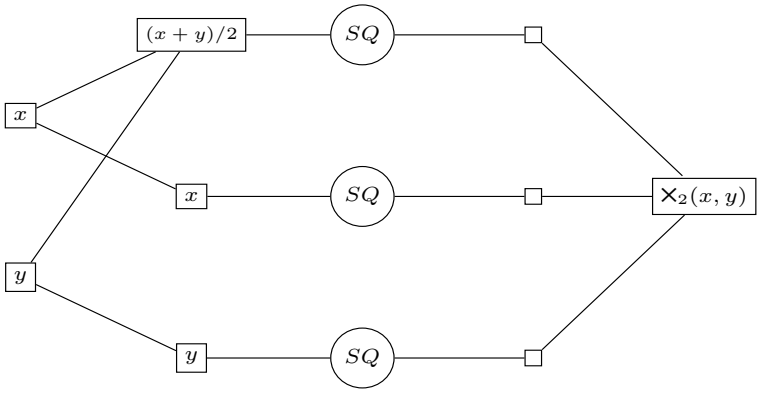

Fig. 2. Construction of $\times_2$. Clearly, $\times_2$ has two more hidden layers than SQ. On each layer the number of neurons is at most three times the number of neurons on each layer of SQ, which is 12.

Each of the functions $(x+y)/2, x, y$ can be realized by a network with one hidden layer. Let SQ denote the network function in Proposition 1. Then we get that for any $0 \le x, y \le 1$,

$$\left|2SQ\left(\frac{x+y}{2}\right) - \frac{1}{2}SQ(x) - \frac{1}{2}SQ(y) - xy\right| \le 4^{-m},$$

and

$$-4^{-m} \le 2SQ\left(\frac{x+y}{2}\right) - \frac{1}{2}SQ(x) - \frac{1}{2}SQ(y) \le 1 + 4^{-m}.$$

Based on the above inequality, we can define

$$\times_2(x,y) = \frac{2SQ\left(\frac{x+y}{2}\right) - \frac{1}{2}SQ(x) - \frac{1}{2}SQ(y) + 4^{-m}}{1 + 2 \times 4^{-m}},$$

which will be guaranteed to take values in $[0,1]$. Moreover, for any $0 \le x, y \le 1$,

$$\left|\times_2(x,y) - xy\right| \le \frac{4 \times 4^{-m}}{1 + 2 \times 4^{-m}} \le 4^{-m+1}.$$

Compared with SQ, $\times_2$ has two additional hidden layers with two inputs and at most 12 nodes in each hidden layer; see Fig. 2. Proof is complete. □

Proposition 3. *For any integers $m \ge 1$ and $s \ge 2$, there exists a neural network function $\times_s$ with $(s-1)(2m+3)-1$ hidden layers and $10+s$ nodes in each hidden layer such that for all $x_1, x_2, \ldots, x_s \in [0,1]$, $0 \le \times_s(x_1, x_2, \ldots, x_s) \le 1$. Moreover, if $|\widetilde{x}_j - x_j| \le \delta$ with $\widetilde{x}_j \in [0,1]$ for $j = 1, 2, \ldots, s$, then*

$$\left|\times_s(\widetilde{x}_1, \widetilde{x}_2, \ldots, \widetilde{x}_s) - \prod_{j=1}^{s} x_j\right| \le (s-1)4^{-m+1} + s\delta.$$

Proof of Proposition 3. Let $\delta_m = 4^{-m+1}$. Here we only prove the case when $s = 3$, and the case for $s > 3$ can be proved inductively. First we apply $\times_2$ to x_1, x_2 and then apply $\times_2$ to $\times_2(x_1, x_2), x_3$. By triangle inequality, we have

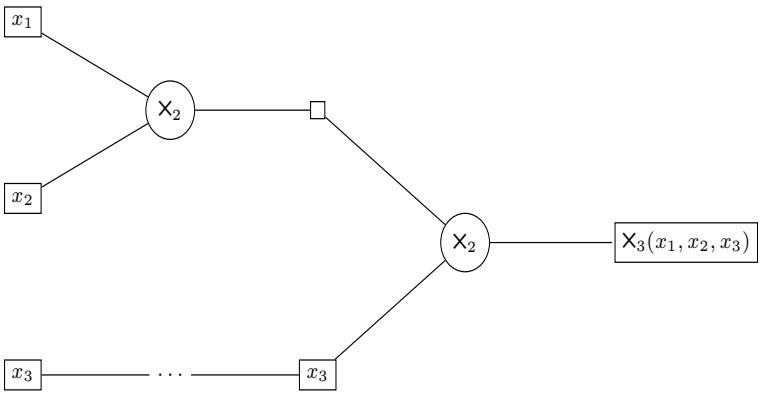

Fig. 3. Construction of $\times_s$ with $s = 3$. $\times_3$ links two $\times_2$ structures sequentially and adds one more hidden layer in the mid. The number of neurons on each hidden layer of $\times_3$ is at most 1 plus the number of neurons on each hidden layer of $\times_2$, which is 13.

$$\left|\times_2\left(\times_2(x_1,x_2),x_3\right)-x_1x_2x_3\right| \le \left|\times_2\left(\times_2(x_1,x_2),x_3\right)-\times_2(x_1,x_2)x_3\right| + \left|\times_2(x_1,x_2)x_3-x_1x_2x_3\right|$$
$$\le 4^{-m+1}+4^{-m+1} \le 2\times 4^{-m+1}.$$

In general, let $\times_s(x_1, x_2, \ldots, x_s) = \times_2(\times_{s-1}(x_1, x_2, \ldots, x_{s-1}), x_s)$ for $s \ge 3$. By induction and triangle inequality, we have

$$\left|\times_s(x_1,x_2,\ldots,x_s)-\prod_{j=1}^{s}x_j\right| \le (s-1)4^{-m+1}.$$

The desired inequality follows from the trivial fact that $|\prod_{i=1}^s \widetilde{x}_i - \prod_{i=1}^s x_i| \le s\delta$. Since we apply neural network $\times_2$ sequentially $(s-1)$ times and there are $(s-2)$ additional hidden layers to store $\times_i(x_1, \ldots, x_i)$ and $x_{i+1}, \ldots, x_s$ for $i = 2, \ldots, s-1$ (See Fig. 3), the total number of hidden layers is $(s-1)(2m+2)+s-2 = (s-1)(2m+3)-1$. Moreover, the number of nodes on each hidden layer is at most $12+s-2 = 10+s$, due to the fact that the first hidden layer has the most number of nodes. Proof is complete. □

Given Proposition 3, we are ready to approximate the kth order univariate B-spline basis $B_{i,k}$. Fixing integer $m \ge 1$, our method is based on the induction formula (4.1) which allows us to start from approximating $B_{i,2}$. Specifically, we approximate $B_{i,2}$ by $\widetilde{B}_{i,2}$ defined as

$$\widetilde{B}_{i,2}(x) = c_1\sigma(x-t_i) + c_2\sigma(x-t_{i+1}) + c_3\sigma(x-t_{i+2}),$$

where

$$c_1 = \frac{1}{t_{i+1}-t_i},\quad c_2 = -\frac{t_{i+2}-t_i}{t_{i+2}-t_{i+1}}c_1,\quad c_3 = -(t_{i+2}-t_i+1)c_1 - (t_{i+2}-t_{i+1}+1)c_2. \tag{4.4}$$

The piecewise linear function $\widetilde{B}_{i,2}$ is exactly a neural network with one hidden layer consisting of three nodes. Suppose that we have constructed $\widetilde{B}_{i,s}(x)$, a neural network approximation of $B_{i,s}$. Next we will approximate $B_{i,s+1}$. For $-k+1 \le i \le M+k-s-1$, we define piecewise linear functions

$$\widetilde{a}_{i,s}(x) = \begin{cases} 0, & \text{if } x < t_i \\ \frac{x-t_i}{t_{i+s}-t_i}, & \text{if } t_i \le x \le t_{i+s} \\ 1, & \text{if } x > t_{i+s} \end{cases},\quad \widetilde{b}_{i,s}(x) = \begin{cases} 1, & \text{if } x < t_{i+1} \\ \frac{t_{i+s+1}-x}{t_{i+s+1}-t_{i+1}}, & \text{if } t_{i+1} \le x \le t_{i+s+1} \\ 0, & \text{if } x > t_{i+s+1} \end{cases}.$$

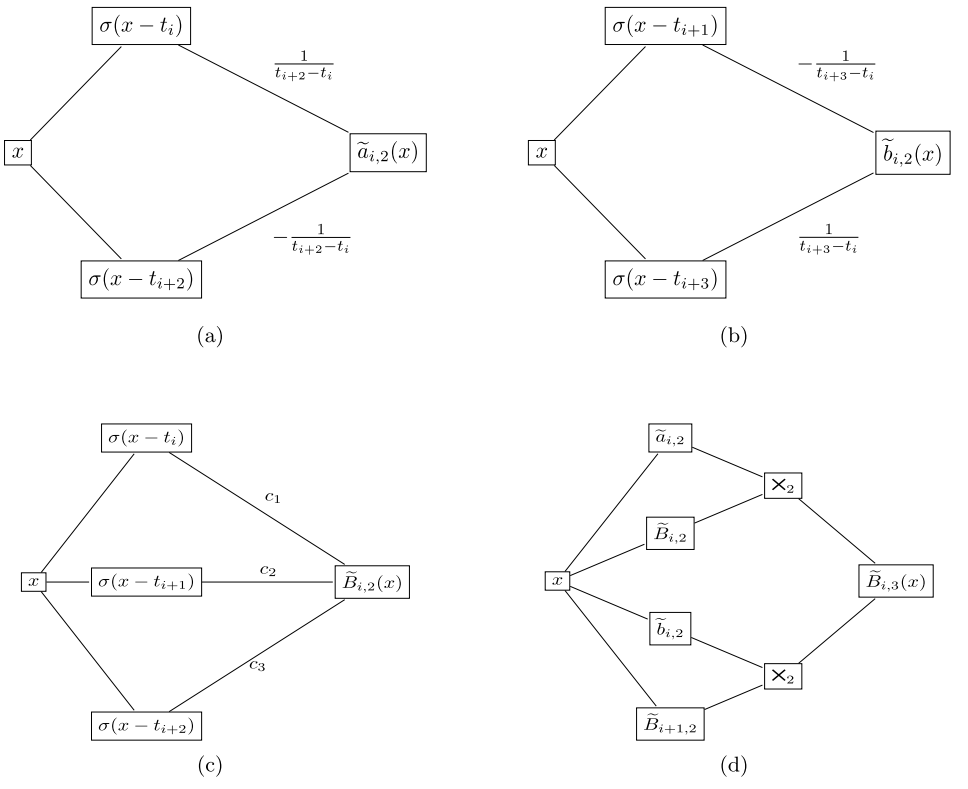

Fig. 4. Construction of $\widetilde{B}_{i,3}$ through induction. (a) and (b) demonstrate the architectures of the networks $\widetilde{a}_{i,2}$ and $\widetilde{b}_{i,2}$. (c) demonstrates the architecture of the network $\widetilde{B}_{i,2}$ with c_1, c_2, c_3 defined in (4.4). (d) demonstrates the induction relationship between $\widetilde{B}_{i,3}$ and $\widetilde{B}_{i,2}$.

In terms of ReLU activation function, we can rewrite the above as $\widetilde{a}_{i,s}(x) = \frac{1}{t_{i+s}-t_i}\sigma(x-t_i) - \frac{1}{t_{i+s}-t_i}\sigma(x-t_{i+s})$ and $\widetilde{b}_{i,s}(x) = -\frac{1}{t_{i+s+1}-t_{i+1}}\sigma(x-t_{i+1}) + \frac{1}{t_{i+s+1}-t_i}\sigma(x-t_{i+s+1}) + 1$, which implies that $\widetilde{a}_{i,s}$ and $\widetilde{b}_{i,s}$ are exactly neural networks with one hidden layer consisting of two nodes (see Fig. 4). For $i = -k+1, \ldots, M+k-s-2$, we define

$$\widetilde{B}_{i,s+1}(x) = \frac{\times_2(\widetilde{a}_{i,s}(x), \widetilde{B}_{i,s}(x)) + \times_2(\widetilde{b}_{i,s}(x), \widetilde{B}_{i+1,s}(x)) + 2\times 4^{-m+1} + \frac{8^s}{7}4^{-m}}{1 + 4\times 4^{-m+1} + \frac{8^s}{14}4^{-m+1}}, \quad \text{for } x \in [0,1].$$

The 'seemingly strange' normalizing constant forces $\widetilde{B}_{i,s+1}(x)$ to take values in $[0,1]$. We repeat the above steps until we reach the construction of $\widetilde{B}_{i,k}$ (see Fig. 4 for an illustration of such induction). We then approximate $B_{i,k}$ by $\widetilde{B}_{i,k}$.

Finally, let us count the number of nodes in each hidden layer of $\widetilde{B}_{i,k}$. Suppose $\widetilde{B}_{i,k}$ has W_k nodes in each hidden layer. Since $\widetilde{B}_{i,2} \in \mathcal{NN}(1,(1,3,1))$ for and $\widetilde{a}_{i,s}, \widetilde{b}_{i,s} \in \mathcal{NN}(1,(1,2,1))$ for all i,s, we know $W_2 = 3$. By Fig. 4(d) and Proposition 2, we show that $W_3 \le \max\{2\times 12, 2\times(2+W_2)\} \le 2W_2 + 28$. By induction, we have that

$$W_k \le 2W_{k-1} + 28 \le 2^{k-2}(W_2+28) - 28 \le 2^{k+3} \tag{4.5}$$

We next approximate the tensor product B-spline basis $D_{\mathbf{i},k}(\mathbf{x}) = \prod_{j=1}^d B_{i_j,k}(x_j)$ by

$$\widetilde{D}_{\mathbf{i},k}(\mathbf{x}) = \times_d(\widetilde{B}_{i_1,k}(x_1), \widetilde{B}_{i_2,k}(x_2), \ldots, \widetilde{B}_{i_d,k}(x_d)), \text{ for each } \mathbf{i} = (i_1,\ldots,i_d) \in \Gamma.$$

Finally, parallelizing $\widetilde{D}_{\mathbf{i},k}(\mathbf{x}), \mathbf{i} \in \Gamma$ according to (4.2), we construct $\widehat{f}_{\text{net}}$ as

$$\widehat{f}_{\text{net}}(\mathbf{x}) = \sum_{\mathbf{i}\in\Gamma} \widehat{b}_{\mathbf{i}} \widetilde{D}_{\mathbf{i},k}(\mathbf{x}), \quad x \in \Omega, \tag{4.6}$$

where the coefficients $\widehat{b}_{\mathbf{i}}$'s are obtained in (4.3).

In comparing (4.2) with (4.6), if we can show that $D_{\mathbf{i},k}$ and $\widetilde{D}_{\mathbf{i},k}$ are close enough, and $\widehat{b}_{\mathbf{i}}$'s are uniformly bounded, then one can expect that $\widehat{f}_{\text{net}}$ performs similarly to $\widehat{f}_{\text{pilot}}$. A rich class of statistical results in literature enable us to efficiently analyze $\widehat{f}_{\text{pilot}}$. In the rest of our analysis, we focus on cardinal B-splines for convenience.

4.3. Approximation error to B-spline basis

The goal of this subsection is to study the differences between $D_{\mathbf{i},k}$'s and $\widetilde{D}_{\mathbf{i},k}$'s. Let $\mathbf{i}_1, \mathbf{i}_2, \ldots, \mathbf{i}_q$ be the elements of Γ, where $q = (M+k-1)^d$ is the total number of tensor product spline basis functions. For simplicity, we define

$$\begin{aligned}
\mathbf{B}_k(x) &= (B_{-k+1,k}(x), B_{-k+2,k}(x), \ldots, B_{M-1,k}(x))^T \in \mathbb{R}^{M-k+1},\\
\mathbf{D}_k(\mathbf{x}) &= (D_{\mathbf{i}_1,k}(\mathbf{x}), D_{\mathbf{i}_2,k}(\mathbf{x}), \ldots, D_{\mathbf{i}_q,k}(\mathbf{x}))^T \in \mathbb{R}^q,\\
\widetilde{\mathbf{B}}_k(x) &= (\widetilde{B}_{-k+1,k}(x), \widetilde{B}_{-k+2,k}(x), \ldots, \widetilde{B}_{M-1,k}(x))^T \in \mathbb{R}^{M-k+1},\\
\widetilde{\mathbf{D}}_k(\mathbf{x}) &= (\widetilde{D}_{\mathbf{i}_1,k}(\mathbf{x}), \widetilde{D}_{\mathbf{i}_2,k}(\mathbf{x}), \ldots, \widetilde{D}_{\mathbf{i}_q,k}(\mathbf{x}))^T \in \mathbb{R}^q.
\end{aligned}$$

Lemmas 1 and 2 bound the approximation errors of $\widetilde{\mathbf{B}}_k(\cdot)$ and $\widetilde{\mathbf{D}}_k(\cdot)$.

Lemma 1. *Given integers $k, M, m \geq 2$ and knots $t_{-k+1} < t_{-k+2} < \ldots < t_0 < t_1 < \ldots < t_M < t_{M+1} < \ldots < t_{M+k-1}$ such that $t_0 = 0, t_M = 1$, there exists a $\widetilde{\mathbf{B}}_k \in \mathcal{NN}(k(2m+3), (1, 2^{k+4}(M+2k), \ldots, 2^{k+4}(M+2k), M+k-1))$ taking values in $[0,1]$, such that*

$$\sup_{x\in[0,1]} \|\widetilde{\mathbf{B}}_k(x) - \mathbf{B}_k(x)\|_\infty \leq \frac{8^k}{14} 4^{-m}.$$

Proof of Lemma 1. First we will approximate $B_{i,2}$, the linear B-spline, using ReLU neural network. Review that for $i = -k+1, \ldots, M+k-3$,

$$B_{i,2}(x) = \begin{cases} \frac{x-t_i}{t_{i+1}-t_i}, & \text{if } x \in [t_i, t_{i+1}] \\ \frac{t_{i+2}-x}{t_{i+2}-t_{i+1}}, & \text{if } x \in [t_{i+1}, t_{i+2}] \\ 0, & \text{elsewhere} \end{cases}.$$

It is easily verified that $B_{i,2}(x) = c_1\sigma(x-t_i) + c_2\sigma(x-t_{i+1}) + c_3\sigma(x-t_{i+2})$, where

$$c_1 = \frac{1}{t_{i+1}-t_i}, \quad c_2 = -\frac{t_{i+2}-t_i}{t_{i+2}-t_{i+1}} c_1, \quad c_3 = -(t_{i+2}-t_i+1)c_1 - (t_{i+2}-t_{i+1}+1)c_2.$$

This implies that $B_{i,2}$ is exactly a ReLU neural network (hence, $\widetilde{B}_{i,2} = B_{i,2}$) with approximation error $\delta_2 = \sup_{x\in[0,1]} |\widetilde{B}_{i,2}(x) - B_{i,2}(x)| = 0$ for all $-k+1 \leq i \leq M+k-3$. Trivially, $B_{i,2}$ takes values in $[0,1]$.

Suppose that we have constructed a neural network approximation $\widetilde{B}_{i,s}$ of $B_{i,s}$ with approximation error $\delta_s = \sup_{x\in[0,1]} |\widetilde{B}_{i,s}(x) - B_{i,s}(x)|$. Moreover, $0 \leq \widetilde{B}_{i,s}(x) \leq 1$ for all $x \in [0,1]$.

Now we will approximate $B_{i,s+1}$. By definition B-splines, we have

$$B_{i,s+1}(x) = \frac{x-t_i}{t_{i+s}-t_i} B_{i,s}(x) + \frac{t_{i+s+1}-x}{t_{i+s+1}-t_{i+1}} B_{i+1,s}(x). \tag{4.7}$$

Let us recall the previously defined piecewise linear functions:

$$a_{i,s}(x)=\begin{cases}0, & \text{if } x<t_i\\ \frac{x-t_i}{t_{i+s}-t_i}, & \text{if } t_i\le x\le t_{i+s}\\ 0, & \text{if } x>t_{i+s}\end{cases},\quad \widetilde{a}_{i,s}(x)=\begin{cases}0, & \text{if } x<t_i\\ \frac{x-t_i}{t_{i+s}-t_i}, & \text{if } t_i\le x\le t_{i+s}\\ 1, & \text{if } x>t_{i+s}\end{cases}.$$

Notice that the first term of the right side of (4.7) is $a_{i,s}B_{i,s}$, which can be approximated by $\times_2(\widetilde{a}_{i,s},\widetilde{B}_{i,s})$. Clearly, $\widetilde{a}_{i,s}(x)=\frac{1}{t_{i+s}-t_i}\sigma(x-t_i)+\sigma(x-t_{i+s})$, which also can be expressed as a ReLU neural network. Moreover, for any $x\in[0,1]$, it follows by Proposition 3 that

$$\begin{aligned}&\left|\times_2(\widetilde{a}_{i,s}(x),\widetilde{B}_{i,s}(x))-a_{i,s}(x)B_{i,s}(x)\right|\\ \le&\left|\times_2(\widetilde{a}_{i,s}(x),\widetilde{B}_{i,s}(x))-\widetilde{a}_{i,s}(x)\widetilde{B}_{i,s}(x)\right|+\left|a_{i,s}(x)B_{i,s}(x)-\widetilde{a}_{i,s}(x)\widetilde{B}_{i,s}(x)\right|\\ \le&\,4^{-m+1}+B_{i,s}(x)\left|a_{i,s}(x)-\widetilde{a}_{i,s}(x)\right|+\widetilde{a}_{i,s}(x)\left|B_{i,s}(x)-\widetilde{B}_{i,s}(x)\right|\\ \le&\,4^{-m+1}+0+\delta_s,\end{aligned}\tag{4.8}$$

where the last inequality follows by the fact that $B_{i,s}$ is supported on $[t_i,t_{i+s}]$. Similarly, let us recall

$$b_{i,s}(x)=\begin{cases}0, & \text{if } x<t_{t+1}\\ \frac{t_{i+s+1}-x}{t_{i+s+1}-t_{i+1}}, & \text{if } t_{i+1}\le x\le t_{i+s+1}\\ 0, & \text{if } x>t_{i+s+1}\end{cases},\quad \widetilde{b}_{i,s}(x)=\begin{cases}1, & \text{if } x<t_{i+1}\\ \frac{t_{i+s+1}-x}{t_{i+s+1}-t_{i+1}}, & \text{if } t_{i+1}\le x\le t_{i+s+1}\\ 0, & \text{if } x>t_{i+s+1}\end{cases}.$$

Notice that the second term of the right side of (4.7) is $b_{i,s}B_{i+1,s}$. Similar to (4.8) we have, for any $x\in[0,1]$,

$$\left|\times_2(\widetilde{b}_{i,s}(x),\widetilde{B}_{i+1,s}(x))-b_{i,s}(x)B_{i+1,s}(x)\right|\le 4^{-m+1}+\delta_s.$$

Now let us recursively define

$$\widetilde{B}_{i,s+1}(x)=\frac{\times_2(\widetilde{a}_{i,s}(x),\widetilde{B}_{i,s}(x))+\times_2(\widetilde{b}_{i,s}(x),\widetilde{B}_{i+1,s}(x))+2\times 4^{-m+1}+2\delta_s}{1+4\times 4^{-m+1}+4\delta_s},$$

which is a ReLU neural network taking values in $[0,1]$. It is not difficult to verify that for any $x\in[0,1]$,

$$\left|\widetilde{B}_{i,s+1}(x)-B_{i,s+1}(x)\right|\le\frac{8\times 4^{-m+1}+8\delta_s}{1+4\times 4^{-m+1}+4\delta_s}\le 8\times 4^{-m+1}+8\delta_s.$$

Taking supremum on the left we get $\delta_{s+1}\le 8\times 4^{-m+1}+8\delta_s$. Using $\delta_2=0$, we can conclude $\delta_s\le\frac{8^s}{14}4^{-m}-\frac{32}{7}4^{-m}\le\frac{8^s}{14}4^{-m}$ for $2\le s\le k$. Deploy $\widetilde{B}_{i,k}$ parallelly to construct the network $\widetilde{\mathbf{B}}_k$.

To count the number hidden layers, we first notice that $\widetilde{B}_{i,2}\in\mathcal{NN}(1,(1,3,1))$ for and $\widetilde{a}_{i,s},\widetilde{b}_{i,s}\in\mathcal{NN}(1,(1,2,1))$ for all i,s by its construction right below Proposition 3. Moreover, from $\widetilde{B}_{i,2}$ to $\widetilde{B}_{i,k}$, we used the network $\times_2$ $k-2$ times. Therefore, by Proposition 2, the number of hidden layers is at most $(2m+2)(k-2)+k-2+1$, which is bounded by $(2m+3)k$. Since in each hidden layer, at most we have $M+2k-3$ different $\widetilde{B}_{i,s}$'s, $\widetilde{a}_{i,s}$'s and $\widetilde{b}_{i,s}$'s for $s=2,\dots,k$. So by (4.5), at most, we have $(2^{k+3}+4)(M+2k)\le 2^{k+4}(M+2k)$ nodes in each hidden layer. The proof is complete. □

Lemma 2. *Given integers $k, M, m \geq 2$ and knots $t_{-k+1} < t_{-k+2} < \ldots < t_0 < t_1 < \ldots < t_M < t_{M+1} < \ldots < t_{M+k-1}$ with $t_0 = 0, t_M = 1$, there exists a $\widetilde{\mathbf{D}}_k \in \mathcal{NN}((2m+3)(k+d-1), (d, 2^{k+4}d(M+2k)^d, \ldots, 2^{k+4}d(M+2k)^d, (M+k-1)^d))$ such that*

$$\sup_{\mathbf{x}\in\Omega} \left\| \widetilde{\mathbf{D}}_k(\mathbf{x}) - \mathbf{D}_k(\mathbf{x}) \right\|_\infty \leq [4(d-1) + 8^k]4^{-m}.$$

Furthermore, each element of $\widetilde{\mathbf{D}}_k$ is in $[0, 1]$.

Proof of Lemma 2. Let $\widetilde{\mathbf{B}}_k(x_1), \widetilde{\mathbf{B}}_k(x_1), \ldots, \widetilde{\mathbf{B}}_k(x_d)$ be the neural networks provided in Lemma 1, which satisfy $|\widetilde{B}_{i,k}(x) - B_{i,k}(x)| \leq \delta_m$, where $\delta_m = 8^k 4^{-m}/14$. For each $(i_1, i_2, \ldots, i_d) \in \{-k+1, -k+2, \ldots, 1, 2, \ldots, M-1\}^d$, we apply the product network $\times_d$ given in Proposition 3 to $(\widetilde{B}_{i_1,k}(x_1), \widetilde{B}_{i_2,k}(x_2), \ldots, \widetilde{B}_{i_d,k}(x_d))$. According to Proposition 3, we have

$$\begin{aligned}\left| \times_d(\widetilde{B}_{i_1,k}(x_1), \widetilde{B}_{i_2,k}(x_2), \ldots, \widetilde{B}_{i_d,k}(x_d)) - \prod_{j=1}^d B_{i_j,k}(x_j) \right| &\leq (d-1)4^{-m+1} + d\delta_m \\ &\leq [4(d-1) + 8^k]4^{-m}.\end{aligned}$$

Now we deploy $\times_d(\widetilde{B}_{i_1,k}(x_1), \widetilde{B}_{i_2,k}(x_2), \ldots, \widetilde{B}_{i_d,k}(x_d))$ parallelly to construct the network $\widetilde{\mathbf{D}}_k$. Since we apply neural network X_d to output of $\widetilde{\mathbf{B}}_k$, so the total number of hidden layers is at most $k(2m+3)+1+(d-1)(2m+3)-1 \leq (2m+3)(d+k)$. Since we parallelly apply $q = (M+k-1)^d$ product networks $\times_d$, the number nodes in each hidden layer is bounded $(10+d)q$, which is further bounded by $d2^{k+4}(M+2k)^d$. This completes the proof. □

In Eckle and Schmidt-Hieber [6], the authors compare neural network methods with multivariate adaptive regression splines (MARS) by showing that any function expressed by MARS can be approximated by a sparse ReLU neural network with an arbitrarily small error. In contrast, Lemma 1 provides a quantitative error bound (in terms of network architecture) for fully connected ReLU neural network approximation of the spline basis. Soon after our work, Kohler et al. [17] independently obtain a relevant result about a quantitative connection between MARS and sparse neural network under smooth activation function.

To end this subsection, let us calculate the number of hidden layers and number of nodes in each hidden layer for $\widehat{f}_{\text{net}}$ defined in (4.6). Notice that to construct $\widehat{f}_{\text{net}}$, we only need to add one more hidden layer to aggregate $\widetilde{D}_{\mathbf{i},k}(\mathbf{x})$ and the coefficients $\widehat{b}_{\mathbf{i}}$. As a consequence, for any integers $k, M, m \geq 2$, we can construct a network $\widehat{f}_{\text{net}}$ such that

$$\widehat{f}_{\text{net}} \in \mathcal{F}(L, \mathbf{p}(T)), \quad \text{with } L = (2m+3)(k+d)+1 \text{ and } T = 2^{k+4}d(M+2k)^d. \tag{4.9}$$

By Proposition 3, we expect $\widehat{f}_{\text{net}} \approx \widehat{f}_{\text{pilot}}$ when $m \to \infty$ (or equivalently $L \to \infty$).

4.4. Asymptotic properties of the pilot estimator

In this subsection, we study the convergence rate of the pilot estimator in (4.2) and the bound of coefficients in (4.3). Let us define $\Phi = (\mathbf{D}_k(\mathbf{X}_1), \ldots, \mathbf{D}_k(\mathbf{X}_n))^T \in \mathbb{R}^{n\times q}$ and $\mathbf{Y} = (Y_1, \ldots Y_n)^T$. Therefore, the coefficients in (4.3) can be expressed as $\widehat{C} = (\Phi^T\Phi)^{-1}\Phi^T\mathbf{Y}$, where the invertibility of the matrix $\Phi^T\Phi$ is guaranteed by Lemma 6 below. Moreover, we denote $\Theta_n = \{g(\mathbf{x}) | g(\mathbf{x}) = V^T\mathbf{D}_k(\mathbf{x}) \text{ for } V \in \mathbb{R}^q\}$ as the linear space spanned by the tenor product B-spline basis $\mathbf{D}_k$'s. An additional assumption is to obtain the desired results, which is stated as follows.

Assumption A2. The knots $\{t_i, i = -k+1, \ldots, M+k-1\}$ have constant separation $h = M^{-1}$. In the theoretical analysis, we require $M \to \infty$ and $h \to 0$.

Remark. Assumption A2 can be relaxed to $\max_i(t_{i+1} - t_i)/\min_i(t_{i+1} - t_i) \leq c$ for some constant $c > 0$, under which one needs to redefine the separation $h = \max_i(t_{i+1} - t_i)$. Results in this section continue to hold. This is a standard assumption for B-spline literature; see Huang [12].

Based on Assumption A2, we can delivery some preliminary lemmas about the B-spline basis. In particular, Lemma 3 quantifies the approximation error of splines; Lemma 4 indicates the equivalence of the norms $\|\cdot\|_n$ and $\|\cdot\|_{L^2}$; Lemma 5 studies the upper and lower bounds of the eigenvalues for the tensor product B-spline basis matrix.

Lemma 3. *For any $f \in \Lambda^\beta(F, \Omega)$, suppose that Assumption A2 is satisfied with some integer $k \geq \beta$. There exists a real sequence $c_{\mathbf{i}}$ such that $\sup_{\mathbf{x}\in\Omega} \left| \sum_{\mathbf{i}\in\Gamma} c_{\mathbf{i}} D_{\mathbf{i},k}(\mathbf{x}) - f(\mathbf{x}) \right| \leq A_f h^\beta$ and $|c_{\mathbf{i}}| \leq A_f$ for all $\mathbf{i} \in \Gamma$. Here $A_f > 0$ is a constant only relying on F, β, k and $\|f\|_{\sup}$. Moreover, it holds that $\sup_{f\in\Lambda^\beta(F,\Omega)} A_f < \infty$, where the upper bound only depends on F, β and k.*

The proof of Lemma 3 requires borrowing some definition from Györfi et al. [10]. Hence, we defer its proof to the Appendix.

Lemma 4. *Suppose Assumptions A1 and A2 hold with some integer $k \geq \max(\beta, 2)$. Moreover, if the sequence h in Assumption A2 satisfies $h = o(1)$ and $\log(h^{-1}) = o(nh^d)$, then*

$$\sup_{g\in\Theta_n} \left| \frac{\|g\|_n^2}{\|g\|_{L^2}^2} - 1 \right| = o_P(1).$$

Proof of Lemma 4. This is Lemma 2.3 in Huang [13]. □

Lemma 5. *Suppose Assumptions A1 and A2 hold with some integer $k \geq \max(\beta, 2)$. Let us define matrix $\mathbf{B} = \int_\Omega \mathbf{D}_k(\mathbf{x})\mathbf{D}_k^T(\mathbf{x})Q(\mathbf{x})d\mathbf{x}$. Then the eigenvalues of $\mathbf{B}$ satisfy that*

$$a_1 h^d \leq \lambda_{\min}(\mathbf{B}) \leq \lambda_{\max}(\mathbf{B}) \leq a_2 h^d,$$

where $0 < a_1 \leq a_2 < \infty$ are constants relying on k and density function Q.

Proof of Lemma 5. It follows from de Boor [2, page 155] that for some constant $\lambda > 1$ depending on k, we have

$$\lambda^{-1}h \leq \lambda_{\min}\left(\int_0^1 \mathbf{B}_k(x)\mathbf{B}_k^T(x)dx\right) \leq \lambda_{\max}\left(\int_0^1 \mathbf{B}_k(x)\mathbf{B}_k^T(x)dx\right) \leq \lambda h.$$

Notice that $\mathbf{D}_k(\mathbf{x})\mathbf{D}_k^T(\mathbf{x}) = \otimes_{j=1}^d \mathbf{B}_k(x_j)\mathbf{B}_k^T(x_j)$ for any $\mathbf{x} = (x_1, x_2, \ldots, x_d)^T \in [0,1]^d$. Here $\otimes$ is the outer product operator. It follows that

$$\int_{[0,1]^d} \mathbf{D}_k(\mathbf{x})\mathbf{D}_k^T(\mathbf{x})d\mathbf{x} = \otimes_{j=1}^d \int_0^1 \mathbf{B}_k(x_j)\mathbf{B}_k^T(x_j)dx_j.$$

By the property of tensor product of matrix, we have

$$\lambda_{\max}\bigg(\int\limits_{[0,1]^d}\mathbf{D}_k(\mathbf{x})\mathbf{D}_k^T(\mathbf{x})d\mathbf{x}\bigg)=\lambda_{\max}^d\bigg(\int\limits_0^1\mathbf{B}_k(x)\mathbf{B}_k^T(x)dx\bigg)\le\lambda^d h^d,$$

$$\lambda_{\min}\bigg(\int\limits_{[0,1]^d}\mathbf{D}_k(\mathbf{x})\mathbf{D}_k^T(\mathbf{x})d\mathbf{x}\bigg)=\lambda_{\min}^d\bigg(\int\limits_0^1\mathbf{B}_k(x)\mathbf{B}_k^T(x)dx\bigg)\ge\lambda^{-d} h^d.$$

By Assumption A1, there exists a constant $c>1$ such that $c^{-1}\int g(\mathbf{x})d\mathbf{x}\le\int g(\mathbf{x})Q(\mathbf{x})d\mathbf{x}\le c\int g(\mathbf{x})d\mathbf{x}$ for any integrable g, which leads to

$$\begin{aligned}C^T\bigg(\int\limits_{[0,1]^d}\mathbf{D}_k(\mathbf{x})\mathbf{D}_k^T(\mathbf{x})Q(\mathbf{x})d\mathbf{x}\bigg)C^T&=\int\limits_{[0,1]^d}|C^T\mathbf{D}_k(\mathbf{x})|^2Q(\mathbf{x})d\mathbf{x}\\&\le c\int\limits_{[0,1]^d}|C^T\mathbf{D}_k(\mathbf{x})|^2d\mathbf{x}\\&\le c\lambda^d h^d,\ \text{for all } C\in\mathbb{R}^q.\end{aligned}$$

Therefore, we have $\lambda_{\max}(\mathbf{B})\le a_2h^d$ with $a_2=c\lambda^d$. Similarly, we can show that the lower bound is valid with $a_1=a_2^{-1}$. Proof is complete. □

To proceed, we need to define the following event

$$\begin{aligned}\Omega_n=\big\{&a_1h^d/2\le\lambda_{\min}(n^{-1}\Phi^T\Phi)\le\lambda_{\max}(n^{-1}\Phi^T\Phi)\le 2a_2h^d\big\}\\&\cap\big\{\|g\|_{L^2}^2/2\le\|g\|_n^2\le 2\|g\|_{L^2}^2,\ \text{for all } g\in\Theta_n\big\},\end{aligned}\tag{4.10}$$

where a_1,a_2 are the constants introduced in Lemma 5. The following lemma reveals the probability of Ω_n approaches one as n diverges, which suggests we can focus our analysis on the event Ω_n.

Lemma 6. *Suppose Assumptions A1 and A2 hold with some integer $k\ge\max(\beta,2)$. Moreover, if the sequence h in Assumption A2 satisfies $h=o(1)$ and $\log(h^{-1})=o(nh^d)$, then it follows that $\lim_{n\to\infty}P(\Omega_n)=1$.*

Proof of Lemma 6. Notice that $n^{-1}\Phi^T\Phi=\sum_{i=1}^n\mathbf{D}_k(\mathbf{X}_i)\mathbf{D}_k^T(\mathbf{X}_i)/n$. Let $\widehat{\mathbf{B}}=n^{-1}\Phi^T\Phi$ and $\mathbf{B}=\int_\Omega\mathbf{D}_k(\mathbf{x})\mathbf{D}_k^T(\mathbf{x})Q(\mathbf{x})d\mathbf{x}$. It follows from Lemma 4 that

$$\begin{aligned}\sup_{u\in\mathbb{R}^q}\bigg|\frac{u^T\widehat{\mathbf{B}}u}{u^T\mathbf{B}u}-1\bigg|&=\sup_{u\in\mathbb{R}^q}\bigg|\frac{\sum_{i=1}^n|u^T\mathbf{D}_k(\mathbf{X}_i)|^2/n}{\int_\Omega|u^T\mathbf{D}_k(\mathbf{x})|^2Q(\mathbf{x})d\mathbf{x}}-1\bigg|\\&=\sup_{g\in\Theta_n}\bigg|\frac{\|g\|_n^2}{\|g\|_{L^2}^2}-1\bigg|=o_P(1).\end{aligned}$$

So the event

$$K_n=\bigg\{\sup_{u\in\mathbb{R}^q}\bigg|\frac{u^T\widehat{\mathbf{B}}u}{u^T\mathbf{B}u}-1\bigg|\le\min(a_2,a_1/2)\bigg\}$$

has probability approaching one. By Lemma 5, on the event K_n, it follows that

$$\sup_{\|u\|_2=1}|u^T\widehat{\mathbf{B}}u|\le\sup_{\|u\|_2=1}|u^T\mathbf{B}u|+\sup_{\|u\|_2=1}|u^T\widehat{\mathbf{B}}u-u^T\mathbf{B}u|$$

$$\leq a_2 h^d + \sup_{\|u\|_2=1} \left| \frac{u^T \widehat{\mathbf{B}} u}{u^T \mathbf{B} u} - 1 \right| \sup_{\|u\|_2=1} |u^T \mathbf{B} u|$$
$$\leq 2a_2 h^d.$$

Similarly, we can show $\inf_{\|u\|_2=1} |u^T A u| \geq a_1 h^d/2$, on the event K_n. Above argument and Lemma 4 together complete the proof. □

Based on the above lemmas, we are ready to provide the main result in this subsection, which provides the convergence rate of the pilot estimator and the bound of $\widehat{C}$.

Lemma 7. *Suppose Assumptions A1 and A2 hold with some integer $k \geq \max(\beta, 2)$. Moreover, if the sequence h in Assumption A2 satisfies $h = o(1)$ and $\log(h^{-1}) = o(nh^d)$, then it follows that*

$$\sup_{f_0 \in \Lambda^\beta(F,\Omega)} \mathbb{E}_{f_0} \left\{ \|\widehat{f}_{pilot} - f_0\|_{L^2}^2 \middle| \mathbb{X} \right\} = O_P\left(h^{2\beta} + \frac{1}{nh^d} \right)$$

and

$$\sup_{f_0 \in \Lambda^\beta(F,\Omega)} \mathbb{E}_{f_0} \left(\widehat{C}^T \widehat{C} \middle| \mathbb{X} \right) = O_P(h^{-d}).$$

Proof of Lemma 7. For any $f_0 \in \Lambda^\beta(F,\Omega)$, let $\mathbf{f}_0 = (f_0(\mathbf{X}_1), \ldots, f_0(\mathbf{X}_n))^T$. Also let $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T$, $\widehat{\mathbf{f}}_{\text{pilot}} = (\widehat{f}_{\text{pilot}}(\mathbf{X}_1), \ldots, \widehat{f}_{\text{pilot}}(\mathbf{X}_n))^T$. According to Lemma 3 and by $k \geq \beta$, there exists a $C = (c_1, c_2, \ldots, c_q)^T \in \mathbb{R}^q$ such that for any $\mathbf{x} \in \Omega$, $|C^T \mathbf{D}_k(\mathbf{x}) - f_0(\mathbf{x})| \leq A_{f_0} h^\beta$. For simplicity, we further define $f^*(\mathbf{x}) = C^T \mathbf{D}_k(\mathbf{x})$ and $\mathbf{f}^* = (f^*(\mathbf{X}_1), \ldots, f^*(\mathbf{X}_n))^\top$.

Notice that on the event Ω_n, $\Phi^T \Phi$ is invertible. The least square algorithm (4.3) implies the following holds on event Ω_n:

$$\begin{aligned} \widehat{\mathbf{f}}_{\text{pilot}} = \Phi(\Phi^T\Phi)^{-1}\Phi^T \mathbf{Y} &= \Phi(\Phi^T\Phi)^{-1}\Phi^T(\Phi C + \mathbf{E} + \boldsymbol{\epsilon}) \\ &= \Phi C + \Phi(\Phi^T\Phi)^{-1}\Phi^T\mathbf{E} + \Phi(\Phi^T\Phi)^{-1}\Phi^T\boldsymbol{\epsilon} \\ &= \mathbf{f}^* + \Phi(\Phi^T\Phi)^{-1}\Phi^T\mathbf{E} + \Phi(\Phi^T\Phi)^{-1}\Phi^T\boldsymbol{\epsilon}, \end{aligned} \tag{4.11}$$

where $\mathbf{E} = (E_1, E_2, \ldots, E_n)^T \in \mathbb{R}^n$ with $E_i = f_0(\mathbf{X}_i) - C^T \mathbf{D}_k(\mathbf{X}_i) = f_0(\mathbf{X}_i) - f^*(\mathbf{X}_i)$. Furthermore, the above equation and Lemma 3 together imply that

$$\begin{aligned} \|\widehat{f}_{\text{pilot}} - f^*\|_n^2 &= \frac{1}{n}(\widehat{\mathbf{f}}_{\text{pilot}} - \mathbf{f}^*)^T(\widehat{\mathbf{f}}_{\text{pilot}} - \mathbf{f}^*) \\ &\leq \frac{2}{n}\mathbf{E}^T\Phi(\Phi^T\Phi)^{-1}\Phi^T\mathbf{E} + \frac{2}{n}\boldsymbol{\epsilon}\Phi(\Phi^T\Phi)^{-1}\Phi^T\boldsymbol{\epsilon} \\ &\leq 2A_{f_0}^2 h^{2\beta} + \frac{2}{n}\boldsymbol{\epsilon}^T\Phi(\Phi^T\Phi)^{-1}\Phi^T\boldsymbol{\epsilon}. \end{aligned}$$

By the fact that $\widehat{f}_{\text{pilot}} - f^* \in \Theta_n$, it holds on event Ω_n that

$$\|\widehat{f}_{\text{pilot}} - f^*\|_{L^2}^2 \leq 2\|\widehat{f}_{\text{pilot}} - f^*\|_n^2,$$

and

$$\mathbb{E}_{f_0}\left(\boldsymbol{\epsilon}^T\Phi(\Phi^T\Phi)^{-1}\Phi^T\boldsymbol{\epsilon} \middle| \mathbb{X} \right) = \text{Tr}\left(\Phi(\Phi^T\Phi)^{-1}\Phi^T \right) = q = (M+k-1)^d \leq 2^d h^{-d},$$

which further implies that

$$\mathbb{E}_{f_0}\left(\|\widehat{f}_{\text{pilot}} - f^*\|_{L^2}^2 \middle| \mathbb{X}\right) \leq 2\mathbb{E}_{f_0}\left(\|\widehat{f}_{\text{pilot}} - f^*\|_n^2 \middle| \mathbb{X}\right) \leq 4A_{f_0}^2 h^{2\beta} + \frac{2^{d+2}}{nh^d}. \tag{4.12}$$

By simple algebra, the above inequality implies that

$$\begin{aligned}\mathbb{E}_{f_0}\left(\|\widehat{f}_{\text{pilot}} - f_0\|_{L^2}^2 \middle| \mathbb{X}\right) &\leq 2\mathbb{E}_{f_0}\left(\|\widehat{f}_{\text{pilot}} - f^*\|_{L^2}^2 \middle| \mathbb{X}\right) + 2\mathbb{E}_{f_0}\left(\|f^* - f_0\|_{L^2}^2 \middle| \mathbb{X}\right) \\ &\leq 8A_{f_0}^2 h^{2\beta} + \frac{2^{d+3}}{nh^d} + 2\|f^* - f_0\|_{\sup}^2 \\ &= \frac{2^{d+3}}{nh^d} + 10A_{f_0}^2 h^{2\beta}, \text{ uniformly for all } f_0 \in \Lambda^\beta(F, \Omega).\end{aligned}$$

Finally, the first statement follows by the uniform boundedness of A_{f_0} over $f_0 \in \Lambda^\beta(F, \Omega)$ in Lemma 3 and $\mathbb{P}(\Omega_n) \to 1$ in Lemma 6.

Let us prove the second statement. According to Lemma 5, it follows that

$$\begin{aligned}\|\widehat{f} - g^*\|_{L^2}^2 &= (\widehat{C} - C)^T \int \mathbf{D}_k(\mathbf{x})\mathbf{D}_k^T(\mathbf{x})Q(\mathbf{x})d\mathbf{x}(\widehat{C} - C) \\ &\geq a_1 h^d (\widehat{C} - C)^T(\widehat{C} - C),\end{aligned}$$

where $a_1 > 0$ is the constant in Lemma 5. Taking conditional expectation and by (4.12), on event Ω_n, we have

$$\begin{aligned}\mathbb{E}_{f_0}\left((\widehat{C} - C)^T(\widehat{C} - C)\middle| \mathbb{X}\right) &\leq \mathbb{E}_{f_0}\left(\|\widehat{f}_{\text{pilot}} - f^*\|_{L^2}^2 \middle| \mathbb{X}\right) \\ &\leq a_1^{-1} 2^{d+2} A_{f_0}^2 \left(h^{2\beta-d} + \frac{1}{nh^{2d}}\right), \text{ uniformly for all } f_0 \in \Lambda^\beta(F, \Omega),\end{aligned}$$

which further leads to

$$\begin{aligned}\mathbb{E}_{f_0}\left(\widehat{C}^T\widehat{C}\middle| \mathbb{X}\right) &\leq 2\mathbb{E}_{f_0}\left((\widehat{C} - C)^T(\widehat{C} - C)\middle| \mathbb{X}\right) + 2C^T C \\ &\leq a_1^{-1} 2^{d+3} A_{f_0}^2 \left(h^{2\beta-d} + \frac{1}{nh^{2d}}\right) + 2qA_{f_0}^2 \\ &\leq a_1^{-1} 2^{d+3} A_{f_0}^2 \left(h^{2\beta-d} + \frac{1}{nh^{2d}}\right) + 2^{d+1} h^{-d} A_{f_0}^2 \\ &\leq a_1^{-1} 2^{d+3} A_{f_0}^2 \left(h^{2\beta-d} + \frac{1}{nh^{2d}} + h^{-d}\right) \\ &\leq a_1^{-1} 2^{d+4} A_{f_0}^2 h^{-d}, \text{ uniformly for all } f_0 \in \Lambda^\beta(F, \Omega),\end{aligned}$$

where the last inequality holds by the fact $h^{2\beta} + n^{-1}h^{-d} = o(1)$. Finally, the second statement follows by the uniform boundedness of A_{f_0} over $f_0 \in \Lambda^\beta(F, \Omega)$ in Lemma 3 and $\mathbb{P}(\Omega_n) \to 1$ in Lemma 6. Proof is complete. □

4.5. Approximation error to the pilot estimator

The following Lemma 8 is the main technical result of this paper, based on which Theorem 1 will be proved.

Lemma 8. *Suppose Assumptions A1 and A2 hold with some integer $k \geq \max(\beta, 2)$ and diverging sequence M. Let m be diverging with respect to sample size n and $F > 0$ be a fixed constant. If $M^d \log(M) = o(n)$, then the network function $\widehat{f}_{net} \in \mathcal{F}(L, \mathbf{p}(T))$, with $L = (2m+3)(k+d)+1$ and $T = 2^{k+4}d(M+2k)^d$ satisfies*

$$\sup_{f_0 \in \Lambda^\beta(F,\Omega)} \mathbb{E}_{f_0} \left\{ \sup_{\mathbf{x}\in\Omega} |\widehat{f}_{net}(\mathbf{x}) - \widehat{f}_{pilot}(\mathbf{x})|^2 \Big| \mathbb{X} \right\} = O_P(M^{2d} 4^{-2m}).$$

Here the O_P is in the sense of diverging m, M, n.

Proof of Lemma 8. By the notation in the proof of Lemma 7 and (4.11), we have

$$\begin{aligned}
\widehat{\mathbf{f}}_{\text{pilot}} = \Phi(\Phi^T\Phi)^{-1}\Phi^T\mathbf{Y} &= \Phi(\Phi^T\Phi)^{-1}\Phi^T(\Phi C + \mathbf{E} + \boldsymbol{\epsilon}) \\
&= \Phi C + \Phi(\Phi^T\Phi)^{-1}\Phi^T\mathbf{E} + \Phi(\Phi^T\Phi)^{-1}\Phi^T\boldsymbol{\epsilon} \\
&= \mathbf{f}_0 - (I - \Phi(\Phi^T\Phi)^{-1}\Phi^T)\mathbf{E} + \Phi(\Phi^T\Phi)^{-1}\Phi^T\boldsymbol{\epsilon}.
\end{aligned}$$

It follows from (4.2), (4.3) and (4.6) that $\widehat{f}_{\text{pilot}}(\mathbf{x}) = \widehat{C}^T\mathbf{D}_k(\mathbf{x})$ and $\widehat{f}_{\text{net}}(\mathbf{x}) = \widehat{C}^T\widetilde{\mathbf{D}}_k(\mathbf{x})$. Therefore, for any $\mathbf{x} \in \Omega$, we have

$$\begin{aligned}
|\widehat{f}_{\text{pilot}}(\mathbf{x}) - \widehat{f}_{\text{net}}(\mathbf{x})|^2 &= \left\| \widehat{C}^T \left(\mathbf{D}_k(\mathbf{x}) - \widetilde{\mathbf{D}}_k(\mathbf{x}) \right) \right\|_2^2 \\
&= \widehat{C}^T\widehat{C} \left(\mathbf{D}_k(\mathbf{x}) - \widetilde{\mathbf{D}}_k(\mathbf{x}) \right)^T \left(\mathbf{D}_k(\mathbf{x}) - \widetilde{\mathbf{D}}_k(\mathbf{x}) \right) \\
&\leq q\widehat{C}^T\widehat{C} \sup_{\mathbf{x}\in[0,1]^d} \left\| \mathbf{D}_k(\mathbf{x}) - \widetilde{\mathbf{D}}_k(\mathbf{x}) \right\|_\infty^2 \leq q\widehat{C}^T\widehat{C}[4(d-1) + 8^k]^2 4^{-2m},
\end{aligned}$$

where the last inequality follows from Lemma 2. Following Lemma 7 and the fact $q = |\Gamma| = (M+k-1)^d \asymp h^{-d}$, we have

$$\begin{aligned}
\sup_{f_0 \in \Lambda^\beta(F,\Omega)} \mathbb{E}_{f_0} \left(\sup_{\mathbf{x}\in\Omega} |\widehat{f}_{\text{pilot}}(\mathbf{x}) - \widehat{f}_{\text{net}}(\mathbf{x})|^2 \Big| \mathbb{X} \right) &\leq q[4(k-1) + 8k]^2 4^{-2m} \sup_{f_0 \in \Lambda(F,\Omega)} \mathbb{E}\left(\widehat{C}^T\widehat{C} | \mathbb{X} \right) \\
&= O_P(h^{-2d} 4^{-2m}), \qquad (4.13)
\end{aligned}$$

which completes the proof by noticing that $M \asymp h^{-1}$. □

To the end of this section, let us complete the proof of Theorem 1. Combining Lemmas 7 and 8, we have

$$\begin{aligned}
&\inf_{\widehat{f}\in\mathcal{F}(L,\mathbf{p}(T))} \sup_{f_0\in\Lambda^\beta(F,\Omega)} \mathbb{E}_{f_0} \left(\|\widehat{f} - f_0\|_{L^2}^2 | \mathbb{X} \right) \\
\leq &\sup_{f_0\in\Lambda^\beta(F,\Omega)} \mathbb{E}_{f_0} \left(\|\widehat{f}_{\text{net}} - f_0\|_{L^2}^2 | \mathbb{X} \right) \\
\leq &\, 2 \sup_{f_0\in\Lambda^\beta(F,\Omega)} \mathbb{E}_{f_0} \left(\|\widehat{f}_{\text{net}} - \widehat{f}_{\text{pilot}}\|_{L^2}^2 | \mathbb{X} \right) + 2 \sup_{f_0\in\Lambda^\beta(F,\Omega)} \mathbb{E}_{f_0} \left(\|\widehat{f}_{\text{pilot}} - f_0\|_{L^2}^2 | \mathbb{X} \right) \\
= &\, O_P\left(M^{-2\beta} + \frac{M^d}{n} \right) + O_P(M^{2d} 4^{-2m})
\end{aligned}$$

$$= O_P\left(T^{-\frac{2\beta}{d}} + \frac{T}{n} + T^2 4^{-\frac{L}{k+d}}\right)$$

where the fact that $M \asymp h^{-1}$, $L = (2m+3)(k+d)+1$ and $T = 2^{k+4}d(M+2k)^d$ is used. We would like to comment that Theorem 1 does not rely on Assumption A2, as we only need such $\widehat{f}_{\text{net}}$ exists.

5. Asymptotic distribution and optimal testing

In this subsection, we derive the asymptotic distribution for $\widehat{f}_{\text{net}}$ and a corresponding hypothesis testing procedure. Let us recall that the network function constructed in (4.9) satisfies

$$\widehat{f}_{\text{net}} \in \mathcal{F}(L, \mathbf{p}(T)), \quad \text{with } L = (2m+3)(k+d)+1 \text{ and } T = 2^{k+4}d(M+2k)^d,$$

where k is the order of tensor product B-spline basis, d is the dimension of explanatory variable $\mathbf{X}$, $M = h^{-1}$ is the inverse of knots separation distance, m is an integer characterizing the number of hidden layers of the network. All the results in this subsection are discussed when m, M, n diverge while assuming k, d are fixed constant.

Theorem 2 below establishes a pointwise asymptotic distribution for $\widehat{f}_{\text{net}}$.

Theorem 2. *Under the Assumptions A1 and A2, if* $k \geq \max(\beta, 2)$, $n^{\frac{1}{2\beta+d}} = o(M)$, $M^d \log(M) = o(n)$ *and* $nM^d = o(16^m)$, *then for any fixed point* $\mathbf{x} \in \Omega$, *we have*

$$\frac{\widehat{f}_{net}(\mathbf{x}) - f_0(\mathbf{x})}{\sqrt{\mathbf{D}_k^T(\mathbf{x})(\Phi^T\Phi)^{-1}\mathbf{D}_k(\mathbf{x})}} \xrightarrow{D} N(0,1),$$

where $\Phi = (\mathbf{D}_k(\mathbf{X}_1), \mathbf{D}_k(\mathbf{X}_2), \ldots, \mathbf{D}_k(\mathbf{X}_n))^T \in \mathbb{R}^{n\times q}$ *with* $q = (M+k-1)^d$.

Proof of Theorem 2. By (4.9) and Assumption A2, we know $M = h^{-1}$ and

$$\widehat{f}_{\text{net}} \in \mathcal{F}(L, \mathbf{p}(T)), \quad \text{with } L = (2m+3)(k+d)+1 \text{ and } T = 2^{k+4}d(M+2k)^d.$$

So the rate conditions are equivalent to $hn^{\frac{1}{2\beta+d}} = o(1)$, $\log(h^{-1}) = o(nh^d)$ and $n^{1/2}h^{-d/2} = o(4^m)$.

For fixed $\mathbf{x} \in \Omega$, let $V(\mathbf{x}) = \mathbf{D}_k^T(\mathbf{x})(\Phi^T\Phi)^{-1}\mathbf{D}_k(\mathbf{x})$. By Huang [13, Theorems 3.1 and 5.2], it follows that

$$\frac{\widehat{f}_{\text{pilot}}(\mathbf{x}) - f_0(\mathbf{x})}{\sqrt{V(\mathbf{x})}} \xrightarrow{D} N(0,1). \tag{5.1}$$

It is well known that the tensor product B-spline basis satisfies $\sum_{s=1}^q D_{\mathbf{i}_s,k}(\mathbf{x}) = 1$ for all $\mathbf{x} \in \Omega$ (e.g., see Section 15 in Györfi et al. [10]). Given a point $\mathbf{x} \in \Omega$, let us denote $\Gamma_{\mathbf{x}} = \{\mathbf{i} \in \Gamma | D_{\mathbf{i},k}(\mathbf{x}) > 0\}$. By the construction of $D_{\mathbf{i},k}$, there are only k^d basis functions among $D_{\mathbf{i}_1,k}(\mathbf{x}), \ldots, D_{\mathbf{i}_q,k}(\mathbf{x})$ with positive values, while the rest are all zero. Hence, it follows that $|\Gamma_{\mathbf{x}}| = k^d$. The above fact implies that $\sum_{\mathbf{i}\in\Gamma_{\mathbf{x}}} D_{\mathbf{i},k}(\mathbf{x}) = 1$ and $\mathbf{D}_k^T(\mathbf{x})\mathbf{D}_k(\mathbf{x}) = \sum_{\mathbf{i}\in\Gamma_{\mathbf{x}}} D_{\mathbf{i},k}^2(\mathbf{x}) \geq |\Gamma_{\mathbf{x}}|^{-1} = k^{-d}$, where the equality holds when $D_{\mathbf{i},k}(\mathbf{x}) = |\Gamma_{\mathbf{x}}|^{-1}$ for all $\mathbf{i} \in \Gamma_{\mathbf{x}}$.

Lemma 6 implies that with probability approaching 1, we have

$$\begin{aligned} V(\mathbf{x}) &= \mathbf{D}_k^T(\mathbf{x})(\Phi^T\Phi)^{-1}\mathbf{D}_k(\mathbf{x}) \\ &\geq \lambda_{\min}((\Phi^T\Phi)^{-1})\mathbf{D}_k(\mathbf{x})^T\mathbf{D}_k(\mathbf{x}) \\ &= \frac{1}{\lambda_{\max}(\Phi^T\Phi)}\mathbf{D}_k(\mathbf{x})^T\mathbf{D}_k(\mathbf{x}) \end{aligned}$$

$$\geq \frac{1}{2a_2 nh^d}\mathbf{D}_k(\mathbf{x})^T\mathbf{D}_k(\mathbf{x}) \geq \frac{1}{2a_2 k^d nh^d},$$

where a_2 is the constant (4.10). By Lemma 8 we get that $|\widehat{f}_{\text{pilot}}(\mathbf{x}) - \widehat{f}_{\text{net}}(\mathbf{x})|^2 = O_P(h^{-2d}4^{-2m})$. Therefore,

$$\frac{\widehat{f}_{\text{pilot}}(\mathbf{x}) - \widehat{f}_{\text{net}}(\mathbf{x})}{\sqrt{V(\mathbf{x})}} = O_P(n^{1/2}h^{-d/2}4^{-m}) = o_P(1). \tag{5.2}$$

Theorem 2 follows by (5.1) and (5.2). This completes the proof. □

In practice, it is often of interest to test whether Y_i and $\mathbf{X}_i$ are statistically independent, equivalently, to test f_0 is constant. In what follows, we consider an elementary hypothesis testing problem: $H_0 : f_0 = 0$ vs. $H_1 : f \neq 0$. In general, one can subtract the constant from f_0, or if the constant is unknown, subtract $\bar{Y}$ from f_0, and test the difference equals zero. Consider a test statistic $T_n = \|\widehat{f}_{\text{net}}\|_n^2$, where $\|f\|_n^2 = \sum_{i=1}^n f(\mathbf{x}_i)^2/n$ is the empirical norm. It should be mentioned that T_n relies on m, M since $\widehat{f}_{\text{net}}$ does. The following Theorem 3 is a byproduct of Lemma 8, which derives null distribution of T_n and analyzes its power under a sequence of local alternatives.

Theorem 3. *Under the Assumptions A1 and A2, if* $k \geq \max(\beta, 2)$, $n^2M^d = O(16^m)$ *and* $M \asymp n^{\frac{2}{4\beta+d}}$, *then the following hold:*

(i) *Under* $H_0 : f_0 = 0$, *it follows that*

$$\frac{nT_n - q}{\sqrt{2q}} \xrightarrow{D} N(0,1), \tag{5.3}$$

where $q = (M + k - 1)^d$.

(ii) *For any* $\delta > 0$, *there exists a* $C_\delta > 0$ *such that, under* $H_1 : f = f_0$ *with* $\|f_0\|_n \geq C_\delta n^{-\frac{2\beta}{4\beta+d}}$, *it holds that*

$$\mathbb{P}\left(\left|\frac{nT_n - q}{\sqrt{2q}}\right| > z_{\alpha/2}\right) \geq 1 - \delta, \tag{5.4}$$

where $z_{\alpha/2}$ *is the* $1 - \alpha/2$ *upper percentile of standard normal variable.*

Part (5.3) of Theorem 3 suggests a testing rule at significance α: reject H_0 if and only if

$$\left|\frac{nT_n - q}{\sqrt{2q}}\right| \geq z_{\alpha/2}.$$

Part (5.4) of Theorem 3 says that the power of T_n is at least $1 - \delta$ provided that the null and alternative hypotheses are separated by $C_\delta n^{-\frac{2\beta}{4\beta+d}}$ in terms of $\|\cdot\|_n$-norm. The separation rate is optimal in the sense of Ingster [14].

Proof of Theorem 3. The proof consists of two steps. The first step is to establish the asymptotic distribution of the test statistic based $\widehat{f}_{\text{pilot}}$, while the second step is to show that the test statistic T_n has the same limiting distribution. By (4.9) and Assumption A2, we know $M = h^{-1}$*and*

$$\widehat{f}_{\text{net}} \in \mathcal{F}(L, \mathbf{p}(T)), \quad \text{with } L = (2m+3)(k+d)+1 \text{ and } T = 2^{k+4}d(M+2k)^d.$$

So the rate conditions are equivalent to $nh^{-d/2}4^{-m} = o(1)$ and $h \asymp n^{-\frac{2}{4\beta+d}}$.

Step 1: Using the notation in the proof of Lemma 8 and by (4.11), we have

$$\widehat{\mathbf{f}}_{\text{pilot}} = \Phi(\Phi^T\Phi)^{-1}\Phi^T\mathbf{f}_0 + \Phi(\Phi^T\Phi)^{-1}\Phi^T\boldsymbol{\epsilon}.$$

Under $H_0 : f_0 = 0$, it follows that $\widehat{\mathbf{f}}_{\text{pilot}}^T\widehat{\mathbf{f}}_{\text{pilot}} = \boldsymbol{\epsilon}^T\Phi(\Phi^T\Phi)^{-1}\Phi^T\boldsymbol{\epsilon}$ and

$$\widehat{\mathbf{f}}_{\text{pilot}}^T\widehat{\mathbf{f}}_{\text{pilot}}|\mathbb{X} \sim \chi^2(q),$$

where we used the fact that ϵ_i are i.i.d. normal and is free of $\mathbb{X}$. Since $q = (M+k-1)^d \asymp h^{-d} \to \infty$, we conclude from central limit theorem that

$$\frac{\widehat{\mathbf{f}}_{\text{pilot}}^T\widehat{\mathbf{f}}_{\text{pilot}} - q}{\sqrt{2q}} \xrightarrow{D} N(0,1). \tag{5.5}$$

Suppose that f_0 satisfies $\|f_0\|_n \geq C_\delta\gamma_n$ with $\gamma_n = n^{-\frac{2\beta}{4\beta+d}}$ for some C_δ large enough. Then it follows that

$$\widehat{\mathbf{f}}_{\text{pilot}}^T\widehat{\mathbf{f}}_{\text{pilot}} = \mathbf{f}_0^T\Phi(\Phi^T\Phi)^{-1}\Phi^T\mathbf{f}_0 + 2\mathbf{f}_0^T\Phi(\Phi^T\Phi)^{-1}\Phi^T\boldsymbol{\epsilon} + \boldsymbol{\epsilon}^T\Phi(\Phi^T\Phi)^{-1}\Phi^T\boldsymbol{\epsilon} \equiv S_1 + 2S_2 + S_3.$$

By simple algebra, we show that

$$\begin{aligned}\mathbf{f}_0^T(I - \Phi(\Phi^T\Phi)^{-1}\Phi^T)\mathbf{f}_0 &= (\Phi C + \mathbf{E})^T(I - \Phi(\Phi^T\Phi)^{-1}\Phi^T)(\Phi C + \mathbf{E}) \\ &= \mathbf{E}^T(I - \Phi(\Phi^T\Phi)^{-1}\Phi^T)\mathbf{E} \\ &\leq \mathbf{E}^T\mathbf{E} \leq A_{f_0}^2 nh^{2\beta}.\end{aligned}$$

As a consequence it follows that

$$S_1 = \mathbf{f}_0^T\mathbf{f}_0 - \mathbf{f}_0^T(I - \Phi(\Phi^T\Phi)^{-1}\Phi^T)\mathbf{f}_0 \geq C_\delta^2 n\gamma_n^2 - A_{f_0}^2 nh^{2\beta} = C_\delta^2 n^{\frac{d}{4\beta+d}} - A_{f_0}^2 nh^{2\beta}.$$

Since $h = M^{-1} \asymp n^{-\frac{2}{4\beta+d}}$, it follows that $nh^{2\beta} \asymp n^{\frac{d}{4\beta+d}}$. If we choose $C_\delta > 0$ large enough, it implies that $S_1 = \frac{1}{2}C_\delta^2 n^{\frac{d}{4\beta+d}}$, which leads to

$$\frac{S_1}{\sqrt{2q}} \geq \frac{1}{2\sqrt{2q}}C_\delta^2 n^{\frac{d}{4\beta+d}} \asymp \frac{1}{2\sqrt{2}}C_\delta^2 n^{\frac{d}{4\beta+d}} h^{\frac{d}{2}} \asymp \frac{1}{2\sqrt{2}}C_\delta^2 \quad \text{and} \quad \sqrt{\frac{S_1}{2q}} \to 0.$$

Here the condition $q = (M+k-1)^d \asymp h^{-d}$ is used. So $\sqrt{\frac{S_1}{2q}} \leq \frac{1}{4C_\delta}\frac{S_1}{\sqrt{2q}}$ for n large enough. Taking conditional expectation, we have

$$\mathbb{P}\left(|S_2|^2 > C_\delta^2 S_1|\mathbb{X}\right) = P(|Z| > C_\delta) \leq \delta,$$

where Z is standard normal random variable and the last inequality holds with large C_δ. Therefore, we have that

$$\begin{aligned}&\mathbb{P}\left(\left|\frac{\widehat{\mathbf{f}}_{\text{pilot}}^T\widehat{\mathbf{f}}_{\text{pilot}} - q}{\sqrt{2q}}\right| \leq Z_{\alpha/2}\right) \\ &= \mathbb{P}\left(\left|\frac{S_3 - q}{\sqrt{2q}} + \frac{S_1}{\sqrt{2q}} + \frac{2S_2}{2q}\right| \leq Z_{\alpha/2}\right) \\ &\leq \mathbb{P}\left(\left|\frac{S_3 - q}{\sqrt{2q}} + \frac{S_1}{\sqrt{2q}} + \frac{2S_2}{\sqrt{2q}}\right| \leq Z_{\alpha/2}, |S_2| \leq C_\delta\sqrt{S_1}\right) + P\left(|S_2| > C_\delta\sqrt{S_1}\right).\end{aligned} \tag{5.6}$$

By the choice of C_δ, the second term in (5.6) is bounded by δ, while the first term yields following inequality:

$$\begin{aligned}
&\mathbb{P}\left(\left|\frac{S_3-q}{\sqrt{2q}}+\frac{S_1}{\sqrt{2q}}+\frac{2S_2}{\sqrt{2q}}\right|\le Z_{\alpha/2},|S_2|\le C_\delta\sqrt{S_1}\right)\\
&=\mathbb{P}\left(-Z_{\alpha/2}-\frac{S_1}{\sqrt{2q}}-\frac{2S_2}{\sqrt{2q}}\le\frac{S_3-q}{\sqrt{2q}}\le Z_{\alpha/2}-\frac{S_1}{\sqrt{2q}}-\frac{2S_2}{\sqrt{2q}},|S_2|\le C_\delta\sqrt{S_1}\right)\\
&\le\mathbb{P}\left(-Z_{\alpha/2}-\frac{S_1}{\sqrt{2q}}-\frac{2C_\delta\sqrt{S_1}}{\sqrt{2q}}\le\frac{S_3-q}{\sqrt{2q}}\le Z_{\alpha/2}-\frac{S_1}{\sqrt{2q}}+\frac{2C_\delta\sqrt{S_1}}{\sqrt{2q}},|S_2|\le C_\delta\sqrt{S_1}\right)\\
&\le\mathbb{P}\left(-Z_{\alpha/2}-\frac{3S_1}{2\sqrt{2q}}\le\frac{S_3-q}{\sqrt{2q}}\le Z_{\alpha/2}-\frac{S_1}{2\sqrt{2q}}\right)\\
&\le\mathbb{P}\left(\frac{S_3-q}{\sqrt{2q}}\le Z_{\alpha/2}-\frac{C_\delta^2}{2\sqrt{2}}\right).
\end{aligned}$$

Combining above and taking limit on both sides, it follows that

$$\lim_{n\to\infty}\mathbb{P}\left(\left|\frac{\widehat{\mathbf{f}}_{\text{pilot}}^T\widehat{\mathbf{f}}_{\text{pilot}}-q}{\sqrt{2q}}\right|\le Z_{\alpha/2}\right)\le\mathbb{P}\left(Z\le Z_{\alpha/2}-\frac{C_\delta^2}{2\sqrt{2}}\right)\le\delta. \tag{5.7}$$

Step 2: Observe that

$$\frac{n\|\widehat{f}_{\text{net}}\|_n^2-q}{\sqrt{2q}}=\frac{n\|\widehat{f}_{\text{pilot}}\|_n^2-q}{\sqrt{2q}}+\frac{n\|\widehat{f}_{\text{net}}\|_n^2-n\|\widehat{f}_{\text{pilot}}\|_n^2}{\sqrt{2q}}. \tag{5.8}$$

By Lemma 7 and Lemma 8, both $\|\widehat{f}_{\text{net}}-\widehat{f}_{\text{pilot}}\|_n$ and $\|\widehat{f}_{\text{pilot}}-f_0\|_n$ are $O_P(1)$, and we have

$$\begin{aligned}
|\|\widehat{f}_{\text{net}}\|_n^2-\|\widehat{f}_{\text{pilot}}\|_n^2|&=|\|\widehat{f}_{\text{net}}\|_n-\|\widehat{f}_{\text{pilot}}\|_n|\times\left(\|\widehat{f}_{\text{net}}\|_n+\|\widehat{f}_{\text{pilot}}\|_n\right)\\
&\le\|\widehat{f}_{\text{net}}-\widehat{f}_{\text{pilot}}\|_n\times\left(\|\widehat{f}_{\text{net}}-\widehat{f}_{\text{pilot}}\|_n+2\|\widehat{f}_{\text{pilot}}\|_n\right)\\
&\le\|\widehat{f}_{\text{net}}-\widehat{f}_{\text{pilot}}\|_n\times\left(\|\widehat{f}_{\text{net}}-\widehat{f}_{\text{pilot}}\|_n+2\|\widehat{f}_{\text{pilot}}-f_0\|_n+2\|f_0\|_n\right)\\
&=\|\widehat{f}_{\text{net}}-\widehat{f}_{\text{pilot}}\|_n\times O_P(1)\\
&=O_P(h^{-d}4^{-m}).
\end{aligned}$$

Therefore, the second term in (5.8) is of order $O_P(nh^{-d}4^{-m}q^{-1/2})=O_P(nh^{-d/2}4^{-m})=o_P(1)$, where we have used the fact $q=(M+k-1)^d\asymp h^{-d}$. The result then follows by (5.5) and (5.7). This completes the proof. □

6. Network approximation to additive model

The optimal rate in Theorem 1 suffers from the 'curse' of dimensionality. In this section, we show that this issue can be addressed when f_0 has an additive structure. Specifically, let us consider the following function space:

$$\Lambda_+^{\boldsymbol{\beta}}(F,\Omega)=\left\{f:\Omega\to\mathbb{R}|\ f(\mathbf{x})=a+\sum_{j=1}^d g_j(x_j)\text{ with }g_j\in\Lambda^{\beta_j}(F,[0,1])\text{ and }\int_0^1 g_j(x)dx=0\right\},$$

where $F>0$ is the radius, and $\boldsymbol{\beta}=(\beta_1,\ldots,\beta_d)\in(0,\infty)^d$ are the degrees of smoothness for g_j's. Clearly, any $f\in\Lambda_+^{\boldsymbol{\beta}}(F,\Omega)$ has an expression $f(\mathbf{x})=a+\sum_{j=1}^d g_j(x_j)$ with the jth additive component belonging

to the ball of univariate β_j-Hölder functions with radius F. Moreover, the constraint $\int_0^1 g_{j,0}(x)dx = 0$ is to avoid identifiability issue.

Theorem 4. *Let Assumption A1 be satisfied. Suppose that $L \to \infty$, $T \to \infty$ and $T \log T = o(n)$ as $n \to \infty$, then for any fixed constant $F > 0$ and vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d) \in (0, \infty)^d$, it follows that*

$$\inf_{\widehat{f} \in \mathcal{F}(L, \mathbf{p}(T))} \sup_{f_0 \in \Lambda_+^{\boldsymbol{\beta}}(F, \Omega)} \mathbb{E}_{f_0} \left(\|\widehat{f} - f_0\|_{L^2}^2 \middle| \mathbb{X} \right) = O_P \left(\frac{1}{T^{2\beta^*}} + \frac{T}{n} + \frac{T^2}{2^{\frac{L}{1+k}}} \right),$$

where $\beta^* = \min_{1 \le j \le d} \beta_j$, k is the smallest integer satisfying $k \ge \max(\beta_1, \ldots, \beta_d, 2)$, and the O_P is in the sense that T, L, n are diverging. Hence, if $T \asymp n^{\frac{1}{2\beta^*+1}}$ and $n^{\frac{2\beta^*+2}{2\beta^*+1}} = O(2^{\frac{L}{1+k}})$, then

$$\inf_{\widehat{f} \in \mathcal{F}(L, \mathbf{p}(T))} \sup_{f_0 \in \Lambda_+^{\boldsymbol{\beta}}(F, \Omega)} \mathbb{E}_{f_0} \left(\|\widehat{f} - f_0\|_{L^2}^2 \middle| \mathbb{X} \right) = O_P \left(n^{-\frac{2\beta^*}{2\beta^*+1}} \right).$$

The rate $n^{-\frac{2\beta^*}{2\beta^*+1}}$ in Theorem 4 is optimal in nonparametric additive estimation. When $\beta_1 = \cdots = \beta_d = \beta$, the rate simply becomes $n^{-\frac{2\beta}{2\beta+1}}$ whose optimality has been proved by Stone [25]. Otherwise, the optimal rate relies on the least order of smoothness of the d univariate functions.

The rest part of this section is devoted to proving Theorem 4. Throughout we keep in mind that the true regression function f_0 admits an additive expression

$$f_0(\mathbf{x}) = f_0(x_1, \ldots, x_d) = \alpha_0 + g_{1,0}(x_1) + \ldots + g_{d,0}(x_d),$$

where α_0 is an unknown constant. Before proving the theorem, let us settle down some notation. For $j = 1, 2, \ldots, d$, given integers $M_j, k_j \ge 2$ and knots $t_{-k_j+1,j} < t_{-k_j+2,j} < \ldots < t_{0,j} < t_{1,j} < \ldots < t_{M_j,j} < t_{M_j+1,j} < \ldots < t_{M_j+k_j+1,j}$ with $t_{0,j} = 0, t_{M_j,0} = 1$, let $\mathbf{B}_{k_j,j}(x) \in \mathbb{R}^{M_j+k_j-1}$ denote the vector of univariate B-spline basis functions (with respect to variable x_j). Since the collection of these univariate B-spline basis does not form a basis on the additive function space due to the sum-to-one condition, we instead use the following polynomial spline basis to approximate the additive components $g_{j,0}$'s:

$$\mathbf{P}_{k_j,j}(x) = \left(x, x^2, \ldots, x^{k_j-1}, (x - t_{1,j})_+^{k_j-1}, \ldots, (x - t_{M_j-1,j})_+^{k_j-1} \right)^T \in \mathbb{R}^{M_j+k_j-2}, j = 1, \ldots, d.$$

The central idea is the approximation $f_0(x_1, \ldots, x_d) \approx a + \sum_{j=1}^d W_j^T \mathbf{P}_{k_j,j}(x_j)$ for some constants $a \in \mathbb{R}$ and $W_j \in \mathbb{R}^{M_j+k_j-2}$. By least square estimation, an estimator of f_0 is

$$\widehat{f}_{\text{pilot}}(x_1, \ldots, x_d) = \widehat{a} + \sum_{j=1}^d \widehat{f}_j(x_j) \text{ with } \widehat{f}_j(x) = \widehat{W}_j^T \mathbf{P}_{k_j,j}(x).$$

If we define the centralized estimator $\widehat{g}_j(x) = \widehat{f}_j(x) - \int_0^1 \widehat{f}_j(u)du$, then it turns out to be a consistent estimator of $g_{j,0}$; see Lemma 12, and we have

$$\widehat{f}_{\text{pilot}}(x_1, \ldots, x_d) = \widehat{\alpha} + \sum_{j=1}^d \widehat{g}_j(x_j) \text{ with } \widehat{\alpha} = \widehat{a} + \sum_{j=1}^d \int_0^1 \widehat{f}_j(u)du. \tag{6.1}$$

Note that $\mathbf{B}_{k_j,j}$ is the B-spline basis. So $\widehat{g}_j$ can be written as $\widehat{C}_j^T \mathbf{B}_{k_j,j}(x)$ for some $\widehat{C}_j \in \mathbb{R}^{M_j+k_j-1}$, we define a neural network estimator $\widetilde{g}_j(x) = \widehat{C}_j^T \widetilde{\mathbf{B}}_{k_j,j}(x)$ for $j = 1, \ldots, d$ and

$$\widehat{f}_{\text{net}}(\mathbf{x}) = \widehat{\alpha} + \sum_{j=1}^{d} \widetilde{g}_j(x_j). \tag{6.2}$$

By similar argument as (4.9), for any integers $k_j, M_j, m \geq 2$, we can construct the network satisfying

$$\widehat{f}_{\text{net}} \in \mathcal{F}(L, \mathbf{p}(T)) \quad \text{with } L = (2m+3) \max_{1\leq j\leq d}(k_j+1) + 1 \text{ and } T = \sum_{j=1}^{d} 2^{k_j+4}(M_j + 2k_j). \tag{6.3}$$

Moreover, the following notation plays a similar role as that in the proof of Theorem 1:

$$\begin{aligned}
q_+ &= 1 + \sum_{j=1}^{d}(M_j + k_j - 2),\\
\mathbf{P}(\mathbf{x}) &= (1, \mathbf{P}_{k_1,1}^T(x_1), \mathbf{P}_{k_2,2}^T(x_2), \ldots, \mathbf{P}_{k_d,d}^T(x_d))^T \in \mathbb{R}^{q_+},\\
\Phi_+ &= (\mathbf{P}(\mathbf{X}_1), \mathbf{P}(\mathbf{X}_2), \ldots, \mathbf{P}(\mathbf{X}_n))^T \in \mathbb{R}^{n\times q_+},\\
\Theta_n^+ &= \{f(\mathbf{x}) | f(\mathbf{x}) = a + \sum_{j=1}^{d} g_j(x_j) \text{ with } a \in \mathbb{R}, g_j(x) = b_j^T \mathbf{P}_{k_j,j}(x),\\
&\qquad \int_0^1 g_j(x)dx = 0 \text{ for some } b_j \in \mathbb{R}^{M_j+k_j-2} \text{ and } j = 1, \ldots, d\},\\
\Omega_n^+ &= \{\|g\|_{L^2}^2/2 \leq \|g\|_n^2 \leq 2\|g\|_{L^2}^2, \text{ for all } g \in \Theta_n^+\}.
\end{aligned} \tag{6.4}$$

To handle the additive model, we introduce a new norm of a function g as $\|g\|^2 = \int_\Omega g^2(\mathbf{x})dx$. We would like to comment that another norm used in previous sections is $\|g\|_{L^2}^2 = \int_\Omega g^2(\mathbf{x})Q(\mathbf{x})dx$, which are equivalent to $\|\cdot\|$ under Assumption A1. Finally, we will need the following assumption during the proof, which is in the similar spirit of Assumption A2.

Assumption A3. For $j = 1, \ldots, d$, the order of B-spline satisfies $k_j \geq \beta_j$, and the knots $\{t_{i,j}, i = -k_j + 1, \ldots, M_j + k_j + 1\}$ are equally separated by constant $h_j = M_j^{-1}$. In the analysis, we need $M_j \to \infty$ and $h_j \to 0$ for all $j = 1 \ldots, d$.

Proposition 4. *Suppose that g_0 is a constant function and g_1 is a measurable function satisfying $\int_\Omega g_1(\mathbf{x})d\mathbf{x} = 0$. Moreover, $\|g_1\|_{sup} \leq K\|g_1\|$ for some constant $K > 0$. Then $\|g_0 + g_1\|_{sup} \leq (K+2)\|g_0 + g_1\|$.*

Proof of Proposition 4. Observe that for any constant function g_0, we have $\|g_1\| = \|g_1 + g_0\| = \|g_1\|^2 + g_0^2$. Moreover, Assumption A1 leads to that, for some $c > 1$ and all g with $\|g\|_{L^2} < \infty$, it holds that $c^{-1}\|g\|^2 \leq \|g\|_{L^2}^2 \leq c\|g\|^2$. Therefore, we have

$$\begin{aligned}
\|g_0 + g_1\|_{\text{sup}} &\leq \|g_0\|_{\text{sup}} + \|g_1\|_{\text{sup}}\\
&\leq \|g_0\| + K\|g_1\|\\
&\leq \|g_0 + g_1\| + \|g_1\| + K\|g_1 + g_0\|\\
&\leq \|g_0 + g_1\| + \|g_1 + g_0\| + K\|g_1 + g_0\|\\
&\leq (K+2)\|g_0 + g_1\|.
\end{aligned}$$

Proof is complete. □

Lemma 9. *Suppose Assumptions A1 and A3 hold with integers $k_j \geq \max(\beta_j, 2)$. Moreover, if the sequences in Assumption A3 satisfies $nh_j^2 \to \infty$ and $h_j \to 0$ for each $j = 1, 2, \ldots, d$, then the following holds*

$$\sup_{g \in \Theta_n^+} \left| \frac{\|g\|_n^2}{\|g\|_{L^2}^2} - 1 \right| = o_P(1),$$

where Θ_n^+ is the function space defined in (4.10). As a consequence, it follows that $\mathbb{P}(\Omega_n^+) \to 1$. Here Ω_n^+ is the event defined in (4.10).

Proof of Lemma 9. Let $g(\mathbf{x}) = \sum_{j=1}^d g_j(x_j)$, where g_j satisfies $\int_0^1 g_j(x)dx = 0$ for $j = 1, \ldots, d$. By DeVore and Lorentz [5, Theorem 5.1.2] we get that $\|g_j\|_{\sup} \leq A_j \|g_j\|$ with $A_j \asymp h_j^{-1/2}$. Direct examination shows that

$$\|g\|_{\sup} \leq \sum_{j=1}^d \|g_j\|_{\sup} \leq \sum_{j=1}^d A_j \|g_j\| \leq \left(\sum_{j=1}^d A_j^2 \right)^{1/2} \left(\sum_{j=1}^d \|g_j\|^2 \right)^{1/2} \leq \left(\sum_{j=1}^d A_j^2 \right)^{1/2} \left(c_d \|g\|^2 \right)^{1/2},$$

where the last inequality follows from Lemma 3.6 of Stone [26] and c_d is a constant depending on d only. Applying Proposition 4 and by Assumption A1, we obtain that

$$\|f\|_{\sup} \leq \left(\left(c_d \sum_{j=1}^d A_j^2 \right)^{1/2} + 2 \right) \|f\|_2 \leq c \left(\left(c_d \sum_{j=1}^d A_j^2 \right)^{1/2} + 2 \right) \|f\|_{L^2}^2, \text{ for all } f \in \Theta_n^+.$$

The dimension of Θ_n^+, $q_+ \leq \sum_{j=1}^d (M_j + k_j - 1) + 1 \asymp \sum_{j=1}^d h_j^{-1}$. Therefore, by Lemma 2.3 in Huang [13] and rate conditions given, we prove the result. □

Lemma 10. *Suppose Assumptions A1 and A3 hold with integers $k_j \geq \max(\beta_j, 2)$. Moreover, if the sequences in Assumption A3 satisfies $nh_j^2 \to \infty$ and $h_j \to 0$ for each $j = 1, 2, \ldots, d$, then on event Ω_n^+, $\Phi_+^T \Phi_+$ is invertible.*

Proof of Lemma 10. Let $\widehat{\mathbf{B}} = n^{-1} \Phi_+^T \Phi_+$ and $\mathbf{B} = \int \mathbf{P}(\mathbf{x}) \mathbf{P}(\mathbf{x})^T Q(\mathbf{x}) d\mathbf{x}$. For $g(\mathbf{x}) = u^T \mathbf{P}(\mathbf{x})$, we have $u^T \widehat{\mathbf{B}} u = \|g\|_n^2$ and $u^T \mathbf{B} u = \|g\|^2$. On event Ω_n^+, since $\mathbf{B}$ is positive definite, $\widehat{\mathbf{B}}$ is also positive definite. Proof is complete. □

Lemma 11. *Suppose Assumptions A1 and A3 hold with integers $k_j \geq \max(\beta_j, 2)$. Moreover, if the sequences in Assumption A3 satisfies $nh_j^2 \to \infty$ and $h_j \to 0$ for each $j = 1, 2, \ldots, d$, then the following holds uniformly for all $f_0 \in \Lambda_+^{\boldsymbol{\beta}}(F, \Omega)$ on event Ω_n^+:*

$$\mathbb{E}_{f_0} \left(\|\widehat{f}_{pilot} - f_0\|_{L^2}^2 \middle| \mathbb{X} \right) \leq 2^{d+3} \sum_{j=1}^d A_{g_{j,0}}^2 h_j^{2\beta_j} + \frac{8q_+}{n}.$$

Proof of Lemma 11. For any $f_0 \in \Lambda_+^{\boldsymbol{\beta}}(F, \Omega)$, let $\mathbf{f}_0 = (f_0(\mathbf{X}_1), \ldots, f_0(\mathbf{X}_n))^T$, $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T$, and $\widehat{\mathbf{f}}_{\text{pilot}} = (\widehat{f}_{\text{pilot}}(\mathbf{X}_1), \ldots, \widehat{f}_{\text{pilot}}(\mathbf{X}_n))^T$. According to Lemma 3 and the condition $k_j \geq \max(\beta_j, 2)$ for $j = 1, \ldots, d$, there exists a vector $W \in \mathbb{R}^{q_+}$ such that $\sup_{\mathbf{x} \in \Omega} |W^T \mathbf{P}(\mathbf{x}) - f_0(\mathbf{x})| \leq \sum_{j=1}^d A_{g_{j,0}} h_j^{\beta_j}$, where the constant $A_{g_{j,0}}$ satisfies $\sup_{g_{j,0} \in \Lambda^{\beta_j}(F,[0,1])} A_{g_{j,0}} < \infty$. For simplicity, we further define $f^*(\mathbf{x}) = W^T \mathbf{P}(\mathbf{x})$ and $\mathbf{f}^* = (f^*(\mathbf{X}_1), \ldots, f^*(\mathbf{X}_n))^\top$.

By Lemma 10 and similar argument in (4.11), it follows on event Ω_n^+ that

$$\begin{aligned}\widehat{\mathbf{f}}_{\text{pilot}} &= \Phi_+(\Phi_+^T\Phi_+)^{-1}\Phi_+^T\mathbf{Y}\\ &= \Phi_+(\Phi_+^T\Phi_+)^{-1}\Phi_+^T(\Phi_+W+\mathbf{E}+\boldsymbol{\epsilon})\\ &= \Phi_+W+\Phi_+(\Phi_+^T\Phi_+)^{-1}\Phi_+^T\mathbf{E}+\Phi_+(\Phi_+^T\Phi_+)^{-1}\Phi_+^T\boldsymbol{\epsilon}\\ &= \mathbf{f}^*+\Phi_+(\Phi_+^T\Phi_+)^{-1}\Phi_+^T\mathbf{E}+\Phi_+(\Phi_+^T\Phi_+)^{-1}\Phi_+^T\boldsymbol{\epsilon},\end{aligned}$$

where $\mathbf{E}=\mathbf{f}_0-\mathbf{f}^*$. As a consequence, we have

$$\begin{aligned}\|\widehat{f}_{\text{pilot}}-f^*\|_n^2 &= \frac{1}{n}(\widehat{\mathbf{f}}_{\text{pilot}}-\mathbf{f}^*)^T(\widehat{\mathbf{f}}_{\text{pilot}}-\mathbf{f}^*)\\ &\le \frac{2}{n}\mathbf{E}^T\Phi_+(\Phi_+^T\Phi_+)^{-1}\Phi_+^T\mathbf{E}+\frac{2}{n}\boldsymbol{\epsilon}^T\Phi_+(\Phi_+^T\Phi_+)^{-1}\Phi_+^T\boldsymbol{\epsilon}\\ &\le \frac{2}{n}\mathbf{E}^T\mathbf{E}+\frac{2}{n}\boldsymbol{\epsilon}^T\Phi_+(\Phi_+^T\Phi_+)^{-1}\Phi_+^T\boldsymbol{\epsilon}\\ &\le 2(\sum_{j=1}^d A_{g_{j,0}}h_j^{\beta_j})^2+\frac{2}{n}\boldsymbol{\epsilon}^T\Phi_+(\Phi_+^T\Phi_+)^{-1}\Phi_+^T\boldsymbol{\epsilon}\\ &\le 2^d\sum_{j=1}^d A_{g_{j,0}}^2h_j^{2\beta_j}+\frac{2}{n}\boldsymbol{\epsilon}^T\Phi_+(\Phi_+^T\Phi_+)^{-1}\Phi_+^T\boldsymbol{\epsilon},\end{aligned}$$

where we use the fact that $\Phi_+^T\Phi_+$ is invertible on Ω_n^+ by Lemma 10. By independence of $\boldsymbol{\epsilon}$ and Φ_+, it follows that on event Ω_n^+,

$$\mathbb{E}_{f_0}\left(\boldsymbol{\epsilon}^T\Phi_+(\Phi_+^T\Phi_+)^{-1}\Phi_+^T\boldsymbol{\epsilon}\middle|\mathbb{X}\right)=\text{Tr}\left(\Phi_+(\Phi_+^T\Phi_+)^{-1}\Phi_+^T\right)=q_+.$$

Combining the above two inequalities and using the definition of Ω_N^+, we show that

$$\mathbb{E}_{f_0}\left(\|\widehat{f}_{\text{pilot}}-f^*\|_{L^2}^2\middle|\mathbb{X}\right)\le 2\mathbb{E}_{f_0}\left(\|\widehat{f}_{\text{pilot}}-f^*\|_n^2\middle|\mathbb{X}\right)\le 2^{d+1}\sum_{j=1}^d A_{g_{j,0}}^2h_j^{2\beta_j}+\frac{4q_+}{n},$$

which further implies that

$$\begin{aligned}\mathbb{E}_{f_0}\left(\|\widehat{f}_{\text{pilot}}-f_0\|_{L^2}^2\middle|\mathbb{X}\right) &\le 2\mathbb{E}_{f_0}\left(\|\widehat{f}_{\text{pilot}}-f^*\|_{L^2}^2\middle|\mathbb{X}\right)+2\mathbb{E}_{f_0}\left(\|f_0-f^*\|_{L^2}^2\middle|\mathbb{X}\right)\\ &\le 2^{d+2}\sum_{j=1}^d A_{g_{j,0}}^2h_j^{2\beta_j}+\frac{8q_+}{n}+2^{d+1}\sum_{j=1}^d A_{g_{j,0}}^2h_j^{2\beta_j}\\ &\le 2^{d+3}\sum_{j=1}^d A_{g_{j,0}}^2h_j^{2\beta_j}+\frac{8q_+}{n}.\end{aligned}$$

Proof is complete. □

Proposition 5. *Under Assumption A1, if $g(\mathbf{x})=a+\sum_{j=1}^d g_j(x_j)$ with $\int_0^1 g_j(x)dx=0$, then it follows that $\|g\|_{L^2}^2\ge a_3^d(a^2+\sum_{j=1}^d\|g_j\|_{L^2}^2)$, where the constant $a_3>0$ only relies on the density Q.*

Proof of Proposition 5. This is a direct consequence of Lemma 3.1 in Stone [26] and Assumption A1. □

Lemma 12. *Suppose Assumptions A1 and A3 hold with integers $k_j \geq \max(\beta_j, 2)$. Moreover, if the sequences in Assumption A3 satisfies $nh_j^2 \to \infty$ and $h_j \to 0$ for each $j = 1, 2, \ldots, d$, then the following statement hold uniformly for all $f_0 \in \Lambda_+^{\beta}(F, \Omega)$ on event Ω_n^+:*

$$\mathbb{E}_{f_0}\left(\|\widehat{g}_j - g_{j,0}\|_{L^2}^2 | \mathbb{X}\right) \leq a_4 \sum_{s=1}^{d} A_{g_{s,0}}^2 h_s^{2\beta_s} + \frac{a_4 q_+}{n}, \quad \text{for } j = 1, 2, \ldots, d,$$

and

$$\mathbb{E}_{f_0}(|\widehat{\alpha} - \alpha_0|^2 | \mathbb{X}) \leq a_4 \sum_{s=1}^{d} A_{g_{s,0}}^2 h_s^{2\beta_s} + \frac{a_4 q_+}{n},$$

where $\widehat{\alpha}$ is the estimated coefficient defined in (6.1), and $a_4 > 0$ is an absolute constant relying on the density function Q and d.

Proof of Lemma 12. Recall $\widehat{f}_{\text{pilot}}(\mathbf{x}) = \widehat{a} + \sum_{j=1}^{d} \widehat{f}_j(x_j) = \widehat{\alpha} + \sum_{j=1}^{d} \widehat{g}_j(x_j)$, where $\widehat{\alpha} = \widehat{a} + \sum_{j=1}^{d} \int_0^1 \widehat{f}_j(u)du$ and $\widehat{g}_j(x) = \widehat{f}_j(x) - \int_0^1 \widehat{f}_j(u)du$. By Assumption A1 there exists a constant $c > 1$ such that for any g, $c^{-1} \int_\Omega g(\mathbf{x})d\mathbf{x} \leq \int_\Omega g(\mathbf{x})Q(\mathbf{x})d\mathbf{x} \leq c \int_\Omega g(\mathbf{x})d\mathbf{x}$. By Proposition 5 we have

$$\begin{aligned}
\|\widehat{f}_{\text{pilot}} - f_0\|_{L^2}^2 &= \|\widehat{\alpha} - \alpha_0 + \sum_{j=1}^{d}(\widehat{g}_j - g_{j,0})\|_{L^2}^2 \\
&\geq c^{-1}\|\widehat{\alpha} - \alpha_0 + \sum_{j=1}^{d}(\widehat{g}_j - g_{j,0})\|_{L^2}^2 \\
&\geq c^{-1} a_3^d \left(|\widehat{\alpha} - \alpha_0|^2 + \sum_{j=1}^{d} \|\widehat{g}_j - g_{j,0}\|_{L^2}^2 \right) \\
&\geq c^{-2} a_3^d \left(|\widehat{\alpha} - \alpha_0|^2 + \sum_{j=1}^{d} \|\widehat{g}_j - g_{j,0}\|_{L^2}^2 \right),
\end{aligned}$$

where a_3 is the constant in Proposition 5. By Lemma 11 and the above inequality, on event Ω_n^+, the following holds for any $f_0 \in \Lambda_+^{\boldsymbol{\beta}}(F, \Omega)$:

$$\mathbb{E}_{f_0}\left(\|\widehat{g}_j - g_{j,0}\|_{L^2}^2 | \mathbb{X}\right) \leq c^2 a_3^{-d} \mathbb{E}_{f_0}\left(\|\widehat{f}_{\text{pilot}} - f_0\|_{L^2}^2 | \mathbb{X}\right) \leq c^2 a_3^{-d} 2^{d+3} \sum_{s=1}^{d} A_{g_{s,0}}^2 h_s^{2\beta_s} + \frac{8c^2 a_3^{-d} q_+}{n},$$

for $j = 1, 2, \ldots, d$, and

$$\mathbb{E}(|\widehat{\alpha} - \alpha_0|^2 | \mathbb{X}) \leq c^2 a_3^{-d} 2^{d+3} \sum_{s=1}^{d} A_{g_{s,0}}^2 h_s^{2\beta_s} + \frac{8c^2 a_3^{-d} q_+}{n}.$$

Therefore, the desired results follow with $a_4 = c^2 a_3^{-d} 2^{d+3}$. Proof is complete. □

Given previous Lemmas, we are ready to prove Theorem 4. By Lemma 3, it holds that

$$\sup_{x \in [0,1]} |C_j^T \mathbf{B}_{k_j,j}(x) - g_{j,0}(x)| \leq A_{g_{j,0}} h_j^{\beta_j}$$

for some $C_j \in \mathbb{R}^{M_j+k_j-1}$ with $\|C_j\|_\infty \le A_{g_{j,0}}$. Let $g_j^* = C_j^T \mathbf{B}_{k_j,j}$ for $j = 1, \dots, d$. Recall that $\widehat{g}_j$ can be written as $\widehat{C}_j^T \mathbf{B}_{k_j,j}(x)$ for some $\widehat{C}_j \in \mathbb{R}^{M_j+k_j-1}$ and the neural network approximating the additive component is $\widetilde{g}_j(x) = \widehat{C}_j^T \widetilde{\mathbf{B}}_{k_j,j}(x)$ according to (6.2).

By Lemma 12, for any $f_0 \in \Lambda_+^{\boldsymbol{\beta}}(F, \Omega)$ we have

$$\begin{aligned}
\mathbb{E}_{f_0}\left(\|\widehat{g}_j - g_{j,0}\|_{L^2}^2 | \mathbb{X}\right) &\le a_4 \sum_{s=1}^{d} A_{g_{s,0}}^2 h_s^{2\beta_s} + \frac{a_4 q_+}{n}, \text{ for } j = 1, 2, \dots, d, \\
\mathbb{E}_{f_0}\left(|\widehat{\alpha} - \alpha_0|^2 | \mathbb{X}\right) &\le a_4 \sum_{s=1}^{d} A_{g_{s,0}}^2 h_s^{2\beta_s} + \frac{a_4 q_+}{n},
\end{aligned} \tag{6.5}$$

where a_4 is the constant in Lemma 12. By Lemma 5, for every $j = 1, \dots, d$ we have

$$\begin{aligned}
a_1 h_j (\widehat{C}_j - C_j)^T (\widehat{C}_j - C_j) &\le \int |\widehat{C}_j^T \mathbf{B}_{k_j,j}(x_j) - C_j^T \mathbf{B}_{k_j,j}(x_j)|^2 Q(\mathbf{x}) d\mathbf{x} \\
&= \|\widehat{g}_j - g_j^*\|^2 \\
&\le 2\|\widehat{g}_j - g_{j,0}\|_{L^2}^2 + 2\|g_j^* - g_{j,0}\|_{L^2}^2,
\end{aligned}$$

which further implies that the following holds on Ω_n^+:

$$\begin{aligned}
\mathbb{E}_{f_0}\left(\widehat{C}_j^T \widehat{C}_j | \mathbb{X}\right) &\le 2C_j^T C_j + 2\mathbb{E}_{f_0}\left((\widehat{C}_j - C_j)^T (\widehat{C}_j - C_j) | \mathbb{X}\right) \\
&\le 2q_+ A_{g_{j,0}}^2 + 4a_1^{-1} h_j^{-1} \mathbb{E}_{f_0}\left(\|\widehat{g}_j - g_{j,0}\|_{L^2}^2 | \mathbb{X}\right) + 4a_1^{-1} h_j^{-1} \mathbb{E}_{f_0}\left(\|g_j^* - g_{j,0}\|_{L^2}^2 | \mathbb{X}\right) \\
&\le 2q_+ A_{g_{j,0}}^2 + 4a_1^{-1} h_j^{-1} \left(a_4 \sum_{s=1}^{d} A_{g_{s,0}}^2 h_s^{2\beta_s} + \frac{a_4 q_+}{n}\right) + 4a_1^{-1} h_j^{-1} A_{g_{j,0}}^2 h_j^{2\beta_j} \\
&\le (2A_{g_{j,0}}^2 + 4a_1^{-1} a_4)\left(q_+ + \frac{q_+}{n h_j}\right) + 4a_1^{-1}(a_4 + 1) h_j^{-1} \sum_{s=1}^{d} A_{g_{s,0}}^2 h_s^{2\beta_s} \\
&\le a_6 \left(\sum_{v=1}^{d} A_{g_{v,0}}^2 + a_1^{-1} a_4\right)\left(q_+ + h_j^{-1} \sum_{s=1}^{d} h_s^{2\beta_s}\right),
\end{aligned}$$

with $a_6 = 8 + 8a_1^{-1}(a_4 + 1)$. In the last inequality we have used $nh_j \to \infty$. Recall $\widetilde{g}_j = \widehat{C}_j^T \widetilde{\mathbf{B}}_{k_j,j}(x)$. Therefore, Lemma 1 implies that the following holds on event Ω_n^+: $\frac{8^k}{14} 4^{-m}$

$$\begin{aligned}
\mathbb{E}_{f_0}(\|\widetilde{g}_j - \widehat{g}_j\|_{L^2}^2 | \mathbb{X}) &= \mathbb{E}_{f_0}\left(\|\widehat{C}_j^T \widetilde{\mathbf{B}}_{k_j,j} - \widehat{C}_j^T \mathbf{B}_{k_j,j}\|_{L^2}^2 | \mathbb{X}\right) \\
&\le (M_j + k_j - 1) \mathbb{E}_{f_0}\left(\widehat{C}_j^T \widehat{C_j} | \mathbb{X}\right) \sup_{x \in [0,1]} \|\widetilde{\mathbf{B}}_{k_j,j}(x) - \mathbf{B}_{k_j,j}(x)\|_\infty^2 \\
&\le a_6 \left(\sum_{v=1}^{d} A_{g_{v,0}}^2 + a_1^{-1} a_4\right) 64^{k_j+1} (M_j + k_j - 1) \left(q_+ + h_j^{-1} \sum_{s=1}^{d} h_s^{2\beta_s}\right) 16^{-m}.
\end{aligned}$$

By the above inequality and (6.5), on event Ω_n^+, we have

$$\begin{aligned}
\mathbb{E}_{f_0}(\|\widetilde{g}_j - g_{j,0}\|_{L^2}^2 | \mathbb{X}) &\le 2\mathbb{E}_{f_0}(\|\widetilde{g}_j - \widehat{g}_j\|_{L^2}^2 | \mathbb{X}) + 2\mathbb{E}_{f_0}(\|\widehat{g}_j - g_{j,0}\|_{L^2}^2 | \mathbb{X}) \\
&\le 2a_6 \left(\sum_{v=1}^{d} A_{g_{v,0}}^2 + a_1^{-1} a_4\right) 64^{k_j+1} (M_j + k_j - 1) \left(q_+ + h_j^{-1} \sum_{s=1}^{d} h_s^{2\beta_s}\right) 16^{-m}
\end{aligned}$$

$$+2a_4\sum_{s=1}^{d}A_{g_{s,0}}^2 h_s^{2\beta_s}+\frac{a_4 q_+}{n}$$

As a consequence, on event Ω_n^+, it follows that

$$\begin{aligned}\mathbb{E}_{f_0}(\|\widehat{f}_{\text{net}}-f_0\|_{L^2}^2|\mathbb{X}) &\le 2^d\mathbb{E}(|\widehat{\alpha}-\alpha_0|^2|\mathbb{X})+2^d\sum_{j=1}^{d}\mathbb{E}(\|\widetilde{g}_j-g_{j,0}\|_{L^2}^2|\mathbb{X})\\ &\le 2^d a_4\left(\sum_{s=1}^{d}A_{g_{s,0}}^2 h_s^{2\beta_s}+\frac{q_+}{n}\right)\\ &\quad+2^{d+1}a_6\sum_{j=1}^{d}\left(\sum_{v=1}^{d}A_{g_{v,0}}^2+a_1^{-1}a_4\right)64^{k_j+1}(M_j+k_j-1)\left(q_++h_j^{-1}\sum_{s=1}^{d}h_s^{2\beta_s}\right)16^{-m}\\ &\quad+2^{d+1}a_4 d\sum_{s=1}^{d}A_{g_{s,0}}^2 h_s^{2\beta_s}+\frac{a_4 d q_+}{n}.\end{aligned}$$

Since $q_+=\sum_{j=1}^{d}(M_j+k_j-1)\asymp\sum_{j=1}^{d}M_j$, $h_j\asymp M_j^{-1}$ and $\sup_{g_{j,0}\in\Lambda^{\beta_j}(F,[0,1])}A_{g_{j,0}}<\infty$ by Lemma 3, taking supremum of the above inequality leads to

$$\sup_{f_0\in\Lambda_+^{\boldsymbol{\beta}}(F,\Omega)}\mathbb{E}_{f_0}(\|\widehat{f}_{\text{net}}-f_0\|_{L^2}^2|\mathbb{X})=O_P\left(\sum_{j=1}^{d}M_j^{-2\beta_j}\right)+O_P\left(\sum_{j=1}^{d}\frac{M_j}{n}\right)+O_P\left(\sum_{j=1}^{d}M_j^2 4^{-2m}\right).$$

Using (6.3), we know $L=(2m+3)\max_{1\le j\le d}(k_j+1)+1$ and $T=\sum_{j=1}^{d}2^{k_j+4}(M_j+2k_j)$. The above inequality further leads to

$$\begin{aligned}\inf_{\widehat{f}\in\mathcal{F}(L,\mathbf{p}(T))}\sup_{f_0\in\Lambda_+^{\boldsymbol{\beta}}(F,\Omega)}\mathbb{E}_{f_0}(\|\widehat{f}_{\text{net}}-f_0\|_{L^2}^2|\mathbb{X}) &\le \sup_{f_0\in\Lambda_+^{\boldsymbol{\beta}}(F,\Omega)}\mathbb{E}_{f_0}(\|\widehat{f}_{\text{net}}-f_0\|_{L^2}^2|\mathbb{X})\\ &=O_P\left(T^{-2\beta_*}+\frac{T}{n}+T^2 4^{-2m}\right).\end{aligned}$$

We can always choose $k_j=\lfloor\beta\rfloor_j+1$ for $j=1,\dots,d$. Therefore the integer $k\ge\max(\beta_1,\dots,\beta_d,2)$ implies $k\ge\max(k_1,\dots,k_d,2)$ and $L=(2m+3)\max_{1\le j\le d}(k_j+1)+1\le 2m(k+1)+3(k+1)+1$. Substituting m with L, we complete the proof.

Acknowledgment

The authors would like to thank the Editor and an anonymous reviewer for their constructive suggestions that have led to a significant improvement in the manuscript. Zuofeng Shang acknowledges supports by NSF DMS-1764280 and DMS-1821157.

Appendix A

A.1. Proof of Lemma 3

In this subsection, we provide the proof of Lemma 3. For simplicity, we consider the case with $d=2$. The extension to the scenario with $d>2$ can be done similarly.

Given integers $k, M\ge 2$ and knots $t_{-k+1}<t_{-k+2}<\ldots<t_0<t_1<\ldots<t_M<t_{M+1}<\ldots<t_{M+k-1}$ with $t_0=0, t_M=1$. Since $d=2$, we can relabel the tensor product B-spline basis as $B_{i,k}(x_1)B_{j,k}(x_2)$, for

$(x_1, x_2)^T \in \Omega$ and $i, j = -k+1, \ldots, M-1$. We would like to comment that the basis is denoted as $D_{\mathbf{i},k}$ in previous section. As a consequence, the function space spanned by $B_{i,k}(x_1)B_{j,k}(x_2)$ is defined as

$$\Theta_n = \{f(\mathbf{x}) = \sum_{i=-k+1}^{M-1} \sum_{j=-k+1}^{M-1} c_{ij} B_{i,k}(x_1) B_{j,k}(x_2) | c_{ij} \in \mathbb{R} \text{ and } \mathbf{x} = (x_1, x_2)^T \in \Omega\}.$$

Let us borrow some definition from the Section 15.1 in Györfi et al. [10]. Let $\mathcal{C}$ be the collection of continuous function supported on Ω. A linear operator $\kappa : \mathcal{C} \to \Theta_n$ is called a quasi interpolant if

$$\kappa f(\mathbf{x}) = \sum_{i=-k+1}^{M-1} \sum_{j=-k+1}^{M-1} \kappa_{ij}(f) B_{i,k}(x_1) B_{j,k}(x_2),$$

where $\kappa_{ij}(f)$ is a constant depending only on the values of f in $[t_i, t_{i+k}) \times [t_j, t_{j+k})$. Moreover, κ is said to have order k if $\kappa f = f$ for all polynomial f with the degrees of x_1 and x_2 not greater than $k-1$.

Lemma A.1 *(Theorem 15.2 of Györfi et al. [10]). Given integers $k, M \geq 2$ and knots $t_{-k+1} < t_{-k+2} < \ldots < t_0 < t_1 < \ldots < t_M < t_{M+1} < \ldots < t_{M+k-1}$ with $t_0 = 0, t_M = 1$. There exists a quasi interpolant $\kappa : \mathcal{C} \to \Theta_n$ with order k such that*

$$|\kappa_{ij}(f)| \leq L_k \sup_{\mathbf{x} \in [t_i, t_{i+k}) \times [t_j, t_{j+k})} |f(\mathbf{x})|.$$

Here L_k is a constant depending only on k but not on the knots.

We are ready to prove Lemma 3. Suppose $f \in \Lambda^\beta(F, \Omega)$. For fixed $\mathbf{u} \in [t_i, t_{i+k}) \times [t_j, t_{j+k})$, let us define the following local Taylor polynomial:

$$p_{\mathbf{u}}(\mathbf{x}) = \sum_{|\boldsymbol{\alpha}| \leq \lfloor \beta \rfloor} \partial^{\boldsymbol{\alpha}} f(\mathbf{u}) \frac{(\mathbf{x} - \mathbf{u})^{\boldsymbol{\alpha}}}{\boldsymbol{\alpha}!} \quad \text{for } \mathbf{x} \in [t_i, t_{i+k}) \times [t_j, t_{j+k}).$$

By Taylor's theorem, it follows that

$$f(\mathbf{x}) = \sum_{|\boldsymbol{\alpha}| < \lfloor \beta \rfloor} \partial^{\boldsymbol{\alpha}} f(\mathbf{u}) \frac{(\mathbf{x} - \mathbf{u})^{\boldsymbol{\alpha}}}{\boldsymbol{\alpha}!} + \sum_{|\boldsymbol{\alpha}| = \lfloor \beta \rfloor} \frac{\lfloor \beta \rfloor}{\boldsymbol{\alpha}!} (\mathbf{x} - \mathbf{u})^{\boldsymbol{\alpha}} \int_0^1 (1-t)^{\lfloor \beta \rfloor - 1} \partial^{\boldsymbol{\alpha}} f(\mathbf{u} + t(\mathbf{x} - \mathbf{u})) dt.$$

Suppose $\mathbf{u} = (u_1, u_2)^T, \mathbf{x} = (x_1, x_2)^T \in [t_i, t_{i+k}) \times [t_j, t_{j+k})$, then Assumption A2 implies that $\|\mathbf{u} - \mathbf{x}\| \leq \sqrt{2(kh)^2} \leq 2kh$. Let us consider two cases of β.

Case 1: If $\lfloor \beta \rfloor = 0$, then $p_{\mathbf{u}}(\mathbf{x}) = f(\mathbf{u})$. By the definition of $\Lambda^\beta(F, \Omega)$, it follows that

$$|f(\mathbf{x}) - p_{\mathbf{u}}(\mathbf{x})| = |f(\mathbf{x}) - f(\mathbf{u})| \leq F \|\mathbf{x} - \mathbf{u}\|^\beta \leq F(2k)^\beta h^\beta.$$

Case 2: If $\lfloor \beta \rfloor \geq 1$, then $\int_0^1 (1-t)^{\lfloor \beta \rfloor - 1} dt = 1/\lfloor \beta \rfloor$. Therefore, we have

$$\begin{aligned} |f(\mathbf{x}) - p_{\mathbf{u}}(\mathbf{x})| &\leq \sum_{|\boldsymbol{\alpha}| = \lfloor \beta \rfloor} \left| \frac{\lfloor \beta \rfloor}{\boldsymbol{\alpha}!} (\mathbf{x} - \mathbf{u})^{\boldsymbol{\alpha}} \right| \int_0^1 (1-t)^{\lfloor \beta \rfloor - 1} \left| \partial^{\boldsymbol{\alpha}} f(\mathbf{u} + t(\mathbf{x} - \mathbf{u})) - \partial^{\boldsymbol{\alpha}} f(\mathbf{u}) \right| dt \\ &\leq \sum_{|\boldsymbol{\alpha}| = \lfloor \beta \rfloor} \lfloor \beta \rfloor \, |x_1 - u_1|^{\alpha_1} |x_2 - u_2|^{\alpha_2} \int_0^1 (1-t)^{\lfloor \beta \rfloor - 1} F \|t(\mathbf{x} - \mathbf{u})\|^{\beta - \lfloor \beta \rfloor} dt \end{aligned}$$

$$\begin{aligned}
&\leq F\lfloor\beta\rfloor \sum_{|\boldsymbol{\alpha}|=\lfloor\beta\rfloor} (kh)^{\alpha_1}(kh)^{\alpha_2}\|\mathbf{x}-\mathbf{u}\|^{\beta-\lfloor\beta\rfloor}\int_0^1 (1-t)^{\lfloor\beta\rfloor-1}dt\\
&\leq F\sum_{|\boldsymbol{\alpha}|=\lfloor\beta\rfloor} k^{\lfloor\beta\rfloor}h^{\lfloor\beta\rfloor}(2kh)^{\beta-\lfloor\beta\rfloor}\\
&\leq F(\lfloor\beta\rfloor+1)^2(2k)^{\beta}h^{\beta}.
\end{aligned}$$

Combining the above two cases, we show that

$$|f(\mathbf{x})-p_{\mathbf{u}}(\mathbf{x})|\leq F(\lfloor\beta\rfloor+1)^2(2k)^{\beta}h^{\beta},$$

for all $\mathbf{x},\mathbf{u}\in[t_i,t_{i+k})\times[t_j,t_{j+k})$ and $f\in\Lambda^{\beta}(F,\Omega)$. Since the operator κ is linear, and $p_u(\mathbf{x})$ is a polynomial with degrees of x_1 and x_2 not greater than $\lfloor\beta\rfloor$. Since κ is an interpolant with order k by Lemma A.1, and $p_{\mathbf{u}}$ is a polynomial with degree at most $\lfloor\beta\rfloor$, the condition $k\geq\beta$ implies that $k-1\geq\lfloor\beta\rfloor$. As a consequence, it follows that

$$\begin{aligned}
|\kappa[f(\mathbf{x})-p_{\mathbf{u}}(\mathbf{x})]| &= |\sum_{i=-k+1}^{M-1}\sum_{j=-k+1}^{M-1}\kappa_{ij}(f-p_{\mathbf{u}})B_{i,k}(x_1)B_{j,k}(x_2)|\\
&\leq \sum_{i=-k+1}^{M-1}\sum_{j=-k+1}^{M-1}|\kappa_{ij}(f-p_{\mathbf{u}})|B_{i,k}(x_1)B_{j,k}(x_2)\\
&\leq \sup_{-k+1\leq i\leq M-1}\sup_{-k+1\leq j\leq M-1}|\kappa_{ij}(f-p_{\mathbf{u}})|\\
&\leq \sup_{-k+1\leq i\leq M-1}\sup_{-k+1\leq j\leq M-1}L_k\sup_{\mathbf{v}\in[t_i,t_{i+k})\times[t_j,t_{j+k})}|f(\mathbf{v})-p_{\mathbf{u}}(\mathbf{v})|\\
&\leq L_kF(\lfloor\beta\rfloor+1)^2(2k)^{\beta}h^{\beta}\quad\text{for all }\mathbf{x}\in\Omega.
\end{aligned}$$

Combining the above inequality, we conclude that

$$\begin{aligned}
|\kappa f(\mathbf{x})-f(\mathbf{x})| &\leq |\kappa f(\mathbf{x})-p_{\mathbf{u}}(\mathbf{x})|+|p_{\mathbf{u}}(\mathbf{x})-f(\mathbf{x})|\\
&= |\kappa f(\mathbf{x})-\kappa p_{\mathbf{u}}(\mathbf{x})|+|p_{\mathbf{u}}(\mathbf{x})-f(\mathbf{x})|\\
&= |\kappa[f(\mathbf{x})-p_{\mathbf{u}}(\mathbf{x})]|+|p_{\mathbf{u}}(\mathbf{x})-f(\mathbf{x})|\\
&\leq (L_k+1)F(\lfloor\beta\rfloor+1)^2(2k)^{\beta}h^{\beta}\quad\text{for all }\mathbf{x}\in\Omega.
\end{aligned}$$

Notice that $\kappa f(\mathbf{x})=\sum_{i=-k+1}^{M-1}\sum_{j=-k+1}^{M-1}\kappa_{ij}(f)B_{i,k}(x_1)B_{j,k}(x_2)=\sum_{\mathbf{i}\in\Gamma}c_{\mathbf{i}}D_{\mathbf{i},k}(\mathbf{x})$, where $c_{\mathbf{i}}$'s is the sequence $\kappa_{ij}(f)$'s after relabeling. Using Lemma A.1 again, we show that $|c_{\mathbf{i}}|\leq L_k\|f(\mathbf{x})\|_{\sup}$. Clearly, we can choose

$$A_f=(L_k+1)F(\lfloor\beta\rfloor+1)^2(2k)^{\beta}+L_k\|f(\mathbf{x})\|_{\sup},$$

which satisfies $\sup_{f\in\Lambda^{\beta}(F,\Omega)}A_f\leq(L_k+1)F(\lfloor\beta\rfloor+1)^2(2k)^{\beta}+L_kF<\infty$. The proof is complete.

A.2. Index of symbols

- k: the smallest integer satisfying $k\geq\max(\beta,2)$ for Theorems 1–3 and $k\geq\max(\beta_1,\ldots,\beta_d,2)$ for Theorem 4.
- m: a diverging auxiliary variable for the number of hidden layers, which is related to the construction of network product operator; see Lemma 2, (4.9) and (6.3).

- M: a diverging auxiliary variable for the number of nodes in each hidden layer, which is related to the number of knots for B-spline basis; see Section 4.1 and (4.9).
- h: $h = M^{-1}$, knots separation distance; see Assumption A2.
- q: $q = (M + k - 1)^d$, number of tensor product B-spline basis functions; see Section 4.1.
- Θ_n: a function space spanned by tensor product B-spline basis; see Section 4.4.
- Ω_n: an event with probability approaching one; see (4.10).
- a_1, a_2: universal constants relying on k and the density Q; see Lemma 5.
- a_3: a universal constant relying on the density Q; see Proposition 5.
- a_4: a universal constant relying on d and the density Q; see Lemma 12.
- k_j: a fixed constant indicating the order of B-spline basis for additive model, which requires $k_j \geq \beta_j$; see Assumption A3.
- M_j: a diverging auxiliary variable for the number of nodes in each hidden layer for additive model, which is related to the number of knots for B-spline basis; see Assumption A3 and (6.3).
- h_j: $h_j = M_j^{-1}$, knots separation distance for additive model; see Assumption A3.
- q_+: $q = 1 + \sum_{j=1}^{d}(M_j + k_j - 2)$, number of B-spline basis functions for additive model; see (6.4).
- Θ_n^+: function space spanned by B-spline basis for additive model; see (6.4).
- Ω_n^+: an event with probability approaching one; see (6.4).

References

[1] M. Bianchini, F. Scarselli, On the complexity of neural network classifiers: a comparison between shallow and deep architectures, IEEE Trans. Neural Netw. Learn. Syst. 25 (8) (2014) 1553–1565.
[2] C. de Boor, A Practical Guide to Splines, Springer Verlag, New York, 1978.
[3] O. Delalleau, Y. Bengio, Shallow vs. deep sum-product networks, in: Advances in Neural Information Processing Systems, 2011, pp. 666–674.
[4] L. Deng, J. Li, J.-T. Huang, K. Yao, D. Yu, F. Seide, M.L. Seltzer, G. Zweig, X. He, J.D. Williams, et al., Recent advances in deep learning for speech research at Microsoft, in: ICASSP, vol. 26, 2013, p. 64.
[5] R.A. DeVore, G.G. Lorentz, Constructive Approximation, vol. 303, Springer Science & Business Media, 1993.
[6] K. Eckle, J. Schmidt-Hieber, A comparison of deep networks with ReLU activation function and linear spline-type methods, Neural Netw. 110 (2019) 232–242.
[7] R. Eldan, O. Shamir, The power of depth for feedforward neural networks, in: Conference on Learning Theory, 2016, pp. 907–940.
[8] M.H. Farrell, T. Liang, S. Misra, Deep neural networks for estimation and inference, Econometrica (2021), forthcoming.
[9] Y. Gal, Z. Ghahramani, A theoretically grounded application of dropout in recurrent neural networks, in: Advances in Neural Information Processing Systems, 2016, pp. 1019–1027.
[10] L. Györfi, M. Kohler, A. Krzyzak, H. Walk, A Distribution-Free Theory of Nonparametric Regression, Springer Science & Business Media, 2006.
[11] M. Hamers, M. Kohler, Nonasymptotic bounds on the L_2 error of neural network regression estimates, Ann. Inst. Stat. Math. 58 (1) (2006) 131–151.
[12] J. Huang, Projection estimation in multiple regression with application to functional anova models, Ann. Stat. 26 (1) (1998) 242–272.
[13] J. Huang, Local asymptotics for polynomial spline regression, Ann. Stat. 31 (5) (2003) 1600–1635.
[14] Y.I. Ingster, Asymptotically minimax hypothesis testing for nonparametric alternatives. I, II, III, Math. Methods Stat. 2 (2) (1993) 85–114.
[15] M. Kohler, A. Krzyżak, Adaptive regression estimation with multilayer feedforward neural networks, J. Nonparametr. Stat. 17 (8) (2005) 891–913.
[16] M. Kohler, A. Krzyżak, Nonparametric regression based on hierarchical interaction models, IEEE Trans. Inf. Theory 63 (3) (2017) 1620–1630.
[17] M. Kohler, A. Krzyzak, S. Langer, A comparison of deep networks with ReLU activation function and linear spline-type methods, Preprint, 2019.
[18] M. Kohler, J. Mehnert, Analysis of the rate of convergence of least squares neural network regression estimates in case of measurement errors, Neural Netw. 24 (3) (2011) 273–279.
[19] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436.
[20] S. Liang, R. Srikant, Why Deep Neural Networks for Function Approximation?, 2017.
[21] R. Liu, Z. Shang, G. Cheng, On deep instrumental variables estimate, arXiv:2004.14954, 2020.
[22] G.F. Montufar, R. Pascanu, K. Cho, Y. Bengio, On the number of linear regions of deep neural networks, in: Advances in Neural Information Processing Systems, 2014, pp. 2924–2932.
[23] M. Raghu, B. Poole, J. Kleinberg, S. Ganguli, J.S. Dickstein, On the expressive power of deep neural networks, in: Proceedings of the 34th International Conference on Machine Learning, vol. 70, 2017, pp. 2847–2854, JMLR. org.

[24] J. Schmidt-Hieber, Nonparametric regression using deep neural networks with ReLU activation function, Ann. Stat. 48 (4) (2020) 1875–1897.
[25] C.J. Stone, Additive regression and other nonparametric models, Ann. Stat. 13 (2) (1985) 689–705.
[26] C.J. Stone, The use of polynomial splines and their tensor products in multivariate function estimation, Ann. Stat. 22 (1) (1994) 118–171.
[27] T. Suzuki, Adaptivity of deep ReLU network for learning in Besov and mixed smooth Besov spaces: optimal rate and curse of dimensionality, in: International Conference on Learning Representations, 2019.
[28] M. Telgarsky, Benefits of depth in neural networks, arXiv preprint, arXiv:1602.04485, 2016.
[29] J. Wan, D. Wang, S.C.H. Hoi, P. Wu, J. Zhu, Y. Zhang, J. Li, Deep learning for content-based image retrieval: a comprehensive study, in: Proceedings of the 22nd ACM International Conference on Multimedia, ACM, 2014, pp. 157–166.
[30] S. Wang, G. Cao, Z. Shang, Estimation of the mean function of functional data via deep neural networks, Stat 10 (1) (2021) e393.
[31] D. Yarotsky, Error bounds for approximations with deep ReLU networks, Neural Netw. 94 (2017) 103–114.
[32] D. Yarotsky, Optimal approximation of continuous functions by very deep ReLU networks, in: Conference on Learning Theory, 2018.