A Deep Learning Framework for Scalar-on-Function Models

Ruiqi Liu
Department of Mathematics and Statistics
Texas Tech University
Lubbock, TX 79409, USA
ruiqliu@ttu.edu

Suneel Babu Chatla
Department of Mathematical Sciences
University of Texas at El Paso
El Paso, TX 79968, USA
sbchatla@utep.edu

Abstract

We propose a deep learning framework for scalar-on-function regression with possibly multiple functional predictors observed on irregular grids. The method represents each functional covariate by finitely many basis coefficients computed from the observed trajectories and fits a deep neural network using these coefficients as features. We establish a general nonasymptotic excess risk bound for the neural network estimator that decomposes the error into statistical, truncation, discretization, and network approximation terms. We then specialize the theory to two important settings. (i) For continuous functional models, an equal-width neural network estimator achieves a nearly minimax optimal rate up to a log-logarithm factor. (ii) For functional multiple-index models, we design a bottleneck-equal-width neural network estimator that exploits the low-dimensional index-link structure and attains a nearly minimax optimal rate up to a logarithm factor. The results provide theoretical support for neural network-based scalar-on-function regression and highlight the importance of architecture design.

1 Introduction

Scalar-on-function regression relates a scalar response to one or more functional predictors, such as curves, trajectories, or spectra (Kokoszka and Reimherr, 2017). These predictors are typically observed at finitely many time points on grids that may be irregular and vary across subjects. Retaining only finitely many basis coefficients introduces truncation error; estimating these coefficients from grid observations introduces numerical integration error. Linear functional models can miss nonlinear effects and interactions (Wang et al., 2024b).

Deep neural networks (DNNs) provide efficient methods for estimating nonlinear relationships between variables (Raghu et al., 2017; Lu et al., 2017; Schmidt-Hieber, 2020). For finite-dimensional covariates, their statistical properties have been studied under frequentist (Bauer and Kohler, 2019; Schmidt-Hieber, 2020; Kohler and Langer, 2021; Kim et al., 2021; Zhang et al., 2024) and Bayesian frameworks (Polson and Ročková, 2018; Chérief-Abdellatif, 2020; Bai et al., 2020). For scalar-on-function regression, Thind et al. (2023) incorporate functional predictors into a neural network, and Yao et al. (2021) learn an adaptive basis representation. Shi et al. (2025) obtain a nearly minimax rate, up to a log-logarithm factor, using a kernel-embedding DNN with a prescribed sparsity structure. Zhou et al. (2026) study generalized functional linear models using population principal component scores. Their analysis assumes fully observed covariates and does not account for discretization or score estimation error. Beyond scalar-on-function regression, DNN methods have also been studied for function-on-scalar regression (Wu et al., 2023; Luo and Qi, 2023), function-on-function regression (Luo and Qi, 2024), and functional classification (Wang et al., 2024a).

Building on these developments, we study DNN-based scalar-on-function regression with discretely observed functional predictors. Our main contributions are summarized as follows.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

(i) We propose a deep learning framework for general scalar-on-function regression models and derive a unified upper bound for the excess risk that explicitly decomposes the overall error into four components: statistical error, truncation error, discretization error, and DNN approximation error. Our approach first represents each functional covariate using its leading Fourier coefficients, which are numerically computed from discretely observed trajectories, and then uses these coefficients as inputs to a DNN. The proposed framework is highly flexible: it accommodates fully nonlinear regression functionals with one or multiple functional covariates, allows both fixed and random sampling schemes, and permits a broad class of network architectures without imposing a specific network structure.

(ii) We specialize the general theory to two important models. First, following Meister (2016) and Shi et al. (2025), we consider a continuous functional model that assumes continuity of the regression functional. We show that an equal-width fully connected DNN estimator achieves a nearly minimax optimal convergence rate up to a log-logarithm factor. Unlike the sparse DNN estimator of Shi et al. (2025), which requires optimization over a prescribed sparsity-constrained network class, our estimator is easy to implement because only the network width and depth, and the number of Fourier coefficients need to be tuned. Second, following Chen et al. (2011), we consider the functional multiple-index model. By employing a tailored network architecture that explicitly exploits the index-link structure, we show that the proposed estimator achieves a nearly minimax-optimal convergence rate, up to a logarithm factor. Establishing this optimality is nontrivial and relies critically on both the carefully designed network architecture and sharp DNN approximation results. Importantly, the specifically designed network estimator is also straightforward to implement in practice, as only a small number of architectural parameters need to be tuned, including the bottleneck size, network depth and width, and the number of retained Fourier coefficients.

(iii) Another important contribution of our work is a systematic analysis of the discretization error arising from discretely observed functional covariates. In practice, functional predictors are typically observed only at a finite number of grid points, so the Fourier coefficients used as DNN inputs must be approximated through numerical integration. We derive explicit bounds for the resulting coefficient estimation error and quantify how this error propagates to the prediction error of the DNN estimators. Our analysis characterizes how the discretization error depends jointly on the observation frequency, the smoothness of the functional covariates, the regularity of the basis functions, and the sampling scheme of the observation grids. In particular, we allow different grid designs and show how the spacing and stochastic properties of the sampling points affect the resulting error bounds.

The rest of the paper is organized as follows. Section 2 presents the proposed estimator and a general excess risk bound. Sections 3 and 4 develop the theory for continuous functional and functional multiple-index models, respectively. Section 5 reports numerical results, and Section 6 concludes with a discussion of limitations. Technical proofs and additional simulation results are provided in Appendices A and B, respectively.

2 General Framework

Let $Y \in \mathbb{R}$ be a scalar response and $\{\boldsymbol{X}(t) = (X_1(t), \ldots, X_q(t))^\top \in \mathbb{R}^q, t \in \mathcal{T} := [0,1]\}$ a q-dimensional functional covariate. Assume that $(Y, \boldsymbol{X})$ satisfies

$$Y = F_0(\boldsymbol{X}) + \epsilon,$$

where $F_0 : \prod_{j=1}^q L_2([0,1]) \to \mathbb{R}$ is unknown and $\epsilon \in \mathbb{R}$ is noise. The i.i.d. sample $\{(Y_i, \boldsymbol{X}_i)\}_{i=1}^n$ has covariates $\boldsymbol{X}_i(t) = (X_{i1}(t), \ldots, X_{iq}(t))^\top$ observed at M_i time points $T_i = \{0 = t_{i0} < t_{i1} < \cdots < t_{i,M_i-1} < t_{i,M_i} = 1\}$. The goal is to estimate the functional $F_0(\boldsymbol{X})$ based on the observations $\{(Y_i, \boldsymbol{X}_i(t)), t \in T_i\}_{i=1}^n$.

2.1 Deep Neural Network

A fully connected rectified linear unit (ReLU) DNN with L hidden layers (depth L) is a function $f : \mathbb{R}^v \to \mathbb{R}$ of the form

$$f(\boldsymbol{z}) = A_{L+1}\, \sigma\Big(A_L\, \sigma\big(\cdots \sigma(A_1 \boldsymbol{z} + b_1) \cdots\big) + b_L\Big) + b_{L+1}, \qquad \boldsymbol{z} \in \mathbb{R}^v.$$

where $\sigma(x) = (x)_+ = \max\{x, 0\}$ acts componentwise. The L hidden layers have widths $(W_1, \ldots, W_L)$; by convention, we set $W_0 = v$ and $W_{L+1} = 1$. The weight matrices and bias vectors satisfy $A_k \in \mathbb{R}^{W_k \times W_{k-1}}$ and $b_k \in \mathbb{R}^{W_k}$ for $k = 1, \ldots, L+1$. The total number of parameters is

$$S = \sum_{k=1}^{L+1} W_k \, (W_{k-1} + 1).$$

For fixed depth L and widths $(W_1, \ldots, W_L)$, define

$$\mathcal{F}(S, L, C_{\mathcal{F}}) = \left\{ f_\theta : \mathbb{R}^v \to \mathbb{R} \;\middle|\; \theta = \{(A_k, b_k)\}_{k=1}^{L+1}, \sup_{\boldsymbol{z} \in \mathbb{R}^v} |f_\theta(\boldsymbol{z})| \le C_{\mathcal{F}} \right\},$$

the ReLU network class with parameters θ and output bound $C_{\mathcal{F}} > 0$, which can be enforced by clipping. Two architectures are considered.

Example 1 (Equal-width DNN)**.** An equal-width (EW) DNN satisfies $W_1 = \cdots = W_L = W$; see Figure 1(a). Its parameter count is

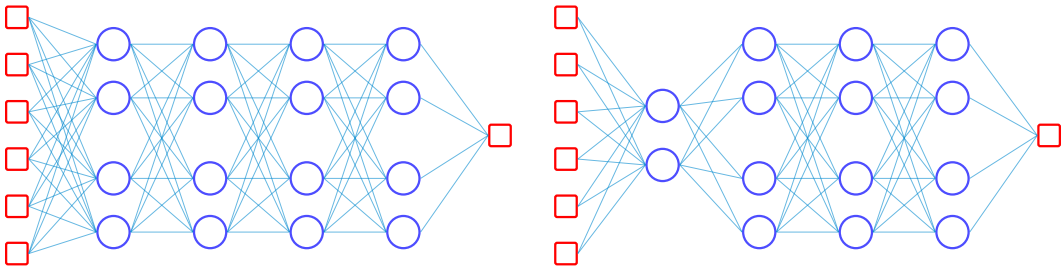
$$S = W(v+1) + (L-1)W(W+1) + (W+1),$$

depending only on v, W, and L. We use $\mathcal{F}_{\text{EW}}(W, L, C_{\mathcal{F}})$ to denote this class.

Example 2 (Bottleneck-equal-width DNN)**.** A bottleneck-equal-width (BEW) DNN uses width W after a first layer of width $W_1 \le W$; see Figure 1(b). Its parameter count is

$$S = W_1(v+1) + (L-2)W(W+1) + W(W_1+1) + (W+1).$$

The first layer compresses the v-dimensional input, as in autoencoders (Bank et al., 2023), while the remaining $L-1$ layers of width W model nonlinear compositions. We use $\mathcal{F}_{\text{BEW}}(W_1, W, L, C_{\mathcal{F}})$ to denote this class.



(a) Equal-width DNN. (b) Bottleneck-equal-width DNN.

Figure 1: Comparison of an equal-width DNN and a bottleneck-equal-width DNN.

2.2 Functional Deep Network Estimation

Let $\{\varphi_k\}_{k=1}^{\infty}$ be an orthonormal basis of $L_2(\mathcal{T})$. For fully observed $X_{ij}(t)$ with $i = 1, \ldots, n$ and $j = 1, \ldots, q$, we have the basis expansion:

$$X_{ij}(t) = \sum_{k=1}^{\infty} \xi_{ijk} \, \varphi_k(t), \text{ with } \xi_{ijk} = \langle X_{ij}, \varphi_k \rangle_{L_2(\mathcal{T})} := \int_{\mathcal{T}} X_{ij}(t) \, \varphi_k(t) \, dt.$$

For discretely observed $X_{ij}(t)$, we estimate these coefficients by the trapezoidal rule:

$$\widehat{\xi}_{ijk} = \sum_{s=0}^{M_i} u_{is} X_{ij}(t_{is}) \varphi_k(t_{is}), \text{ with } u_{is} = \begin{cases} t_{i1}/2 & \text{if } s = 0, \\ (t_{i,s+1} - t_{i,s-1})/2 & \text{if } s = 1, \ldots, M_i - 1, \\ (t_{iM_i} - t_{i,M_i-1})/2 & \text{if } s = M_i. \end{cases}$$

Retaining the first D coefficients from each component, define the exact coefficient vector and its estimate by

$$\boldsymbol{\xi}_i = \left(\xi_{i11}, \ldots, \xi_{i1D}, \ldots, \xi_{iq1}, \ldots, \xi_{iqD}\right)^\top \in \mathbb{R}^{qD},$$
$$\widehat{\boldsymbol{\xi}}_i = \left(\widehat{\xi}_{i11}, \ldots, \widehat{\xi}_{i1D}, \ldots, \widehat{\xi}_{iq1}, \ldots, \widehat{\xi}_{iqD}\right)^\top \in \mathbb{R}^{qD}.$$

Since $\boldsymbol{\xi}_i$ is not avaiable, we use $\widehat{\boldsymbol{\xi}}_i$ as the network input and estimate $F_0(\boldsymbol{X})$ by

$$\widehat{f} = \underset{f \in \mathcal{F}(S,L,C_{\mathcal{F}})}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \Big(Y_i - f(\widehat{\boldsymbol{\xi}}_i) \Big)^2 . \tag{2.1}$$

For out-of-sample evaluation, assume that $\{(\boldsymbol{X}_i^*, \widehat{\boldsymbol{\xi}}_i^*)\}_{i=1}^n$ are mutually independent and independent of the training sample, with $(\boldsymbol{X}_i^*, \widehat{\boldsymbol{\xi}}_i^*)$ distributed as $(\boldsymbol{X}_i, \widehat{\boldsymbol{\xi}}_i)$ for each $i = 1, \ldots, n$. Define the averaged excess risk of $\widehat{f}$ by

$$\mathcal{E}(\widehat{f}) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{\mathcal{D}} \left(\big| \widehat{f}(\widehat{\boldsymbol{\xi}}_i^*) - F_0(\boldsymbol{X}_i^*) \big|^2 \right), \tag{2.2}$$

where $\mathbb{E}_{\mathcal{D}}$ denotes expectation over $(\boldsymbol{X}_i^*, \widehat{\boldsymbol{\xi}}_i^*)$, conditional on the training sample used to fit $\widehat{f}$. The average in (2.2) accounts for subject-specific sampling grids, which may give $\widehat{\boldsymbol{\xi}}_i^*$ different distributions across i. When the sampling schemes are identical, the averaged risk reduces to the usual population excess risk, $\mathcal{E}(\widehat{f}) = \mathbb{E}_{\mathcal{D}}(|\widehat{f}(\widehat{\boldsymbol{\xi}}^*) - F_0(X^*)|^2)$; see Cai and Yuan (2012). We impose the following assumptions.

Assumption A1. *(i) There exists a constant $C_\epsilon > 0$ such that $\mathbb{E}(e^{\epsilon t} \mid \boldsymbol{X}) \le e^{C_\epsilon t^2/2}$ for all $t \in \mathbb{R}$ almost surely.*

(ii) The clipping parameter satisfies $C_{\mathcal{F}} \ge 2|F_0(\boldsymbol{X})|$ almost surely.

(iii) The sampling grids $T_1, \ldots, T_n$ are mutually independent, and the collection $\{T_i\}_{i=1}^n$ is independent of the entire data sample $\{(Y_i, \boldsymbol{X}_i)\}_{i=1}^n$.

Assumption A1(i) assumes conditionally sub-Gaussian noise ϵ given $\boldsymbol{X}$. This is a standard assumption in nonparametric and high-dimensional regression (e.g., see Wainwright, 2019), and it enables sharp concentration bounds for the empirical process and multiplier terms that arise in our analysis. Assumption A1(ii) requires the output bound $C_{\mathcal{F}}$ to be sufficiently large relative to the true regression values. Since $C_{\mathcal{F}} \ge 2|F_0(\boldsymbol{X})|$ almost surely, clipping leaves $F_0(\boldsymbol{X})$ unchanged Schmidt-Hieber (2020); Farrell et al. (2021). Assumption A1(iii) assumes independence between the observation grids and the raw data $\{(Y_i, \boldsymbol{X}_i)\}_{i=1}^n$. Similar conditions are also imposed in Cai and Yuan (2012).

Let $\Pi_D(\boldsymbol{X}) = \big(\Pi_D(X_1), \ldots, \Pi_D(X_q)\big)^\top$ denote componentwise D-term truncation of $\boldsymbol{X} = (X_1, \ldots, X_q)^\top$. Specifically, for $j = 1, \ldots, q$, define

$$\Pi_D(X_j) = \sum_{k=1}^{D} \xi_{jk}\, \varphi_k,$$

where $\xi_{jk} = \langle X_j, \varphi_k \rangle_{L_2(\mathcal{T})}$ are the Fourier coefficients. Let

$$\boldsymbol{\xi} = \big(\xi_{11}, \ldots, \xi_{1D}, \ldots, \xi_{q1}, \ldots, \xi_{qD}\big)^\top \in \mathbb{R}^{qD}$$

collect all truncated coefficients. Define $f_D : \mathbb{R}^{qD} \to \mathbb{R}$ by

$$f_D(\boldsymbol{\xi}) = F_0\big(\Pi_D(\boldsymbol{X})\big).$$

The function f_D represents F_0 on the qD-dimensional coefficient space. Using the above assumptions and notation, we now state the general excess risk bound.

Theorem 1. *Suppose that Assumption A1 holds. If $S \le n^3$, then with probability at least $1 - C\log(n)/n$, it holds that*

$$\mathcal{E}(\widehat{f}) \le C \left\{ \frac{SL}{n} \big(\log(n)\big)^3 + \mathcal{E}_{tr} + \mathcal{E}_{dis} + \mathcal{E}_{app} \right\}.$$

Here $C > 0$ is a constant depending on $C_{\mathcal{F}}$ and C_ϵ, and

$$\mathcal{E}_{tr} = \mathbb{E}\left(|F_0(\boldsymbol{X}) - f_D(\boldsymbol{\xi})|^2 \right), \quad \mathcal{E}_{dis} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left(\Big| f_D(\boldsymbol{\xi}_i) - f_D(\widehat{\boldsymbol{\xi}}_i) \Big|^2 \right),$$

$$\mathcal{E}_{app} = \inf_{f \in \mathcal{F}(S,L,C_{\mathcal{F}})} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left(\Big| f_D(\widehat{\boldsymbol{\xi}}_i) - f(\widehat{\boldsymbol{\xi}}_i) \Big|^2 \right).$$

Theorem 1 bounds the risk of $\widehat{f}$ by four error terms. The statistical error $SL(\log(n))^3/n$ depends on the network complexity SL. The truncation error $\mathcal{E}_{\mathrm{tr}}$ comes from replacing $\boldsymbol{X}$ by its D-term projection $\Pi_D(\boldsymbol{X})$, and the discretization error $\mathcal{E}_{\mathrm{dis}}$ comes from using $\widehat{\boldsymbol{\xi}}_i$ instead of $\boldsymbol{\xi}_i$. The term $\mathcal{E}_{\mathrm{app}}$ is the best approximation error of f_D over $\mathcal{F}(S, L, C_{\mathcal{F}})$. To obtain explicit rates, we will bound these terms under additional assumptions on F_0, $\boldsymbol{X}$, and the sampling schemes in the two models below.

3 Application I – Continuous Functional Model

We specialize Theorem 1 to continuous regression functionals. We first introduce the required notation and additional assumptions.

For a positive constant m, the m-Sobolev space is defined as

$$\mathcal{H}^m(\mathcal{T}) = \left\{ f = \sum_{k=1}^{\infty} a_k \varphi_k \in L_2(\mathcal{T}) : \sum_{k=1}^{\infty} a_k^2 k^{2m} < \infty \right\}.$$

The corresponding Sobolev norm is equivalent to the weighted coefficient norm; that is, there exist constants $0 < c_m < C_m < \infty$, depending only on m, such that $c_m \sum_{k=1}^{\infty} a_k^2 k^{2m} \le \|f\|^2_{\mathcal{H}^m(\mathcal{T})} \le C_m \sum_{k=1}^{\infty} a_k^2 k^{2m}$ for all $f \in \mathcal{H}^m(\mathcal{T})$.

Assumption B1. *There exist constants $C_\gamma > 0$ and $0 < \gamma \le 1$ such that, for any $\boldsymbol{X}(t) = (X_1(t), \ldots, X_q(t))^\top$ and $\widetilde{\boldsymbol{X}}(t) = (\widetilde{X}_1(t), \ldots, \widetilde{X}_q(t))^\top$, it holds that*

$$\left| F_0(\boldsymbol{X}) - F_0(\widetilde{\boldsymbol{X}}) \right| \le C_\gamma \left(\sum_{j=1}^{q} \|X_j - \widetilde{X}_j\|^2_{L_2(\mathcal{T})} \right)^{\gamma/2}.$$

Assumption B2. *For each $j = 1, \ldots, q$, the sample path X_j belongs to $\mathcal{H}^{m_1}(\mathcal{T})$ almost surely for some constant $m_1 > 3/2$. Moreover, it holds that $\max_{1 \le j \le q} \mathbb{E}\big(\|X_j\|^2_{\mathcal{H}^{m_1}(\mathcal{T})}\big) < \infty$.*

Assumption A2. *Let ξ_{jk} and ξ_{jl} be the two Fourier coefficients of the functional predictor X_j, and let ρ_{jkl} denote their correlation coefficient. Assume the following conditions hold.*

- *(i)* $\max_{1 \le j \le q} \sup_{k \ge 1} \sum_{l=1}^{\infty} |\rho_{jkl}| < \infty$.
- *(ii) There is a constant $C_\xi > 0$ such that $\mathbb{P}(|\xi_{jk} k^{m_1}| > t) \le C_\xi \exp(-t^2/C_\xi^2)$ for all $t > 0$ and $k \ge 1, j = 1, \ldots, q$.*

Assumption A3. $\sup_{k \ge 1} \|\varphi_k\|_\infty + \sup_{k \ge 1} k^{-1} \|\varphi_k^{(1)}\|_\infty < \infty$, *where $\|\cdot\|_\infty$ denotes the supremum norm on $[0, 1]$.*

Assumption A4. *For each subject i, recall the observation times are $0 = t_{i0} < t_{i1} < \cdots < t_{i,M_i-1} < t_{iM_i} = 1$. We assume that there exist constants $C_M > c_M > 0$ such that the sampling design satisfies one of the following conditions:*

- *(i) (Fixed Design) The grid is quasi-uniform, i.e., $c_M M_i^{-1} \le |t_{ij} - t_{i,j-1}| \le C_M M_i^{-1}$ for all i, j.*
- *(ii) (Random Design) The interior points $0 < t_{i1} < \cdots < t_{i,M_i-1} < 1$ are the order statistics of i.i.d. draws from $[0, 1]$ with density ω_i satisfying $c_M \le \omega_i(t) \le C_M$ for all $t \in [0, 1]$.*

Assumption B1 is Hölder continuity in the product L_2 norm: $\gamma = 1$ gives Lipschitz continuity, while $0 < \gamma < 1$ allows weaker regularity, as in Meister (2016); Shi et al. (2025). Assumption B2 requires finite second moments of the Sobolev norms, whereas Meister (2016) assumes these norms are uniformly bounded almost surely. The condition $m_1 > 3/2$ is used to control the discretization error. For fully observed predictors ($M_i = \infty$), this error vanishes, and the condition can be relaxed to $m_1 > 0$.

Assumption A2(i) allows dependence among the coefficients, provided their absolute correlations are uniformly summable. This condition is used to control the discretization error and can be omitted for fully observed predictors. Condition A2(ii) requires uniformly sub-Gaussian tails for the rescaled coefficients, a stronger requirement than the sub-exponential condition in Zhou et al. (2026). For

example, a possibly noncentered Gaussian process with independent Fourier coefficients satisfies Assumption A2 under Assumption B2. Assumption A3 imposes bounds on the basis functions and their first derivatives to control the numerical error in estimating the coefficients. In particular, the trigonometric basis satisfies this assumption. Assumption A4 controls the numerical-integration error $\mathcal{E}_{\text{dis}}$. Fixed spacings are of order M_i^{-1}; random sampling densities $\omega_i(\cdot)$ bounded away from zero and infinity limit large gaps in $\{t_{ij}\}_{j=0}^{M_i}$. Related sampling schemes appear in Cai and Yuan (2011).

For the conditional functional model, we fit the equal-width class from Example 1 by least squares:

$$\widehat{f}_{\text{EW}} = \underset{f \in \mathcal{F}_{\text{EW}}(W,L,C_{\mathcal{F}})}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \Big(Y_i - f(\widehat{\boldsymbol{\xi}}_i) \Big)^2. \tag{3.1}$$

Theorem 2. *Suppose that Assumptions A1-A4, B1, and B2 hold. Assume that the width $W = 3^{qD+3}(qD + N_1 + 1)$ and depth $L = 12N_2 + 15 + 2qD$ for some positive integers $N_1, N_2 \to \infty$. If $D, W, L \le n$, then with probability at least $1 - C\log(n)/n$, it holds that*

$$\mathcal{E}\left(\widehat{f}_{\text{EW}}\right) \le C\Bigg\{ \frac{9^{qD}(qD + N_1 + 1)^2(N_2 + qD)^2}{n} \big(\log(n)\big)^3 + \underbrace{D^{-2m_1\gamma}}_{\mathcal{E}_{tr}} + \underbrace{\frac{D^{3\gamma}}{n} \sum_{i=1}^{n} M_i^{-2\gamma}}_{\mathcal{E}_{dis}} + \underbrace{D(N_1 N_2)^{-\frac{4\gamma}{qD}} (\log(n))^2}_{\mathcal{E}_{app}} \Bigg\}.$$

Here $C > 0$ is a constant free of n, W, D, L, and M_i. In particular, if the functional covariates are fully observed ($M_i = \infty$), and

$$N_2 \gtrsim D \asymp \frac{\log(n)}{(\log\log(n))^2}, \quad N_1 N_2 \asymp \left(\frac{nD}{9^{qD}\log(n)} \right)^{\frac{qD}{2\gamma + 2qD}},$$

then it holds with probability at least $1 - C\log(n)/n$ that

$$\mathcal{E}\left(\widehat{f}_{\text{EW}}\right) \le C\left(\log(n)\right)^{-2m_1\gamma} \left(\log\log(n)\right)^{4m_1\gamma},$$

where $C > 0$ is a constant free of n, W, D, L, and M_i.

The four terms in Theorem 2 follow the general error decomposition. Since $S \asymp qDW + W^2L$ (Example 1), the statistical term has order $SL(\log(n))^3/n$, yielding the stated bound upon substituting the chosen width and depth. The truncation error depends on the predictor smoothness m_1 and the continuity exponent γ. The discretization error depends on both D and grid resolution: retaining more coefficients can increase this error unless the grids become denser. The approximation term depends on γ and the input dimension qD. For fixed D, increasing the width or depth reduces the approximation bound but increases statistical complexity. For fully observed predictors ($M_i = \infty$), the stated tuning yields a convergence rate of $\big(\log(n)/(\log\log(n))^2\big)^{-2m_1\gamma}$, which is minimax optimal up to a log-logarithm factor, as shown by Theorem 4 in the Appendix. Comparable rates for kernel and sparse-DNN estimators are established in Meister (2016); Shi et al. (2025).

Another advantage of the proposed estimator $\widehat{f}_{\text{EW}}$ is its practical implementability. It employs an equal-width, fully connected network architecture and requires tuning only three hyperparameters, D, W, and L. In comparison, the DNN estimator studied in Shi et al. (2025) is theoretically defined through optimization over a sparsity-constrained class of neural networks with a prescribed sparsity level, which is difficult to implement directly using standard DNN training procedures.[1]

4 Application II – Functional Multiple-index Model

In this section, we specialize the general framework to the functional multiple-index model. This structural assumption restricts the form of F_0 and leads to a polynomial excess risk bound. Specifically,

[1]The numerical studies in Shi et al. (2025) use equal-width and fully connected DNN architectures, which are the same as $\widehat{f}_{\text{EW}}$.

we assume that

$$F_0(\boldsymbol{X}) = h_0\Big(\underbrace{\underbrace{U_{11},\ldots,U_{1d_1}}_{\substack{d_1 \text{ indices}\\ \text{from } X_1}},\ldots,\underbrace{U_{q1},\ldots,U_{qd_q}}_{\substack{d_q \text{ indices}\\ \text{from } X_q}}}_{\text{total } d = \sum_{j=1}^{q} d_j \text{ indices}}\Big), \text{ with } U_{jk} = \langle X_j, \beta_{jk}^0\rangle_{L_2(\mathcal{T})},$$

where the index functions β_{jk}^0 and link $h_0 : \mathbb{R}^d \to \mathbb{R}$ are unknown. Each X_j contributes d_j scalar projections. The model includes functional single-index regression when $q = d_1 = 1$ (Balasubramanian et al., 2025), functional linear regression when h_0 is linear (Yuan and Cai, 2010), and generalized functional linear regression when h_0 and the error distribution are known (Zhou et al., 2026).

Using the BEW class in Example 2, we define

$$\widehat{f}_{\mathrm{BEW}} = \underset{f\in\mathcal{F}_{\mathrm{BEW}}(W_1,W,L,C_{\mathcal{F}})}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n}\Big(Y_i - f(\widehat{\boldsymbol{\xi}}_i)\Big)^2. \tag{4.1}$$

The BEW architecture is specifically designed for functional multiple-index models because it enforces a low-dimensional index bottleneck (approximating U_{jk}'s) followed by a flexible nonlinear map (approximating h_0), thereby aligning with the model's two-stage structure. To analyze the BEW DNN estimator, we introduce the following additional assumptions.

Assumption C1. *The link function $h_0 : \mathbb{R}^d \to \mathbb{R}$ satisfies that $\sup_{\boldsymbol{x}\in\mathbb{R}^d} |h_0(\boldsymbol{x})| \le C_{\mathcal{F}}/2$. Moreover, there is an integer $p \ge 1$ such that $\sup_{\boldsymbol{x}\in\mathbb{R}^d} |\partial^{\boldsymbol{\alpha}} h_0(x)| < \infty$ for all multi-indices $\boldsymbol{\alpha}$ with $|\boldsymbol{\alpha}| \le p$.*

Assumption C2. *For $j = 1,\ldots,q$, the j-th functional predictor has the expansion $X_j(t) = \sum_{k=1}^{\infty}\xi_{jk}\varphi_k(t)$ with $\mu_{X_j}(t) = \mathbb{E}(X_j(t))$. Assume that $\max_{1\le j\le q}\sup_{k\ge 1} Var(\xi_{jk})k^{2m_1} < \infty$ and $\max_{1\le j\le q}\|\mu_{X_j}\|_{\mathcal{H}^{m_1}(\mathcal{T})} < \infty$ for some constant $m_1 > 3/2$.*

Assumption C3. *For each $j = 1,\ldots,q, k = 1,\ldots,d_j$, the index function satisfies $\beta_{jk}^0 \in \mathcal{H}^{m_2}(\mathcal{T})$ for some constant $m_2 \ge 1$. Moreover, it holds that $\max_{1\le j\le q}\max_{1\le k\le d_j}\|\beta_{jk}^0\|_{\mathcal{H}^{m_2}(\mathcal{T})} < \infty$*

Assumption C1 imposes two standard regularity conditions on the unknown link function. The uniform bound $\sup_{\boldsymbol{x}\in\mathbb{R}^d}|h_0(\boldsymbol{x})| \le C_{\mathcal{F}}/2$ is consistent with Assumption A1(ii) and ensures that output clipping leaves the true regression function unchanged. Bounded derivatives up to order p provide the smoothness needed to control the neural network approximation error $\mathcal{E}_{\text{app}}$ (Schmidt-Hieber, 2020; Farrell et al., 2021).

Assumption C2 controls the mean and stochastic components of each predictor separately. It requires the mean function $\mu_{X_j}(t)$ to belong to $\mathcal{H}^{m_1}(\mathcal{T})$ and the variances of the Fourier coefficients to decay at least as fast as k^{-2m_1}, as is standard in functional linear regression (Cai and Yuan, 2011, 2012). Assumption B2 imposes the stronger requirement of a finite expected squared Sobolev norm, which implies Assumption C2. Essentially, the parameter m_1 has different interpretations under the two assumptions: under Assumption C2, it primarily characterizes the decay of the Fourier coefficient variances, whereas under Assumption B2, it describes the Sobolev smoothness of the random sample paths.

Assumption C3 imposes Sobolev smoothness on the index functions β_{jk}^0, analogous to Assumption B2 for the predictors. The smoothness indices m_1 and m_2 jointly determine the order of the truncation error $\mathcal{E}_{\text{tr}}$. A weaker condition assumes $\beta_{jk}(t) = \sum_{s=1}^{\infty} b_{jks}\varphi_s(t)$ with $|b_{jks}| \lesssim s^{-m_2}$ (Cai and Hall, 2006; Zhou et al., 2026). This coefficientwise decay condition allows a larger class of index functions and yields a slower convergence rate; see the discussion following Theorem 3. Finally, the requirement $m_2 \ge 1$ is to ensure a nearly sharp DNN approximation error bound.

Theorem 3. *Suppose that Assumptions A1-A4 and C1-C3 hold. Assume that $W_1 \geq 2d, W \to \infty$, and $L \to \infty$ such that $D, W, L \leq n$. Then it holds with probability at least $1 - C\log(n)/n$ that*

$$\mathcal{E}\left(\widehat{f}_{\mathrm{BEW}}\right) \leq C\Bigg\{\frac{(DW_1 + W_1W + W^2L)L}{n}\big(\log(n)\big)^3 + \underbrace{D^{-2(m_1+m_2)}}_{\mathcal{E}_{tr}} + \underbrace{\left(\sum_{k\leq D} k^{1-m_2}\right)^2 \cdot \frac{1}{n}\sum_{i=1}^{n} M_i^{-2}}_{\mathcal{E}_{dis}} + \underbrace{(WL)^{-\frac{4p}{d}}\left(\log(n)\right)^{4p+\frac{8p}{d}}}_{\mathcal{E}_{app}}\Bigg\}, \tag{4.2}$$

where $C > 0$ is a constant free of n, W_1, W, L, D and M_i. As a consequence, if the functional covariates are fully observed ($M_i = \infty$), $W_1 \asymp 1$, $L \asymp \log(n)$, $D \asymp n^{1/(2(m_1+m_2)+1)}$, and $W \asymp n^{d/(4p+2d)}$, then it holds with probability at least $1 - C\log(n)/n$ that

$$\mathcal{E}\left(\widehat{f}_{\mathrm{BEW}}\right) \leq C\left(n^{-\frac{2(m_1+m_2)}{2(m_1+m_2)+1}} + n^{-\frac{2p}{2p+d}}\right)(\log(n))^{\delta}, \tag{4.3}$$

where $C > 0$ is a constant free of n, W_1, W, L, D, and $\delta = \max(5, 4p + 8p/d)$.

The four terms in (4.2) correspond to the general error decomposition. The parameter count $S \asymp DW_1 + W_1W + W^2L$ in Example 2 yields a statistical term of order $SL(\log(n))^3/n$. The combined smoothness $m_1 + m_2$ governs the truncation error. Similar to Theorem 2, the discretization error depends on D and grid resolution. The smoothness p and input dimension d of the link function determine the approximation error. Notably, a fixed bottleneck width $W_1 \geq 2d$ suffices to represent all indices. Increasing W_1 raises statistical complexity without improving the approximation bound.

With fully observed covariates ($M_i = \infty$), the tuning choices in Theorem 3 balance the errors. The resulting rate contains two polynomial terms: $n^{-2(m_1+m_2)/(2(m_1+m_2)+1)}$ for estimating the functional indices (Cai and Yuan, 2012), and $n^{-2p/(2p+d)}$ for estimating the d-dimensional link (Tsybakov, 2009). The slower term determines the overall rate. These two polynomial terms describe different restrictions on prediction accuracy. A smoother link reduces the second term, but cannot remove the difficulty of estimating the functional indices. Conversely, faster coefficient decay and smoother index functions reduce the first term, while estimation of the unknown d-dimensional link remains necessary. Notably, our convergence rate in (4.3) is nearly minimax optimal for the functional multiple-index model, up to a logarithm factor, as established by Theorem 5 in the Appendix.

A closely related work is that of Zhou et al. (2026), who study a generalized functional linear model with a known link. They obtain a rate of $n^{-(2m_1+2m_2-1)/(2m_1+2m_2)}$, up to a logarithm factor, for excess risk under a weaker regularity condition on β_{jk}^0 (see the discussion of Assumption C3). Their rate contains only one polynomial term because the link need not be estimated.

Remark 1 (Benefit of the BEW architecture)**.** *Setting $W_1 = W$ in the BEW class recovers the EW class of Example 1. Theorem 3 therefore gives the following bound with probability at least $1 - C\log(n)/n$:*

$$\mathcal{E}\left(\widehat{f}_{\mathrm{EW}}\right) \lesssim \frac{(DW + W^2L)L}{n}\big(\log(n)\big)^3 + D^{-2(m_1+m_2)} + \left(\sum_{k\leq D} k^{1-m_2}\right)^2 \frac{1}{n}\sum_{i=1}^{n} M_i^{-2} + (WL)^{-\frac{4p}{d}}\left(\log(n)\right)^{4p+\frac{8p}{d}}.$$

The input-layer contribution to the statistical term increases from order D to order DW, while the other three error terms remain the same.

For fully observed functional covariates ($M_i = \infty$), the discretization term vanishes. Optimizing the remaining terms with $D \asymp n^{\kappa/(2(m_1+m_2))}$, $W \asymp n^{\kappa d/(4p)}$, and $L \asymp \log(n)$ yields

$$\mathcal{E}\left(\widehat{f}_{\mathrm{EW}}\right) \lesssim \left(n^{-\frac{4p(m_1+m_2)}{4p(m_1+m_2)+d(m_1+m_2)+2p}} + n^{-\frac{2p}{2p+d}}\right)\big(\log(n)\big)^{\delta}, \tag{4.4}$$

where $\delta = \max(5, 4p + 8p/d)$ *and*

$$\kappa = \min\left\{\frac{4p(m_1+m_2)}{4p(m_1+m_2)+d(m_1+m_2)+2p}, \frac{2p}{2p+d}\right\}.$$

The EW and BEW bounds in (4.4) *and* (4.3)*, respectively, have the same polynomial rate when* $2p \leq d(m_1+m_2)$*, whereas the EW bound decays at a strictly slower polynomial rate when* $2p > d(m_1+m_2)$*. This difference stems from the number of parameters in the input layer: for a fixed bottleneck width* $W_1 \geq 2d$*, this number is of order* D *for BEW and* DW *for EW. The numerical results in Section 5 and Appendix B further illustrate the potential predictive advantages of the BEW architecture.*

5 Numerical Studies

Summary of Simulation Studies. To provide a concise overview, we summarize the main simulation findings here and report the full numerical results and corresponding tables in Section B. The simulation study compares the proposed EW and BEW architectures with the approach proposed in Zhou et al. (2026) (hereafter, referred to as FPC) under varying sample sizes, grid resolutions, functional-covariate generation mechanisms, and interaction strengths. Two main conclusions emerge. First, all methods improve as the sample size and grid resolution increase, but their relative performance depends strongly on the complexity of the regression structure. When no interaction is present, EW is generally competitive and often achieves the lowest prediction error. In contrast, when nonlinear interactions are introduced, BEW increasingly outperforms EW and FPC approaches, particularly for larger sample sizes and denser functional observations. Second, the sensitivity analyses highlight the importance of architectural regularization. BEW performs best with a relatively narrow bottleneck and a small-to-moderate basis dimension, whereas unnecessarily large values of W_1 or D can increase prediction error. Overall, the results support the theoretical motivation for BEW: its additional flexibility is most beneficial when the regression surface contains nonlinear interactions, while the simpler EW architecture remains advantageous in less complex settings.

5.1 Data Application

Appliances Energy Data We analyze the Appliances Energy Prediction data from the UCI Machine Learning Repository (Candanedo, 2017; Candanedo et al., 2017), collected over approximately 4.5 months in a low-energy residence in Belgium. Appliance energy use and indoor temperature and humidity were recorded every 10 minutes. We treat each complete day as one observation, yielding functional trajectories on 144 equally spaced time points. Four functional predictors are used:

$$X_{i1}(t) = T_{1i}(t), \qquad X_{i2}(t) = T_{2i}(t), \qquad X_{i3}(t) = RH_{1i}(t), \qquad X_{i4}(t) = RH_{2i}(t),$$

where T_1, RH_1 correspond to the kitchen area and T_2, RH_2 to the living room. Letting $A_i(t)$ denote the daily appliance-energy trajectory, the scalar response is the daily 95th percentile,

$$Y_i = Q_{0.95}\left\{A_i(t) : t \in [0,1]\right\},$$

which targets high-consumption behavior. Prediction accuracy is assessed using an out-of-bag (OOB) bootstrap: bootstrap multiplicities are retained in the training sample, validation and OOB samples contain unique observations, and the response is standardized using the training sample only. The same splits are used for EW, BEW, and FPC. The average OOB MSE values across 200 replications are reported in Table 1.

Diffusion Tensor Imaging Data We also consider the DTI2 data from the `refund` package (Goldsmith et al., 2011, 2012; Swihart et al., 2014). The data contain repeated diffusion tensor imaging measurements together with Paced Auditory Serial Addition Test (PASAT) scores. We use mean diffusivity profiles along the corpus callosum and parallel diffusivity profiles along the right corticospinal tract as the two functional predictors,

$$X_{i1}(t) = \text{corpus callosum profile}, \qquad X_{i2}(t) = \text{right corticospinal tract profile},$$

with PASAT score as the scalar response. The two profiles are observed at 93 and 55 locations, respectively. To obtain independent observations, we retain the first complete visit for each subject and exclude incomplete cases. The functional domains are rescaled to $[0,1]$. Performance is evaluated using the same OOB bootstrap scheme as above, with PASAT standardized using the training sample only. The MRI/DTI data were collected at Johns Hopkins University and the Kennedy-Krieger Institute. The average OOB MSEs over 200 replications are reported in Table 1.

Table 1: Out-of-bag prediction performance for the UCI Appliances and DTI datasets across 200 replications. Entries are mean OOB MSE with standard deviation in parentheses; the median OOB MSE is also reported. Results for DTI are on the standardized PASAT scale.

	UCI Appliances		DTI	
Method	Mean MSE (Std.)	Median MSE	Mean MSE (Std.)	Median MSE
EW	1.0867 (0.2579)	1.0410	1.3211 (0.7094)	1.1196
BEW	**1.0736** (0.2866)	**1.0366**	**1.2780** (0.6394)	**1.1099**
FPC	1.1327 (0.2905)	1.0993	1.3120 (0.6884)	1.1193

Results The OOB results across 200 replications are summarized in Table 1. For the Appliances data, BEW achieves the lowest mean MSE (1.0736) and median MSE (1.0366), compared with 1.0867 and 1.0410 for EW and 1.1327 and 1.0993 for FPC, respectively. A similar pattern is observed for DTI, where BEW attains the lowest mean OOB MSE (1.2780) and median OOB MSE (1.1099), again improving on both EW and FPC. Although the differences are modest, the results are consistent across both applications and suggest that the additional flexibility of BEW can yield a small but systematic predictive gain in real-data settings. This finding is also in line with the simulation results, where BEW becomes increasingly competitive, and often achieves the lowest test MSE, as the regression structure becomes more complex through stronger nonlinear interactions and as the sample size and grid resolution increase.

6 Conclusion and Limitations

This paper develops a deep learning framework for scalar-on-function regression. The proposed method first represents functional covariates by finitely many basis coefficients and then fits a ReLU DNN using the resulting finite-dimensional features. We establish a general nonasymptotic excess risk bound that decomposes the error into statistical, truncation, discretization, and network approximation terms. We further apply the general theory to two important settings: continuous functional models and functional multiple-index models. With appropriately designed architectures, the proposed DNN estimators achieve nearly minimax optimal convergence rates in both settings.

Several limitations remain. First, our sampling design assumes that, for each subject i, all functional components of $\boldsymbol{X}_i$ are observed on the same subject-specific grid T_i. In some applications, different components X_{ij} may instead be observed on different grids T_{ij}, with different resolutions or missingness patterns. The proposed method can in principle be extended to this setting by estimating the basis coefficients of each component on its own grid. However, doing so would substantially complicate the notation and the discretization error analysis, since the error would need to be tracked separately over both subjects and functional components. Second, in the functional multiple-index model, the individual index functions are generally not identifiable without additional structural constraints. For functional single-index models, identifiability can be achieved under suitable restrictions; see, for example, Chen et al. (2011); Jiang and Wang (2011). For the more general multiple-index setting considered here, however, transformations of the index functions can often be absorbed into a corresponding transformation of the link function, leaving the overall regression functional unchanged. Therefore, our theory should be interpreted primarily as providing prediction guarantees for estimating $F_0(\boldsymbol{X})$, rather than as guaranteeing recovery of the true index functions. Developing identifiable formulations and inference procedures for functional multiple-index models is an important direction for future work. Finally, our analysis relies on a conditionally sub-Gaussian error assumption. Relaxing this condition to accommodate heavy-tailed noise would broaden the applicability of the proposed framework and is another important topic for future research.

References

Allard, T. and Bölcskei, H. (2026). Metric entropy of ellipsoids in Banach spaces: Techniques and precise asymptotics. *Journal of Functional Analysis*, 290(10):111400.

Bai, J., Song, Q., and Cheng, G. (2020). Efficient variational inference for sparse deep learning with theoretical guarantee. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 466–476. Curran Associates, Inc.

Balakrishnan, N. and Nevzorov, V. B. (2003). *A Primer on Statistical Distributions*. John Wiley & Sons, Inc., Hoboken, New Jersey.

Balasubramanian, K., Müller, H.-G., and Sriperumbudur, B. K. (2025). Functional linear and single-index models: A unified approach via Gaussian Stein identity. *Bernoulli*, 31(2):973–1006.

Bank, D., Koenigstein, N., and Giryes, R. (2023). Autoencoders. In Rokach, L., Maimon, O., and Shmueli, E., editors, *Machine Learning for Data Science Handbook: Data Mining and Knowledge Discovery Handbook*, pages 353–374. Springer International Publishing, Cham, third edition.

Bartlett, P. L., Bousquet, O., and Mendelson, S. (2005). Local Rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537.

Bartlett, P. L., Harvey, N., Liaw, C., and Mehrabian, A. (2019). Nearly-tight VC-dimension and pseudodimension bounds for piecewise linear neural networks. *Journal of Machine Learning Research*, 20(63):1–17.

Bauer, B. and Kohler, M. (2019). On deep learning as a remedy for the curse of dimensionality in nonparametric regression. *The Annals of Statistics*, 47(4):2261–2285.

Cai, T. T. and Hall, P. (2006). Prediction in functional linear regression. *The Annals of Statistics*, 34(5):2159–2179.

Cai, T. T. and Yuan, M. (2011). Optimal estimation of the mean function based on discretely sampled functional data: Phase transition. *The Annals of Statistics*, 39(5):2330–2355.

Cai, T. T. and Yuan, M. (2012). Minimax and adaptive prediction for functional linear regression. *Journal of the American Statistical Association*, 107(499):1201–1216.

Candanedo, L. (2017). Appliances energy prediction. UCI Machine Learning Repository. Dataset.

Candanedo, L. M., Feldheim, V., and Deramaix, D. (2017). Data driven prediction models of energy use of appliances in a low-energy house. *Energy and Buildings*, 140:81–97.

Chen, D., Hall, P., and Müller, H.-G. (2011). Single and multiple index functional regression models with nonparametric link. *The Annals of Statistics*, 39(3):1720–1747.

Chérief-Abdellatif, B.-E. (2020). Convergence rates of variational inference in sparse deep learning. In Daumé, III, H. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1831–1842. PMLR.

Farrell, M. H., Liang, T., and Misra, S. (2021). Deep neural networks for estimation and inference. *Econometrica*, 89(1):181–213.

Goldsmith, J., Bobb, J., Crainiceanu, C. M., Caffo, B., and Reich, D. (2011). Penalized functional regression. *Journal of Computational and Graphical Statistics*, 20(4):830–851.

Goldsmith, J., Crainiceanu, C. M., Caffo, B., and Reich, D. (2012). Longitudinal penalized functional regression for cognitive outcomes on neuronal tract measurements. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 61(3):453–469.

Jiang, C.-R. and Wang, J.-L. (2011). Functional single index models for longitudinal data. *The Annals of Statistics*, 39(1):362–388.

Kim, Y., Ohn, I., and Kim, D. (2021). Fast convergence rates of deep neural networks for classification. *Neural Networks*, 138:179–197.

Kohler, M. and Langer, S. (2021). On the rate of convergence of fully connected deep neural network regression estimates. *The Annals of Statistics*, 49(4):2231–2249.

Kokoszka, P. and Reimherr, M. (2017). *Introduction to Functional Data Analysis*. Chapman and Hall/CRC, first edition.

Kosorok, M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference*. Springer Series in Statistics. Springer, New York, NY, first edition.

Lu, J., Shen, Z., Yang, H., and Zhang, S. (2021). Deep network approximation for smooth functions. *SIAM Journal on Mathematical Analysis*, 53(5):5465–5506.

Lu, Z., Pu, H., Wang, F., Hu, Z., and Wang, L. (2017). The expressive power of neural networks: A view from the width. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30, pages 6231–6239. Curran Associates, Inc.

Luo, R. and Qi, X. (2023). Nonlinear function-on-scalar regression via functional universal approximation. *Biometrics*, 79(4):3319–3331.

Luo, R. and Qi, X. (2024). General nonlinear function-on-function regression via functional universal approximation. *Journal of Computational and Graphical Statistics*, 33(2):578–587.

Meister, A. (2016). Optimal classification and nonparametric regression for functional data. *Bernoulli*, 22(3):1729–1744.

Mohri, M. and Muñoz Medina, A. (2014). Learning theory and algorithms for revenue optimization in second-price auctions with reserve. In Xing, E. P. and Jebara, T., editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32(1) of *Proceedings of Machine Learning Research*, pages 262–270, Beijing, China. PMLR.

Ozaki, Y., Watanabe, S., and Yanase, T. (2025). Optunahub: A platform for black-box optimization. *arXiv preprint arXiv:2510.02798*.

Polson, N. G. and Ročková, V. (2018). Posterior concentration for sparse deep learning. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 31, pages 930–941. Curran Associates, Inc.

Raghu, M., Poole, B., Kleinberg, J., Ganguli, S., and Sohl-Dickstein, J. (2017). On the expressive power of deep neural networks. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2847–2854. PMLR.

Rao, J. S. and Sobel, M. (1980). Incomplete Dirichlet integrals with applications to ordered uniform spacings. *Journal of Multivariate Analysis*, 10(4):603–610.

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

Schreuder, N. (2020). Bounding the expectation of the supremum of empirical processes indexed by Hölder classes. *Mathematical Methods of Statistics*, 29(1):76–86.

Shen, Z., Yang, H., and Zhang, S. (2020). Deep network approximation characterized by number of neurons. *Communications in Computational Physics*, 28(5):1768–1811.

Shi, Z., Fan, J., Song, L., Zhou, D.-X., and Suykens, J. A. (2025). Nonlinear functional regression by functional deep neural network with kernel embedding. *Journal of Machine Learning Research*, 26(284):1–49.

Swihart, B. J., Goldsmith, J., and Crainiceanu, C. M. (2014). Restricted likelihood ratio tests for functional effects in the functional linear model. *Technometrics*, 56(4):483–493.

Thind, B., Multani, K., and Cao, J. (2023). Deep learning with functional inputs. *Journal of Computational and Graphical Statistics*, 32(1):171–180.

Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, New York, NY, first edition.

Wainwright, M. J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, volume 48 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.

Wang, S., Huang, Y., and Cao, G. (2024a). Review on functional data classification. *WIREs Computational Statistics*, 16(1):e1638.

Wang, S., Zhang, W., Cao, G., and Huang, Y. (2024b). Functional data analysis using deep neural networks. *WIREs Computational Statistics*, 16(4):e70001.

Wu, S., Beaulac, C., and Cao, J. (2023). Neural networks for scalar input and functional output. *Statistics and Computing*, 33(5):118.

Yao, J., Mueller, J., and Wang, J.-L. (2021). Deep learning for functional data analysis with adaptive basis layers. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 11898–11908. PMLR.

Yuan, M. and Cai, T. T. (2010). A reproducing kernel Hilbert space approach to functional linear regression. *The Annals of Statistics*, 38(6):3412–3444.

Zhang, Z., Shi, L., and Zhou, D.-X. (2024). Classification with deep neural networks and logistic loss. *Journal of Machine Learning Research*, 25(125):1–117.

Zhou, H., Hong, J.-S., Ding, X., and Wang, J.-L. (2026). Deep neural network regression with functional covariates. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, pages 164978–164998. PMLR.

A Technical Proofs

This appendix collects the main mathematical derivations used to prove the theorems in the paper. We begin by introducing some notation. For a function class $\mathcal{F}$, define its (conditional) Rademacher complexity by

$$\mathrm{Rad}_n(\mathcal{F}) = \mathbb{E}\left(\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \vartheta_i f(\widehat{\boldsymbol{\xi}}_i) \;\middle|\; \widehat{\boldsymbol{\xi}}_1, \ldots, \widehat{\boldsymbol{\xi}}_n \right),$$

where $\{\vartheta_i\}_{i=1}^n$ are i.i.d. Rademacher random variables. For any measurable function f, define the empirical and population L_2 norms

$$\|f\|_n^2 = \frac{1}{n}\sum_{i=1}^n f^2(\widehat{\boldsymbol{\xi}}_i), \qquad \|f\|^2 = \mathbb{E}\left(\|f\|_n^2\right).$$

Moreover, for a normed space $(\mathcal{F}, \|\cdot\|_{\mathcal{F}})$, let $N(\epsilon, \mathcal{F}, \|\cdot\|_{\mathcal{F}})$ denote the ϵ-covering number of $\mathcal{F}$ under $\|\cdot\|_{\mathcal{F}}$.

A.1 Preliminary Lemmas

Lemma 1. *Let $0 = t_0 < t_1 < \cdots < t_{M-1} < t_M = 1$ be an increasing sequence, and let $g : [0,1] \to \mathbb{R}$ be a differentiable function. Then it follows that*

$$\left| \int_0^1 g(t)dt - \sum_{s=0}^{M-1} \frac{t_{s+1}-t_s}{2}\left(g(t_s)+g(t_{s+1})\right) \right| \le \frac{1}{4}\sum_{s=0}^{M-1}(t_{s+1}-t_s)^2 \sup_{x\in[0,1]}|g'(x)|,$$

$$\left| \sum_{s=0}^{M-1} \frac{t_{s+1}-t_s}{2}\left(g(t_s)+g(t_{s+1})\right) \right| \le \sup_{x\in[0,1]}|g(x)|.$$

Proof. Let $w = (a+b)/2$ with $a < b$. Integration by parts shows that

$$\int_a^b (t-w)g'(t)dt = (t-w)g(t)\big|_a^b - \int_a^b g(t)dt = \frac{b-a}{2}\left(g(a)+g(b)\right) - \int_a^b g(t)dt.$$

As a consequence, we have

$$\left| \int_a^b g(t)dt - \frac{b-a}{2}\left(g(a)+g(b)\right) \right| \le \sup_{x\in[a,b]}|g'(x)| \int_a^b |t-w|dt = \frac{(b-a)^2}{4}\sup_{x\in[a,b]}|g'(x)|.$$

By the preceding display, we have

$$\begin{aligned} &\left| \int_0^1 g(t)dt - \sum_{s=0}^{M-1} \frac{t_{s+1}-t_s}{2}\left(g(t_s)+g(t_{s+1})\right) \right| \\ &\le \sum_{s=0}^{M-1} \left| \int_{t_s}^{t_{s+1}} g(t)dt - \frac{t_{s+1}-t_s}{2}\left(g(t_s)+g(t_{s+1})\right) \right| \le \frac{1}{4}\sum_{s=0}^{M-1}(t_{s+1}-t_s)^2 \sup_{x\in[0,1]}|g'(x)|. \end{aligned}$$

Hence, we complete the proof of the first inequality.

Let $w_0 = (t_1 - t_0)/2, w_M = (t_M - t_{M-1})/2$, and $w_s = (t_{s+1} - t_{s-1})/2$ for $1 \le s \le M-1$. Then we have $\sum_{s=0}^M w_s = 1$ and

$$\left| \sum_{s=0}^{M-1} \frac{t_{s+1}-t_s}{2}\left(g(t_s)+g(t_{s+1})\right) \right| = \left| \sum_{s=0}^{M} w_s g(t_s) \right| \le \sup_{x\in[0,1]}|g(x)|,$$

which completes the proof of the second inequality. □

Lemma 2. *Let $\mathcal{G}$ be a class of functions and $\phi : \mathbb{R} \to \mathbb{R}$ be a Lipschitz function with Lipschitz constant K, then $\mathrm{Rad}_n(\phi \circ \mathcal{G}) \le K\,\mathrm{Rad}_n(\mathcal{G})$, where $\phi \circ \mathcal{G} = \{\phi \circ f : f \in \mathcal{G}\}$.*

Proof. This is the famous Talagrand's contraction principle; e.g, see Lemma 8 of Mohri and Muñoz Medina (2014). □

Lemma 3. *Suppose that $Z_1, \ldots, Z_n$ are independent random vectors. Let $\mathcal{G}$ be a class of functions such that $\|g\|_\infty \leq B$ and $\sum_{i=1}^n Var(g(Z_i)) \leq nV$ for all $g \in \mathcal{G}$ and for some constants $B, V > 0$. Then for any $\eta > 0$, with probability at least $1 - 2e^{-\eta}$, it holds that*

$$\sup_{g \in \mathcal{G}} \left| \frac{1}{n} \sum_{i=1}^n g(Z_i) - \frac{1}{n} \sum_{i=1}^n \mathbb{E}\left(g(Z_i)\right) \right| \leq 3\mathbb{E}(\mathrm{Rad}_n(\mathcal{G})) + \sqrt{\frac{2V\eta}{n}} + \frac{4B\eta}{3n}.$$

Proof. This is a direct extension of Theorem 2.1 in Bartlett et al. (2005) to independent but non-identically distributed data; the required modifications are routine. □

Lemma 4. *Suppose that $Z_1, \ldots, Z_n$ are independent random vectors. Let $r > 0$ be a constant, and assume that $\mathcal{G}_{n,r}$ is a class of functions such that $\sup_{f \in \mathcal{G}_{n,r}} \|f\|_n \leq r$ for some $r > 0$. Then it follows that*

$$\mathrm{Rad}_n(\mathcal{G}_{n,r}) \leq \inf_{0 < x < r} \left\{ 4x + 12r \sqrt{\frac{\log N(x, \mathcal{G}_{n,r}, \|\cdot\|_n)}{n}} \right\}.$$

Proof. This is the refined Dudley's bound; e.g., see Theorem 3 of Schreuder (2020). □

Lemma 5. *Suppose that $Z_1, \ldots, Z_n$ are independent random vectors. Let $\mathcal{G}$ be a class of functions, and let v be its VC dimension. If $\max_{1 \leq i \leq n} \sup_{f \in \mathcal{G}} |f(Z_i)| \leq B$ for some constant $B > 0$ almost surely, then it follows that*

$$\log N(x, \mathcal{G}, \|\cdot\|_n) \leq Cv \log\left(\frac{CB}{x}\right), \quad \text{for all } 0 < x \leq B,$$

where $C > 0$ is constant free of v, n, B, and Z_i. As a consequence, we have

$$\log N(x, \mathcal{F}(S, L, C_{\mathcal{F}}), \|\cdot\|_n) \leq CSL \log(S) \log\left(\frac{C}{x}\right),$$

where $C > 0$ is a constant depending on $C_{\mathcal{F}}$.

Proof. The first result is Theorem 9.3 in Kosorok (2008). The second one follows from the first result, Lemma 2, and Theorem 7 in Bartlett et al. (2019). □

Lemma 6. *Let $f_* \in \mathcal{F}(S, L, C_{\mathcal{F}})$ be a fixed deterministic function and $\mathcal{G}_{n,r} = \{f - f_* : f \in \mathcal{F}(S, L, C_{\mathcal{F}}), \|f - f_*\|_n \leq r\}$ with $r > 1/n$. Then it holds that*

$$\mathrm{Rad}_n(\mathcal{G}_{n,r}) \leq C \sqrt{\frac{SL \log(S) \log(n)}{n}} r,$$

where the constant $C > 0$ is depending on $C_{\mathcal{F}}$. Especially, we have

$$\mathrm{Rad}_n(\mathcal{F}(S, L, C_{\mathcal{F}})) \leq C \sqrt{\frac{SL \log(S) \log(n)}{n}},$$

where the constant $C > 0$ is depending on $C_{\mathcal{F}}$. In addition, if $S, L \leq n$, then it holds that

$$\mathrm{Rad}_n(\mathcal{G}_{n,r}) \leq C \sqrt{\frac{SL}{n}} \log(n) r, \quad \mathrm{Rad}_n(\mathcal{F}(S, L, C_{\mathcal{F}})) \leq C \sqrt{\frac{SL}{n}} \log(n),$$

where the constant $C > 0$ is depending on $C_{\mathcal{F}}$. Moreover, assume $S \leq n^3$ and define

$$\mathcal{G}(r) = \{f - f_* : f \in \mathcal{F}(S, L, C_{\mathcal{F}}),\ \|f - f_*\| \leq r\}.$$

If $r \geq \sqrt{SL/n} \log(n)$, then

$$\mathbb{E}\left(\mathrm{Rad}_n(\mathcal{G}(r))\right) \leq Cr \sqrt{\frac{SL}{n}} \log(n),$$

where $C > 0$ depends only on $C_{\mathcal{F}}$.

Proof. By Lemmas 4 and 5, it follows that

$$\begin{aligned}\mathrm{Rad}_n(\mathcal{G}_{n,r}) &\le \inf_{0<x<\min\{r,1\}} \left\{ 4x + 12r\sqrt{\frac{\log N(x, \mathcal{F}(S,L,C_{\mathcal{F}}), \|\cdot\|_n)}{n}} \right\} \\ &\le C \inf_{0<x<\min\{r,1\}} \left\{ x + r\sqrt{\frac{SL\log(S)\log(C/x)}{n}} \right\} \overset{\text{(i)}}{\le} C\sqrt{\frac{SL\log(S)\log(n)}{n}}\, r.\end{aligned}$$

Here $C > 0$ depends only on $C_{\mathcal{F}}$. To justify (i), choose $x = \min\{r, 1\}/n$. Since $n \ge 2$ and $r > 1/n$, we have $n^{-2} < x \le n^{-1}$ and $0 < x < \min\{r, 1\}$, so

$$\log(C/x) \le \log(Cn^2) \le \left(2 + \frac{\log C}{\log 2}\right)\log(n).$$

Moreover, $S \ge 2$ and $L \ge 1$ for the network architectures considered, and hence

$$x \le \frac{r}{n} \le Cr\sqrt{\frac{SL\log(S)\log(n)}{n}}.$$

Substituting these bounds establishes (i) and proves the first inequality.

The global bound follows by applying the first inequality with $r = 2C_{\mathcal{F}} + 1/n$, since $\|f - f_*\|_\infty \le 2C_{\mathcal{F}}$ and $r \le 2C_{\mathcal{F}} + 1/2$ for $n \ge 2$.

To prove the population bound, write $a_n = \sqrt{SL/n}\log(n)$ and define

$$R_r = \sup_{g\in\mathcal{G}(r)} \|g\|_n, \qquad Z_r = \sup_{g\in\mathcal{G}(r)} \left|\|g\|_n^2 - \|g\|^2\right|.$$

Then $R_r^2 \le r^2 + Z_r$ and $R_r \le 2C_{\mathcal{F}}$. If $R_r > 1/n$, then conditionally on the covariates, we have

$$\begin{aligned}\mathrm{Rad}_n(\mathcal{G}(r)) &\overset{\text{(i)}}{\le} \inf_{0<x<R_r} \left\{ 4x + 12R_r\sqrt{\frac{\log N(x, \mathcal{G}(r), \|\cdot\|_n)}{n}} \right\} \\ &\overset{\text{(ii)}}{\le} \frac{C}{n} + CR_r\sqrt{\frac{SL\log(S)\log(n)}{n}} \\ &\overset{\text{(iii)}}{\le} \frac{C}{n} + Ca_nR_r,\end{aligned}$$

where (i) is due to Lemma 4, (ii) follows from Lemma 5 by choosing $x = 1/n$, and (iii) uses $S \le n^3$. If $R_r \le 1/n$, the same bound follows from $\mathrm{Rad}_n(\mathcal{G}(r)) \le R_r$ by Cauchy–Schwarz.

Let $\mathcal{G}^2(r) = \{g^2 : g \in \mathcal{G}(r)\}$. For every $g \in \mathcal{G}(r)$, $\|g^2\|_\infty \le 4C_{\mathcal{F}}^2$ and

$$\frac{1}{n}\sum_{i=1}^n \mathrm{Var}\left(g^2(\widehat{\boldsymbol{\xi}}_i)\right) \le \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left(g^4(\widehat{\boldsymbol{\xi}}_i)\right) \le 4C_{\mathcal{F}}^2\|g\|^2 \le 4C_{\mathcal{F}}^2 r^2.$$

Applying Lemma 3 to the fixed class $\mathcal{G}^2(r)$ with $\eta = 2\log(n)$, it holds with probability at least $1 - 2/n^2$ that

$$\begin{aligned}Z_r &\le 3\mathbb{E}\left(\mathrm{Rad}_n(\mathcal{G}^2(r))\right) + 4C_{\mathcal{F}}r\sqrt{\frac{\log(n)}{n}} + \frac{32C_{\mathcal{F}}^2\log(n)}{3n} \\ &\overset{\text{(i)}}{\le} 12C_{\mathcal{F}}\mathbb{E}\left(\mathrm{Rad}_n(\mathcal{G}(r))\right) + 4C_{\mathcal{F}}r\sqrt{\frac{\log(n)}{n}} + \frac{32C_{\mathcal{F}}^2\log(n)}{3n} \\ &\overset{\text{(ii)}}{\le} \frac{C}{n} + Ca_n\mathbb{E}(R_r) + Cr\sqrt{\frac{\log(n)}{n}} + \frac{C\log(n)}{n} \\ &\overset{\text{(iii)}}{\le} Ca_n\mathbb{E}(R_r) + Cra_n,\end{aligned}$$

where (i) is due to Lemma 2, since squaring is $4C_{\mathcal{F}}$-Lipschitz on $[-2C_{\mathcal{F}}, 2C_{\mathcal{F}}]$, (ii) follows from the preceding Rademacher bound, and (iii) uses $r \ge a_n$. Taking expectations gives

$$\begin{aligned}\mathbb{E}(R_r^2) &\overset{\text{(i)}}{\le} r^2 + Ca_n\mathbb{E}(R_r) + Cra_n + \frac{8C_{\mathcal{F}}^2}{n^2} \\ &\overset{\text{(ii)}}{\le} r^2 + Ca_n\sqrt{\mathbb{E}(R_r^2)} + Cra_n + \frac{8C_{\mathcal{F}}^2}{n^2} \overset{\text{(iii)}}{\le} \frac{1}{2}\mathbb{E}(R_r^2) + Cr^2,\end{aligned}$$

where (i) uses $R_r^2 \le r^2 + Z_r$ on the concentration event and $R_r^2 \le 4C_{\mathcal{F}}^2$ on its complement, whose probability is at most $2/n^2$, (ii) is due to Jensen's inequality, and (iii) follows from Young's inequality, $a_n \le r$, and $1/n^2 \le a_n^2$. Thus $\mathbb{E}(R_r) \le Cr$, and the conditional Rademacher bound yields

$$\mathbb{E}\left(\mathrm{Rad}_n(\mathcal{G}(r))\right) \le \frac{C}{n} + Ca_n\mathbb{E}(R_r) \le Cra_n,$$

where the last inequality uses $1/n \le a_n^2 \le ra_n$. This proves the last statement. □

Lemma 7. *For any $0 < A \le B$, it holds that*

$$\int_0^A \sqrt{\log\left(\frac{B}{x}\right)}dx \le A\sqrt{\log\left(\frac{eB}{A}\right)}.$$

In particular, there exists a universal constant $C > 0$ such that

$$\int_0^A \sqrt{\log\left(\frac{B}{x}\right)}dx \le CA\sqrt{\log\left(\frac{B}{A}\right)}, \qquad 0 < A \le B/e.$$

Proof. By Cauchy–Schwarz, we have

$$\begin{aligned}\left(\int_0^A \sqrt{\log\left(\frac{B}{x}\right)}dx\right)^2 &\le A\int_0^A \log\left(\frac{B}{x}\right)dx \\ &= A^2\left\{1 + \log\left(\frac{B}{A}\right)\right\} = A^2\log\left(\frac{eB}{A}\right).\end{aligned}$$

Taking square roots yields the first inequality. If $A \le B/e$, then $\log(B/A) \ge 1$ and hence $\log(eB/A) \le 2\log(B/A)$, which gives the second inequality with $C = \sqrt{2}$. This completes the proof. □

Lemma 8. *Assume that $\sup_{k\ge1}\|\varphi_k\|_\infty < \infty$. If $m \ge 1$, then for all $f \in \mathcal{H}^m(\mathcal{T})$, it holds that $\|f\|_\infty \le C\|f\|_{\mathcal{H}^m(\mathcal{T})}$, where the constant $C > 0$ depending on m and the basis functions.*

Proof. Let $f = \sum_{k=1}^\infty a_k\varphi_k$. By the uniformly boundedness of φ_k, it holds that

$$|f(t)| = \left|\sum_{k=1}^\infty a_k\varphi_k(t)\right| \le C\sum_{k=1}^\infty |a_k|.$$

By Cauchy-Schwarz, it holds that

$$\sum_{k=1}^\infty |a_k| = \sum_{k=1}^\infty |a_k|k^m k^{-m} \le \left(\sum_{k=1}^\infty a_k^2 k^{2m}\right)^{1/2}\left(\sum_{k=1}^\infty k^{-2m}\right)^{1/2} \le C\|f\|_{\mathcal{H}^m(\mathcal{T})}.$$

Combining the preceding two displays, we complete the proof. □

Lemma 9. *Assume $X(t) = \sum_{k=1}^\infty A_k\varphi_k(t)$ and $\beta(t) = \sum_{k=1}^\infty b_k\varphi_k(t)$. Let $0 = t_0 < t_1 < \ldots < t_{M-1} < t_M = 1$ be a grid on $[0,1]$ and define $\widehat{A}_k = \sum_{i=0}^M u_i X(t_i)\varphi_k(t_i)$ with $u_0 = t_1/2, u_M = (t_M - t_{M-1})/2$ and $u_i = (t_{i+1} - t_{i-1})/2$ for $i = 1, \ldots, M-1$. Moreover, assume there are constants $C_U > 0, m_1 > 3/2$ and $m_2 \ge 1$ such that the following conditions hold:*

- *(i)* $\sup_{k\ge1}\|\varphi_k\|_\infty + \sup_{k\ge1} k^{-1}\|\varphi_k'\|_\infty \le C_U$.
- *(ii) Let ρ_{ij} be the correlation coefficient between A_i and A_j. It satisfies that $\sup_{i\ge1}\sum_{j=1}^\infty |\rho_{ij}| \le C_U$.*
- *(iii) For $\sigma_k^2 = Var(A_k)$, it satisfies that $\sigma_k^2 \le C_U k^{-2m_1}$ for all $k \ge 1$.*
- *(iv) For $\mu_k = \mathbb{E}(A_k)$, it satisfies that $\sum_{k=1}^\infty \mu_k^2 k^{2m_1} \le C_U^2$.*
- *(v) $A_k k^{m_1}$ is C_U-sub-Gaussian.*
- *(vi) $\sum_{k=1}^\infty b_k^2 k^{2m_2} \le C_U^2$.*

Under Conditions (i)-(iv), it holds that

$$\mathbb{E}\left(\left|\widehat{A}_k - A_k\right|^2\right) \leq Ck^2\left(\sum_{i=1}^{M}(t_i - t_{i-1})^2\right)^2,$$

where $C > 0$ is a constant depending on m_1 and C_U. Under Conditions (i)-(v), it holds that

$$\mathbb{P}\left(|\widehat{A}_k| \geq t\right) \leq 2\exp(-t^2/C), \text{ for all } t > 0,$$

where $C > 0$ is a constant depending on m_1 and C_U. Under Conditions (i)-(vi), it holds that

$$\mathbb{E}\left(\left|\langle X - \Pi_D(X), \beta\rangle_{L_2(\mathcal{T})}\right|^2\right) \leq CD^{-2(m_1+m_2)},$$

where $C > 0$ is a constant depending on m_1, m_2, and C_U.

Proof. We will prove the three inequalities one by one.

Step 1: By direct examination, we have

$$\widehat{A}_k = \sum_{i=0}^{M} u_i X(t_i)\varphi_k(t_i) = \sum_{i=0}^{M} u_i \sum_{j=1}^{\infty} A_j\varphi_j(t_i)\varphi_k(t_i) = \sum_{j=1}^{\infty} A_j \sum_{i=0}^{M} u_i\varphi_j(t_i)\varphi_k(t_i),$$
$$A_k = \int_0^1 X(t)\varphi_k(t)dt = \int_0^1 \sum_{j=1}^{\infty} A_j\varphi_j(t)\varphi_k(t)dt = \sum_{j=1}^{\infty} A_j \int_0^1 \varphi_j(t)\varphi_k(t)dt.$$

For each pair j and k, we define

$$R_{jk} := \sum_{i=0}^{M} u_i\varphi_j(t_i)\varphi_k(t_i) - \int_0^1 \varphi_j(t)\varphi_k(t)dt.$$

By Lemma 1, we have

$$|R_{jk}| \leq \frac{\sup_{t\in[0,1]} |\varphi_j'(t)\varphi_k(t) + \varphi_j(t)\varphi_k'(t)|}{4} \sum_{i=1}^{M}(t_i - t_{i-1})^2 \overset{\text{(i)}}{\leq} C_U^2(j+k)\sum_{i=0}^{M}(t_i - t_{i-1})^2$$
$$:= C(j+k)S_M$$

where (i) is due to Condition (i). As a consequence, it follows that

$$\widehat{A}_k - A_k = \sum_{j=1}^{\infty} A_j R_{jk} = \sum_{j=1}^{\infty} \mu_j R_{jk} + \sum_{j=1}^{\infty}(A_j - \mu_j)R_{jk}.$$

By direct examination, we have

$$\left|\sum_{j=1}^{\infty} \mu_j R_{jk}\right|^2 \leq \left(\sum_{j=1}^{\infty} j^{2m_1}\mu_j^2\right)\left(\sum_{j=1}^{\infty} j^{-2m_1}|R_{jk}|^2\right)$$
$$= CC_U^2 S_M^2 \sum_{j=1}^{\infty} j^{-2m_1}(j+k)^2 \overset{\text{(i)}}{\leq} C(1+k)^2 S_M^2,$$

where (i) holds since $m_1 > 3/2$. Let $C_\rho = \sup_{i\geq 1}\sum_{j=1}^{\infty}|\rho_{ij}| \leq C_U$ by Condition (ii), and we have

$$\begin{aligned}
\mathbb{E}\left(\left|\sum_{j=1}^{\infty}(A_j-\mu_j)R_{jk}\right|^2\right) &= \sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\mathrm{Cov}(A_i,A_j)R_{ik}R_{jk} \\
&\leq C\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}|\rho_{ij}|\sqrt{\mathrm{Var}(A_i)\mathrm{Var}(A_j)}(i+k)(j+k)S_M^2 \\
&\overset{\text{(i)}}{\leq} CS_M^2\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}|\rho_{ij}|i^{-m_1}j^{-m_1}(i+k)(j+k) \\
&\leq CS_M^2\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}|\rho_{ij}|\times\frac{i^{-2m_1}(i+k)^2+j^{-2m_1}(j+k)^2}{2} \\
&\leq CC_\rho S_M^2\sum_{i=1}^{\infty}i^{-2m_1}(i+k)^2 \overset{\text{(ii)}}{\leq} C(1+k)^2S_M^2,
\end{aligned}$$

where (i) is from Conditions (iii), and (ii) holds since $m_1 > 3/2$. Combining the previous three displays, we prove the first inequality.

Step 2: By direct examination, we have

$$\widehat{A}_k = \sum_{j=1}^{\infty}A_j r_{jk},$$

where $r_{jk} = \sum_{i=0}^{M}u_i\varphi_j(t_i)\varphi_k(t_i)$ satisfies $|r_{jk}| \leq C_U^2$ due to Condition (i). Using the sub-Gaussianity of A_j in Condition (v), we have

$$\left\|\widehat{A}_k\right\|_{\psi_2} \leq C_U^2\sum_{j=1}^{\infty}\|A_j\|_{\psi_2} \overset{\text{(i)}}{\leq} C_U^3\sum_{j=1}^{\infty}j^{-m_1} \overset{\text{(ii)}}{\leq} C,$$

where $\|\cdot\|_{\psi_2}$ is the Orlicz norm with $\psi_2(x) = e^{x^2}-1$, (i) uses the relation between Orlicz norm and sub-Gaussianity, and (ii) holds since $m_1 > 3/2$. Using the relation between Orlicz norm and sub-Gaussianity, we complete the proof of the second inequality.

Step 3: By direct examination, we have

$$\langle X-\Pi_D(X),\beta\rangle_{L_2(\mathcal{T})} = \sum_{k>D}A_kb_k = \sum_{k>D}\mu_kb_k + \sum_{k>D}(A_k-\mu_k)b_k.$$

Similar analysis to those in Step 1 gives

$$\begin{aligned}
\left|\sum_{k>D}\mu_kb_k\right|^2 &\leq \left(\sum_{k>D}k^{2m_2}b_k^2\right)\left(\sum_{k>D}k^{-2m_2}\mu_k^2\right) \\
&\leq D^{-2(m_1+m_2)}\left(\sum_{k>D}k^{2m_2}b_k^2\right)\left(\sum_{k>D}k^{2m_1}\mu_k^2\right) \overset{\text{(i)}}{\leq} C_U^4D^{-2(m_1+m_2)},
\end{aligned}$$

where (i) uses Condition (iv) and the additional condition on b_k. Similarly, we have

$$\begin{aligned}
\mathbb{E}\left(\left|\sum_{k>D}(A_k-\mu_k)b_k\right|^2\right) &= \sum_{k>D}\sum_{l>D} b_k b_l \mathrm{Cov}(A_k, A_l) \\
&\le \sum_{k>D}\sum_{l>D} |b_k||b_l||\rho_{kl}|\sqrt{\mathrm{Var}(A_k)\mathrm{Var}(A_l)} \\
&\overset{\text{(i)}}{\le} C_U \sum_{k>D}\sum_{l>D} k^{-m_1}|b_k| l^{-m_1}|b_l||\rho_{kl}| \\
&\le C_U \sum_{k>D}\sum_{l>D} \frac{k^{-2m_1}b_k^2 + l^{-2m_1}b_l^2}{2} \times |\rho_{kl}| \\
&\overset{\text{(ii)}}{\le} C_U^2 \sum_{k>D} k^{-2m_1} b_k^2 \\
&\le C_U^2 D^{-2(m_1+m_2)} \sum_{k>D} k^{2m_2} b_k^2 \overset{\text{(iii)}}{\le} C_U^4 D^{-2(m_1+m_2)}.
\end{aligned}$$

where (i) follows from Condition (iii), (ii) uses Condition (ii), and (iii) follows from Condition (vi). Since the centered term has mean zero, combining the preceding bounds gives

$$\mathbb{E}\left(\left|\langle X - \Pi_D(X), \beta\rangle_{L_2(\mathcal{T})}\right|^2\right) \le 2C_U^4 D^{-2(m_1+m_2)}.$$

This proves the third inequality. □

Lemma 10. *Assume $X(t) = \sum_{k=1}^{\infty} A_k\varphi_k(t)$ and $\beta(t) = \sum_{k=1}^{\infty} b_k\varphi_k(t)$. Let $0 = t_0 < t_1 < \ldots < t_{M-1} < t_M = 1$ be a grid on $[0,1]$ and define $\widehat{A}_k = \sum_{i=0}^{M} u_i X(t_i)\varphi_k(t_i)$ with $u_0 = t_1/2, u_M = (t_M - t_{M-1})/2$ and $u_i = (t_{i+1} - t_{i-1})/2$ for $i = 1, \ldots, M-1$. Moreover, assume there are constants $C_U > 0$ and $m_1 > 3/2$ such that the following conditions hold:*

- *(i)* $\sup_{k\ge1}\|\varphi_k\|_\infty + \sup_{k\ge1} k^{-1}\|\varphi_k'\|_\infty \le C_U$.
- *(ii) Let ρ_{ij} be the correlation coefficient between A_i and A_j. It satisfies that* $\sup_{i\ge1}\sum_{j=1}^{\infty}|\rho_{ij}| \le C_U$.
- *(iii)* $\mathbb{E}(\|X\|^2_{\mathcal{H}^{m_1}(\mathcal{T})}) \le C_U^2$.
- *(iv) $A_k k^{m_1}$ is C_U-sub-Gaussian.*

Under Conditions (i)-(iv), it holds that

$$\begin{aligned}
\mathbb{E}\left(\left|\widehat{A}_k - A_k\right|^2\right) &\le Ck^2\left(\sum_{i=1}^{M}(t_i - t_{i-1})^2\right)^2, \\
\mathbb{P}\left(|\widehat{A}_k| \ge t\right) &\le 2\exp(-t^2/C), \\
\mathbb{E}\left(\|X - \Pi_D(X)\|^2_{L_2(\mathcal{T})}\right) &\le CD^{-2m_1},
\end{aligned}$$

where $C > 0$ is a constant depending on m_1 and C_U.

Proof. By the norm equivalence in Section 3, we have

$$\sum_{k=1}^{\infty} k^{2m_1} A_k^2 \le c_{m_1}^{-1}\|X\|^2_{\mathcal{H}^{m_1}(\mathcal{T})}.$$

Let $\mu_k = \mathbb{E}(A_k)$ and $\sigma_k^2 = \mathrm{Var}(A_k)$. Taking expectations and using Condition (iii), we obtain

$$\sigma_k^2 + \mu_k^2 = \mathbb{E}(A_k^2) \le k^{-2m_1}\mathbb{E}\left(\sum_{j=1}^{\infty} j^{2m_1}A_j^2\right) \le c_{m_1}^{-1}C_U^2 k^{-2m_1}.$$

Moreover, Jensen's inequality and Tonelli's theorem give

$$\sum_{k=1}^{\infty} k^{2m_1}\mu_k^2 \le \sum_{k=1}^{\infty} k^{2m_1}\mathbb{E}(A_k^2) = \mathbb{E}\left(\sum_{k=1}^{\infty} k^{2m_1}A_k^2\right) \le c_{m_1}^{-1}C_U^2.$$

Thus, Conditions (iii) and (iv) of Lemma 9 hold with C_U replaced by $C_U' = \max\{1, C_U, c_{m_1}^{-1}C_U^2\}$. Together with Conditions (i), (ii), and (iv), these bounds allow us to apply Lemma 9 with C_U' to obtain the first two inequalities. The resulting constants depend only on m_1 and C_U.

For the third inequality, Parseval's identity and the same norm equivalence yield

$$\begin{aligned}\mathbb{E}\left(\|X - \Pi_D(X)\|_{L_2(\mathcal{T})}^2\right) &= \mathbb{E}\left(\sum_{k>D} A_k^2\right) \\ &\le D^{-2m_1}\mathbb{E}\left(\sum_{k>D} k^{2m_1}A_k^2\right) \\ &\le c_{m_1}^{-1}D^{-2m_1}\mathbb{E}\left(\|X\|_{\mathcal{H}^{m_1}(\mathcal{T})}^2\right) \le c_{m_1}^{-1}C_U^2 D^{-2m_1},\end{aligned}$$

where the last inequality follows from Condition (iii). This completes the proof. □

Lemma 11. *Suppose $\beta(t) = \sum_{k=1}^{\infty} b_k\varphi_k(t)$ satisfies that $\|\beta\|_{\mathcal{H}^{m_2}(\mathcal{T})} < \infty$. Then it holds that $b_k^2 \le Ck^{-2m_2}$, where the constant $C > 0$ depends on $\|\beta\|_{\mathcal{H}^{m_2}(\mathcal{T})}$ and m_2.*

Proof. The lower bound in the norm equivalence from Section 3 gives

$$k^{2m_2}b_k^2 \le \sum_{j=1}^{\infty} j^{2m_2}b_j^2 \le c_{m_2}^{-1}\|\beta\|_{\mathcal{H}^{m_2}(\mathcal{T})}^2.$$

Hence, $b_k^2 \le Ck^{-2m_2}$ with $C = 1 + c_{m_2}^{-1}\|\beta\|_{\mathcal{H}^{m_2}(\mathcal{T})}^2$, which depends only on m_2 and $\|\beta\|_{\mathcal{H}^{m_2}(\mathcal{T})}$. This completes the proof. □

Lemma 12. *Let $X_1, \ldots, X_{M-1}$ be i.i.d. random variables on $[0,1]$ with density g satisfying $c_g \le g(x) \le C_g$ for all $x \in [0,1]$ and for some constants $C_g > c_g > 0$. Let $X_{(1)} \le \cdots \le X_{(M-1)}$ be the order statistics, and set $X_{(0)} = 0$ and $X_{(M)} = 1$. Then*

$$\mathbb{E}\left(\sum_{s=1}^{M}\left(X_{(s)} - X_{(s-1)}\right)^2\right) \le \frac{C}{M}, \quad \mathbb{E}\left(\left|\sum_{s=1}^{M}\left(X_{(s)} - X_{(s-1)}\right)^2\right|^2\right) \le \frac{C}{M^2},$$

where the constant $C > 0$ depends only on c_g and C_g.

Proof. Let $\{U_{(i)}\}_{i=1}^{M-1}$ be the order statistics of i.i.d. uniform draws from $[0,1]$. For convention, we also define $U_{(0)} = 0$ and $U_{(M)} = 1$. Let $\Delta_i = U_{(i)} - U_{(i-1)}$ for $i = 1, \ldots, M$, and it is well-known (e.g., see Rao and Sobel, 1980) that

$$(\Delta_1, \ldots, \Delta_M) \sim \text{Dir}(1, \ldots, 1),$$

where Dir standards for the Dirichlet distribution. By the product moments formula for Dirichlet (e.g., see Balakrishnan and Nevzorov, 2003), it follows that

$$\begin{aligned}\mathbb{E}(\Delta_i^2) &= \frac{\Gamma(M)}{\Gamma(M+2)} \times \frac{\Gamma(3)}{\Gamma(1)} = \frac{2}{M(M+1)}, \\ \mathbb{E}(\Delta_i^4) &= \frac{\Gamma(M)}{\Gamma(M+4)} \times \frac{\Gamma(5)}{\Gamma(1)} = \frac{24}{M(M+1)(M+2)(M+3)}, \\ \mathbb{E}(\Delta_i^2\Delta_j^2) &= \frac{\Gamma(M)}{\Gamma(M+4)} \times \frac{\Gamma^2(3)}{\Gamma^2(1)} = \frac{4}{M(M+1)(M+2)(M+3)}, \text{ for } i \ne j.\end{aligned}$$

As a consequence, we have

$$\begin{aligned}\mathbb{E}\left(\sum_{s=1}^{M}\Delta_s^2\right) &= M\mathbb{E}(\Delta_1^2) \le \frac{2}{M},\\ \mathbb{E}\left(\left|\sum_{s=1}^{M}\Delta_s^2\right|^2\right) &= M\mathbb{E}(\Delta_1^4) + M(M-1)\mathbb{E}(\Delta_1^2\Delta_2^2) \le \frac{28}{M^2}.\end{aligned} \tag{1.1}$$

Let G be the cdf associated with g. The density bounds imply that G is continuous and strictly increasing on $[0,1]$. Thus, the variables $U_i = G(X_i)$ are i.i.d. uniform on $[0,1]$, and their ordered spacings satisfy (1.1). Moreover, $U_{(i)} = G(X_{(i)})$ for $i = 0,\dots,M$, including the endpoints. Consequently, for each $i = 1,\dots,M$, we have

$$\Delta_i = G(X_{(i)}) - G(X_{(i-1)}) = \int_{X_{(i-1)}}^{X_{(i)}} g(t)dt.$$

Since $c_g \le g(t) \le C_g$, it follows that

$$\frac{\Delta_i}{C_g} \le X_{(i)} - X_{(i-1)} \le \frac{\Delta_i}{c_g}.$$

Combining the preceding display with (1.1), we complete the proof □

A.2 Proofs of Results in Section 2

Lemma 13. *Suppose that Assumption A1 holds and let $f_* \in \mathcal{F}(S,L,C_{\mathcal{F}})$ be a fixed deterministic function. For any $t > 0$, it follows that*

$$\mathbb{P}\left(\sup_{\substack{f\in\mathcal{F}(S,L,C_{\mathcal{F}})\\ \|f-f_*\|_n\le\delta}}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\left(f(\widehat{\boldsymbol{\xi}}_i) - f_*(\widehat{\boldsymbol{\xi}}_i)\right)\right| > t\right) \le 2\exp\left(-\frac{nt^2}{C\delta^2 SL\log(S)\log(C/\delta)}\right).$$

Here $C > 0$ is a constant depending on C_ϵ and $C_{\mathcal{F}}$. As a consequence, if $S \le n^3$ and $\delta_n \ge 1/n$, then it holds for all $t > 0$ that

$$\mathbb{P}\left(\sup_{\substack{f\in\mathcal{F}(S,L,C_{\mathcal{F}})\\ \|f-f_*\|_n\le\delta_n}}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\left(f(\widehat{\boldsymbol{\xi}}_i) - f_*(\widehat{\boldsymbol{\xi}}_i)\right)\right| > t\right) \le 2\exp\left(-\frac{nt^2}{C\delta_n^2 SL\big(\log(n)\big)^2}\right).$$

Here $C > 0$ is a constant depending on C_ϵ and $C_{\mathcal{F}}$.

Proof. Since $\|f - f_*\|_\infty \le 2C_{\mathcal{F}}$ for all $f \in \mathcal{F}(S,L,C_{\mathcal{F}})$, Lemmas 5 and 7 together imply that

$$\begin{aligned}\int_0^\delta \sqrt{\log(1 + N(x,\mathcal{F}(S,L,C_{\mathcal{F}}),\|\cdot\|_n))}dx &\le \int_0^\delta C\sqrt{SL\log(S)}\sqrt{\log\left(\frac{C}{x}\right)}dx\\ &\le C\delta\sqrt{SL\log(S)\log(C/\delta)} := J(S,L),\end{aligned}$$

where $C > 0$ is a constant depending on C_ϵ and $C_{\mathcal{F}}$. By Assumption A1(iii), T_i is independent of $(\boldsymbol{X}_i, \epsilon_i)$. Together with Assumption A1(i), this gives

$$\mathbb{E}\left(e^{t\epsilon_i} \mid \boldsymbol{X}_i, T_i\right) = \mathbb{E}\left(e^{t\epsilon_i} \mid \boldsymbol{X}_i\right) \le e^{C_\epsilon t^2/2},$$

for each i and all $t \in \mathbb{R}$, almost surely. Since the data pairs are i.i.d. and the sampling grids satisfy Assumption A1(iii), the noise variables are independent conditional on $\{(\boldsymbol{X}_i, T_i)\}_{i=1}^n$. Applying standard chaining for sub-Gaussian processes (e.g., see Lemma 8.1 and Theorem 8.4 in Kosorok, 2008), we obtain

$$\begin{aligned}\mathbb{P}\left(\sup_{\substack{f\in\mathcal{F}(S,L,C_{\mathcal{F}})\\ \|f-f_*\|_n\le\delta}}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\left(f(\widehat{\boldsymbol{\xi}}_i) - f_*(\widehat{\boldsymbol{\xi}}_i)\right)\right| > t\right) &\le 2\exp\left(-\frac{nt^2}{J^2(S,L)}\right)\\ &\le 2\exp\left(-\frac{nt^2}{C\delta^2 SL\log(S)\log(C/\delta)}\right),\end{aligned}$$

for all $t > 0$. Hence, the proof is complete. □

Lemma 14. *Suppose that Assumption A1 holds, and let $f_* \in \mathcal{F}(S, L, C_{\mathcal{F}})$ be a fixed deterministic function. If $S \le n^3$, then with probability at least $1 - C\log(n)/n$, it holds that*

$$\|\widehat{f} - f_*\|_n^2 \le C\left\{\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left(\left|F_0(\boldsymbol{X}_i) - f_*(\widehat{\boldsymbol{\xi}}_i)\right|^2\right) + \frac{SL}{n}\big(\log(n)\big)^3\right\},$$

where $C > 0$ is a constant depending on $C_{\mathcal{F}}$ and C_ϵ.

Proof. Throughout the proof, $C > 0$ denotes a generic constant depending only on $C_{\mathcal{F}}$ and C_ϵ. By the definition of the estimator, we have

$$\frac{1}{n}\sum_{i=1}^n \left(Y_i - \widehat{f}(\widehat{\boldsymbol{\xi}}_i)\right)^2 \le \frac{1}{n}\sum_{i=1}^n \left(Y_i - f_*(\widehat{\boldsymbol{\xi}}_i)\right)^2.$$

By simple algebra and the Cauchy–Schwarz inequality, the preceding display leads to

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n \left(\widehat{f}(\widehat{\boldsymbol{\xi}}_i) - f_*(\widehat{\boldsymbol{\xi}}_i)\right)^2 \\
&\le \frac{2}{n}\sum_{i=1}^n \epsilon_i\left(\widehat{f}(\widehat{\boldsymbol{\xi}}_i) - f_*(\widehat{\boldsymbol{\xi}}_i)\right) + \frac{2}{n}\sum_{i=1}^n \left(F_0(\boldsymbol{X}_i) - f_*(\widehat{\boldsymbol{\xi}}_i)\right)\left(\widehat{f}(\widehat{\boldsymbol{\xi}}_i) - f_*(\widehat{\boldsymbol{\xi}}_i)\right) \\
&\le \frac{2}{n}\sum_{i=1}^n \epsilon_i\left(\widehat{f}(\widehat{\boldsymbol{\xi}}_i) - f_*(\widehat{\boldsymbol{\xi}}_i)\right) + 2\sqrt{\frac{1}{n}\sum_{i=1}^n \left(F_0(\boldsymbol{X}_i) - f_*(\widehat{\boldsymbol{\xi}}_i)\right)^2}\,\|\widehat{f} - f_*\|_n. \qquad (1.2)
\end{aligned}$$

Since $|F_0(\boldsymbol{X}_i)|, |f_*(\widehat{\boldsymbol{\xi}}_i)| \le C_{\mathcal{F}}$, Bernstein's inequality implies that, with probability at least $1 - 1/n$, we have

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n \left(F_0(\boldsymbol{X}_i) - f_*(\widehat{\boldsymbol{\xi}}_i)\right)^2 \\
&\le \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left(\left|F_0(\boldsymbol{X}_i) - f_*(\widehat{\boldsymbol{\xi}}_i)\right|^2\right) + C\sqrt{\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left(\left|F_0(\boldsymbol{X}_i) - f_*(\widehat{\boldsymbol{\xi}}_i)\right|^2\right) \times \frac{\log(n)}{n}} + \frac{C\log(n)}{n} \\
&\le \frac{C}{n}\sum_{i=1}^n \mathbb{E}\left(\left|F_0(\boldsymbol{X}_i) - f_*(\widehat{\boldsymbol{\xi}}_i)\right|^2\right) + \frac{C\log(n)}{n} := a_n^2. \qquad (1.3)
\end{aligned}$$

Here the constant $C \ge 1$ depends only on $C_{\mathcal{F}}$, and a_n is deterministic. Let E_0 be the event that (1.3) holds.

For each $1/n \le r \le 2C_{\mathcal{F}}$, applying the first inequality of Lemma 13 with $S \le n^3$ and $r \ge 1/n$ yields, with probability at least $1 - 2/n$,

$$\sup_{\substack{f \in \mathcal{F}(S,L,C_{\mathcal{F}}) \\ \|f - f_*\|_n \le r}} \left|\frac{1}{n}\sum_{i=1}^n \epsilon_i\left(f(\widehat{\boldsymbol{\xi}}_i) - f_*(\widehat{\boldsymbol{\xi}}_i)\right)\right| \le \frac{C_0}{2}\sqrt{\frac{SL}{n}}\big(\log(n)\big)^{3/2} r,$$

where $C_0 \ge 1$ is a constant depending only on $C_{\mathcal{F}}$ and C_ϵ. Let $E(r)$ denote this event. On $E_0 \cap E(r)$, if $\|\widehat{f} - f_*\|_n \le r$, then (1.2) yields

$$\|\widehat{f} - f_*\|_n^2 \le C_0\sqrt{\frac{SL}{n}}\big(\log(n)\big)^{3/2} r + 2a_n\|\widehat{f} - f_*\|_n. \qquad (1.4)$$

Let us define the critical radius

$$r_* := 2C_0\sqrt{\frac{SL}{n}}\big(\log(n)\big)^{3/2} + 4a_n.$$

If $r_* \ge C_{\mathcal{F}}$, then $\|\widehat{f} - f_*\|_n \le 2C_{\mathcal{F}} \le 2r_*$ deterministically. Hence, assume $r_* < C_{\mathcal{F}}$. Define

$$K = \lceil \log_2(2C_{\mathcal{F}}/r_*) \rceil, \qquad r_k = \min\{2^{k+1} r_*, 2C_{\mathcal{F}}\}, \quad k = 1, \dots, K.$$

Since $r_* \geq n^{-1/2} \geq 1/n$, we have $1/n \leq r_k \leq 2C_{\mathcal{F}}$. If $2^k r_* \leq \|\widehat{f} - f_*\|_n \leq 2^{k+1} r_*$, then $\|\widehat{f} - f_*\|_n \leq r_k$. Thus, on $E_0 \cap E(r_k)$, (1.4) implies that

$$\|\widehat{f} - f_*\|_n^2 \leq C_0 \sqrt{\frac{SL}{n}} \big(\log(n)\big)^{3/2} 2^{k+1} r_* + 2a_n \|\widehat{f} - f_*\|_n.$$

The preceding display further leads to

$$\begin{aligned} r_* &\leq C_0 \sqrt{\frac{SL}{n}} \big(\log(n)\big)^{3/2} 2^{1-k} + a_n 2^{2-k} \\ &\leq C_0 \sqrt{\frac{SL}{n}} \big(\log(n)\big)^{3/2} + 2a_n = \frac{r_*}{2}, \end{aligned}$$

which is impossible since $r_* > 0$. Hence, for each $k = 1, \ldots, K$, we deduce that

$$\mathbb{P}\left(2^k r_* \leq \|\widehat{f} - f_*\|_n \leq 2^{k+1} r_*, E_0\right) \leq \mathbb{P}\left(E(r_k)^c\right) \leq \frac{2}{n}.$$

Since $\|\widehat{f} - f_*\|_n \leq 2C_{\mathcal{F}}$, we have

$$\left\{\|\widehat{f} - f_*\|_n \geq 2r_*\right\} \subset \bigcup_{k=1}^{K} \left\{2^k r_* \leq \|\widehat{f} - f_*\|_n \leq 2^{k+1} r_*\right\}.$$

As a consequence, it follows that

$$\begin{aligned} \mathbb{P}\left(\|\widehat{f} - f_*\|_n \geq 2r_*\right) &\leq \mathbb{P}\left(\|\widehat{f} - f_*\|_n \geq 2r_*, E_0\right) + \mathbb{P}(E_0^c) \\ &\leq \frac{2K+1}{n} \leq \frac{C\log(n)}{n}, \end{aligned}$$

where the last inequality holds since $r_* \geq n^{-1/2}$ implies $K \leq C\log(n)$. Finally, by the definitions of r_* and a_n, with probability at least $1 - C\log(n)/n$,

$$\|\widehat{f} - f_*\|_n^2 \leq 4r_*^2 \leq C\left\{\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left(\left|F_0(\boldsymbol{X}_i) - f_*(\widehat{\boldsymbol{\xi}}_i)\right|^2\right) + \frac{SL}{n}\big(\log(n)\big)^3\right\}.$$

Hence, the proof is complete. □

Lemma 15. *Suppose that Assumption A1 holds, and let $f_* \in \mathcal{F}(S, L, C_{\mathcal{F}})$ be a fixed deterministic function. If $S \leq n^3$, then with probability at least $1 - C\log(n)/n$, it holds that*

$$\|\widehat{f} - f_*\|^2 \leq C\left\{\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left(\left|F_0(\boldsymbol{X}_i) - f_*(\widehat{\boldsymbol{\xi}}_i)\right|^2\right) + \frac{SL}{n}\big(\log(n)\big)^3\right\},$$

where $C > 0$ is a constant depending on $C_{\mathcal{F}}$ and C_ϵ.

Proof. Throughout the proof, $C > 0$ denotes a generic constant depending only on $C_{\mathcal{F}}$ and C_ϵ. By Lemma 14, we may choose a sufficiently large constant $C_0 \geq 1$ such that, with probability at least $1 - C_0\log(n)/n$,

$$\|\widehat{f} - f_*\|_n^2 \leq C_0\left\{\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left(\left|F_0(\boldsymbol{X}_i) - f_*(\widehat{\boldsymbol{\xi}}_i)\right|^2\right) + \frac{SL}{n}\big(\log(n)\big)^3\right\} := r_*^2. \tag{1.5}$$

Let E_0 be the event that (1.5) holds. Since $\|\widehat{f} - f_*\| \leq 2C_{\mathcal{F}}$, the conclusion follows directly if $4r_* \geq 2C_{\mathcal{F}}$. Hence, assume $4r_* < 2C_{\mathcal{F}}$.

For $r > 0$, define the fixed population classes

$$\mathcal{G}(r) = \{f - f_* : f \in \mathcal{F}(S, L, C_{\mathcal{F}}),\ \|f - f_*\| \leq r\}, \qquad \mathcal{G}^2(r) = \{g^2 : g \in \mathcal{G}(r)\}.$$

For every $g \in \mathcal{G}(r)$, we have $\|g\|_\infty \leq 2C_{\mathcal{F}}$ and

$$\frac{1}{n}\sum_{i=1}^{n} \mathrm{Var}\left(g^2(\widehat{\boldsymbol{\xi}}_i)\right) \leq \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left(g^4(\widehat{\boldsymbol{\xi}}_i)\right) \leq 4C_{\mathcal{F}}^2 \|g\|^2 \leq 4C_{\mathcal{F}}^2 r^2.$$

Moreover, since $r_* \geq \sqrt{SL/n}\log(n)$, Lemmas 2 and 6 imply that

$$\mathbb{E}\left(\mathrm{Rad}_n\left(\mathcal{G}^2(2^{k+1}r_*)\right)\right) \leq C2^{k+1}r_*\sqrt{\frac{SL}{n}}\log(n), \qquad k \geq 2.$$

Applying Lemma 3 to the fixed class $\mathcal{G}^2(2^{k+1}r_*)$ with $\eta = \log(n)$, for each $k \geq 2$ there is an event E_k with $\mathbb{P}(E_k) \geq 1 - 2/n$ on which

$$\sup_{g \in \mathcal{G}(2^{k+1}r_*)} \left|\|g\|^2 - \|g\|_n^2\right| \leq C2^{k+1}r_*\sqrt{\frac{SL}{n}}\log(n) + C2^{k+1}r_*\sqrt{\frac{\log(n)}{n}} + \frac{C\log(n)}{n}. \quad (1.6)$$

Suppose that $2^k r_* \leq \|\widehat{f} - f_*\| \leq 2^{k+1}r_*$ for some $k \geq 2$. On $E_0 \cap E_k$, the preceding display gives

$$\begin{aligned}
4^k r_*^2 \leq \|\widehat{f} - f_*\|^2 &\leq \|\widehat{f} - f_*\|_n^2 + C2^{k+1}r_*\sqrt{\frac{SL}{n}}\log(n) + C2^{k+1}r_*\sqrt{\frac{\log(n)}{n}} + \frac{C\log(n)}{n} \\
&\overset{\text{(i)}}{\leq} r_*^2 + \frac{1}{2}4^k r_*^2 + \frac{CSL}{n}\big(\log(n)\big)^2 + \frac{C\log(n)}{n} \\
&\overset{\text{(ii)}}{\leq} 2r_*^2 + \frac{1}{2}4^k r_*^2,
\end{aligned}$$

where (i) follows from (1.5) and Young's inequality, and (ii) holds by taking C_0 sufficiently large, since $SL \geq 1$. The preceding display implies $4^k \leq 4$, which contradicts $k \geq 2$. Thus,

$$\mathbb{P}\left(2^k r_* \leq \|\widehat{f} - f_*\| \leq 2^{k+1}r_*,\ E_0\right) \leq \mathbb{P}(E_k^c) \leq \frac{2}{n}, \qquad k \geq 2.$$

Since $\|\widehat{f} - f_*\| \leq 2C_{\mathcal{F}}$, we have

$$\{\|\widehat{f} - f_*\| \geq 4r_*\} \subset \bigcup_{k=2}^{K}\{2^k r_* \leq \|\widehat{f} - f_*\| \leq 2^{k+1}r_*\},$$

where $K = \lceil \log_2(2C_{\mathcal{F}}/r_*)\rceil$. Therefore,

$$\begin{aligned}
\mathbb{P}\left(\|\widehat{f} - f_*\| \geq 4r_*\right) &\leq \mathbb{P}(E_0^c) + \sum_{k=2}^{K}\mathbb{P}\left(2^k r_* \leq \|\widehat{f} - f_*\| \leq 2^{k+1}r_*,\ E_0\right) \\
&\leq \frac{C_0\log(n)}{n} + \frac{2(K-1)}{n} \leq \frac{C\log(n)}{n},
\end{aligned}$$

where the last inequality follows from $r_* \geq \sqrt{C_0/n}$ and hence $K \leq C\log(n)$. Consequently, $\|\widehat{f} - f_*\|^2 \leq 16r_*^2$ with the stated probability. Substituting the definition of r_* completes the proof. □

Proof of Theorem 1. Define

$$f_* = \operatorname*{arginf}_{f \in \mathcal{F}(S,L,C_{\mathcal{F}})} \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left(\left|f_D(\widehat{\boldsymbol{\xi}}_i) - f(\widehat{\boldsymbol{\xi}}_i)\right|^2\right),$$

which is a deterministic function. Using the basic inequality $(a+b)^2 \leq 2a^2 + 2b^2$, we have

$$\begin{aligned}
\mathbb{E}\left(\left|F_0(\boldsymbol{X}_i^*) - f_*(\widehat{\boldsymbol{\xi}}_i^*)\right|^2\right) &\overset{\text{(i)}}{=} \mathbb{E}\left(\left|F_0(\boldsymbol{X}_i) - f_*(\widehat{\boldsymbol{\xi}}_i)\right|^2\right) \\
&\leq 4\mathbb{E}\left(|F_0(\boldsymbol{X}_i) - f_D(\boldsymbol{\xi}_i)|^2\right) + 4\mathbb{E}\left(\left|f_D(\boldsymbol{\xi}_i) - f_D(\widehat{\boldsymbol{\xi}}_i)\right|^2\right) \\
&\quad + 4\mathbb{E}\left(\left|f_D(\widehat{\boldsymbol{\xi}}_i) - f_*(\widehat{\boldsymbol{\xi}}_i)\right|^2\right), \quad (1.7)
\end{aligned}$$

where (i) holds since $(\widehat{\boldsymbol{\xi}}_i^*, \boldsymbol{X}_i^*)$ has the same distribution as $(\widehat{\boldsymbol{\xi}}_i, \boldsymbol{X}_i)$. Similarly, we have

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}}\left(\left|\widehat{f}(\widehat{\boldsymbol{\xi}}_i^*) - F_0(\boldsymbol{X}_i^*)\right|^2\right) &\le 2\mathbb{E}_{\mathcal{D}}\left(\left|\widehat{f}(\widehat{\boldsymbol{\xi}}_i^*) - f_*(\widehat{\boldsymbol{\xi}}_i^*)\right|^2\right) + 2\mathbb{E}_{\mathcal{D}}\left(\left|f_*(\widehat{\boldsymbol{\xi}}_i^*) - F_0(\boldsymbol{X}_i^*)\right|^2\right) \\
&\overset{\text{(i)}}{=} 2\mathbb{E}_{\mathcal{D}}\left(\left|\widehat{f}(\widehat{\boldsymbol{\xi}}_i^*) - f_*(\widehat{\boldsymbol{\xi}}_i^*)\right|^2\right) + 2\mathbb{E}\left(\left|f_*(\widehat{\boldsymbol{\xi}}_i^*) - F_0(\boldsymbol{X}_i^*)\right|^2\right) \\
&\overset{\text{(ii)}}{\le} 2\mathbb{E}_{\mathcal{D}}\left(\left|\widehat{f}(\widehat{\boldsymbol{\xi}}_i^*) - f_*(\widehat{\boldsymbol{\xi}}_i^*)\right|^2\right) + 8\mathbb{E}\left(|F_0(\boldsymbol{X}_i) - f_D(\boldsymbol{\xi}_i)|^2\right) \\
&\quad + 8\mathbb{E}\left(\left|f_D(\boldsymbol{\xi}_i) - f_D(\widehat{\boldsymbol{\xi}}_i)\right|^2\right) + 8\mathbb{E}\left(\left|f_D(\widehat{\boldsymbol{\xi}}_i) - f_*(\widehat{\boldsymbol{\xi}}_i)\right|^2\right),
\end{aligned}$$

where (i) holds since f_* and F_0 are deterministic, and (ii) is due to (1.7). Taking average over i and using the definition of $\|\cdot\|$ and $\mathcal{E}(\widehat{f})$, we have

$$\begin{aligned}
\mathcal{E}(\widehat{f}) \le 2\|\widehat{f} - f_*\|^2 &+ \frac{8}{n}\sum_{i=1}^n \mathbb{E}\left(|F_0(\boldsymbol{X}_i) - f_D(\boldsymbol{\xi}_i)|^2\right) + \frac{8}{n}\sum_{i=1}^n \mathbb{E}\left(\left|f_D(\boldsymbol{\xi}_i) - f_D(\widehat{\boldsymbol{\xi}}_i)\right|^2\right) \\
&+ \frac{8}{n}\sum_{i=1}^n \mathbb{E}\left(\left|f_D(\widehat{\boldsymbol{\xi}}_i) - f_*(\widehat{\boldsymbol{\xi}}_i)\right|^2\right).
\end{aligned}$$

Finally, using Lemma 15 and (1.7) again, the preceding display leads to the desired result. □

A.3 Proofs of Results in Section 3

Proof of Theorem 2. Throughout the proof, we use $C > 0$ to denote a generic constant that is independent of n, W, D, L, and M_i. We divide the proof into five steps.

Step 1: For a neural network $f \in \mathcal{F}_{\text{EW}}(W, L, C_{\mathcal{F}}) : \mathbb{R}^{qD} \to \mathbb{R}$, it has $S = W(qD+1) + (L-1)W(W+1) + (W+1)$ parameters. Therefore, by the rate conditions given, we have

$$S \le C(qDW + W^2L) \le CW^2L \le C9^{qD}(qD + N_1 + 1)^2(N_2 + qD),$$

which further implies that

$$SL \le C9^{qD}(qD + N_1 + 1)^2(N_2 + qD)^2. \tag{1.8}$$

Step 2: By the definition f_D, we have

$$f_D(\boldsymbol{\xi}) = F_0\left(\Pi_D(X_1), \ldots, \Pi_D(X_q)\right).$$

Using Assumption B1, we have

$$|F_0(\boldsymbol{X}) - f_D(\boldsymbol{\xi})|^2 \le C_\gamma^2\left(\sum_{j=1}^q \|X_j - \Pi_D(X_j)\|_{L_2(\mathcal{T})}^2\right)^\gamma.$$

Taking expectation and using Assumptions A2, A3, and B2, we have

$$\begin{aligned}
\mathcal{E}_{\text{tr}} &= \mathbb{E}\left(|F_0(\boldsymbol{X}) - f_D(\boldsymbol{\xi})|^2\right) \\
&\le C_\gamma^2\mathbb{E}\left\{\left(\sum_{j=1}^q \|X_j - \Pi_D(X_j)\|_{L_2(\mathcal{T})}^2\right)^\gamma\right\} \\
&\overset{\text{(i)}}{\le} C_\gamma^2\mathbb{E}^\gamma\left\{\sum_{j=1}^q \|X_j - \Pi_D(X_j)\|_{L_2(\mathcal{T})}^2\right\} \overset{\text{(ii)}}{\le} CD^{-2m_1\gamma},
\end{aligned} \tag{1.9}$$

where (i) holds since $\mathbb{E}(|X|^\gamma) \le \mathbb{E}^\gamma(|X|)$ when $0 < \gamma \le 1$, and (ii) uses Lemma 10.

Step 3: By the definition f_D, we have

$$\begin{aligned}
f_D(\boldsymbol{\xi}_i) &= F_0\left(\Pi_D(X_{i1}), \ldots, \Pi_D(X_{iq})\right), \\
f_D(\widehat{\boldsymbol{\xi}}_i) &= F_0\left(\Pi_D(\widehat{X}_{i1}), \ldots, \Pi_D(\widehat{X}_{iq})\right),
\end{aligned}$$

where $\Pi_D(\widehat{X}_{ij}) = \sum_{k\le D}\widehat{\xi}_{ijk}\varphi_k$. Hence, it follows that

$$\begin{aligned}\left|f_D(\boldsymbol{\xi}_i) - f_D(\widehat{\boldsymbol{\xi}}_i)\right|^2 &= \left|F_0\left(\Pi_D(X_{i1}),\ldots,\Pi_D(X_{iq})\right) - F_0\left(\Pi_D(\widehat{X}_{i1}),\ldots,\Pi_D(\widehat{X}_{iq})\right)\right|^2 \\ &\overset{\text{(i)}}{\le} C_\gamma^2\left(\sum_{j=1}^q\left\|\Pi_D(X_{ij}) - \Pi_D(\widehat{X}_{ij})\right\|_{L_2(\mathcal{T})}^2\right)^\gamma \\ &= C_\gamma^2\left(\sum_{j=1}^q\sum_{k\le D}|\xi_{ijk} - \widehat{\xi}_{ijk}|^2\right)^\gamma.\end{aligned}$$

where (i) follows from Assumption B1. By Assumptions A2, A3, A4, and B2, we conclude that

$$\mathbb{E}\left(|\xi_{ijk} - \widehat{\xi}_{ijk}|^2\right) \le \frac{Ck^2}{M_i^2}, \text{ for all } i,j,k.$$

Here Assumption A1(iii) allows us to apply Lemma 10 conditionally on T_i; for the random design, we then average over T_i using Lemma 12. As a consequence, we have

$$\begin{aligned}\mathbb{E}\left(\left|f_D(\boldsymbol{\xi}_i) - f_D(\widehat{\boldsymbol{\xi}}_i)\right|^2\right) &\le C_\gamma^2\mathbb{E}\left\{\left(\sum_{j=1}^q\sum_{k\le D}|\xi_{ijk} - \widehat{\xi}_{ijk}|^2\right)^\gamma\right\} \\ &\le C_\gamma^2\mathbb{E}^\gamma\left\{\sum_{j=1}^q\sum_{k\le D}|\xi_{ijk} - \widehat{\xi}_{ijk}|^2\right\} \\ &\le C\left(\sum_{j=1}^q\sum_{k\le D}\frac{k^2}{M_i^2}\right)^\gamma \le CD^{3\gamma}M_i^{-2\gamma},\end{aligned}$$

which further implies that

$$\mathcal{E}_{\text{dis}} = \frac{1}{n}\sum_{i=1}^n\mathbb{E}\left(\left|f_D(\boldsymbol{\xi}_i) - f_D(\widehat{\boldsymbol{\xi}}_i)\right|^2\right) \le CD^{3\gamma}\frac{1}{n}\sum_{i=1}^n M_i^{-2\gamma}. \tag{1.10}$$

Step 4: Define

$$\overline{f}_D(\boldsymbol{v}) = \max\left\{-\frac{C_\mathcal{F}}{2}, \min\left\{f_D(\boldsymbol{v}), \frac{C_\mathcal{F}}{2}\right\}\right\},$$

which is obtained by clipping f_D to $[-C_\mathcal{F}/2, C_\mathcal{F}/2]$. It follows that

$$\begin{aligned}\left|f_D(\widehat{\boldsymbol{\xi}}_i) - \overline{f}_D(\widehat{\boldsymbol{\xi}}_i)\right| &\overset{\text{(i)}}{=} \left(\left|f_D(\widehat{\boldsymbol{\xi}}_i)\right| - \frac{C_\mathcal{F}}{2}\right)_+ \\ &\overset{\text{(ii)}}{\le} \left(\left|f_D(\widehat{\boldsymbol{\xi}}_i)\right| - |F_0(\boldsymbol{X}_i)|\right)_+ \\ &\overset{\text{(iii)}}{\le} \left|f_D(\widehat{\boldsymbol{\xi}}_i) - F_0(\boldsymbol{X}_i)\right| \le \left|f_D(\widehat{\boldsymbol{\xi}}_i) - f_D(\boldsymbol{\xi}_i)\right| + |f_D(\boldsymbol{\xi}_i) - F_0(\boldsymbol{X}_i)|,\end{aligned} \tag{1.11}$$

where (i) is the clipping identity, (ii) follows from $|F_0(\boldsymbol{X}_i)| \le C_\mathcal{F}/2$ under Assumption A1(ii), and (iii) follows from the reverse triangle inequality.

Moreover, Lemma 9, together with Assumptions A2, A3, A4, and B2, yields

$$\mathbb{P}\left(\left|\widehat{\xi}_{ijk}\right| \le R, \text{ for all } j = 1,\ldots,q, k = 1,\ldots,D,\right) \ge 1 - 2qDe^{-R^2/C}, \text{ for all } R > 0.$$

According to Lemma 18, there exists a DNN $\phi \in \mathcal{F}_{\text{EW}}(W, L, C_\mathcal{F})$ with $W = 3^{qD+3}(qD + N_1 + 1)$ and $L = 12N_2 + 15 + 2qD$ such that

$$\begin{aligned}\mathbb{E}\left(\left|\overline{f}_D(\widehat{\boldsymbol{\xi}}_i) - \phi(\widehat{\boldsymbol{\xi}}_i)\right|^2\right) &\le \mathbb{E}\left(\left|\overline{f}_D(\widehat{\boldsymbol{\xi}}_i) - \phi(\widehat{\boldsymbol{\xi}}_i)\right|^2 I\left(\widehat{\boldsymbol{\xi}}_i \in [-R,R]^{qD}\right)\right) + 4C_\mathcal{F}^2\mathbb{P}\left(\widehat{\boldsymbol{\xi}}_i \notin [-R,R]^{qD}\right) \\ &\overset{\text{(i)}}{\le} CqDR^{2\gamma}(N_1N_2)^{-\frac{4\gamma}{qD}} + CqDe^{-\frac{R^2}{C}},\end{aligned}$$

where (i) uses preceding display and the approximation property of ϕ in Lemma 18. Setting $R = C\log(n)$ in the preceding display, we have

$$\begin{aligned}
\mathcal{E}_{\text{app}} &\le \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left(\left|f_D(\widehat{\boldsymbol{\xi}}_i) - \phi(\widehat{\boldsymbol{\xi}}_i)\right|^2\right) \\
&\le \frac{2}{n}\sum_{i=1}^{n}\mathbb{E}\left(\left|f_D(\widehat{\boldsymbol{\xi}}_i) - \overline{f}_D(\widehat{\boldsymbol{\xi}}_i)\right|^2\right) + \frac{2}{n}\sum_{i=1}^{n}\mathbb{E}\left(\left|\overline{f}_D(\widehat{\boldsymbol{\xi}}_i) - \phi(\widehat{\boldsymbol{\xi}}_i)\right|^2\right) \\
&\overset{\text{(i)}}{\le} 4\mathcal{E}_{\text{dis}} + 4\mathcal{E}_{\text{tr}} + C(\log(n))^2 qD(N_1N_2)^{-\frac{4\gamma}{qD}} + \frac{CqD}{n}, \qquad (1.12)
\end{aligned}$$

where (i) follows by squaring and averaging (1.11).

Combining Theorem 1 with (1.8)-(1.12), we complete the proof of the first statement.

Step 5: When $M_i = \infty$, the discretization error vanishes. Combining the bounds obtained in the preceding steps gives

$$\begin{aligned}
\mathcal{E}\left(\widehat{f}_{\text{EW}}\right) &\le C\left\{\frac{9^{qD}(qD+N_1+1)^2(N_2+qD)^2}{n}\big(\log(n)\big)^3 + D^{-2m_1\gamma} + \big(\log(n)\big)^2 D(N_1N_2)^{-\frac{4\gamma}{qD}}\right\} \\
&:= C(I_1 + I_2 + I_3).
\end{aligned}$$

Since $N_2 \gtrsim D$ and $N_1 \ge 1$, we have $N_2 + qD \le CN_2$ and $qD + N_1 + 1 \le CDN_1$. Under the stated choice of N_1N_2, it follows that

$$\begin{aligned}
I_1 &\overset{\text{(i)}}{\le} C\frac{9^{qD}D^2\big(\log(n)\big)^3}{n}\left(\frac{nD}{9^{qD}\log(n)}\right)^{\frac{qD}{\gamma+qD}} \overset{\text{(ii)}}{\le} Cn^{-\frac{\gamma}{\gamma+qD}}\big(\log(n)\big)^5, \\
I_3 &\overset{\text{(i)}}{\le} CD\big(\log(n)\big)^2\left(\frac{nD}{9^{qD}\log(n)}\right)^{-\frac{2\gamma}{\gamma+qD}} \overset{\text{(ii)}}{\le} Cn^{-\frac{2\gamma}{\gamma+qD}}\big(\log(n)\big)^3,
\end{aligned}$$

where (i) uses the prescribed choice of N_1N_2 and the preceding bounds, and (ii) follows from $D \le \log(n)$, $qD \ge \gamma$, and the bounds $9^{\gamma qD/(\gamma+qD)} \le 9^{\gamma}$ and $9^{2\gamma qD/(\gamma+qD)} \le 9^{2\gamma}$. Consequently, we have

$$\begin{aligned}
I_1 + I_3 &\overset{\text{(i)}}{\le} Cn^{-\frac{\gamma}{2qD}}\big(\log(n)\big)^5 \\
&= C\exp\left(-\frac{\gamma\log(n)}{2qD} + 5\log\log(n)\right) \overset{\text{(ii)}}{\le} C\big(\log(n)\big)^{-2m_1\gamma},
\end{aligned}$$

where (i) uses $\gamma + qD \le 2qD$, and (ii) follows from $D \asymp \log(n)/(\log\log(n))^2$. Using the choice of D, we have Finally,

$$I_2 = D^{-2m_1\gamma} \asymp \big(\log(n)\big)^{-2m_1\gamma}\big(\log\log(n)\big)^{4m_1\gamma}.$$

Combining these estimates establishes the second statement and completes the proof. □

Theorem 4. *For fixed $R > 0$, define*

$$\begin{aligned}
\mathcal{M}(R) = \Bigg\{ F : \prod_{j=1}^{q} L_2([0,1]) \to \mathbb{R} \;\Bigg|\; &\left|F(\boldsymbol{X}) - F(\widetilde{\boldsymbol{X}})\right| \le R\left(\sum_{j=1}^{q}\|X_j - \widetilde{X}_j\|^2_{L_2(\mathcal{T})}\right)^{\gamma/2}, \\
&|F(\boldsymbol{X})| \le R \text{ for all } \boldsymbol{X}, \widetilde{\boldsymbol{X}} \in \prod_{j=1}^{q} L_2([0,1])\Bigg\},
\end{aligned}$$

$$\begin{aligned}
\mathcal{Q}_1(R) = \Bigg\{ Q \;\Bigg|\; &(\boldsymbol{X},\epsilon) \sim Q, \quad \mathbb{E}_Q(\epsilon \mid \boldsymbol{X}) = 0, \quad \mathbb{E}_Q(\epsilon^2 \mid \boldsymbol{X}) \le 1, \\
&\max_{1\le j\le q}\mathbb{E}_Q\left(\|X_j\|^2_{\mathcal{H}^{m_1}(\mathcal{T})}\right) \le R^2\Bigg\}.
\end{aligned}$$

Let $F \in \mathcal{M}(R)$ and $(\boldsymbol{X}, \epsilon) \sim Q \in \mathcal{Q}_1(R)$. Assume that $\{(Y_i, \boldsymbol{X}_i)\}_{i=1}^n$ are i.i.d. observations from the model $Y = F(\boldsymbol{X}) + \epsilon$. Then it holds that

$$\liminf_{n\to\infty} \inf_{\widehat{F}} \sup_{\substack{F\in\mathcal{M}(R)\\ Q\in\mathcal{Q}_1(R)}} \mathbb{E}_{F,Q}\Big\{\mathcal{E}_Q(\widehat{F}, F)\Big\}\big(\log(n)\big)^{2m_1\gamma} > 0.$$

Here $\mathbb{E}_{F,Q}$ denotes expectation over the training sample generated under F and Q, and the infimum is taken over all estimators based on these data.

Proof. Let $\epsilon \sim N(0,1)$ and for $b > 0$, define

$$\begin{aligned}
E_b &= \Big\{X = \sum_{k=1}^{\infty} a_k\varphi_k : \sum_{k=1}^{\infty} a_k^2 k^{2m_1} \le b^2\Big\},\\
\otimes_{i=1}^q E_b &= \left\{\boldsymbol{X} = (X_1, \ldots, X_q)^\top : X_i \in E_b, \text{ for } i = 1, \ldots, q\right\},\\
d(\boldsymbol{X}, \widetilde{\boldsymbol{X}}) &= \left(\sum_{j=1}^q \|X_j - \widetilde{X}_j\|_{L_2(\mathcal{T})}^2\right)^{1/2}, \text{ for } \boldsymbol{X}, \widetilde{\boldsymbol{X}} \in \otimes_{i=1}^q E_b.
\end{aligned}$$

We can choose b sufficiently small so that $\|X\|_{\mathcal{H}^{m_1}(\mathcal{T})} \le R$ for every $X \in E_b$. Therefore, we have

$$(\boldsymbol{X}, \epsilon) \sim Q \in \mathcal{Q}_1(R), \quad \text{for all } X \in \otimes_{i=1}^q E_b. \tag{1.13}$$

Corollary 16 of Allard and Bölcskei (2026) gives $\log N(\delta, E_b, \|\cdot\|_{L_2}) \asymp \delta^{-1/m_1}$ as $\delta \downarrow 0$. Since q is fixed, restricting to one coordinate gives the lower bound for $\log N(\delta, \otimes_{i=1}^q E_b, d)$, while product covers with coordinate radii $\delta/\sqrt{q}$ give the upper bound. Hence, it follows that

$$\log N\left(\delta, \otimes_{i=1}^q E_b, d\right) \asymp \delta^{-1/m_1}.$$

In particular, the design $\{Q : (\boldsymbol{X}, \epsilon) \sim Q, \boldsymbol{X} \in \otimes_{i=1}^q E_b, \epsilon \sim N(0,1)\}$ satisfies the conditions of Theorem 3.2(a) of Meister (2016). According to this theorem and (1.13), there exists a sequence of distributions $Q_n \in \mathcal{Q}_1(R)$ such that

$$\liminf_{n\to\infty} \inf_{\widehat{F}} \sup_{F\in\mathcal{M}(R)} \mathbb{E}_{F,Q_n}\Big\{\mathcal{E}_{Q_n}(\widehat{F}, F)\Big\}\big(\log(n)\big)^{2m_1\gamma} \ge c,$$

where $c > 0$ is a constant free of n. Taking the supremum over $Q \in \mathcal{Q}_1(R)$, we complete the proof. □

A.4 Proofs of Results in Section 4

Proof of Theorem 3. Using Theorem 1, we divide the proof into four steps.

Step 1: For a deep neural network $f \in \mathcal{F}_{\mathrm{BEW}}(W_1, W, L, C_{\mathcal{F}}) : \mathbb{R}^{qD} \to \mathbb{R}$, it has $S = qDW_1 + W_1(W+1) + (L-2)W^2 + LW + 1$ parameters. Therefore, we have

$$S \le C(DW_1 + WW_1 + W^2L), \tag{1.14}$$

for some constant $C > 0$ free of n, W, D, L and M_i.

Step 2: By definition of the model, we have

$$\begin{aligned}
F_0(\boldsymbol{X}) &= h_0\left(\langle X_1, \beta_{11}^0\rangle_{L_2(\mathcal{T})}, \ldots, \langle X_q, \beta_{qd_q}^0\rangle_{L_2(\mathcal{T})}\right),\\
f_D(\boldsymbol{\xi}) &= h_0\left(\langle \Pi_D(X_1), \beta_{11}^0\rangle_{L_2(\mathcal{T})}, \ldots, \langle \Pi_D(X_q), \beta_{qd_q}^0\rangle_{L_2(\mathcal{T})}\right).
\end{aligned}$$

Using Assumption C1, we have

$$|F_0(\boldsymbol{X}) - f_D(\boldsymbol{\xi})|^2 \le C\sum_{j=1}^q \sum_{k=1}^{d_j} \left|\langle X_j - \Pi_D(X_j), \beta_{jk}^0\rangle_{L_2(\mathcal{T})}\right|^2,$$

where $C > 0$ is a Lipschitz constant. By Assumptions A2, A3, C3, and C2 and using Lemma 9, the preceding display leads to

$$\mathcal{E}_{\text{tr}} = \mathbb{E}\left(|F_0(\boldsymbol{X}) - f_D(\boldsymbol{\xi})|^2\right) \leq CD^{-2(m_1+m_2)}, \tag{1.15}$$

where $C > 0$ is a universal constant free of n, W, D, L and M_i.

Step 3: By the definition f_D, we have

$$f_D(\boldsymbol{\xi}_i) = h_0\left(\langle \Pi_D(X_{i1}), \beta_{11}^0\rangle_{L_2(\mathcal{T})}, \ldots, \langle \Pi_D(X_{iq}), \beta_{qd_q}^0\rangle_{L_2(\mathcal{T})}\right),$$
$$f_D(\widehat{\boldsymbol{\xi}}_i) = h_0\left(\langle \Pi_D(\widehat{X}_{i1}), \beta_{11}^0\rangle_{L_2(\mathcal{T})}, \ldots, \langle \Pi_D(\widehat{X}_{iq}), \beta_{qd_q}^0\rangle_{L_2(\mathcal{T})}\right),$$

where $\Pi_D(\widehat{X}_{ij}) = \sum_{k\leq D} \widehat{\xi}_{ijk}\varphi_k$. Hence, it follows that

$$\begin{aligned}\left|f_D(\boldsymbol{\xi}_i) - f_D(\widehat{\boldsymbol{\xi}}_i)\right|^2 &\overset{\text{(i)}}{\leq} C\sum_{j=1}^{q}\sum_{s=1}^{d_j}\left|\langle \Pi_D(X_{ij}) - \Pi_D(\widehat{X}_{ij}), \beta_{js}^0\rangle_{L_2(\mathcal{T})}\right|^2 \\ &\overset{\text{(ii)}}{\leq} C\sum_{j=1}^{q}\sum_{s=1}^{d_j}\left(\sum_{k\leq D}|\xi_{ijk} - \widehat{\xi}_{ijk}|k^{-m_2}\right)^2.\end{aligned}$$

where (i) follows from Assumption C1, and (ii) is due to Assumption C3 and Lemma 11. Taking expectation and using Assumptions A2, A3, and C2, we have

$$\mathbb{E}(|\xi_{ijk} - \widehat{\xi}_{ijk}|^2) \leq \frac{Ck^2}{M_i^2}, \text{ for all } i, j, k.$$

Here Assumption A1(iii) allows us to apply Lemma 9 conditionally on T_i; for the random design, we then average over T_i using Lemma 12. As a consequence, we have

$$\begin{aligned}\mathbb{E}\left(\left|f_D(\boldsymbol{\xi}_i) - f_D(\widehat{\boldsymbol{\xi}}_i)\right|^2\right) &\leq C\sum_{j=1}^{q}\sum_{s=1}^{d_j}\mathbb{E}\left\{\left(\sum_{k\leq D}|\xi_{ijk} - \widehat{\xi}_{ijk}|k^{-m_2}\right)^2\right\} \\ &\overset{\text{(i)}}{\leq} C\sum_{j=1}^{q}\sum_{s=1}^{d_j}\left(\sum_{k\leq D}\sqrt{\mathbb{E}\left(|\xi_{ijk} - \widehat{\xi}_{ijk}|^2\right)}\,k^{-m_2}\right)^2 \\ &\leq CM_i^{-2}\left(\sum_{k\leq D}k^{1-m_2}\right)^2.\end{aligned}$$

where $C > 0$ is a universal constant free of n, W, D, L and M_i, and (i) is due to Minkowski's inequality. Therefore, it holds that

$$\mathcal{E}_{\text{dis}} \leq C\left(\sum_{k\leq D}k^{1-m_2}\right)^2\frac{1}{n}\sum_{i=1}^{n}M_i^{-2}. \tag{1.16}$$

where $C > 0$ is a constant free of n, W, D, L and M_i.

Step 4: By Lemma 9 and Assumptions A2, A3, A4, and C2, we have

$$\mathbb{P}\left(\left|\widehat{\xi}_{ijk}\right| \leq R, \text{ for all } j = 1, \ldots, q, k = 1, \ldots, D, \right) \geq 1 - 2qDe^{-R^2/C}, \text{ for all } R > 0.$$

where $C > 0$ is a constant free of n, W, D, L and M_i. Therefore, Lemma 21 implies that

$$\begin{aligned}\mathbb{E}\left(\left|f_D(\widehat{\boldsymbol{\xi}}_i) - f(\widehat{\boldsymbol{\xi}}_i)\right|^2\right) &\leq \mathbb{E}\left(\left|f_D(\widehat{\boldsymbol{\xi}}_i) - f(\widehat{\boldsymbol{\xi}}_i)\right|^2 I\left(\widehat{\boldsymbol{\xi}}_i \in [-R, R]^{qD}\right)\right) + 4C_{\mathcal{F}}^2\mathbb{P}\left(\widehat{\boldsymbol{\xi}}_i \notin [-R, R]^{qD}\right) \\ &\leq C\left(\log(D)R\right)^{2p}\left(\frac{\log(W)\log(L)}{WL}\right)^{\frac{4p}{d}} + CqDe^{-\frac{R^2}{C}},\end{aligned}$$

where $C > 0$ is a constant free of n, W, D, L and M_i. Setting $R = C\log(n)$ and taking average over i, we have

$$\mathcal{E}_{\text{app}} \leq C\left(\log(D)\log(n)\right)^{2p}\left(\frac{\log(W)\log(L)}{WL}\right)^{\frac{4p}{d}} + \frac{CqD}{n}. \tag{1.17}$$

Finally, combining Theorem 1 and (1.14)-(1.17), we complete the proof. □

Theorem 5. *For fixed $R > 0$, define*

$$\begin{aligned}
\mathcal{H}(R) &= \Big\{h : \mathbb{R}^d \to \mathbb{R} \;\Big|\; \max_{|\boldsymbol{\alpha}|\leq p}\sup_{\boldsymbol{u}\in\mathbb{R}^d}\left|\partial^{\boldsymbol{\alpha}}h(\boldsymbol{u})\right| \leq R\Big\},\\
\mathcal{Q}_2(R) &= \Big\{Q \;\Big|\; (\boldsymbol{X},\epsilon)\sim Q, \quad \mathbb{E}_Q(\epsilon \mid \boldsymbol{X}) = 0, \quad \mathbb{E}_Q(\epsilon^2 \mid \boldsymbol{X}) \leq 1,\\
&\qquad X_j = \sum_{k=1}^{\infty}\xi_{jk}\varphi_k \text{ for } 1\leq j\leq q,\\
&\qquad \max_{1\leq j\leq q}\|\mathbb{E}_Q(X_j)\|_{\mathcal{H}^{m_1}(\mathcal{T})} \leq R, \quad \max_{1\leq j\leq q}\sup_{k\geq 1}\mathit{Var}_Q(\xi_{jk})k^{2m_1} \leq R^2\Big\},\\
\mathcal{B}(R) &= \Big\{\boldsymbol{\beta} = (\beta_{11},\ldots,\beta_{qd_q}) : \mathcal{T}\to\mathbb{R}^d \;\Big|\; \max_{1\leq j\leq q}\max_{1\leq k\leq d_j}\|\beta_{jk}\|_{\mathcal{H}^{m_2}(\mathcal{T})} \leq R\Big\}.
\end{aligned}$$

Let $h \in \mathcal{H}(R)$, $(\boldsymbol{X},\epsilon) \sim Q \in \mathcal{Q}_2(R)$, and $\boldsymbol{\beta} \in \mathcal{B}(R)$. Assume that $\{(Y_i, \boldsymbol{X}_i)\}_{i=1}^n$ are i.i.d. observations from the model

$$Y = F_{h,\boldsymbol{\beta}}(\boldsymbol{X}) + \epsilon := h(U_{11},\ldots,U_{1d_1},\ldots,U_{q1},\ldots,U_{qd_q}) + \epsilon,$$

where $U_{jk} = \langle X_j, \beta_{jk}\rangle_{L_2(\mathcal{T})}$. Then it holds that

$$\liminf_{n\to\infty}\inf_{\widehat{F}}\sup_{\substack{h\in\mathcal{H}(R)\\ Q\in\mathcal{Q}_2(R)\\ \boldsymbol{\beta}\in\mathcal{B}(R)}}\mathbb{E}_{h,\boldsymbol{\beta},Q}\Big\{\mathcal{E}_Q(\widehat{F}, F_{h,\boldsymbol{\beta}})\Big\}\left(n^{-\frac{2(m_1+m_2)}{2(m_1+m_2)+1}} + n^{-\frac{2p}{2p+d}}\right)^{-1} > 0.$$

Here $\mathbb{E}_{h,\boldsymbol{\beta},Q}$ denotes expectation over the training sample generated under h, $\boldsymbol{\beta}$, and Q, while $\mathcal{E}_Q$ is the population excess risk conditional on this sample. The infimum is taken over all estimators based on these data.

Proof. We divide the proof into two steps.

Step 1 Define

$$s = m_1 + m2, \quad M = \lceil n^{1/(2s+1)}\rceil \asymp n^{1/(2s+1)}.$$

Consider the following specification:

$$\epsilon \sim N(0,1), \quad X_1(t) = X_2(t) = \ldots = X_q(t) = \sum_{k=1}^{\infty}\xi_k\varphi_k(t),$$

where $\{\xi_k \sim N(0, c_X^2 k^{-2m_1})\}_{k=1}^{\infty}$ are independent for some constant $c_X > 0$, and ϵ is independent of these coefficients. If we choose c_X sufficiently small, then the distribution Q_0 of $(\boldsymbol{X},\epsilon)$ belongs to $\mathcal{Q}_2(R)$. The Gaussian series converges in $L_2(\mathcal{T})$ almost surely because $m_1 > 1/2$ and $\sum_{k=1}^{\infty}\mathbb{E}(\xi_k^2) < \infty$. Throughout Step 1, the L_2 norm of a regression functional is taken with respect to this common distribution of $\boldsymbol{X}$.

For each $\boldsymbol{w} = (w_1,\ldots,w_M)^\top \in \{-1,1\}^M$, define

$$\beta_{\boldsymbol{w}}(t) = a_M\sum_{k=1}^{M}w_k\varphi_{M+k}(t), \text{ with } a_M = c_\beta M^{-m_2-\frac{1}{2}},$$

where $c_\beta > 0$ is a constant. Since $M \leq M+k \leq 2M$, we have

$$\|\beta_{\boldsymbol{w}}\|^2_{\mathcal{H}^{m_2}(\mathcal{T})} \asymp a_M^2\sum_{k=1}^{M}(M+k)^{2m_2} \asymp c_\beta^2.$$

Therefore, if we define $\boldsymbol{\beta}_{\boldsymbol{w}} = (\beta_{\boldsymbol{w}}, 0, \ldots, 0)$, then it follows that $\boldsymbol{\beta}_{\boldsymbol{w}} \in \mathcal{B}(R)$ if we choose c_β sufficiently small.

Let $U_{\boldsymbol{w}} = \langle X_1, \beta_{\boldsymbol{w}} \rangle_{L_2(\mathcal{T})} = a_M \sum_{k=1}^M w_k \xi_{M+k}$, which is a Gaussian random variable. Noting that

$$\mathbb{E}(U_{\boldsymbol{w}}^2) = a_M^2 \sum_{k=1}^M c_X^2 (M+k)^{-2m_1} \asymp M^{-2s} \asymp n^{-\frac{2s}{2s+1}},$$

the Gaussian tail inequality implies that

$$\mathbb{P}(|U_{\boldsymbol{w}}| > 1) \le \mathbb{E}\left(U_{\boldsymbol{w}}^2 I(|U_{\boldsymbol{w}}| > 1)\right) \le C e^{-cM^{2s}} \le C e^{-cn^{2s/(2s+1)}}, \tag{1.18}$$

where $c, C > 0$ are two constants free of $M, n, \boldsymbol{w}$.

Let $\widetilde{h}(\boldsymbol{u}) = c_h \rho(u_1)$ for $\boldsymbol{u} = (u_1, \ldots, u_d)^\top \in \mathbb{R}^d$. Here $c_h > 0$ is a constant, and $\rho : \mathbb{R} \to \mathbb{R}$ is an infinitely differentiable function such that $\max_{0 \le k \le p} \sup_{u \in \mathbb{R}} \left|\rho^{(k)}(u)\right| \le C_\rho$ and $\rho(u) = u$ when $|u| \le 1$ (can be constructed using smooth cut-off). Therefore, it follows

$$\max_{|\boldsymbol{\alpha}| \le p} \sup_{\boldsymbol{u} \in \mathbb{R}^d} \left|\partial^{\boldsymbol{\alpha}} \widetilde{h}(\boldsymbol{u})\right| \le c_h C_\rho,$$

which implies that $\widetilde{h} \in \mathcal{H}(R)$ if we choose c_h sufficiently small.

Now let us define $F_{\boldsymbol{w}}(\boldsymbol{X}) = \widetilde{h}(U_{\boldsymbol{w}}, 0, \ldots, 0) = c_h \rho(U_{\boldsymbol{w}})$ and $G_{\boldsymbol{w}}(\boldsymbol{X}) = c_h U_{\boldsymbol{w}}$. Then direct examination implies that

$$\begin{aligned}
\|F_{\boldsymbol{w}} - G_{\boldsymbol{w}}\|_{L_2}^2 &= c_h^2 \mathbb{E}\left(|\rho(U_{\boldsymbol{w}}) - U_{\boldsymbol{w}}|^2 I(|U_{\boldsymbol{w}}| > 1)\right) \\
&\overset{\text{(i)}}{\le} 2c_h^2 \left(C_\rho^2 \mathbb{P}(|U_{\boldsymbol{w}}| > 1) + \mathbb{E}\left(U_{\boldsymbol{w}}^2 I(|U_{\boldsymbol{w}}| > 1)\right)\right) \overset{\text{(ii)}}{\le} C e^{-cM^{2s}},
\end{aligned}$$

where (i) uses the boundedness of $\rho(\cdot)$, and (ii) is from (1.18). This bound holds uniformly over $\boldsymbol{w} \in \{-1, 1\}^M$. For any $\boldsymbol{w}, \boldsymbol{w}' \in \{-1, 1\}^M$, define their Hamming distance by

$$H(\boldsymbol{w}, \boldsymbol{w}') = \sum_{k=1}^M I(w_k \ne w_k').$$

Independence and centering of the coefficients imply that

$$\begin{aligned}
\|G_{\boldsymbol{w}} - G_{\boldsymbol{w}'}\|_{L_2}^2 &= c_h^2 a_M^2 \mathbb{E}\left(\left|\sum_{k=1}^M (w_k - w_k') \xi_{M+k}\right|^2\right) \\
&= 4c_h^2 a_M^2 c_X^2 \sum_{k : w_k \ne w_k'} (M+k)^{-2m_1} \\
&\asymp M^{-2s-1} H(\boldsymbol{w}, \boldsymbol{w}'),
\end{aligned}$$

where the comparison constants are independent of $M, \boldsymbol{w}, \boldsymbol{w}'$, since $M \le M + k \le 2M$. Since $\sup_{u \in \mathbb{R}} |\rho'(u)| \le C_\rho$, we also have

$$\|F_{\boldsymbol{w}} - F_{\boldsymbol{w}'}\|_{L_2}^2 \le C_\rho^2 \|G_{\boldsymbol{w}} - G_{\boldsymbol{w}'}\|_{L_2}^2.$$

For the lower bound, the triangle inequality yields, for some constant $c_1 > 0$ independent of $M, \boldsymbol{w}, \boldsymbol{w}'$,

$$\begin{aligned}
\|F_{\boldsymbol{w}} - F_{\boldsymbol{w}'}\|_{L_2}^2 &\ge \frac{1}{2} \|G_{\boldsymbol{w}} - G_{\boldsymbol{w}'}\|_{L_2}^2 - 2\|F_{\boldsymbol{w}} - G_{\boldsymbol{w}}\|_{L_2}^2 - 2\|F_{\boldsymbol{w}'} - G_{\boldsymbol{w}'}\|_{L_2}^2 \\
&\ge c_1 M^{-2s-1} H(\boldsymbol{w}, \boldsymbol{w}') - 4C e^{-cM^{2s}}.
\end{aligned}$$

For distinct vertices, $H(\boldsymbol{w}, \boldsymbol{w}') \ge 1$. Since $s > 0$, there is an integer M_0, independent of $\boldsymbol{w}, \boldsymbol{w}'$, such that $4C e^{-cM^{2s}} \le (c_1/2) M^{-2s-1}$ for all $M \ge M_0$. Thus, setting $c_* = c_1/2$, we obtain

$$\|F_{\boldsymbol{w}} - F_{\boldsymbol{w}'}\|_{L_2}^2 \ge c_* M^{-2s-1} H(\boldsymbol{w}, \boldsymbol{w}'), \qquad \boldsymbol{w}, \boldsymbol{w}' \in \{-1, 1\}^M, \quad M \ge M_0. \tag{1.19}$$

The case $\boldsymbol{w} = \boldsymbol{w}'$ holds trivially because both sides are zero.

Next, we consider the model

$$Y = F_{\boldsymbol{w}}(\boldsymbol{X}) + \epsilon, \text{ with } \epsilon \sim N(0,1).$$

Let $P_{\boldsymbol{w}}^{(n)}$ be the joint distribution of n independent samples generated from the above model. For neighboring vertices $\boldsymbol{w}, \boldsymbol{w}'$ with $H(\boldsymbol{w}, \boldsymbol{w}') = 1$, the common design distribution and the independent $N(0,1)$ noise give

$$KL\left(P_{\boldsymbol{w}}^{(n)}, P_{\boldsymbol{w}'}^{(n)}\right) = \frac{n}{2}\|F_{\boldsymbol{w}} - F_{\boldsymbol{w}'}\|_{L_2}^2 \le 2c_h^2 C_\rho^2 c_\beta^2 c_X^2 n M^{-2s-1} \le \frac{1}{8}. \tag{1.20}$$

Here $nM^{-2s-1} \le 1$, and the last inequality follows by choosing the fixed constants c_X, c_β, c_h sufficiently small, while retaining membership in $\mathcal{Q}_2(R), \mathcal{B}(R), \mathcal{H}(R)$, respectively. Applying Assouad's lemma (see, e.g., Tsybakov, 2009, Section 2.7.2) with the separation bound (1.19) and the KL bound (1.20) gives

$$\liminf_{n\to\infty} \inf_{\widehat{F}} \sup_{\boldsymbol{w}\in\{-1,1\}^M} \mathbb{E}_{\widetilde{h},\boldsymbol{\beta}_{\boldsymbol{w}},Q_0}\Big\{\mathcal{E}_{Q_0}(\widehat{F}, F_{\boldsymbol{w}})\Big\} n^{\frac{2s}{2s+1}} \ge c, \tag{1.21}$$

where $c > 0$ is a constant independent of n.

Step 2: Let $\{U_{jk}, k = 1, \ldots, d_j, j = 1, \ldots, q\}$ be i.i.d. uniform random variables over $[0,1]$, and let $\epsilon \sim N(0,1)$ be independent of them. For fixed constants $c_U, c_B > 0$, consider the submodel

$$\begin{aligned} X_j(t) &= c_U \sum_{k=1}^{d_j} U_{jk}\varphi_k(t), \text{ for } j = 1, \ldots, q, \\ \beta_{jk}(t) &= c_B \varphi_k(t), \text{ for } k = 1, \ldots, d_j, j = 1, \ldots, q. \end{aligned} \tag{1.22}$$

Because each expansion has finitely many terms, choosing c_U, c_B sufficiently small ensures that the distribution $\widetilde{Q}$ of $(\boldsymbol{X}, \epsilon)$ belongs to $\mathcal{Q}_2(R)$ and that the collection $\widetilde{\boldsymbol{\beta}}$ in (1.22) belongs to $\mathcal{B}(R)$. The index vector is uniform on the fixed cube $[0, c_U c_B]^d$, and the model becomes

$$Y = F_{h,\widetilde{\boldsymbol{\beta}}}(\boldsymbol{X}) + \epsilon = h(c_U c_B U_{11}, \ldots, c_U c_B U_{qd_q}) + \epsilon.$$

Since the functional covariates and the index vector determine each other, estimating $F_{h,\widetilde{\boldsymbol{\beta}}}$ under this submodel is equivalent to estimating the p-smooth link function h under a uniform design on $[0, c_U c_B]^d$. The classical expected-risk lower bound for p-smooth regression on a fixed d-dimensional cube (e.g., Tsybakov, 2009) implies that

$$\liminf_{n\to\infty} \inf_{\widehat{F}} \sup_{h\in\mathcal{H}(R)} \mathbb{E}_{h,\widetilde{\boldsymbol{\beta}},\widetilde{Q}}\Big\{\mathcal{E}_{\widetilde{Q}}(\widehat{F}, F_{h,\widetilde{\boldsymbol{\beta}}})\Big\} n^{\frac{2p}{2p+d}} \ge c, \tag{1.23}$$

where $c > 0$ is a constant independent of n.

Since both submodels lie in the parameter classes of the theorem, combining (1.21) and (1.23) with $\max\{x, y\} \ge (x+y)/2$, and choosing a common sufficiently small constant $c > 0$, we have

$$\liminf_{n\to\infty} \inf_{\widehat{F}} \sup_{\substack{h\in\mathcal{H}(R)\\ Q\in\mathcal{Q}_2(R)\\ \boldsymbol{\beta}\in\mathcal{B}(R)}} \mathbb{E}_{h,\boldsymbol{\beta},Q}\Big\{\mathcal{E}_Q(\widehat{F}, F_{h,\boldsymbol{\beta}})\Big\}\Big(n^{-\frac{2s}{2s+1}} + n^{-\frac{2p}{2p+d}}\Big)^{-1} \ge \frac{c}{2} > 0.$$

Nothing that $s = m_1 + m_2$, we complete the proof. □

A.5 Network Approximation Lemmas

Let $\sigma(u) = \max\{u, 0\}$ denote the ReLU activation function. A scalar-output depth-L DNN can be written as

$$g(x) = A_L\,\sigma\Big(A_{L-1}\,\sigma\big(\cdots\sigma(A_1 x + b_1)\cdots\big) + b_{L-1}\Big) + b_L,$$

where $x \in \mathbb{R}^d$, $A_1 \in \mathbb{R}^{p_1\times d}$, $A_\ell \in \mathbb{R}^{p_\ell \times p_{\ell-1}}$ for $\ell = 2, \ldots, L-1$, $A_L \in \mathbb{R}^{1\times p_{L-1}}$, and $b_\ell \in \mathbb{R}^{p_\ell}$ for $\ell = 1, \ldots, L-1$, with $b_L \in \mathbb{R}$. We introduce two common network structures that will be used often.

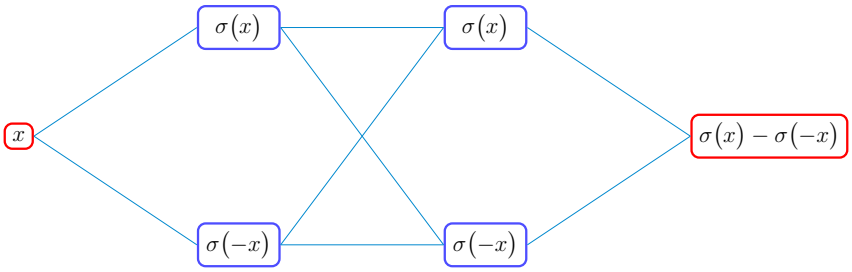


Figure 2: Example of passing values to next layer.

(i) Passing values: Using the identity $x = \sigma(x) - \sigma(-x)$ for all $x \in \mathbb{R}$, we can pass any value to next layers by adding two additional nodes. See Figure 2.

(ii) Network composition: Consider the network:

$$g(\boldsymbol{x}) = A_L\,\sigma\Big(A_{L-1}\,\sigma\big(\cdots\sigma(A_1\boldsymbol{x}+b_1)\cdots\big)+b_{L-1}\Big)+b_L, \qquad \boldsymbol{x}\in\mathbb{R}^k.$$

Suppose we want to build the composite map $g\big(h_1(\boldsymbol{z}),\ldots,h_k(\boldsymbol{z})\big)$ from networks $h_1,\ldots,h_k$. If we directly feed $h_j(\boldsymbol{z})$ into the first hidden layer of g through a ReLU gate, then the sign information is lost because $\sigma(h_j(\boldsymbol{z})) \geq 0$. To preserve sign information, for each input coordinate we introduce two additional ReLU nodes:

$$u_j^+(\boldsymbol{z}) := \sigma\big(h_j(\boldsymbol{z})\big), \qquad u_j^-(\boldsymbol{z}) := \sigma\big(-h_j(\boldsymbol{z})\big), \qquad j=1,\ldots,k.$$

Then $h_j(\boldsymbol{z}) = u_j^+(\boldsymbol{z}) - u_j^-(\boldsymbol{z})$, so the original input vector $\boldsymbol{h}(\boldsymbol{z}) = (h_1(\boldsymbol{z}),\ldots,h_k(\boldsymbol{z}))^\top$ can be recovered from the $2k$-dimensional nonnegative vector

$$\widetilde{\boldsymbol{h}}(\boldsymbol{z}) = (u_1^+(\boldsymbol{z}), u_1^-(\boldsymbol{z}),\ldots,u_k^+(\boldsymbol{z}),u_k^-(\boldsymbol{z}))^\top.$$

Let $\widetilde{A}_1 \in \mathbb{R}^{p_1\times 2k}$ be obtained by splitting the columns of A_1 as

$$\widetilde{A}_1 = \big[\,A_1,\,-A_1\,\big],$$

i.e., for each j we replace the j-th column a_j of A_1 by two columns $(a_j,-a_j)$. Then

$$A_1\,\boldsymbol{h}(\boldsymbol{z}) = A_1\,(u^+-u^-) = \widetilde{A}_1\,\widetilde{\boldsymbol{h}}(\boldsymbol{z}),$$

and therefore the composite function admits a ReLU-network representation by using the same layers $(A_2,\ldots,A_L)$ and biases $(b_1,\ldots,b_L)$ as g, but with the modified first layer $\widetilde{A}_1$ and the $2k$-dimensional input $\widetilde{\boldsymbol{h}}(\boldsymbol{z})$:

$$g\big(h_1(\boldsymbol{z}),\ldots,h_k(\boldsymbol{z})\big) = A_L\,\sigma\Big(A_{L-1}\,\sigma\big(\cdots\sigma(\widetilde{A}_1\widetilde{\boldsymbol{h}}(\boldsymbol{z})+b_1)\cdots\big)+b_{L-1}\Big)+b_L.$$

In particular, composing ReLU networks increases the width by at most a factor of 2 at the interface (to carry positive and negative parts), while keeping the depth additive up to this interface layer. See Figure 3.

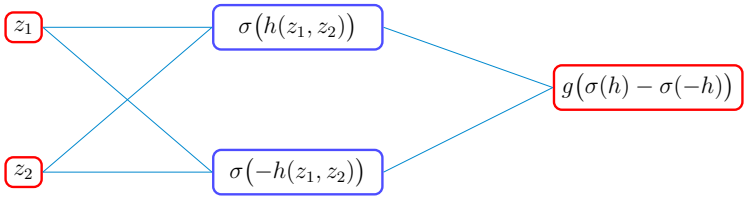


Figure 3: Example of network composition.

Lemma 16. *Given $f \in C([0,1]^d)$, for any $N_1, N_2 \in \mathbb{N}^+$, there exists a function ϕ implemented by an equal-width ReLU FNN with width $W = C_1 \max\{d\lfloor N_1^{1/d} \rfloor, N_1 + 1\}$ and depth $L = 12N_2 + C_2$ such that*

$$\sup_{\boldsymbol{v} \in [0,1]^d} |f(\boldsymbol{v}) - \phi(\boldsymbol{v})| \leq 19\sqrt{d}\, \omega_f \left(N_1^{-2/d} N_2^{-2/d} \right),$$

where $C_1 = 3^{d+3}$ and $C_2 = 14 + 2d$. Here $\omega_f(\delta) = \sup_{\|\boldsymbol{v}_1 - \boldsymbol{v}_2\| \leq \delta} |f(\boldsymbol{v}_1) - f(\boldsymbol{v}_2)|$ is the modulus of continuity.

Proof. This is Theorem 1.1 in Shen et al. (2020). □

Lemma 17. *Given $f \in C(\mathbb{R}^d)$, for any $N_1, N_2 \in \mathbb{N}^+$, there exists a function ϕ implemented by an equal-width ReLU FNN with width $W = C_1 \max\{d\lfloor N_1^{1/d} \rfloor, N_1 + 1\}$ and depth $L = 12N_2 + C_2 + 1$ such that*

$$\sup_{\boldsymbol{v} \in [-R,R]^d} |f(\boldsymbol{v}) - \phi(\boldsymbol{v})| \leq 19\sqrt{d}\, \omega_f \left(2RN_1^{-2/d} N_2^{-2/d} \right),$$

where $C_1 = 3^{d+3}$ and $C_2 = 14 + 2d$. Here $\omega_f(\delta) = \sup_{\|\boldsymbol{v}_1 - \boldsymbol{v}_2\| \leq \delta} |f(\boldsymbol{v}_1) - f(\boldsymbol{v}_2)|$ is the modulus of continuity.

Proof. Using Lemma 16, the proof is similar to that of Lemma 20. Hence, we omit it. □

Lemma 18. *Suppose that Assumptions A1-A4, B1, and B2 hold. Define*

$$\overline{f}_D(\boldsymbol{v}) = \max\left\{ -\frac{C_{\mathcal{F}}}{2}, \min\left\{ f_D(\boldsymbol{v}), \frac{C_{\mathcal{F}}}{2} \right\} \right\},$$

which is obtained by clipping f_D to $[-C_{\mathcal{F}}/2, C_{\mathcal{F}}/2]$. For any $N_1, N_2 \in \mathbb{N}^+$, there exists a function ϕ implemented by an equal-width ReLU FNN with width $W = 3^{qD+3}(qD + N_1 + 1)$ and depth $L = 12N_2 + 15 + 2qD$ such that

$$\sup_{\boldsymbol{v} \in [-R,R]^{qD}} \left| \phi(\boldsymbol{v}) - \overline{f}_D(\boldsymbol{v}) \right| \leq C\sqrt{qD} R^{\gamma} (N_1 N_2)^{-\frac{2\gamma}{qD}},$$

where $C > 0$ is a constant depending on q and γ.

Proof. For simplicity, we consider the case $q = 1$. For any $\boldsymbol{v}, \widetilde{\boldsymbol{v}} \in \mathbb{R}^D$, choose curves X and $\widetilde{X}$ whose projections satisfy

$$\Pi_D(X) = \sum_{k=1}^{D} v_k \varphi_k, \qquad \Pi_D(\widetilde{X}) = \sum_{k=1}^{D} \widetilde{v}_k \varphi_k,$$

where v_k and $\widetilde{v}_k$ are the k-th components of $\boldsymbol{v}$ and $\widetilde{\boldsymbol{v}}$, respectively. By Assumption B1, we obtain

$$\begin{aligned} |\overline{f}_D(\boldsymbol{v}) - \overline{f}_D(\widetilde{\boldsymbol{v}})| &\overset{\text{(i)}}{\leq} |f_D(\boldsymbol{v}) - f_D(\widetilde{\boldsymbol{v}})| \\ &= |F_0(\Pi_D(X)) - F_0(\Pi_D(\widetilde{X}))| \\ &\overset{\text{(ii)}}{\leq} C_\gamma \|\Pi_D(X) - \Pi_D(\widetilde{X})\|_{L_2(\mathcal{T})}^{\gamma} = C_\gamma \|\boldsymbol{v} - \widetilde{\boldsymbol{v}}\|^{\gamma}, \end{aligned}$$

where (i) uses the 1-Lipschitz property of truncation, and (ii) follows from Assumption B1. Hence, we have $\omega_{\overline{f}_D}(\delta) \leq C_\gamma \delta^\gamma$. Applying Lemma 17 to $\overline{f}_D$ gives a ReLU network ϕ with width $W = 3^{qD+3} \max\{qD\lfloor N_1^{\frac{1}{qD}} \rfloor, N_1 + 1\}$ and depth $L = 12N_2 + 15 + 2D$ such that

$$\begin{aligned} \sup_{\boldsymbol{v} \in [-R,R]^D} |\phi(\boldsymbol{v}) - \overline{f}_D(\boldsymbol{v})| &\overset{\text{(i)}}{\leq} 19\sqrt{D}\, \omega_{\overline{f}_D} \left(2R(N_1 N_2)^{-\frac{2}{D}} \right) \\ &\leq C\sqrt{D} R^{\gamma} (N_1 N_2)^{-\frac{2\gamma}{D}}, \end{aligned}$$

where (i) follows from Lemma 17. Finally, noting that $qDN_1^{1/(qD)} \leq qD + N_1$, we have $W \leq 3^{qD+3}(qD + N_1 + 1)$, which completes the proof. □

Lemma 19. *Given a smooth function $f \in C^p\big([0,1]^d\big)$ with $p \in \mathbb{N}^+$, for any $N, L \in \mathbb{N}^+$, there exists a function ϕ implemented by an equal-width ReLU DNN with width $C_1(N+2)\log_2(8N)$ and depth $C_2(L+2)\log_2(4L)+2d$ such that*

$$\|\phi - f\|_{L^\infty([0,1]^d)} \le C_3 \, \|f\|_{C^p([0,1]^d)} \, N^{-2p/d} L^{-2p/d},$$

where $C_1 = 17\, p^{d+1} 3^d d$, $C_2 = 18p^2$, and $C_3 = 85(p+1)^d 8^p$. Here $\|f\|_{C^p(\Omega)} = \max_{|\boldsymbol{\alpha}|\le p} \sup_{\boldsymbol{u}\in\Omega} |\partial^{\boldsymbol{\alpha}} f(\boldsymbol{u})|$.

Proof. This is the main result in Lu et al. (2021). □

Lemma 20. *Given a smooth function $g \in C^p(\mathbb{R}^d)$ with $p \in \mathbb{N}^+$ for any $N, L \in \mathbb{N}^+$, there exists a function ϕ implemented by an equal-width ReLU DNN with width $C_1(N+2)\log_2(8N)$ and depth $C_2(L+2)\log_2(4L)+2d+1$ such that*

$$\|\phi - g\|_{L^\infty([-R,R]^d)} \le C_3 R^p \|g\|_{C^p([-R,R]^d)} (NL)^{-\frac{2p}{d}},$$

where $C_1 = 17 p^{d+1} 3^d d$, $C_2 = 18p^2$, and $C_3 = 340(p+1)^d 8^p$. Here $\|f\|_{C^p(\Omega)} = \max_{|\boldsymbol{\alpha}|\le p} \sup_{\boldsymbol{u}\in\Omega} |\partial^{\boldsymbol{\alpha}} f(\boldsymbol{u})|$.

Proof. Define $f(\boldsymbol{u}) = g(w(\boldsymbol{u}))$ with $w(\boldsymbol{u}) = (2Ru_1 - R, \ldots, 2Ru_d - R)^\top$ for $\boldsymbol{u} = (u_1, \ldots, u_d)^\top \in [0,1]^d$. By Lemma 19, there is a function ψ implemented by a ReLU DNN with width $C_1(N+2)\log_2(8N)$ and depth $C_2(L+2)\log_2(4L)+2d$ such that

$$\sup_{\boldsymbol{u}\in[0,1]^d} |\psi(\boldsymbol{u}) - f(\boldsymbol{u})| \le C_3 \|f\|_{C^p([0,1]^d)} (NL)^{-\frac{2p}{d}}.$$

where $C_1 = 17\, p^{d+1} 3^d d$, $C_2 = 18p^2$, and $C_3 = 85(p+1)^d 8^p$.

Let $\phi(\boldsymbol{x}) = \psi(w^{-1}(\boldsymbol{x}))$ for $\boldsymbol{x} = (x_1, \ldots, x_d)^\top \in [-R,R]^d$, where $w^{-1}(\boldsymbol{x}) = ((x_1 + R)/(2R), \ldots, (x_d + R)/(2R))^\top$ is the inverse map. As a consequence, we have

$$\begin{aligned}\sup_{\boldsymbol{x}\in[-R,R]^d} |\phi(\boldsymbol{x}) - g(\boldsymbol{x})| &= \sup_{\boldsymbol{x}\in[-R,R]^d} |\psi(w^{-1}(\boldsymbol{x})) - f(w^{-1}(\boldsymbol{x}))| \\ &= \sup_{\boldsymbol{u}\in[0,1]^d} |\psi(\boldsymbol{u}) - f(\boldsymbol{u})| \le C_3 \|f\|_{C^p([0,1]^d)} (NL)^{-\frac{2p}{d}}.\end{aligned}$$

By the chain rule, we have

$$\partial^{\boldsymbol{\alpha}} f(\boldsymbol{u}) = (2R)^{|\boldsymbol{\alpha}|} (\partial^{\boldsymbol{\alpha}} g)(w(\boldsymbol{u})),$$

which implies that

$$\begin{aligned}\|f\|_{C^p([0,1]^d)} &= \max_{|\boldsymbol{\alpha}|\le p} \sup_{\boldsymbol{u}\in[0,1]^d} |\partial^{\boldsymbol{\alpha}} f(\boldsymbol{u})| \\ &\le \max_{|\boldsymbol{\alpha}|\le p} (2R)^{|\boldsymbol{\alpha}|} \sup_{\boldsymbol{u}\in[0,1]^d} |(\partial^{\boldsymbol{\alpha}} g)(w(\boldsymbol{u}))| \\ &\le (2R)^p \max_{|\boldsymbol{\alpha}|\le p} \sup_{\boldsymbol{x}\in[-R,R]^d} |(\partial^{\boldsymbol{\alpha}} g)(\boldsymbol{x})| = (2R)^p \|g\|_{C^p([-R,R]^d)}.\end{aligned}$$

Combining the above two upper bounds, we verify the error bound.

Finally, since $d \le C_1$, it is easy to check that $\phi(\boldsymbol{x}) = \psi(w^{-1}(\boldsymbol{x}))$ can be implemented by a ReLU DNN by adding one more layer (with d nodes) before ψ. Hence, it has width $C_1(N+2)\log_2(8N)$ and depth $C_2(L+2)\log_2(4L)+2d+1$. See Figure 4.

□

Lemma 21. *Suppose that Assumptions A1-A4 and C1-C3 hold. Let $W_1 \ge 2d$ and $W = C_1(N+2)\log_2(8N)$. There exists a function f_* implemented by a ReLU FNN with depth $L = C_2(J+2)\log_2(4J)+2d+1$ and width structure $(W_1, W, \ldots, W)$ such that*

$$\sup_{\boldsymbol{v}\in[-R,R]^{qD}} |f_*(\boldsymbol{v}) - f_D(\boldsymbol{v})| \le C\,(\log(D)R)^p\,(NJ)^{-\frac{2p}{d}},$$

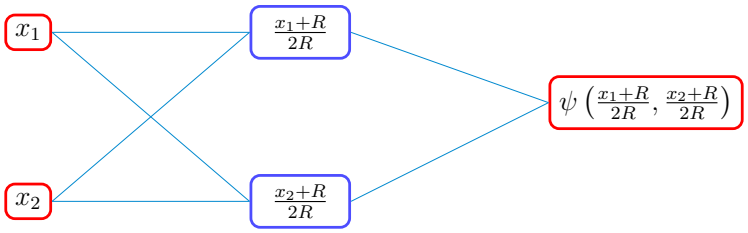


Figure 4: Visualization of DNN in Lemma 20. Example with $d = 2$.

where $C_1 = 17p^{d+1}3^d d$, $C_2 = 18p^2$, and $C > 0$ is a constant depending on p, d. As a consequence, if $W_1 \geq 2d$, $W \to \infty$ and $L \to \infty$, it follows that

$$\sup_{\boldsymbol{v}\in[-R,R]^{qD}} |f_*(\boldsymbol{v}) - f_D(\boldsymbol{v})| \leq C\,(\log(D)R)^p \left(\frac{\log(W)\log(L)}{WL}\right)^{\frac{2p}{d}},$$

where $C > 0$ is a constant depending on p and d.

Proof. Let $\boldsymbol{v} = \big(\boldsymbol{v}_1^\top, \ldots, \boldsymbol{v}_q^\top\big)^\top \in [-R, R]^{qD}$ with $\boldsymbol{v}_j = (v_{j1}, \ldots, v_{jD})^\top \in [-R, R]^D$. Recall that

$$f_D(\boldsymbol{v}) = h_0\big(\Pi_D(U_{11}), \ldots, \Pi_D(U_{1d_1}), \ldots, \Pi_D(U_{q1}), \ldots, \Pi_D(U_{qd_q})\big),$$

where

$$\Pi_D(U_{jk}) = \sum_{s=1}^{D} v_{js}\,\langle\varphi_s, \beta_{jk}^0\rangle_{L_2(\mathcal{T})} = \boldsymbol{a}_{jk}^\top \boldsymbol{v}_j,$$
$$\boldsymbol{a}_{jk} = \big(\langle\varphi_1, \beta_{jk}^0\rangle_{L_2(\mathcal{T})}, \ldots, \langle\varphi_D, \beta_{jk}^0\rangle_{L_2(\mathcal{T})}\big)^\top \in \mathbb{R}^D.$$

Define

$$\boldsymbol{A}_j := \begin{pmatrix} \boldsymbol{a}_{j1}^\top \\ \vdots \\ \boldsymbol{a}_{jd_j}^\top \end{pmatrix} \in \mathbb{R}^{d_j \times D}.$$

Then we can write

$$\begin{pmatrix} \Pi_D(U_{11}) \\ \vdots \\ \Pi_D(U_{1d_1}) \\ \vdots \\ \Pi_D(U_{q1}) \\ \vdots \\ \Pi_D(U_{qd_q}) \end{pmatrix} = \begin{pmatrix} \boldsymbol{A}_1 & 0 & \cdots & 0 \\ 0 & \boldsymbol{A}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \boldsymbol{A}_q \end{pmatrix}_{d\times qD} \begin{pmatrix} \boldsymbol{v}_1 \\ \boldsymbol{v}_2 \\ \vdots \\ \boldsymbol{v}_q \end{pmatrix}.$$

Moreover, Assumption C3 and Lemma 11 imply that

$$\langle\varphi_s,\, \beta_{jk}^0\rangle_{L_2(\mathcal{T})} \in [-Cs^{-m_2}, Cs^{m_2}],$$

which further leads to

$$|\Pi_D(U_{jk})| = |\boldsymbol{a}_{jk}^\top \boldsymbol{v}_j| \leq C\sum_{s=1}^{D} s^{-m_2} R \leq C\sum_{s=1}^{D} s^{-1} R \leq C\log(D)R, \tag{1.24}$$

where we use the condition $m_2 \geq 1$ in Assumption C3. By Lemma 20 and Assumption C1, there exists a function ψ implemented by a ReLU DNN ψ with width $C_1(N+2)\log_2(8N)$ and depth $C_2(J+2)\log_2(4J) + 2d + 1$ such that

$$\|\psi - h_0\|_{L^\infty([-C\log(D)R, C\log(D)R]^d)} \leq C\big(\log(D)R\big)^p (NJ)^{-\frac{2p}{d}},$$

where $C_1 = 17p^{d+1}3^d d$, $C_2 = 18p^2$, and the constant $C > 0$ depends on p, d.

Let $f_*(\boldsymbol{v}) = \psi\big(\boldsymbol{a}_{11}^\top \boldsymbol{v}_1, \ldots, \boldsymbol{a}_{1d_1}^\top \boldsymbol{v}_1, \ldots, \boldsymbol{a}_{q1}^\top \boldsymbol{v}_q, \ldots \boldsymbol{a}_{qd_q}^\top \boldsymbol{v}_q\big)$, and we have

$$\begin{aligned}\|f_*(\boldsymbol{v}) - f_D(\boldsymbol{v})\|_{L^\infty([-R,R]^{qD})} &\overset{\text{(i)}}{\le} \|\psi - h_0\|_{L^\infty([-C\log(D)R, C\log(D)R]^d)} \\ &\le C\big(\log(D)R\big)^p (NJ)^{-\frac{2p}{d}},\end{aligned}$$

where (i) is due to (1.24). Therefore, we verify the error bound.

Finally, it is easy to see the function f_* can be implemented by a ReLU FNN with depth $L = C_2(J+2)\log_2(4J) + 2d + 2$ (add one more layer) and width structure $(W_1, W, \ldots, W)$. Here $W_1 \ge 2d$ and $W = C_1(N+2)\log_2(8N)$; see Figure 5. □

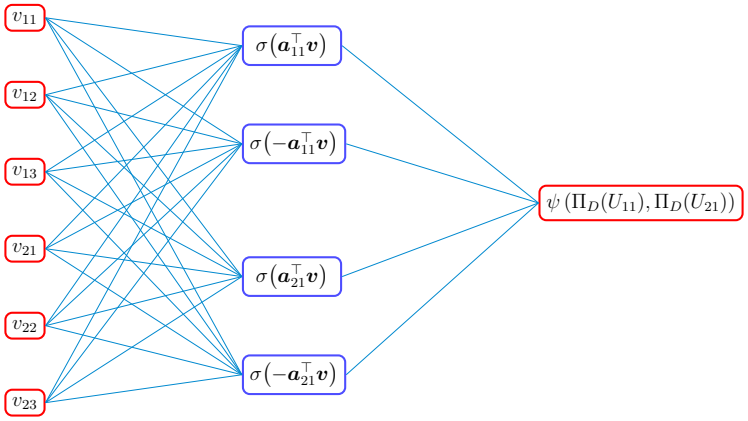


Figure 5: Visualization of DNN in Lemma 21. Example with $q = 2, D = 3, d_1 = d_2 = 1$.

B Simulation Study

We evaluate the predictive performance of the proposed EW and BEW deep learning frameworks for scalar-on-function regression using simulated data. For subject $i = 1, \ldots, n$, we generate $q = 4$ functional predictors $X_{ij}(t)$, $t \in [0, 1]$, and a scalar response

$$Y_i = \mu_i + \varepsilon_i, \qquad \varepsilon_i \sim N(0, 0.1^2),$$

where

$$\mu_i = Z_{i1} + \sin(Z_{i2}) + \alpha \sin(Z_{i1}) \cos(Z_{i2}), \qquad Z_{ir} = \sum_{j=1}^{4} \int_0^1 X_{ij}(t)\beta_{jr}(t)\, dt, \quad r = 1, 2. \tag{2.1}$$

Here, α controls the strength of the interaction between the two latent functional indices. In particular, $\alpha = 0$ corresponds to a model without interaction, whereas larger values of α induce increasingly stronger nonlinear interactions.

Functional covariate generation. We consider two mechanisms for generating the functional predictors through a truncated Fourier expansion,

$$X_{ij}(t) = \sum_{k=1}^{K_x} \xi_{ijk}\phi_k(t), \qquad t \in [0, 1],$$

where $\{\phi_k\}$ denotes the Fourier basis and the random coefficients ξ_{ijk} are Gaussian with mean μ_k and variance σ_k^2. The same generating mechanism is used for each functional covariate.

Under the **CFM** mechanism, the coefficient moments decay sufficiently rapidly so that

$$\sum_{k \geq 1} k^{2m_1} E(\xi_{ijk}^2) < \infty.$$

Specifically, we generate

$$\mu_k = C_\mu k^{-\{m_1 + (1+\delta)/2\}}, \qquad \sigma_k^2 = C_\sigma k^{-(2m_1 + 1 + \delta)}.$$

Thus, both the mean and variance decay rapidly with the basis index k, producing relatively smooth functional trajectories.

Under the **FMI** mechanism, the mean coefficients follow the same decay, while the variance decreases more slowly:

$$\mu_k = C_\mu k^{-\{m_1 + (1+\delta)/2\}}, \qquad \sigma_k^2 = C_\sigma k^{-2m_1}.$$

This construction satisfies

$$\mathrm{Var}(\xi_{ijk}) \leq C k^{-2m_1}, \qquad \sum_{k \geq 1} k^{2m_1} \left|E(\xi_{ijk})\right|^2 < \infty,$$

while allowing a relatively greater contribution from higher-frequency basis components. Consequently, the FMI trajectories are generally less smooth and provide a more challenging functional design than those generated under CFM.

Throughout the simulations, we set

$$m_1 = 2, \qquad K_x = 50, \qquad \delta = 1, \qquad C_\mu = 0.5, \qquad C_\sigma = 1.$$

Coefficient (β) generation. For each index $r \in \{1, 2\}$ and functional predictor $j \in \{1, \ldots, 4\}$, the coefficient function β_{jr} is represented using an orthonormal cosine basis on $[0, 1]$,

$$\phi_k(t) = \sqrt{2}\cos(k\pi t), \qquad k = 1, \ldots, K_\beta,$$

so that

$$\beta_{jr}(t) = \sum_{k=1}^{K_\beta} b_{jr,k}\phi_k(t).$$

The coefficients are chosen to satisfy the polynomial decay

$$|b_{jr,k}| \asymp k^{-\left(m_2+\frac{1}{2}+\delta_\beta\right)},$$

where $m_2 > 0$ controls the smoothness of the coefficient functions and $\delta_\beta > 0$ provides additional decay. Consequently,

$$\sum_{k=1}^{\infty} k^{2m_2} b_{jr,k}^2 < \infty,$$

so that each β_{jr} belongs to the corresponding Sobolev class of order m_2.

To generate distinct coefficient functions while preserving the same smoothness level, deterministic sign patterns are imposed on the coefficient sequences. Specifically,

$$b_{jr,k} = s_{jr,k} k^{-\left(m_2+\frac{1}{2}+\delta_\beta\right)}, \qquad s_{jr,k} \in \{-1, 1\},$$

where the sequence $s_{jr,k}$ varies across the eight (j, r) combinations. These sign patterns change the shapes of the coefficient functions without altering their rate of decay.

Finally, each coefficient sequence is normalized in ℓ_2,

$$\widetilde{b}_{jr,k} = \frac{b_{jr,k}}{\left(\sum_{\ell=1}^{K_\beta} b_{jr,\ell}^2\right)^{1/2}}.$$

By orthonormality of the cosine basis,

$$\|\beta_{jr}\|_{L^2}^2 = \sum_{k=1}^{K_\beta} \widetilde{b}_{jr,k}^2 = 1.$$

Thus, all coefficient functions have the same L^2 norm while remaining distinct in shape. The normalization is performed directly through the basis coefficients, ensuring that the same coefficient functions are used for all subjects and do not depend on subject-specific observation grids. For the coefficient-function generation, we set

$$m_2 = 1, \qquad K_\beta = 50, \qquad \delta_\beta = 0.1,$$

corresponding to moderately smooth coefficient functions represented using 50 basis terms.

Tuning Parameters. In all simulation experiments, tuning parameters are selected using the validation sample. The candidate architecture ranges are guided qualitatively by the theoretical results without requiring knowledge of the unknown smoothness parameters appearing in the asymptotic rates.

The basis dimension D and hidden-layer width W are allowed to increase slowly with the training sample size, whereas the bottleneck width W_1 is searched over a fixed bounded set, consistent with the theoretical requirement $W_1 = O(1)$. For $d = 2$, we consider

$$W_1 \in \{4, 6, 8, 10, 15, 20, 25, 30, 35, 40\}.$$

The basis dimension is selected from

$$D \in \{2, 4, 6, 8, 10, 12, 15, 20, 25, 30, 40, 50\},$$

subject to

$$D \le \min\left\{D_{\max}, \max\left(10, \lceil\sqrt{n_{\mathrm{tr}}}\rceil\right)\right\}.$$

Similarly, the hidden-layer width is selected from

$$W \in \{4, 8, 12, 16, 24, 32, 48, 64, 96, 128\},$$

subject to

$$W \le \max\left\{16, \lceil 4\sqrt{n_{\mathrm{tr}}}\rceil\right\}.$$

The network depth is searched locally around the logarithmic order suggested by the theory. Letting

$$L_0 = \max\left\{2, \lceil\log(n_{\mathrm{tr}})\rceil\right\},$$

we consider

$$L \in \{\max(1, L_0 - 2), \max(1, L_0 - 1), L_0, L_0 + 1, L_0 + 2\}.$$

The remaining optimization parameters, including the learning rate, dropout, and weight decay, are tuned jointly with the architectural parameters by minimizing the validation MSE using Optuna (Ozaki et al., 2025). For FPC, the number of retained functional principal components and the neural-network tuning parameters are selected analogously. The held-out test sample is not used during model fitting or hyperparameter selection.

Evaluation. Functional curves are observed under a random design with grid sizes

$$M \in \{50, 100, 200\},$$

and the functional integrals are approximated using trapezoidal quadrature. We consider

$$n \in \{200, 400, 600, 800, 1000\}$$

and conduct 200 independent replications for each configuration. For each replication, the observations are randomly divided into training, validation, and test samples in proportions $40\%/30\%/30\%$.

DNN hyperparameters are selected using 200 Optuna trials (Ozaki et al., 2025). The networks are trained using the Adam optimizer and mean squared error loss, with early stopping based on validation error. Performance is measured by

$$\mathrm{MSE}_{\mathrm{test}} = |\mathcal{I}_{\mathrm{te}}|^{-1} \sum_{i \in \mathcal{I}_{\mathrm{te}}} \left(f_D(\widehat{\boldsymbol{\xi}}_i) - \widehat{f}(\widehat{\boldsymbol{\xi}}_i) \right)^2,$$

and we report the mean and standard deviation of the test MSE across the 200 replications.

The DNN inputs are constructed as described in Section 2.2. In addition to the proposed EW and BEW methods, we include the DNN estimation method of Zhou et al. (2026) for comparison. Hereafter, we denote this method by FPC.

Results. The results for Model 2.1 with $\alpha \in \{0, 2, 4\}$ are reported in Tables 2–4 and Figure 6. Across all three methods, the test MSE generally decreases as both the sample size n and grid size M increase. When $\alpha = 0$, EW is generally competitive and most often attains the smallest prediction error. In contrast, when $\alpha > 0$, BEW increasingly outperforms EW and FPC as n and M increase. The improvement is particularly evident under stronger interactions and denser functional observations, highlighting the advantage of the BEW architecture for capturing nonlinear interactions among the latent functional indices.

To further examine the behavior of the proposed architecture, we investigate sensitivity to the interaction strength, bottleneck width, and basis dimension. For these experiments, we fix $n = 400$ and $M = 100$ and conduct 200 replications for each configuration.

- **Interaction strength.** Table 5 reports the results as α varies. When $\alpha = 0$, EW attains the smallest average test MSE under both CFM and FMI. For every nonzero value of α considered, however, BEW achieves the lowest average test MSE. Moreover, the advantage of BEW generally becomes more pronounced as the interaction strength increases. This pattern supports the motivation for the BEW architecture, whose additional bottleneck structure provides greater flexibility for capturing nonlinear interactions among the learned functional indices.
- **Bottleneck width W_1.** Fixing $\alpha = 4$, Table 6 examines sensitivity to the bottleneck width W_1. BEW outperforms EW for small to moderate values of W_1, with its best performance attained at relatively small bottleneck widths. Increasing W_1 beyond this range provides no systematic improvement and eventually deteriorates predictive performance. These findings support the use of a low-dimensional bottleneck and are consistent with the theoretical treatment of W_1 as a bounded architectural parameter.
- **Basis dimension D.** With $\alpha = 4$, Table 7 reports the results for varying basis dimension D. BEW performs best for small to moderate values of D, with the smallest average test MSE obtained at $D = 4$. For $D = 2, 4$, and 8, BEW also outperforms EW and FPC. Increasing D beyond this range does not improve prediction and eventually leads to substantially larger test errors. This illustrates the bias–variance trade-off associated with the basis expansion: a moderate basis dimension is sufficient to capture the relevant functional structure, whereas an unnecessarily rich representation increases estimation variability without a corresponding reduction in approximation error.

Overall, the simulation results indicate that EW is effective when the underlying regression structure is relatively simple, whereas BEW becomes increasingly advantageous when nonlinear interactions are present and the available information increases through larger sample sizes or denser functional observations. FPC remains competitive in some settings, but its performance is generally weaker when the regression signal depends strongly on nonlinear interactions among the functional indices.

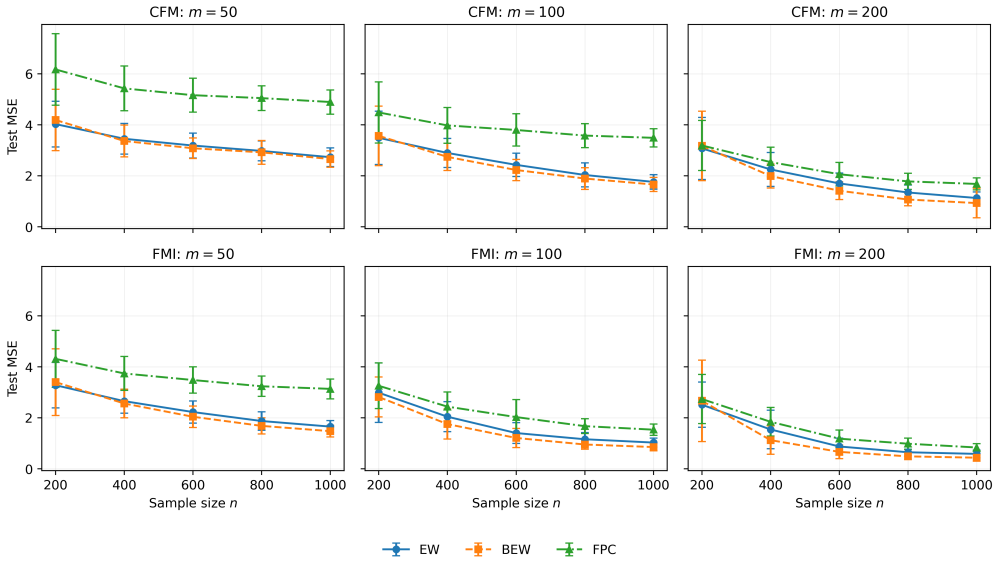


Figure 6: Test MSE, averaged over 200 replications, under the random design with $\alpha = 4.0$ in Model 2.1 for the CFM and FMI mechanisms, across varying sample sizes n and grid sizes M.

Table 2: Test MSE under the random design with $\alpha = 0$ in Model 2.1, across varying sample sizes n and grid sizes M, for the CFM and FMI functional-covariate generation mechanisms. Entries report the mean test MSE, with standard deviations in parentheses, over 200 replications.

		CFM			FMI		
Grid size	n	EW	BEW	FPC	EW	BEW	FPC
50	200	**0.5493** (0.2096)	0.5700 (0.1994)	1.3002 (0.2976)	**0.3308** (0.1927)	0.3543 (0.2468)	0.6107 (0.1737)
	400	**0.4395** (0.0963)	0.4538 (0.0999)	1.1667 (0.1998)	**0.2458** (0.0510)	0.2551 (0.0572)	0.5268 (0.0932)
	600	**0.3820** (0.0644)	0.4021 (0.0618)	1.0807 (0.1522)	**0.2067** (0.0335)	0.2177 (0.0336)	0.4808 (0.0770)
	800	**0.3534** (0.0548)	0.3729 (0.0537)	1.0517 (0.1279)	**0.1925** (0.0338)	0.2069 (0.1456)	0.4643 (0.0655)
	1000	**0.3259** (0.0472)	0.3427 (0.0469)	1.0212 (0.1373)	**0.1787** (0.0284)	0.1829 (0.0284)	0.4485 (0.0564)
100	200	**0.3746** (0.0997)	0.3949 (0.1837)	0.7021 (0.2004)	**0.2071** (0.0544)	0.2114 (0.0579)	0.3034 (0.1339)
	400	**0.2584** (0.0607)	0.2723 (0.0625)	0.6235 (0.1113)	**0.1511** (0.0287)	0.1568 (0.0302)	0.2389 (0.0439)
	600	**0.2303** (0.0417)	0.2367 (0.0427)	0.5785 (0.1056)	**0.1324** (0.0263)	0.1335 (0.0220)	0.2147 (0.0394)
	800	**0.2143** (0.0371)	0.2231 (0.0378)	0.5593 (0.0874)	**0.1201** (0.0220)	0.1228 (0.0224)	0.1953 (0.0334)
	1000	**0.1982** (0.0295)	0.2065 (0.0286)	0.5488 (0.0703)	**0.1096** (0.0183)	0.1110 (0.0379)	0.1886 (0.0274)
200	200	**0.2533** (0.1830)	0.2845 (0.2439)	0.3365 (0.1944)	**0.1554** (0.1991)	0.2020 (0.3496)	0.2125 (0.1860)
	400	**0.1708** (0.0451)	0.1769 (0.0423)	0.2633 (0.0554)	**0.0983** (0.0254)	0.1014 (0.0282)	0.1347 (0.0298)
	600	**0.1336** (0.0259)	0.1381 (0.0280)	0.2398 (0.0424)	0.0834 (0.0164)	**0.0829** (0.0183)	0.1101 (0.0226)
	800	**0.1169** (0.0218)	0.1195 (0.0224)	0.2272 (0.0432)	0.0735 (0.0130)	**0.0713** (0.0181)	0.0981 (0.0218)
	1000	0.1312 (0.2186)	**0.1091** (0.0174)	0.2225 (0.0328)	0.0673 (0.0125)	**0.0646** (0.0174)	0.0897 (0.0159)

Table 3: Test MSE under the random design with $\alpha = 2$ in Model 2.1, across varying sample sizes n and grid sizes M, for the CFM and FMI functional-covariate generation mechanisms. Entries report the mean test MSE, with standard deviations in parentheses, over 200 replications.

		CFM			FMI		
Grid Size	n	EW	BEW	FPC	EW	BEW	FPC
50	200	**1.6333** (0.4289)	1.6468 (0.5450)	2.9409 (0.6580)	**1.1891** (0.4260)	1.2187 (0.5303)	1.7335 (0.4830)
	400	**1.3416** (0.2413)	1.3723 (0.2603)	2.6172 (0.4268)	0.9547 (0.1948)	**0.9154** (0.1713)	1.5364 (0.2542)
	600	**1.2271** (0.1907)	1.2376 (0.1929)	2.4367 (0.3134)	0.8120 (0.1571)	**0.7914** (0.1450)	1.4292 (0.2070)
	800	**1.1421** (0.1487)	1.1585 (0.1571)	2.3780 (0.2650)	0.7409 (0.1608)	**0.6892** (0.1137)	1.3621 (0.1959)
	1000	**1.0550** (0.1438)	1.0712 (0.1260)	2.3361 (0.2388)	0.6411 (0.0983)	**0.6079** (0.0954)	1.3135 (0.1450)
100	200	1.3465 (0.5074)	**1.3288** (0.5502)	1.9411 (0.5148)	**0.9224** (0.2409)	0.9390 (0.4190)	1.1812 (0.4639)
	400	1.0101 (0.2253)	**0.9820** (0.2061)	1.6932 (0.3007)	0.7215 (0.1971)	**0.6334** (0.1751)	0.9018 (0.2004)
	600	0.8578 (0.1602)	**0.8419** (0.1460)	1.5900 (0.2334)	0.5216 (0.1219)	**0.4815** (0.1147)	0.7729 (0.1378)
	800	0.7372 (0.1403)	**0.7202** (0.1216)	1.5507 (0.1970)	0.4384 (0.0752)	**0.3901** (0.0719)	0.6737 (0.1156)
	1000	0.6753 (0.1092)	**0.6610** (0.0922)	1.5118 (0.1861)	0.3908 (0.0541)	**0.3418** (0.0576)	0.6118 (0.0974)
200	200	**1.0498** (0.5450)	1.0994 (0.6164)	1.1814 (0.4492)	**0.8067** (0.5267)	0.8652 (0.8613)	0.8817 (0.4204)
	400	0.7448 (0.1478)	**0.7035** (0.1673)	0.9616 (0.2381)	0.4505 (0.1604)	**0.4086** (0.3253)	0.6117 (0.1489)
	600	0.5988 (0.2846)	**0.5235** (0.1165)	0.8320 (0.1388)	0.2974 (0.0741)	**0.2518** (0.0831)	0.4361 (0.1043)
	800	0.4592 (0.0992)	**0.4027** (0.0880)	0.7612 (0.1284)	0.2442 (0.0526)	**0.1933** (0.0411)	0.3468 (0.0638)
	1000	0.3865 (0.0697)	**0.3581** (0.3154)	0.7008 (0.0964)	0.2156 (0.0321)	**0.1727** (0.0439)	0.2972 (0.0595)

Table 4: Test MSE under the random design with $\alpha = 4$ in Model 2.1, across varying sample sizes n and grid sizes M, for the CFM and FMI functional-covariate generation mechanisms. Entries report the mean test MSE, with standard deviations in parentheses, over 200 replications.

		CFM			FMI		
Grid size	n	EW	BEW	FPC	EW	BEW	FPC
50	200	**4.0254** (0.8963)	4.1863 (1.2012)	6.1676 (1.4000)	**3.2786** (0.8911)	3.4000 (1.3060)	4.3118 (1.1129)
	400	3.4505 (0.5984)	**3.3627** (0.6242)	5.4253 (0.8759)	2.6538 (0.4692)	**2.5617** (0.5613)	3.7391 (0.6657)
	600	3.1858 (0.4893)	**3.0789** (0.4043)	5.1592 (0.6629)	2.2268 (0.4336)	**2.0423** (0.4240)	3.4819 (0.5138)
	800	2.9740 (0.3909)	**2.9186** (0.4628)	5.0443 (0.4868)	1.8748 (0.3590)	**1.6845** (0.3143)	3.2369 (0.3987)
	1000	2.7292 (0.3649)	**2.6554** (0.3189)	4.8926 (0.4741)	1.6565 (0.2342)	**1.4781** (0.2267)	3.1354 (0.3883)
100	200	**3.4887** (1.0468)	3.5654 (1.1637)	4.4823 (1.1955)	2.9854 (1.1673)	**2.8165** (0.7829)	3.2563 (0.8966)
	400	2.8946 (0.5720)	**2.7414** (0.5326)	3.9743 (0.7007)	2.0446 (0.5904)	**1.7543** (0.5890)	2.4392 (0.5711)
	600	2.4288 (0.4578)	**2.2269** (0.4128)	3.7961 (0.6342)	1.4043 (0.4108)	**1.2111** (0.3777)	2.0300 (0.6839)
	800	2.0325 (0.4712)	**1.8926** (0.4270)	3.5766 (0.4693)	1.1610 (0.2606)	**0.9550** (0.1825)	1.6743 (0.2922)
	1000	1.7599 (0.2806)	**1.6583** (0.2766)	3.4886 (0.3579)	1.0307 (0.1791)	**0.8548** (0.1455)	1.5358 (0.2196)
200	200	**3.0742** (1.2175)	3.1722 (1.3617)	3.1874 (0.9819)	**2.5143** (0.8842)	2.6673 (1.5982)	2.7370 (0.9603)
	400	2.2486 (0.6623)	**1.9904** (0.4703)	2.5374 (0.5851)	1.5442 (0.7573)	**1.1253** (0.5565)	1.8445 (0.5609)
	600	1.7000 (0.4097)	**1.4141** (0.3470)	2.0619 (0.4625)	0.8772 (0.3284)	**0.6665** (0.2637)	1.1835 (0.3390)
	800	1.3451 (0.2915)	**1.0696** (0.2483)	1.7805 (0.3168)	0.6508 (0.1204)	**0.4899** (0.1274)	0.9860 (0.2238)
	1000	1.1298 (0.2391)	**0.9305** (0.5788)	1.6807 (0.2348)	0.5873 (0.1521)	**0.4379** (0.1410)	0.8380 (0.1490)

Table 5: Test MSE under varying interaction strength α for the CFM and FMI processes. Results are reported as mean with standard deviation in parentheses over 200 replications. Within each process and interaction level, the best-performing method is shown in bold.

	CFM			FMI		
α	EW	BEW	FPC	EW	BEW	FPC
0	**0.2707** (0.0640)	0.2838 (0.0617)	0.6377 (0.1358)	**0.1572** (0.0320)	0.1635 (0.0383)	0.2396 (0.0525)
2	1.0154 (0.2186)	**0.9830** (0.2052)	1.7400 (0.3415)	0.6880 (0.1923)	**0.6568** (0.3461)	0.9383 (0.2111)
4	2.8488 (0.5112)	**2.7574** (0.7120)	4.0543 (0.6623)	2.0374 (0.5558)	**1.8683** (1.0932)	2.5379 (0.6361)
6	5.8590 (1.2401)	**5.6430** (1.5122)	7.4608 (1.1310)	4.3521 (1.2711)	**3.8084** (1.8428)	5.2320 (1.1392)
8	10.0437 (1.8473)	**9.6664** (2.0573)	12.2034 (1.8637)	7.7307 (2.1362)	**6.8616** (3.0905)	8.7175 (2.0148)

Table 6: Test MSE under varying bottleneck width W_1 for the data generated under CFM process with interaction ($\alpha = 4.0$). Results are reported as mean with standard deviation in parentheses. The best-performing method in each row is shown in bold.

W_1	EW	BEW	FPC
1	2.8488 (0.5112)	**2.7424** (0.9269)	4.0543 (0.6623)
2	2.8439 (0.5015)	**2.6833** (0.7245)	4.0590 (0.6676)
4	2.8488 (0.5112)	**2.7594** (0.6845)	4.0543 (0.6623)
8	2.8488 (0.5112)	**2.6930** (0.5967)	4.0543 (0.6623)
16	2.8488 (0.5112)	**2.8344** (0.5876)	4.0543 (0.6623)
32	**2.8488** (0.5112)	2.8933 (0.5821)	4.0543 (0.6623)

Table 7: Test MSE under varying basis dimension D for the CFM process with interaction ($\alpha = 4.0$). Results are reported as mean with standard deviation in parentheses. The best-performing method in each row is shown in bold.

D	EW	BEW	FPC
2	2.9520 (0.6978)	**2.8488** (0.6311)	4.0543 (0.6623)
4	2.7001 (0.4887)	**2.6166** (0.5270)	4.0543 (0.6623)
8	3.1306 (0.5296)	**3.1035** (0.5990)	4.0543 (0.6623)
16	**3.9845** (0.6421)	4.1659 (0.7043)	4.0543 (0.6623)
32	5.3035 (0.7549)	5.5411 (0.8568)	**4.0543** (0.6623)
64	6.5119 (0.8475)	6.7610 (0.9538)	**4.0543** (0.6623)

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

(i). **Claims**

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: The abstract and introduction accurately characterize the proposed framework. These descriptions are aligned with the methodological development presented in Sections 2-4.

Guidelines:

- The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

(ii). **Limitations**

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: The paper concludes with a discussion of its limitations in Section 6.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

(iii). **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Under Assumptions A1-C3, we provide all mathematical proofs in Appendix A.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

(iv). **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

The experimental setup and evaluation metrics are provided in Appendix B, including the data-generating mechanism, network architectures, training procedure, and hyperparameter choices. The Python packages used to produce the results are also documented there. The paper also reports the network architectures of the proposed method, which further supports reproducibility.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

(v). **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The code and data will be released in our GitHub repository shortly after acceptance.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.

- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

(vi). **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: The evaluation protocol in Appendix B clearly outlines the procedure for train, validation and test splitting, hyperparameter selection with the Python package `Optuna`, and the details about the optimizer and the criterion for early stopping.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

(vii). **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Tables 2–4 in Appendix B report the results as means and standard deviations over 200 independent runs, thereby reflecting statistical variability.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

(viii). **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Section B describes the computing environment used in our experiments. Since the simulation runs were not computationally demanding, we do not report memory usage or execution times.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

(ix). **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: This work focuses on methodological development, theoretical analysis, and evaluation on standard benchmarks. It does not involve human subjects or clearly ethically sensitive applications, and we believe it is consistent with the NeurIPS Code of Ethics.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

(x). **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [No]

Justification: The paper focuses on the technical contributions and does not include a specific discussion of broader positive or negative societal impacts.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

(xi). **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [No]

Justification: The paper focuses on scalar-on-function regression. The proposed model and the data used in the study do not appear to present substantial misuse risks, and therefore do not require release safeguards beyond standard open-source practices.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

(xii). **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [No]

Justification: The paper appropriately acknowledges and cites the sources of existing assets, including baseline methods such as AddFNN in Yao et al. (2021).

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

(xiii). **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The primary new asset introduced by this work is the source code for the proposed method, which will be released as open source. Relevant documentation will be provided together with the code in our GitHub repository.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.

- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

(xiv). **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The research does not involve crowdsourcing experiments or research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

(xv). **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The research does not involve human subjects, therefore IRB approval is not applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

(xvi). **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs were used solely to assist with writing and editing.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.