Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: www.tandfonline.com/journals/uasa20

Deep Regression for Repeated Measurements

Shunxing Yan, Fang Yao & Hang Zhou

To cite this article: Shunxing Yan, Fang Yao & Hang Zhou (2025) Deep Regression for Repeated Measurements, Journal of the American Statistical Association, 120:552, 2461-2472, DOI: 10.1080/01621459.2025.2458344

To link to this article: https://doi.org/10.1080/01621459.2025.2458344

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2025, VOL. 120, NO. 552, 2461–2472: Theory and Methods
https://doi.org/10.1080/01621459.2025.2458344

Deep Regression for Repeated Measurements

Shunxing Yan[a], Fang Yao[a], and Hang Zhou[b]

[a]School of Mathematical Sciences, Center for Statistical Science, Peking University, Beijing, China; [b]Department of Statistics, University of California Davis, Davis, CA

ABSTRACT

Nonparametric mean function regression with repeated measurements serves as a cornerstone for many statistical branches, such as longitudinal/panel/functional data analysis. In this work, we investigate this problem using fully connected deep neural network (DNN) estimators with flexible shapes. A novel theoretical framework allowing arbitrary sampling frequency is established by adopting empirical process techniques to tackle clustered dependence. We then consider the DNN estimators for Hölder target function and illustrate a key phenomenon, the phase transition in the convergence rate, inherent to repeated measurements and its connection to the curse of dimensionality. Furthermore, we study several examples with low intrinsic dimensions, including the hierarchical composition model, low-dimensional support set and anisotropic Hölder smoothness. We also obtain new approximation results and matching lower bounds to demonstrate the adaptivity of the DNN estimators for circumventing the curse of dimensionality. Simulations and real data examples are provided to support our theoretical findings and practical implications. Supplementary materials for this article are available online, including a standardized description of the materials available for reproducing the work.

ARTICLE HISTORY
Received November 2023
Accepted January 2025

KEYWORDS
Clustered dependence; Fully connected ReLU neural networks; Functional data; Intrinsic dimension; Phase transition

1. Introduction

1.1. Literature Review

Repeated measurements collected from a sample of subjects have been extensively investigated in statistics, finding applications across fields like biomedicine, epidemiology, economics, and engineering. These observations typically possess dependence within subjects while being independent across subjects, also referred to as "clustered dependence." Consequently, several statistical branches have emerged, including longitudinal data analysis (Diggle et al. 2002; Weiss 2005; Hedeker and Gibbons 2006), panel data analysis (Chamberlain 1984; Baltagi 2021; Hsiao 2022), and functional data analysis (Ramsay and Silverman 2005; Ferraty and Vieu 2006; Hsing and Eubank 2015). Among these, estimating the mean function is a fundamental problem and serves as a crucial step for subsequent analyses.

Classical nonparametric regression methods (Rice and Rosenblatt 1983; Fan and Gijbels 1992) have addressed this for longitudinal data that typically have finite measurements per subject (i.e., finite sampling frequencies, Brumback and Rice 1998; Lin and Carroll 2000), while cases that allow the sampling frequency to infinity were established in the context of functional data by Cai and Yuan (2011) and Zhang and Wang (2016). This line of research on traditional methods spanning over 20 years reflects the challenges and significance of establishing a unified method with theory for various sampling schemes, especially exploring the effect of sampling frequencies and indicating how densely such measurements are needed to achieve the optimal rate of convergence. Moreover, traditional nonparametric methods are usually for univariate or low-dimensional covariates and suffer from the curse of dimensionality. Hence, estimators based on modern tools like neural networks are desired, given the success achieved by deep learning in recent years.

Deep neural networks (DNNs) with ReLU activation function have rapidly received increasing attention with a vast literature emerging, due to its computational power and approximation ability (Telgarsky 2016; Fan, Ma, and Zhong 2021). Yarotsky (2017), Schmidt-Hieber (2020), Yarotsky (2018), and Kohler and Langer (2021) have established non-asymptotic approximation bounds for sparsely connected or more implementable ReLU DNNs. Recently, Shen (2020), Lu et al. (2021) and Shen, Yang, and Zhang (2022) established the (nearly) optimal rate of approximation of fully connected neural networks in terms of both width and depth. A crucial preponderance of DNNs is the adaptability to various low intrinsic-dimensional structures that may circumvent curse of dimensionality. Without specific structure knowledge, DNNs automatically approximate the composition models (Bauer and Kohler 2019; Schmidt-Hieber 2020; Kohler and Langer 2021) that contain the common semiparametric models, and various smoothness settings (Barron 1993; Suzuki 2018; Suzuki and Nitanda 2021). Another perspective is that the predictor lies on a low-dimensional subset. Shaham, Cloninger, and Coifman (2018) used the 4-layer networks to approximate functions with suitable smoothness on low-dimensional manifolds. Schmidt-Hieber (2019), Chen et al. (2022), and Nakada and Imaizumi (2020) considered DNN estimators for more general smooth

CONTACT Fang Yao fyao@math.pku.edu.cn School of Mathematical Sciences, Center for Statistical Science, Peking University, 5 Yiheyuan Road, Beijing, China.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

functions with manifold or Minkowski dimension sets input, and Cloninger and Klock (2021) showed the adaptivity of DNN estimator to intrinsic dimensionality even beyond the domain for some cases. A comprehensive theory for fully connected DNN estimators of low-dimensional inputs was then derived in Jiao et al. (2023). Recently, Zhang et al. (2023) introduced a novel effective Minkowski dimension determining the regression rate.

Benefiting from the aforementioned advantages, DNNs are gaining attention as an emerging nonparametric approach in various statistical problems. There has been a series of work that successfully established theories of nonparametric DNN regression for independent observations (Suzuki 2018; Bauer and Kohler 2019; Schmidt-Hieber 2020; Farrell, Liang, and Misra 2021; Kohler and Langer 2021), which here is referred to as *cross-sectional* model/setting that measures a single observation for each subject and is distinguished from repeated measurements setting. Sarkar and Panaretos (2022) proposed a novel CovNet to estimate the covariance function for fully and grid-observed functional data, which approximates well and facilitates a direct eigendecomposition. Fan, Gu, and Zhou (2024) studied the Huber DNN estimator for heavy-tail and Zhong, Müller, and Wang (2021, 2022) considered survival models. Recently, some noteworthy works (Fan and Gu 2023; Bhattacharya, Fan, and Mukherjee 2023) considered the high dimensional problem, including factor augmented sparse throughput model and nonparametric interaction models with rigorous theory and matching lower bounds.

1.2. Challenges of DNN Regression for Repeated Measurements

Compared with the cross-sectional setting, the repeated measurements model is composed of n subjects (often referred to as sample size), and the ith individual is measured m_i times. Here we assume $m_i \asymp m$ for convenience, where $a_n \lesssim (\gtrsim) b_n$ indicates $a_n \leq (\geq) C b_n$ for some constant $C > 0$, and $a_n \asymp b_n$ means $a_n \lesssim b_n$ and $b_n \lesssim a_n$. To estimate the mean function, we adopt the pooling strategy that combines observations from all subjects. It works well for both sparse and dense designs and has been widely used in many contexts, such as random effects models, generalized estimating equation approaches (Zeger, Liang, and Albert 1988; Lin and Carroll 2000), and functional data models (Cai and Yuan 2011; Zhang and Wang 2016), among others.

Along with broad applicability and excellent properties of pooling repeated measurements, the clustered dependence structure, especially m can be finite or growing with the sample size n at any rate, brings theoretical difficulties, which renders the techniques and results of the DNN regression based on independent observations (Györfi et al. 2002; Koltchinskii 2006). Different from the kernel methods with closed-form estimates and spline/wavelet based methods with specific-structure basis, the DNN suffers greater challenges. Recently, Wang et al. (2021) studied this problem without generalization error, while a key assumption on the vanishing maximal eigenvalue of covariance matrix (as $m \to \infty$) contradicts its positive lower bound. Therefore, a new theoretical framework is desired for clustered dependent data to which existing empirical process results are not applicable.

Similar to the cross-sectional setting, the repeated measurements model also suffers from the curse of dimensionality. However, the latter behaves differently due to an interesting phenomenon, namely the *phase transition* in the sense that the convergence rate changes when the number of measurements grows from sparse to dense sampling schemes. Here, the curse of dimensionality is accompanied by larger errors under sparse design and the requirement for higher sampling frequency to achieve a faster and even a parametric rate. We shall study this phenomenon for DNN estimators in multi-dimensional and complex clustered structures, which would guide the practical sampling design for improving estimation efficiency in statistical scenarios.

One of our objectives is to circumvent the curse of dimensionality for repeated measurements as in cross-sectional settings. For this purpose, we need accurate characterizations of the DNN approximation ability in various function spaces that are not fully understood in the existing literature. For instance, the hierarchical composition model used in Kohler and Langer (2021) requires that the smoothness of each function is no less than one and the neural network has a specific shape, which limits its applicability. Moreover, there are no results available on the approximation of fully connected ReLU networks to anisotropic smoothness functions. These problems urge us to develop some new approximation results that are meaningful for both cross-sectional and repeated measurements settings.

1.3. Our Contributions

Motivated by the aforementioned challenges and issues, we establish a comprehensive framework for DNN regression in estimating mean functions for repeated measurements. In particular, we exploit a conditioning argument to decouple the randomness between and within subjects, which is the first attempt to adopt and thus allow the use of empirical process tools for general clustered dependence. This can also be used for further theoretical analysis in relevant models and problems. The main contributions of this article are summarized as follows.

1. We develop a series of oracle bounds on prediction error for estimators based on minimizing the empirical least squares loss function in the repeated measurements setting. To the best of our knowledge, it is the first work in this setting and may serve as a cornerstone for developing theories of various statistical models coupled with DNNs.
2. We derive the upper bound on the prediction error using fully connected DNNs (and also regression splines as a by-product) when the true function belongs to the Hölder space, as well as the matching minimax lower bound. It follows that a phase transition phenomenon occurs for DNN estimators, which reveals the impact of dimensionality on the convergence rate in a discrepant way with cross-sectional models.
3. To circumvent the curse of dimensionality, we study several representative low intrinsic-dimensional structures, including the hierarchical composition model, low dimensional input and anisotropic Hölder smoothness, and analyze the

Table 1. A summary of convergence rates for ReLU DNN estimators for various function classes and sampling frequencies.

Function class	Cross-sectional Regression ($m = 1$)	Repeated measures regression ($m \geq 2$)		
		Sparse	Phase transition	Dense
Hölder	$n^{-2s/(2s+d)}$ (Kohler and Langer 2021)	$(nm)^{-2s/(2s+d)}$	$m \asymp n^{d/2s}$	n^{-1}
Hierarchical	$n^{-2\gamma/(2\gamma+1)}$ (Kohler and Langer 2021)	$(nm)^{-2\gamma/(2\gamma+1)}$	$m \asymp n^{1/2\gamma}$	n^{-1}
Manifold Supp	$n^{-2s/(2s+d_{\mathcal{M}})}$ (Jiao et al. 2023)	$(nm)^{-2s/(2s+d_{\mathcal{M}})}$	$m \asymp n^{d_{\mathcal{M}}/2s}$	n^{-1}
Anisotropic	$n^{-2\bar{s}/(2\bar{s}+d)}$ (sparse net Suzuki and Nitanda 2021)	$(nm)^{-2\bar{s}/(2\bar{s}+d)}$	$m \asymp n^{d/2\bar{s}}$	n^{-1}

NOTE: All rates are minimax optimal.
[a] Rates for $m \geq 2$ are established in this work. Here, "Phase transition" indicates the order when phase transition in the convergence rate occurs. "Sparse" refers to sampling frequencies m sparser than the order where phase transition occurs, while "Dense" pertains to those denser.

minimax optimal rates and phase transition phenomena, respectively. We also develop some new results having their own merits even for cross-sectional settings, for example, representing the hierarchical composition model to allow the smoothness smaller than one and deriving a new approximation result for anisotropic Hölder smooth functions.

1.4. Organization

The rest of the article is organized as follows. The model settings, estimators and key oracle inequalities are presented in Section 2, based on which Section 3 presents the convergence analysis for Hölder space. We then provide in Section 4 two useful examples to circumvent the curse of dimensionality. Section 5 illustrates the theoretical findings via simulation experiments and Section 6 presents two real data examples, while further discussion is given in Section 7. The additional theoretical and numerical results, all proofs and auxiliary lemmas are deferred to the supplementary material for space economy.

2. Model Estimation and Oracle Inequalities

2.1. Model and Estimation

Given a random function $Z(x)$ defined on $[0,1]^d$. Let $Z_1(x)$, $Z_2(x), \ldots, Z_n(x)$ be iid realizations. In practice, these realizations cannot be fully observed and are measured for the ith subject at m_i points as $X_{i1}, X_{i2}, \ldots, X_{im_i}$, inevitably contaminated with additive noise. Under random design assumption, we assume that $X_{ij}, 1 \leq i \leq n, 1 \leq j \leq m_i$ are iid from a distribution $\mathcal{P}_X$ which is supported on some $\Omega \subseteq [0,1]^d$ and independent of $Z_i, 1 \leq i \leq n$. Here, Ω is assumed to have a well-defined volume measure and $\mathcal{P}_X$ is absolutely continuous with respect to the volume measure. Let $f^\circ(x) = \mathbb{E}[Z(X)|X = x]$ be the regression function (i.e., the mean structure) of interest. Then the repeated measurement responses follow

$$\begin{aligned} Y_{ij} &= Z_i(X_{ij}) + \epsilon_{ij} \\ &= f^\circ(X_{ij}) + U_i(X_{ij}) + \epsilon_{ij}, \qquad 1 \leq i \leq n, 1 \leq j \leq m_i, \end{aligned} \tag{1}$$

where $U_i(x) = Z_i(x) - f^\circ(x)$ are individual stochastic parts and the ϵ_{ij} are independent noise variables. A key feature of this repeated measurements model is that observations from the same subject are correlated/dependent, while those from different subjects are independent, which is distinct from the cross-sectional setting based on independent observations. In the sequel, to simplify exposition, we assume that each subject is measured m times, that is, $m_1 = \cdots = m_n = m$ (also referred to as sampling frequency).

With the available data $\{(Y_{ij}, X_{ij}) : 1 \leq i \leq n, 1 \leq j \leq m\}$, our target is to estimate the mean function $f^\circ(x)$. We pool observations from all subjects for estimation, and define the nonparametric least square estimator within a suitably chosen function class $\mathcal{F}_{nm}$, minimizing the empirical risk as follows,

$$\hat{f}_{nm} \in \arg\min_{f \in \mathcal{F}_{nm}} \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \big(Y_{ij} - f(X_{ij})\big)^2. \tag{2}$$

To evaluate the accuracy for an estimator $\hat{f}_{nm}$ quantitatively, we introduce the mean squared prediction error

$$\mathcal{E}_{f^\circ}(\hat{f}_{nm}) := \mathbb{E}\left[\int_\Omega (\hat{f}_{nm}(x) - f^\circ(x))^2 d\mathcal{P}_X\right],$$

which is also equal to the excess risk $\mathbb{E}[|Y - \hat{f}_{nm}(X)|^2] - \mathbb{E}[|Y - f^\circ(X)|^2]$.

In (2), the candidate class $\mathcal{F}_{nm}$ plays a central role in estimation procedures for the tradeoff of approximation error and estimation error. In this article, we mainly consider the neural network class. A ReLU feedforward neural network with architecture $(L, (W_L, W_{L-1}, \ldots, W_0))$ can be written as a composition of a series of functions

$$\begin{aligned} f(x) = \mathcal{L}_L \circ \mathrm{ReLU} \circ \mathcal{L}_{L-1} \circ \mathrm{ReLU} \circ \mathcal{L}_{L-2} \circ \mathrm{ReLU} \circ \cdots \\ \circ \mathcal{L}_1 \circ \mathrm{ReLU} \circ \mathcal{L}_0(x), \end{aligned} \tag{3}$$

where $\mathcal{L}_i(x) = L_i x + b_i$ is a multivariate linear function with $L_i \in \mathbb{R}^{W_{i+1} \times W_i}, b_i \in \mathbb{R}^{W_i}, x \in \mathbb{R}^{W_0}$, and the operator ReLU : $\mathbb{R}^{W_i} \to \mathbb{R}^{W_i}$ applies rectified linear unit (ReLU) function $a \mapsto \max\{a, 0\}$ to each component of its input. Then we define

$$\begin{aligned} \mathcal{FNN}(d, L, W) = \Big\{ f : \mathbb{R}^d \to \mathbb{R} \text{ is of the form (3) with} \\ W_0 = d \text{ and } W_i \leq W, 1 \leq i \leq L \Big\}, \end{aligned}$$

and

$$\begin{aligned} \mathcal{FNN}(d, L, W, \beta) &= \mathcal{T}_\beta \mathcal{FNN}(d, L, W) \\ &= \big\{ \mathcal{T}_M f : f \in \mathcal{T}_\beta \mathcal{FNN}(d, L, W) \big\}, \end{aligned}$$

where $\mathcal{T}_\beta f(x) = f(x) I(|f(x)| \leq \beta) + \mathrm{sign}(f(x)) \beta I(|f(x)| > \beta)$ is the truncation operator. Specifically, we call L and W the depth and width of the neural networks, respectively.

2.2. Oracle Inequalities

We first establish some universal bounds on the mean squared prediction error of the proposed estimators without imposing specific $\mathcal{F}_{nm}$. Denote the infinity norm $\|f\|_\infty = \sup_{x\in\Omega}|f(x)|$ for a function f defined on Ω. We make the following assumptions.

Assumption 1 (Regression function and candidate class). There exist some positive number B_1, such that the target function $\|f^\circ\|_\infty \le B_1$ and $\|f\|_\infty \le B_1$ for each $f \in \mathcal{F}_{nm}$.

Assumption 2 (Random process and noise). The random processes $U_i(x)$ are continuous, and there exist nonnegative numbers B_2 and B_3, such that for $1 \le i \le n, 1 \le j \le m$,

$$\mathbb{E}\left[\exp\left\{\left|\frac{U_i(X_{ij})}{B_2}\right|\right\}\right] \le 1, \quad \mathbb{E}\left[\exp\left\{\left|\frac{\epsilon_{ij}}{B_3}\right|\right\}\right] \le 1. \tag{4}$$

Assumption 1 is a standard condition in empirical process and nonparametric regression. For example, it guarantees a probability bound of $\mathcal{E}_{f^\circ}(\hat{f}_{nm})$ by its empirical version. To satisfy this assumption, the candidate functions are often truncated. Remarkably, our later results are non-asymptotic with respect to B_1, so it can be relaxed as an increasing sequence with respect to mn when the exact bound of $\|f^\circ\|_\infty$ is unknown. Assumption 2 that requires sub-exponential random parts is also standard.

Now we specify a condition to characterize the complexity of $\mathcal{F}_{nm}$. Define $\mathcal{F}_{nm}(r) = \left\{f \in \mathcal{F}_{nm} : \int_\Omega \left(f(x) - f^\circ(x)\right)^2 d\mathcal{P}_X \le r\right\}$. We say a function $\phi_{nm}(r)$ is sub-root if and only if $\phi_{nm}(r)/\sqrt{r}$ is nonnegative and nonincreasing. Let the iid Rademacher variables $\{\sigma_{ij}, 1 \le i \le n, 1 \le j \le m\}$ be uniformly chosen from $\{-1, +1\}$. For mathematical rigor, we introduce the definition $\mathbb{E}\sup_{f\in\mathcal{F}} V(f) = \sup\left\{\mathbb{E}\sup_{f\in\mathcal{F}^*} V(f) : \mathcal{F}^* \subset \mathcal{F}, \mathcal{F}^* \text{ is finite}\right\}$ for a random process V when the index set $\mathcal{F}$ is uncountable.

Definition 2.1 (Rademacher fixed point). Let r^*_{nm} be a positive number and $\phi_{nm}(r)$ be a sub-root function for $r \ge r^*_{nm}$. Assume that $\phi_{nm}(r^*_{nm}) \le r^*_{nm}$ and

$$\phi_{nm}(r) \ge \mathbb{E}\left[\sup_{f\in\mathcal{F}_{nm}(r)}\left|\frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\sigma_{ij}\left(f(X_{ij}) - f^\circ(X_{ij})\right)\right|\right]$$

for $r \ge r^*_{nm}$. Then r^*_{nm} is called the Rademacher fixed point of $\mathcal{F}_{nm}$.

The positive number r^*_{nm} is usually referred to as the fixed point of the continuity modulus of the Rademacher average that characterizes the neighborhood complexity of f° in $\mathcal{F}_{nm}$. In the empirical process with independent random variables (Bartlett, Bousquet, and Mendelson 2005; Koltchinskii 2006), the fixed point is found to be of the same order as the convergence rate of the estimators that minimize an empirical loss criterion over a given function class. Our first main result in the next theorem establishes a non-asymptotic bound on the prediction error in the repeated measures model using r^*_{nm}.

Theorem 2.1. Consider the repeated measures model (1) and the estimator $\hat{f}_{nm}$ obtained by (2). Under Assumptions 1 and 2, there is a universal constant c_1 such that

$$\mathcal{E}_{f^\circ}(\hat{f}_{nm}) \le c_1\left(\inf_{f\in\mathcal{F}_{nm}}\int_\Omega \left(f - f^\circ\right)^2 d\mathcal{P}_X + \frac{B_1^2 + B_2^2}{n} + (r^*_{nm} + \frac{1}{nm})(B_1^2 + B_2^2(\log n)^2 + B_3^2(\log nm)^2)\right). \tag{5}$$

The result above can also take optimization error into account using standard techniques, and Remark S.5.1 in the supplementary material discusses the lower and upper bounds when the estimator is not the global minimizer. By Theorem 2.1, the prediction error of an empirical risk minimizer $\hat{f}_{nm}$ can be bounded by the approximation error $\inf_{f\in\mathcal{F}_{nm}}\int_\Omega \left(f - f^\circ\right)^2 d\mathcal{P}_X$ and the estimation error $(B_1^2 + B_2^2)n^{-1} + (r^*_{nm} + (nm)^{-1})(B_1^2 + B_2^2(\log n)^2 + B_3^2(\log nm)^2)$. Compared with the cross-sectional model

$$Y_i = f^\circ(X_i) + \epsilon_i, \qquad 1 \le i \le n, \tag{6}$$

which can be viewed as a special case of (1) that measures a single observation for each subject. Accordingly, letting $m = 1$ and $B_2 = 0$ yields a parallel result for model (6). We remark that the term in (5) involving n^{-1} plays an interesting role in the repeated measurements model, coupling with the terms $\inf_{f\in\mathcal{F}_{nm}}\int_\Omega \left(f - f^\circ\right)^2 d\mathcal{P}_X + (r^*_{nm} + (nm)^{-1})(\log nm)^2$ depending on the total number of observations nm. To minimize the right-hand side (r.h.s.) of (5), we choose the function class $\mathcal{F}_{nm}$ that usually expands with the sample size to balance the tradeoff between two errors. Then the upper bound on the r.h.s. can be shown of the order $n^{-1} + (nm)^{-\alpha}$ up to logarithmic factors for specific problems. This reveals some fundamentally different behavior of nonparametric regression estimators in the repeated measurements model. No matter how large m is, even when the n functions are fully observed, one cannot make the prediction error faster than the order of n^{-1} that is customarily called the parametric rate in terms of sample size.

The logarithmic terms in (5) are caused by technical issues due to the sub-exponential tails in Assumption 2, and will vanish if ones assume bounded variables as in Györfi et al. (2002). A similar result could be also given with sub-Gaussian assumptions and is omitted. It is worth noting that the bound B_1 for the target function f° and $f \in \mathcal{F}_{nm}$ can be relaxed as a sequence that slowly grows with nm, especially when one has no prior information about such functions. For example, we can set $B_1 = \log n$ for the chosen function class $\mathcal{F}_{nm}$ as in (S.3) in the supplementary material, and the convergence rate in the result remains nearly unchanged except for a logarithmic factor. In the sequel, we do not distinguish whether B_1 is finite or increases with mn when no confusion arises. For convenience, we denote $\iota_{nm} = B_1^2 + B_2^2(\log n)^2 + B_3^2(\log nm)^2$.

Next, we extend the result in Theorem 2.1 for several situations of interest. The reason is that the bound (5) uses the Rademacher fixed points for $\mathcal{F}_{nm}$ which is an excellent distribution-dependent tool for measuring the complexity of the candidate function class. However, this can only be computed when the distribution is known, which is often not the case in practice. One proposal is to estimate r^*_{nm} by an empirical version like Theorem 4.2 in Bartlett, Bousquet, and Mendelson (2005) that is data-dependent. Alternatively, a feasible resolution is to derive an upper bound using additional knowledge about the complexity of the class.

Corollary 2.1. Under the same model and assumptions as Theorem 2.1, we have

(i) random/uniform covering numbers:

$$\mathcal{E}_{f^\circ}(\hat{f}_{nm}) \le c_2 \left(\inf_{f\in\mathcal{F}_{nm}} \int_\Omega (f-f^\circ)^2 \, d\mathcal{P}_X + \frac{B_1^2+B_2^2}{n} + \frac{\iota_{nm}}{nm} \left(\mathbb{E}\left[\log \mathcal{N}\left(\frac{B_1}{nm}, \mathcal{F}_{nm}|_{\mathcal{X}}, L^2 \right) \right] + \log nm \right) \right),$$

and the bound in uniform covering number is of the same form, but with the expectation of $\log \mathcal{N}\left((nm)^{-1}B_1, \mathcal{F}_{nm}|_{\mathcal{X}}, L^2\right)$ replaced by its supremum $\log \mathcal{N}_{nm}\left((nm)^{-1}B_1, \mathcal{F}_{nm}, L^2\right)$.

(ii) VC dimension:

$$\mathcal{E}_{f^\circ}(\hat{f}_{nm}) \le c_3 \left(\inf_{f\in\mathcal{F}_{nm}} \int_\Omega (f-f^\circ)^2 \, d\mathcal{P}_X + \frac{B_1^2+B_2^2}{n} + \frac{\mathrm{VCdim}(\mathcal{F}_{nm})\iota_{nm}\log nm}{nm} \right).$$

where c_2 and c_3 are universal constants.

Lastly, we present the case that $\mathcal{F}_{nm}$ is a set of truncated fully connected ReLU feedforward neural networks $\mathcal{FNN}(d, L, W, B_1)$, which demonstrates how the width and depth of the network determine the estimation error explicitly.

Corollary 2.2. Under the same model and assumptions as Theorem 2.1 and set the function class $\mathcal{F}_{nm} = \mathcal{FNN}(d, L, W, B_1)$ with $\log LW \ge 1$, then

$$\mathcal{E}_{f^\circ}(\hat{f}_{nm}) \le c_4 \left(\inf_{f\in\mathcal{F}_{nm}} \int_\Omega (f-f^\circ)^2 \, d\mathcal{P}_X + \frac{B_1^2+B_2^2}{n} + \frac{L^2W^2\log LW\iota_{nm}\log nm}{nm} \right),$$

where c_4 is a positive constant.

Since this result is the first in the repeated measurement setting, we may only compare it with existing results for the cross-sectional setting (6) by letting $m = 1$. First, our error bounds work for sub-exponential noise, while Györfi et al. (2002) assumes the response to be bounded and Schmidt-Hieber (2020) requires the noises to be not heavier than Gaussian. Second, the estimation error in our results is bounded in various ways, including Rademacher fixed point, random/uniform covering numbers and VC dimension. This allows easy extensions for future works, even if $\mathcal{F}_{nm}$ is not a feedforward neural network class. In particular, these results are usually sharper than the functional covering number which leads to the further boundedness requirement of network weights and bias.

3. Hölder Space and Phase Transition

In this section, we apply these oracle inequalities to the situation that the target function belongs to the Hölder space that has general smoothness assumptions and plays an important role in nonparametric regression problems. We first establish a theoretical bound of the DNN estimator considered above, then demonstrate a matching lower bound and analyze its optimal convergence rate, while the phase transition owing to repeated measurements is revealed. Assume $\Omega = [0,1]^d$ in this section for brevity. Recall the definition of Hölder space.

Definition 3.1 (Hölder space $\mathcal{C}^s(\Sigma)$). Let s be a positive number and $\lfloor s \rfloor$ be the largest integer strictly smaller than s. The s-Hölder norm of a function $f : \Sigma \to \mathbb{R}$ is defined as

$$\|f\|_{\mathcal{C}^s} = \sum_{\boldsymbol{\alpha}:|\boldsymbol{\alpha}|<s} \left\|\partial^{\boldsymbol{\alpha}} f\right\|_\infty + \sum_{\boldsymbol{\alpha}:|\boldsymbol{\alpha}|=\lfloor s\rfloor} \sup_{\substack{x,x'\in\Sigma \\ x\neq x'}} \frac{\left|\partial^{\boldsymbol{\alpha}} f(x) - \partial^{\boldsymbol{\alpha}} f(x')\right|}{\|x-x'\|^{s-\lfloor s\rfloor}},$$

where $\partial^\alpha := \partial_1^{\alpha_1}\partial_2^{\alpha_2}\ldots\partial_d^{\alpha_d}$ and $|\boldsymbol{\alpha}| := \alpha_1+\alpha_2+\cdots+\alpha_d$ with $\boldsymbol{\alpha} = (\alpha_1,\ldots,\alpha_d) \in \mathbb{N}^d$. Hölder space $\mathcal{C}^s(\Sigma)$ consists of all functions for which the s-Hölder norm is finite. Typical examples include Lipschitz functions and continuously differentiable functions on a compact set.

Assumption 3 (Hölder smoothness). The target function f° belongs to the Hölder space $\mathcal{C}^s\left([0,1]^d\right)$ for a given positive number s, and $\|f^\circ\|_{\mathcal{C}^s} \le B_4$ for some constant $B_4 > 0$.

The non-asymptotic error bound of the ReLU feedforward neural network estimator is determined by a tradeoff between the estimation error and the approximation error in Corollary 2.2, associated with the depth and width. Combining Corollary 2.2 and the approximation error in Jiao et al. (2023), we arrive at the following result.

Theorem 3.1. Consider the model (1) and the estimator $\hat{f}_{nm}$ obtained by (2). Suppose that Assumptions 1–3 hold, and set $\mathcal{F}_{nm} = \mathcal{FNN}\left(d, c_5 L\log L, c_6 W\log W, B_1\right)$ with c_5, c_6 not depending on L, W. If we specify the neural network in a flexible way satisfying $LW = \lfloor c_7 (nm)^{d/(4s+2d)} (\log nm)^{-4d/(2s+d)} \rfloor$ for some constant c_7 not depending on n and m, then

$$\mathcal{E}_{f^\circ}(\hat{f}_{nm}) \le c_8 (n^{-1} + (nm)^{-2s/(2s+d)} (\log nm)^{16s/(2s+d)}),$$

where c_8 is a constant free of n and m.

Let $\mathcal{P}_{Z,\epsilon}$ be the probability measure for the random function Z and noise ϵ, and $\mathcal{Q}_{Z,\epsilon}$ be a collection of $\mathcal{P}_{Z,\epsilon}$ satisfying some certain conditions. The next theorem provides a lower bound on the minimax convergence rate for repeated measurement regression estimators.

Theorem 3.2. Consider the repeated measures model (1) and assume that $\mathcal{P}_X$ has a uniform distribution on Ω. Under the Assumptions 2 and 3, we have

$$\inf_{\tilde{f}} \sup_{\mathcal{P}_{Z,\epsilon}\in\mathcal{Q}_{Z,\epsilon}} \mathcal{E}_{f^\circ}(\tilde{f}) \ge c_9 (n^{-1} + (nm)^{-2s/(2s+d)}),$$

where the infimum is taken over all possible estimators $\tilde{f}$ based on the available dataset $\left\{\left(Y_{ij}, X_{ij}\right) : 1 \le i \le n, 1 \le j \le m\right\}$, the supremum is taken over all distribution $\mathcal{P}_{Z,\epsilon} \in \mathcal{Q}_{Z,\epsilon}$ satisfying Assumptions 2 and 3, and c_9 is a positive number not depending on n and m.

Combining the two theorems yields that the optimal convergence rate is of the order $O(n^{-1} + (nm)^{-2s/(2s+d)})$ (up to logarithmic factors) for Hölder smoothness index $s \in \mathbb{R}_+$. It matches the existing minimax rate $n^{-2s/(2s+d)}$ in cross-sectional case (Stone 1980; Schmidt-Hieber 2020; Kohler and Langer 2021) by setting $m = 1$, and $n^{-1} + (nm)^{-2s/(2s+1)}$ in repeated measures regression for RKHS with $O(k^{-2s})$ eigendecay (Cai and Yuan 2011) and $n^{-1} + (nm)^{-4/5}$ for local linear regression by assuming $s = 2, d = 1$ (Zhang and Wang 2016). Intriguingly, we observe a phase transition phenomenon in the minimax convergence rate. To see this, when the sampling frequency $m \gtrsim n^{d/2s}$, the rate reaches n^{-1} which is usually called the parametric rate of convergence in terms of the sample size n. After that, no matter how many more repeated measurements are collected, the order n^{-1} cannot be improved, even when the n functions are fully observed which is considered the ideal situation in functional data analysis. On the other hand, When $m \lesssim n^{d/2s}$ that is viewed as the sparse sampling scheme, the convergence is slowed to an of order $(nm)^{-2s/(2s+d)}$ that depends on the total number of observations nm and in accordance to the so-called nonparametric rate. For $d = 1$, the phenomenon was studied in integer Sobolev smoothness (Cai and Yuan 2011) and continuously differentiable functions (Zhang and Wang 2016), as well as in Riemannian manifold space (Shao, Lin, and Yao 2022) and online estimation setting (Yang and Yao 2023).

This phase transition reveals the impact of dimensionality on the convergence rate in two folds. First, when the sample size is n, the optimal rate n^{-1} is not improvable. To achieve this rate, the sampling frequency should be of order $m \asymp n^{d/2s}$. A higher dimension d requires a much larger number of measurements that has to grow exponentially in terms of the dimension d. Second, when the subjects are sparsely sampled, the rate $(nm)^{-2s/(2s+d)}$ would be fairly slow when the dimension is relatively large compared to smoothness, hence, the curse of dimensionality behaves similarly to the cross-sectional setting with nm independent observations. We close the section by mentioning that the same technique can be applied to other commonly used $\mathcal{F}_{nm}$ to establish the associated optimal convergence analysis, which is also illustrated for regression spline estimators in Section S.2 in the supplementary material.

4. Circumventing the Curse of Dimensionality

In modern statistical modeling and machine learning, the prevalence of multi-dimensional data has posed the formidable challenge of the curse of dimensionality, particularly salient for nonparametric regression. Our previous findings indicate that while the convergence of DNN-based estimators in repeated measurements differs from that in cross-sectional context, the curse of dimensionality persists in smoothness class like Hölder space, which requires a much larger sampling frequency to achieve the optimal parametric rate. In this work, we examine several low intrinsic-dimensional function spaces, encompassing the hierarchical composition model, manifold input and anisotropic Hölder smoothness. For brevity, this section only shows the results for the first two cases, while additional results are given in Section S.1.2 in supplementary material. Demonstrably, DNN estimators exhibit adaptability to these structures without prior knowledge and specialized design, thereby mitigating the curse of dimensionality to a certain extent.

4.1. Hierarchical Composition Structure

The classical approach in statistics to circumvent the curse of dimensionality is to make semiparametric model assumptions, such as single index models, generalized additive models and partially linear models, so that faster convergence can be achieved. In fact, these model assumptions can be subsumed within a broader framework of the hierarchical composition structure (Bauer and Kohler 2019; Schmidt-Hieber 2020; Kohler and Langer 2021) that has optimal cross-sectional regression rate $n^{-2\gamma/(2\gamma+1)}$ with some index γ.

Now, we introduce the definition of the Hierarchical composition model $\mathcal{H}^{l,\mathcal{G}}(\Sigma)$ in this article.

Definition 4.1 (Hierarchical composition model $\mathcal{H}^{l,\mathcal{G}}(\Sigma)$). Suppose $\Sigma \subseteq \mathbb{R}^d, l \in \mathbb{N}$ and $\mathcal{G}$ is a l height tree with nodes belonging to $(0,\infty)\times\mathbb{N}$, then the hierarchical composition model $\mathcal{H}^{l,\mathcal{G}}(\Sigma)$ is recursively defined as follows. For $l = 1$ and $\mathcal{G} = ((s,K))$,

$$\mathcal{H}^{l,\mathcal{G}}(\Sigma) = \Big\{ f : \Sigma \to \mathbb{R} : f(x) = g(x_{\pi(1)}, x_{\pi(2)}, \ldots, x_{\pi(K)}), \text{ where } g : \mathbb{R}^K \to \mathbb{R} \text{ satisfies } \|g\|_{\mathcal{C}^s} \le M \text{ and } \pi : [K] \to [d] \Big\}.$$

For $l \ge 2$, given $\mathcal{G} = ((s,K); \mathcal{G}_1, \ldots, \mathcal{G}_K)$ with $(s,K) \in (0,\infty) \times \mathbb{N}$,

$$\mathcal{H}^{l,\mathcal{G}}(\Sigma) = \Big\{ f : \Sigma \to \mathbb{R} : f(x) = g\left(h_1(x), h_2(x), \ldots, h_K(x)\right), \text{ where } g : \mathbb{R}^K \to \mathbb{R} \text{ satisfies } \|g\|_{\mathcal{C}^s} \le M \text{ and } h_k \in \mathcal{H}^{l_k,\mathcal{G}_k}(\Sigma) \text{ with } l_k \le l-1 \text{ for } 1 \le k \le K \Big\}.$$

Here, M is a positive constant and $x_{\pi(j)}$ is the $\pi(j)$th component of x.

Note that our definition of the hierarchical composition model contains more information than that in Kohler and Langer (2021). Due to the difficulty in deriving the convergence upper bound, Theorem 1 in Kohler and Langer (2021) demands that the smoothness of functions in each layer should be no less than 1, which limits its applicability. Therefore, in the above definition, we introduced $\mathcal{G}$ with a tree data structure to store the smoothness and dimension of functions at each level to deal with the transport of sub-smoothness in the composition. Further, we define the intrinsic smoothness-dimension ratio γ by recursion. Initially, let $\gamma = s/K$ for $l = 1$ and $\mathcal{G} = ((s,K))$. For $l \ge 2$, assume the root node of $\mathcal{G}$ is (s,K) and $\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_K$ are sub-trees whose roots are children of (s,K). Then we define

$$\gamma = \min\left(s/K, \min_{1\le k\le K} \gamma_k \cdot \min(1,s) \right),$$

where γ_k are the intrinsic smoothness-dimension ratios of $\mathcal{H}^{l-1,\mathcal{G}_k}(\Sigma)$. Alternatively, the intrinsic smoothness-dimension ratio could be calculated directly. For each node $G = (s_G, K_G)$ of $\mathcal{G}$, define the effective smoothness index as $s_G^* = s \cdot \prod_{(s',K')\in\mathcal{A}(G)} \min(1,s')$, where $\mathcal{A}(G)$ denotes the ancestor set of G in the tree $\mathcal{G}$. One can verify that $\gamma = \min_{G\in\text{node}(\mathcal{G})} s_G^*/K_G$, where the minimum value is taken over all nodes G of tree $\mathcal{G}$.

We derive a new result that characterizes the approximation ability of deep ReLU feedforward neural network to the hierarchical composition function, which sets the stage for studying the error bound and optimal rate of convergence.

Theorem 4.1. Suppose that $f \in \mathcal{H}^{l,\mathcal{G}}(\Omega)$, $\|f\|_\infty \le B_1$ and γ is defined as above. Then, for any $L, W \ge 3$, there exists some neural network $f_{\mathcal{FNN}} \in \mathcal{FNN}\left(d, c_{10}L\log L, c_{11}W\log W, B_1\right)$ satisfying

$$|f_{\mathcal{FNN}}(x) - f(x)| \le c_{12}\,(LW)^{-2\gamma},$$

for every $x \in \Omega$, where c_{10} and c_{11} are constants not depending on L, W, and f, and c_{12} is a constant free of L and W.

Now we assume the target function f° has such a hierarchical structure, and arrive at the following non-asymptotic error bound for repeated measurements model by coupling with Corollary 2.2, which shows how this alleviates the curse of dimensionality.

Assumption 4 (Hierarchical structure). The target function $f^\circ(x)$ belongs to the hierarchical composition model $\mathcal{H}^{l,\mathcal{G}}(\Omega)$ for given l, $\mathcal{G}$, and M.

Theorem 4.2. Consider model (1) and the estimator $\hat{f}_{nm}$ obtained by (2). Suppose that Assumptions 1, 2, and 4 hold, and set $\mathcal{F}_{nm} = \mathcal{FNN}\left(d, c_5 L\log L, c_6 W\log W, B_1\right)$. If we specify the neural network in a flexible way satisfying $LW = \lfloor c_{13}(nm)^{1/(4\gamma+2)}(\log nm)^{-4/(2\gamma+1)}\rfloor$ for some constant c_{13} not depending on n and m, then

$$\mathcal{E}_{f^\circ}(\hat{f}_{nm}) \le c_{14}(n^{-1} + (nm)^{-2\gamma/(2\gamma+1)}(\log nm)^{16\gamma/(2\gamma+1)}),$$

where c_{14} is a constant free of n and m.

The following lower bound is constructed for the repeated measurement setting and the hierarchical composition model.

Theorem 4.3. Consider the repeated measures model (1) and assume that $\mathcal{P}_X$ has a uniform distribution on Ω. Assume that there exists a node $G = (s_G, K_G)$ such that $s_G^*/K_G = \gamma$ and $K_G \le d$. Under the Assumptions 2 and 4, we have

$$\inf_{\tilde{f}} \sup_{\mathcal{P}_{Z,\epsilon} \in \mathcal{Q}_{Z,\epsilon}} \mathcal{E}_{f^\circ}(\tilde{f}) \ge c_{15}(n^{-1} + (nm)^{-2\gamma/(2\gamma+1)}),$$

where the infimum is taken over all possible estimators $\tilde{f}$ based on the available dataset $\left\{\left(Y_{ij}, X_{ij}\right) : 1 \le i \le n, 1 \le j \le m\right\}$, the supremum is taken over all distributions $\mathcal{P}_{Z,\epsilon} \in \mathcal{Q}_{Z,\epsilon}$ satisfying Assumptions 2 and 4, and c_{15} is a positive number not depending on n and m.

The previous two theorems demonstrate that the minimax optimal rate in the hierarchical composition model becomes $O(n^{-1} + (nm)^{-2\gamma/(2\gamma+1)})$ up to some logarithmic factors. That means the intrinsic smoothness-dimension ratio γ, rather than the extrinsic smoothness-dimension ratio s/d, determines the convergence rates under the assumption of the representational function set $\mathcal{H}^{l,\mathcal{G}}(\Omega)$, while one typically has $\gamma < s/d$. Accordingly, the phase transition phenomenon inherent to repeated measurements takes place at the order $m \asymp O(n^{1/2\gamma})$ instead of $O(n^{d/2s})$. Hence, the curse of dimensionality is lessened, and exponentially less sample frequency is required to achieve the parametric rate.

In traditional works, hierarchical structure often appears as model assumptions in one of its specific forms, such as assuming that the true function satisfies a single-index model, additive model, varying coefficient model, or their combinations. Given a particular structure, tailored estimation methods are designed. A closely related work is Horowitz and Mammen (2007) which used the penalized least squares method to estimate the known composition of the additive model. For the adaptivity of traditional tools to unknown hierarchical structures, Schmidt-Hieber (2020) demonstrated that wavelet series do not adapt to unknown hierarchical structures. Suzuki and Nitanda (2021) extended this conclusion on non-adaptivity to more general linear estimators, including local polynomials, RKHS, regression splines, and wavelets. For completeness, we also show that the linear estimators cannot be adaptive to hierarchical models without specific knowledge and designs in the repeated measurements model, even if the true function is of a single index form; see Section S.1.1 in supplementary material for more details.

4.2. Low-Dimensional Support

Considering predictors in a low-dimensional structure is sensible in many applications, such as image processing, natural language processing and bio-medical studies. There is rich literature to study regression on an unknown manifold $\mathcal{M}$, especially for local polynomial methods (Bickel and Li 2007; Aswani, Bickel, and Tomlin 2011; Cheng and Wu 2013). Bickel and Li (2007) revealed that local polynomial regression can adapt to the unknown low dimensional structure and achieve the optimal convergence rate $O(n^{-2s/(2s+d_{\mathcal{M}})})$, Aswani, Bickel, and Tomlin (2011) and Cheng and Wu (2013) learned the manifold structure first and consider further estimation. Recent works in the cross-sectional regression have shown that DNNs are also adaptive to intrinsic structures (e.g., Schmidt-Hieber 2019; Nakada and Imaizumi 2020; Chen et al. 2022; Jiao et al. 2023), classification (e.g., Liu et al. 2021; Zhang et al. 2024), and more advanced scenarios (e.g., Dahal et al. 2022; Xu et al. 2024). The previous regression works are all cross-sectional, and now we consider the repeated measures model when $\Omega = \mathcal{M}$.

Assumption 5 (Manifold support). The predictor X is supported on a compact Riemannian manifold $\mathcal{M}$ that is $d_{\mathcal{M}}$-dimensional and isometrically embedded in $[0,1]^d$ with conditional number $1/\tau$ and a finite area. The target function $f^\circ(x)$ belongs to the Hölder space $\mathcal{C}^s(\mathcal{M})$ for a given positive number s.

In this assumption, the condition number is defined to be $1/\tau$, where τ is the largest number such that, for any $r < \tau$, the open normal bundle about $\mathcal{M}$ of radius r is embedded in $\mathbb{R}^d$. Intuitively, a smaller condition number means the manifold is flatter, and it is easier to approximate by DNNs. Definition 3.1 cannot be directly generalized to manifolds, a detailed definition of $\mathcal{C}^s(\mathcal{M})$ is provided in Section S.1.3 in the supplementary material.

Theorem 4.4. Consider model (1) and the estimator $\hat{f}_{nm}$ obtained by (2). Suppose that Assumptions 1, 2, 3, and 5 hold, set $\mathcal{F}_{nm} = \mathcal{FNN}\left(d, c_{16}L\log L, c_{17}W\log W, B_1\right)$. If we specify the neural network in a flexible way satisfying $LW = \lfloor c_{18}(nm)^{d_{\mathcal{M}}/(4s+2d_{\mathcal{M}})}(\log nm)^{-4d_{\mathcal{M}}/(2s+d_{\mathcal{M}})}\rfloor$ for some constant c_{18} not depending on n and m, then

$$\mathcal{E}_{f^\circ}(\hat{f}_{nm}) \le c_{19}(n^{-1} + (nm)^{-2s/(2s+d_{\mathcal{M}})}(\log nm)^{16s/(2s+d_{\mathcal{M}})}),$$

where c_{19} is a constant independent of n and m.

Combining with the minimax lower bound in Theorem 3.2, it follows that the ReLU DNN estimator attains the optimal minimax rate $O(n^{-1} + (nm)^{-2s/(2s+d_{\mathcal{M}})})$ up to logarithmic factors under the manifold support set assumption. The convergence rate is usually faster than $O(n^{-1} + (nm)^{-2s/(2s+d)})$ since the intrinsic dimension $d_{\mathcal{M}} \le d$. Accordingly, the phase transition takes place at $O(n^{d_{\mathcal{M}}/2s})$ rather than $O(n^{d/2s})$, that is, which hints that substantially less repeated measurements are required to obtain the parametric rate.

We close this section by mentioning that the results developed for the repeated measurements model are readily applicable to other low-dimensional structures like the approximate manifold and Minkowski dimension set. Some error bounds use recent results(Jiao et al. 2023) have been shown in Section S.1.3 in the supplementary material.

5. Simulation Studies

We conduct simulation studies to demonstrate the adaptivity of the DNN estimators and the phase transition phenomenon. Under various unknown structures, the comparisons with RKHS method in Cai and Yuan (2011) and local linear method in Zhang and Wang (2016) are also provided since they are two available common methods of allowing arbitrary sampling frequency m. The numerical behaviors of the regression spline method in Section S.2 of the supplementary material for cases with $d \le 5$ that our computational capacity allows are also shown. For more simulated examples on hierarchical structure with common closed forms and on anisotropic smoothness, see Section S.3 in the supplementary material.

The Adam optimization algorithm is used with a maximum of 200 epochs and early stopping with a patience of 10 for regularization. In all simulations, the d-dimensional covariate X is generated from the uniform distribution on $\Omega \subseteq [0,1]^d$. Recalling the model setting $Y_{ij} = f^\circ(X_{ij}) + U_i(X_{ij}) + \epsilon_{ij}$, we take the individual stochastic part $U_i(x) = \sum_{k=1}^{\infty}\sum_{l=1}^{d}\frac{\sqrt{3}\xi_{ikl}}{\sqrt{d}k}\cos(k\pi x_l)$ with independent scores $\xi_{ikl} \sim N(0, 0.1^2)$, and measurement errors $\epsilon_{ij} \sim N(0, 0.01^2)$. Let $a = 0.3$ and $b = 1.6$, we consider the following cases:

$$\text{Case 1}: f^\circ(x) = \sum_{k_1,\ldots,k_5\ge 1} a^{\max\{k_1,k_2,k_3,k_4,k_5\}} \prod_{l=1}^{5}\cos(2\pi l^{-1}b^{k_l}x_l), \Omega = [0,1]^5,$$

$$\text{Case 2}: f^\circ(x) = \sum_{l=1}^{5}\sum_{k\ge 1} a^k\cos(2\pi l^{-1}b^{k_l}x_l), \Omega = [0,1]^5,$$

$$\text{Case 3}: f^\circ(x) = \sum_{k_1,\ldots,k_5\ge 1} a^{\max\{k_1,k_2,k_3,k_4,k_5\}} \prod_{l=1}^{5}\cos(2\pi l^{-1}b^{k_l}x_l),$$
$$\Omega = \{(\cos(2\pi t_1), \sin(2\pi t_1), \sin(2\pi t_2), t_1, t_2) : t_1, t_2 \in [0,1]\},$$

$$\text{Case 4}: f^\circ(x) = \sin\left(\frac{12\pi}{d(d+1)}\sum_{l=1}^{d} lx_l\right), \Omega = [0,1]^d, d = 10, 20, 30.$$

Cases 1–3 take the form of series expansions, ensuring their appropriate smoothness inversely related with the values of a and b. Case 2 represents an additive function, thus serving as an example of the hierarchical model. In Case 3, the support Ω is an intrinsically two-dimensional manifold. Case 4 contains large-dimensional examples with low intrinsic dimensions due to the hierarchical structure.

For each case, we conduct 50 Monte Carlo runs and the datasets of smaller n and m are subsets of larger ones. A separate iid set, comprising 25% of the training sample size, is used for hyperparameter selection of each method. For DNN estimators, the networks are setting to be $(L, W) = (1, 50), (2, 100), (3, 200), (4, 400), (5, 600), (5, 800)$ with equal width hidden layers. Each example trains all networks with different depths and widths, then we select the model with the minimal validation error. The estimate $\hat{f}$ is then evaluated using empirical mean squared error, computed over an independent set of 10^4 independently and uniformly distributed points. Results of DNN estimators are shown in Figure 1, and the comparisons are in Table 2.

Analyzing the results in Figure 1, we observe that for a fixed value of sample size n, as m increases initially, the error significantly decreases. However, if m continues to increase beyond some points, the gain becomes less significant, which supports our theoretical findings on the phase transition phenomenon. Compared to Case 1 with similar series structures, Case 2 has a more explicit additive structure and thus converges faster. Case 3 and Case 1 share the same f°, differing only in their input support, and the results exhibit substantially improved convergence rates and a more pronounced phase transition phenomenon, aligning well with our theoretical guidance. Case 4 benefits from the composition structure that allows these large dimensional functions to converge with relatively small samples. Nevertheless, the larger d is, the larger m is needed for the same n to obtain the same convergence error, because the coefficients of the nm term depending on d. Table 2 presents a comparison of the performance of the four methods, highlighting the strengths of the DNN estimator in all cases. As expected by theoretical developments, the DNN estimator is preferred over linear estimators when dealing with unknown hierarchical structures in Case 2. With low-dimensional input, the DNN estimator surpasses the local linear estimator, which is also adaptive to unknown manifold structures (Bickel and Li 2007). The superiority of the DNN estimator is also evident in the high-dimensional cases (Case 4).

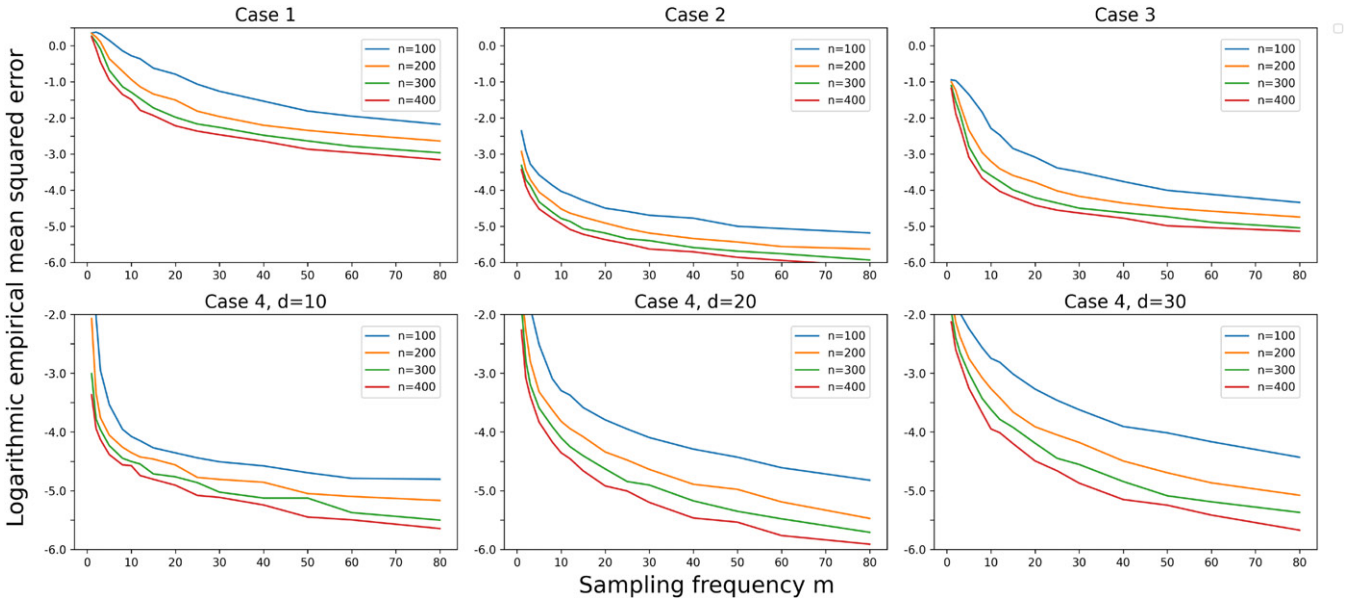


Figure 1. Finite sample mean squared error of the DNN estimators for $n = 100, 200, 300, 400$ and sampling frequencies $m = 1, 2, 3, 5, 8, 10, 12, 15, 20, 25, 30, 40, 50, 60, 80$. Each subplot portrays sampling frequencies (m) on the horizontal axis and logarithmic empirical mean squared error on the vertical axis.

Table 2. Finite sample mean square error of the DNN estimator, regression spline estimator (for cases with $d \leq 5$), RKHS and local linear regression estimators.

	Case 1				Case 2				Case 3			
Sample size n	100		300		100		300		100		300	
Sampling frequency m	10	30	10	30	10	30	10	30	10	30	10	30
DNN	0.7598	0.2839	0.2762	0.1043	0.0178	0.0092	0.0084	0.0045	0.1016	0.0305	0.0275	0.0112
1.5913	0.8543	0.7995	0.3080	0.2105	0.0570	0.0570	0.0333	0.2170	0.0518	0.0464	0.0100	
RKHS	0.8289	0.4630	0.4652	0.2205	0.0264	0.0130	0.0129	0.0069	0.1090	0.0401	0.0390	0.0149
Local linear	1.3112	0.9221	0.8854	0.6147	0.6633	0.4535	0.5883	0.2353	0.2371	0.1395	0.1472	0.1260
	Case 4: $d = 10$				Case 4: $d = 20$				Case 4: $d = 30$			
Sample size n	100		300		100		300		100		300	
Sampling frequency m	10	30	10	30	10	30	10	30	10	30	10	30
DNN	0.0170	0.0110	0.0111	0.0066	0.0370	0.0167	0.0166	0.0074	0.0642	0.0268	0.0269	0.0105
RKHS	0.1996	0.1100	0.1104	0.0573	0.1820	0.1364	0.1363	0.0754	0.1119	0.0967	0.0963	0.0752
Local linear	0.5246	0.3767	0.3988	0.2447	0.4644	0.3049	0.3305	0.2074	0.2644	0.2064	0.2012	0.1790

6. Real Data Applications

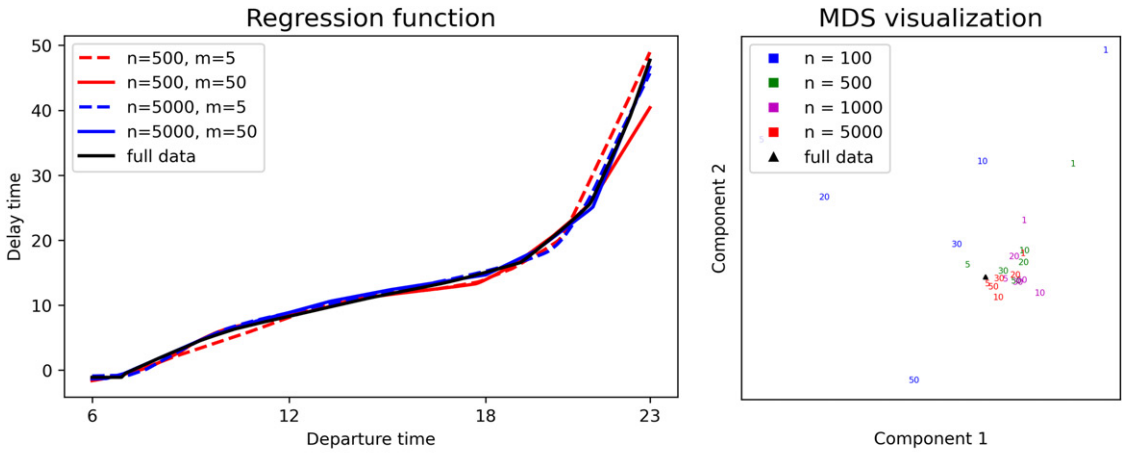
6.1. Airline Delay Example

The airline dataset contains information about commercial flight arrivals and departures in the USA from October 1987 to April 2008, which is publicly available at *https://community.amstat.org/jointscsg-section/dataexpo/dataexpo2009*. With nearly 120 million records involving 304 airports, daily records per airport range from 1 to 1269. Our study focuses on delay durations during peak hours (6:00 to 23:00) in 2008, considering airports with at least 50 daily flights in this period. Assuming data independence across airports and days, we treat daily data for each airport as individual subjects. Time acts as a covariate, while flight delay times constitute measurements. This aligns with our clustered dependent model, as delay times within the same airport on a given day exhibit correlation. There are a total of 8234 subjects in 2008, we divide the training and validation sets in a 4:1 ratio. The network construction is the same as in Simulation. For illustration, we randomly sample 50 measurements from each subject as full sample based on which the estimation is taken as the baseline. To illustrate the proposed method and theory, in a sequentially growing manner, we also train subsamples $n = 100, 500, 1000, 5000$ and sampling frequencies $m = 1, 5, 10, 25, 50$, while the rest data with the sample size $0.25n$ and sampling frequencies m are used as the validation sets. The results are shown in Figure 2.

As depicted in Figure 2(a), estimated functions progressively approach the baseline with larger sample sizes and increased sampling frequencies. Figure 2(b) provides a more intuitive confirmation of our points. When fixing the number of subjects n, increasing m results in shifting toward some sub-center. As n increases, these sub-centers become closer to the baseline obtained from the full data. This provides evidence for the phase transition and certain practical guidance on different roles played by the sample size and sampling frequency.

6.2. PM2.5 Example

The second example concerns the PM2.5 concentration problem with the dataset extracted from Zheng and Chen (2024) available at *https://github.com/FlyHighest/Dynamic-Synthetic-Control*. The dataset includes air pollution and meteorological variables



(a) Estimated functions (b) Multiple-dimensional scaling

Figure 2. (a) Graphs depicting estimated function profiles for distinct sample sizes and sampling frequencies. (b) Multiple-dimensional scaling visualization. The numerical values denote m, the colors represent n (with darker shades indicating larger n values), and the black triangles symbolize the estimation with full data.

Table 3. Mean square testing error of the DNN estimator, RKHS estimator and local linear regression estimator.

	DNN	Linear	Single index	Additive	RKHS	Local linear
Testing error	0.1142	0.4449	0.1541	0.1310	0.1520	0.1691

of 94 monitoring stations in Beijing and nearby provinces (Tianjin, Hebei, Shandong, Shanxi). The variables include pollution concentrations of PM2.5, PM10, SO_2, NO_2, O_3, and CO, wind speed, humidity, dew point temperature, and air pressure. Each station records 72 hourly measurements, with the last 24 taken during Beijing's orange air pollution alert, impacting all stations. In this study, each station is taken as a subject, and we focus on the first 48 records of each station, that is, $m = 48$. The concentrations of SO_2, NO_2, O_3, and CO, meteorological conditions including wind speed, humidity, dew point temperature, and air pressure are taken as covariates, while the logarithmic concentration of PM2.5 is the response. Given the complicated relationship between these variables and PM2.5 (Chen et al. 2017; Zou, Ke, and Zhang 2022), we frame this as an eight-dimensional regression problem.

For comparison, we consider three commonly used statistical models: linear model, single index model (R package "PLSiM-Cpp", Wu et al. 2022) and additive model (R package "gam", Hastie 2023), using working independence correlation during estimation. Meanwhile, the RKHS estimation (Cai and Yuan 2011) and the local linear regression method (Zhang and Wang 2016) are also compared. We randomly select 20% of the subjects as test data to compare the performance of the proposed DNN estimator with alternative methods. Then, the remaining subjects are randomly split into training and validation sets in 4 : 1 ratio. The validation set is used for hyperparameter selections, including early stops, network size, and tuning parameters of other methods. The DNN structures are the same as in the simulation. Now, with $n_{train} = 62$ and $m = 48$, the testing errors of the all methods are shown in Table 3.

From Table 3, the DNN estimator exhibits the lowest testing error. The linear regression model performs poorly due to bias introduced by its overly strong linear assumptions, and the single index model shows improvement. The additive model performs relatively well, indicating a significant proportion of the effects of covariates on PM2.5 can be attributed separately. However, it still underperforms the DNN estimator, suggesting that interactions between predictors are also important and cannot be ignored. The RKHS estimator and local linear estimator outperform both the linear regression and single index models due to no model bias. Nevertheless, they still fall short of the performance of the DNN estimator.

Figure 3 displays the marginal relation between the log PM2.5 concentration and each covariate, with the remaining covariates set to their respective q-quantiles. For most variables, the response exhibits similar trends across different quantiles, suggesting that the function is closely approximated by an additive model. Among the four pollutants, NO_2 and CO are most significantly correlated with PM2.5, mainly due to their similar sources in the considered region, including traffic exhaust and industrial emissions. Higher wind speeds tend to homogenize PM2.5 concentrations. Humidity and dew point temperature have complex effects on PM2.5 levels: they elevate PM2.5 levels by promoting particle growth and secondary aerosol formation, while also bringing precipitation to wash out particulates from the air. Higher air pressure typically means cold and clean air from Siberia, hence, leading to lower PM2.5 levels.

7. Conclusion and Discussion

Clustered dependent regression with arbitrary sampling frequency m is a cornerstone for related statistical fields such as longitudinal and functional data analysis. We made the attempt to adopt and (thus) allow the use of empirical process tools for general clustered dependence problems, obtaining a series of oracle inequalities. These inequalities are new and their proof inspires further studies. Then a phenomenon, called "phase transition," that is repeated measures regression can achieve a parametric rate when increasing the sampling frequency m, imparts a unique influence on the asymptotic performance and the curse of dimensionality. By considering three further exam-

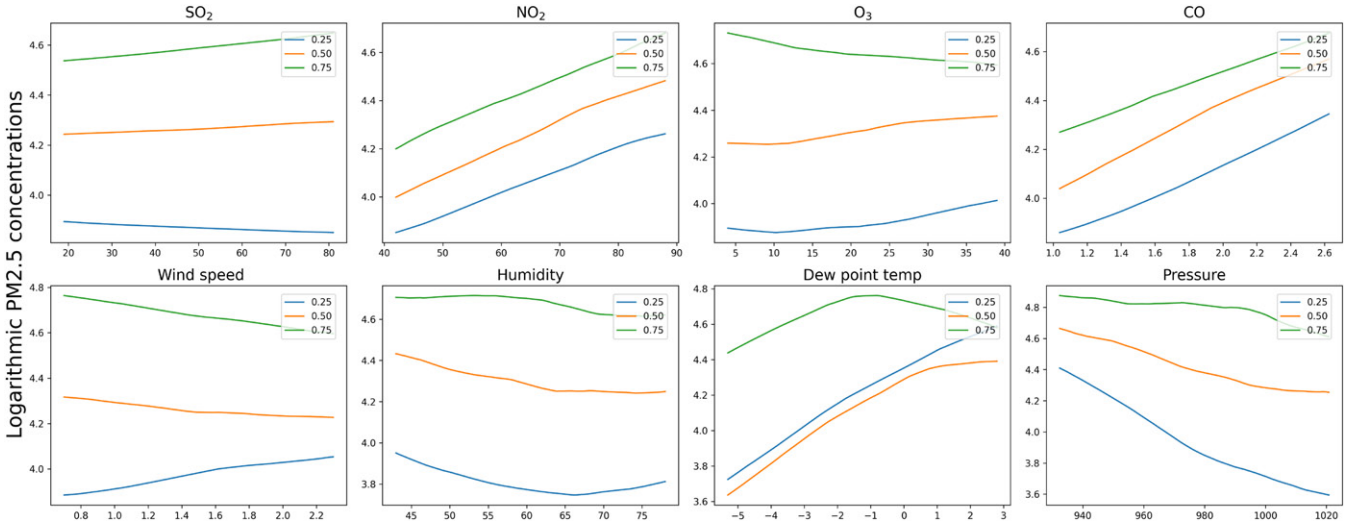


Figure 3. Plots of the estimated function with respect to each covariate. In each subplot, other covariates are held at their 0.25/0.50/0.75 quantile.

ples, we show the proposed estimator can circumvent the curse of dimensionality, such as achieving a parametric rate with lower sampling frequency and getting faster convergence with sparse design.

The neural networks employed in this article are fully connected feedforward neural networks, while other popular networks are certainly of interest, like convolutional neural networks and deep residual networks. More of their properties need to be further understood. Our oracle inequalities distinguish the approximation error and quantify estimation errors in multiple means (e.g., Rademacher complexity and covering numbers) since we hope that our results will still work with new results about network construction and complexity measure in the future. For example, the oracle inequalities could be applied to overparameterized convolutional residual networks with norm constraints using recent advanced results developed in Zhang et al. (2024).

Supplementary Materials

Deep_Regression_for_Repeated_Measurements_supp The additional theoretical results related to Section 3 and Section 4, more simulation examples and implementation details, all proofs and auxiliary lemmas are deferred to the supplementary material for space economy. (.pdf file)

Code and Data Code to implement and reproduce the simulations and real data applications and corresponding output and raw datasets.

Acknowledgments

Shunxing Yan is the first author. Fang Yao is the corresponding author. The authors would like to thank the Editor, the AE and reviewers for their constructive comments.

Disclosure Statement

The authors report there are no competing interests to declare.

Funding

This research is partially supported by the National Key R&D Program of China (No. 2022YFA1003800), the National Natural Science Foundation of China (No. 12292981, 12288101), the New Cornerstone Science Foundation through the Xplorer Prize, the LMAM, the Fundamental Research Funds for the Central Universities, Peking University, and LMEQF.

ORCID

Fang Yao http://orcid.org/0000-0002-8562-6373

References

Aswani, A., Bickel, P., and Tomlin, C. (2011), "Regression on Manifolds: Estimation of the Exterior Derivative," *Annals of Statistics*, 39, 48–81. [2467]

Baltagi, B. H. (2021), *Econometric Analysis of Panel Data* (6th ed.), Cham: Springer. [2461]

Barron, A. R. (1993), "Universal Approximation Bounds for Superpositions of a Sigmoidal Function," *IEEE Transactions on Information Theory*, 39, 930–945. [2461]

Bartlett, P. L., Bousquet, O., and Mendelson, S. (2005), "Local Rademacher Complexities," *The Annals of Statistics*, 33, 1497–1537. [2464]

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, 2261–2285. [2461,2462,2466]

Bhattacharya, S., Fan, J., and Mukherjee, D. (2023), "Deep Neural Networks for Nonparametric Interaction Models with Diverging Dimension," ArXiv preprint. arXiv:2302.05851. [2462]

Bickel, P. J., and Li, B. (2007), "Local Polynomial Regression on Unknown Manifolds," *Lecture Notes-Monograph Series*, 54, 177–186. [2467,2468]

Brumback, B. A., and Rice, J. A. (1998), "Smoothing Spline Models for the Analysis of Nested and Crossed Samples of Curves," *Journal of the American Statistical Association*, 93, 961–976. [2461]

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, 2330–2355. [2461,2462,2466,2468,2470]

Chamberlain, G. (1984), "Panel Data," in *Handbook of Econometrics* (Vol. 2), eds. Z. Griliches and M. D. Intriligator, pp. 1247–1318, Amsterdam: Elsevier. [2461]

Chen, M., Jiang, H., Liao, W., and Zhao, T. (2022), "Nonparametric Regression on Low-Dimensional Manifolds Using Deep ReLU Networks: Function Approximation and Statistical Recovery," *Information and Inference: A Journal of the IMA*, 11, 1203–1253. [2461,2467]

Chen, Z., Cai, J., Gao, B., Xu, B., Dai, S., He, B., and Xie, X. (2017), "Detecting the Causality Influence of Individual Meteorological Factors on Local pm2. 5 Concentration in the Jing-Jin-Ji Region," *Scientific Reports*, 7, 40735. [2470]

Cheng, M.-Y., and Wu, H.-t. (2013), "Local Linear Regression on Manifolds and its Geometric Interpretation," *Journal of the American Statistical Association*, 108, 1421–1434. [2467]

Cloninger, A., and Klock, T. (2021), "A Deep Network Construction that Adapts to Intrinsic Dimensionality Beyond the Domain," *Neural Networks*, 141, 404–419. [2462]

Dahal, B., Havrilla, A., Chen, M., Zhao, T., and Liao, W. (2022), "On Deep Generative Models for Approximation and Estimation of Distributions on Manifolds," in *Advances in Neural Information Processing Systems* (Vol. 35), pp. 0615–10628. [2467]

Diggle, P., Diggle, P. J., Heagerty, P., Liang, K.-Y., Zeger, S., et al. (2002), *Analysis of Longitudinal Data* (2nd ed.), Oxford: Oxford University Press. [2461]

Fan, J., and Gijbels, I. (1992), "Variable Bandwidth and Local Linear Regression Smoothers," *The Annals of Statistics*, 20, 2008–2036. [2461]

Fan, J., and Gu, Y. (2023), "Factor Augmented Sparse Throughput Deep ReLU Neural Networks for High Dimensional Regression," *Journal of the American Statistical Association*, 119, 2680–2694. [2462]

Fan, J., Gu, Y., and Zhou, W.-X. (2024), "How Do Noise Tails Impact on Deep ReLU Networks?" *The Annals of Statistics*, 52, 1845–1871. [2462]

Fan, J., Ma, C., and Zhong, Y. (2021), "A Selective Overview of Deep Learning," *Statistical Science: A Review Journal of the Institute of Mathematical Statistics*, 36, 264–290. [2461]

Farrell, M. H., Liang, T., and Misra, S. (2021), "Deep Neural Networks for Estimation and Inference," *Econometrica*, 89, 181–213. [2462]

Ferraty, F., and Vieu, P. (2006), *Nonparametric Functional Data Analysis: Theory and Practice* (Vol. 76, 1st ed.), New York: Springer-Verlag. [2461]

Györfi, L., Kohler, M., Krzyzak, A., Walk, H., et al. (2002), *A Distribution-Free Theory of Nonparametric Regression* (1st ed.), New York: Springer. [2462,2464,2465]

Hastie, T. (2023), *gam: Generalized Additive Models*, R package version 1.22-3. [2470]

Hedeker, D., and Gibbons, R. D. (2006), *Longitudinal Data Analysis* (1st ed.), Hoboken, NJ: Wiley-Interscience. [2461]

Horowitz, J. L., and Mammen, E. (2007), "Rate-Optimal Estimation for a General Class of Nonparametric Regression Models with Unknown Link Functions," *The Annals of Statistics*, 35, 2589–2619. [2467]

Hsiao, C. (2022), *Analysis of Panel Data* (3rde ed.), Cambridge: Cambridge University Press. [2461]

Hsing, T., and Eubank, R. (2015), *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators* (1st ed.), Chichester, West Sussex: Wiley. [2461]

Jiao, Y., Shen, G., Lin, Y., and Huang, J. (2023), "Deep Nonparametric Regression on Approximate Manifolds: Nonasymptotic Error Bounds with Polynomial Prefactors," *The Annals of Statistics*, 51, 691–716. [2462,2463,2465,2467,2468]

Kohler, M., and Langer, S. (2021), "On the Rate of Convergence of Fully Connected Deep Neural Network Regression Estimates," *The Annals of Statistics*, 49, 2231–2249. [2461,2462,2463,2466]

Koltchinskii, V. (2006), "Local Rademacher Complexities and Oracle Inequalities in Risk Minimization," *The Annals of Statistics*, 34, 2593–2656. [2462,2464]

Lin, X., and Carroll, R. J. (2000), "Nonparametric Function Estimation for Clustered Data When the Predictor is Measured Without/With Error," *Journal of the American statistical Association*, 95, 520–534. [2461,2462]

Liu, H., Chen, M., Zhao, T., and Liao, W. (2021), "Besov Function Approximation and Binary Classification on Low-Dimensional Manifolds Using Convolutional Residual Networks," in *International Conference on Machine Learning*, PMLR, pp. 6770–6780. [2467]

Lu, J., Shen, Z., Yang, H., and Zhang, S. (2021), "Deep Network Approximation for Smooth Functions," *SIAM Journal on Mathematical Analysis*, 53, 5465–5506. [2461]

Nakada, R., and Imaizumi, M. (2020), "Adaptive Approximation and Generalization of Deep Neural Network with Intrinsic Dimensionality," *Journal of Machine Learning Research*, 21, 1–38. [2461,2467]

Ramsay, J., and Silverman, B. (2005), *Functional Data Analysis* (1st ed.), Springer Series in Statistics, New York: , Springer. [2461]

Rice, J., and Rosenblatt, M. (1983), "Smoothing Splines: Regression, Derivatives and Deconvolution," *The Annals of Statistics*, 11, 141–156. [2461]

Sarkar, S., and Panaretos, V. M. (2022), "Covnet: Covariance Networks for Functional Data on Multidimensional Domains," *Journal of the Royal Statistical Society*, Series B, 84, 1785–1820. [2462]

Schmidt-Hieber, J. (2019), "Deep ReLU Network Approximation of Functions on a Manifold," ArXiv preprint. arXiv:1908.00695. [2461,2467]

——— (2020), "Nonparametric Regression Using Deep Neural Networks with ReLU Activation Function," *The Annals of Statistics*, 48, 1875–1897. [2461,2462,2465,2466,2467]

Shaham, U., Cloninger, A., and Coifman, R. R. (2018), "Provable Approximation Properties for Deep Neural Networks," *Applied and Computational Harmonic Analysis*, 44, 537–557. [2461]

Shao, L., Lin, Z., and Yao, F. (2022), "Intrinsic Riemannian Functional Data Analysis for Sparse Longitudinal Observations," *The Annals of Statistics*, 50, 1696–1721. [2466]

Shen, Z. (2020), "Deep Network Approximation Characterized by Number of Neurons," *Communications in Computational Physics*, 28, 1768–1811. [2461]

Shen, Z., Yang, H., and Zhang, S. (2022), "Optimal Approximation Rate of ReLU Networks in Terms of Width and Depth," *Journal de Mathématiques Pures et Appliquées*, 157, 101–135. [2461]

Stone, C. J. (1980), "Optimal Rates of Convergence for Nonparametric Estimators," *The Annals of Statistics*, 8, 1348–1360. [2466]

Suzuki, T. (2018), "Adaptivity of Deep ReLU Network for Learning in Besov and Mixed Smooth Besov Spaces: Optimal Rate and Curse of Dimensionality," ArXiv preprint. arXiv:1810.08033. [2461,2462]

Suzuki, T., and Nitanda, A. (2021), "Deep Learning is Adaptive to Intrinsic Dimensionality of Model Smoothness in Anisotropic Besov Space," in *Advances in Neural Information Processing System* (Vol. 34) (NeurIPS 2021). [2461,2463,2467]

Telgarsky, M. (2016), "Benefits of Depth in Neural Networks," in *Conference on Learning Theory*, PMLR, pp. 1517–1539. [2461]

Wang, S., Cao, G., Shang, Z., and Initiative, A. D. N. (2021), "Estimation of the Mean Function of Functional Data via Deep Neural Networks," *Stat*, 10, e393. [2462]

Weiss, R. E. (2005), *Modeling Longitudinal Data* (1st ed.), New York: Springer-Verlag. [2461]

Wu, S., Zhang, Q., Li, Z., and Liang, H. (2022), *PLSiMCpp: Methods for Partial Linear Single Index Model*, R package version 1.0.4. [2470]

Xu, Z., Ji, X., Chen, M., Wang, M., and Zhao, T. (2024), "Sample Complexity of Neural Policy Mirror Descent for Policy Optimization on Low-Dimensional Manifolds," *Journal of Machine Learning Research*, 25, 1–67. [2467]

Yang, Y., and Yao, F. (2023), "Online Estimation for Functional Data," *Journal of the American Statistical Association*, 118, 1630–1644. [2466]

Yarotsky, D. (2017), "Error Bounds for Approximations with Deep ReLU Networks," *Neural Networks*, 94, 103–114. [2461]

——— (2018), "Optimal Approximation of Continuous Functions by Very Deep Relu Networks," in *Conference on Learning Theory PMLR*, pp. 639–649. [2461]

Zeger, S. L., Liang, K.-Y., and Albert, P. S. (1988), "Models for Longitudinal Data: A Generalized Estimating Equation Approach," *Biometrics*, 44, 1049–1060. [2462]

Zhang, K., Zhang, Z., Chen, M., Takeda, Y., Wang, M., Zhao, T., and Wang, Y.-X. (2024), "Nonparametric Classification on Low Dimensional Manifolds Using Overparameterized Convolutional Residual Networks," ArXiv preprint. arXiv:2307.01649. [2467,2471]

Zhang, X., and Wang, J.-L. (2016), "From Sparse to Dense Functional Data and Beyond," *The Annals of Statistics*, 44, 2281–2321. [2461,2462,2466,2468,2470]

Zhang, Z., Chen, M., Wang, M., Liao, W., and Zhao, T. (2023), "Effective Minkowski Dimension of Deep Nonparametric Regression: Function Approximation and Statistical Theories," in *International Conference on Machine Learning*, PMLR, pp. 40911–40931. [2462]

Zheng, X., and Chen, S. X. (2024), "ynamic Synthetic Control Method for Evaluating Treatment Effects in Auto-Regressive Processes," *Journal of the Royal Statistical Society*, Series B, 86, 155–176. [2469]

Zhong, Q., Müller, J. W., and Wang, J.-L. (2021), "Deep Extended Hazard Models for Survival Analysis," in *Advances in Neural Information Processing Systems* (Vol. 34), pp. 15111–15124, Curran Associates, Inc. [2462]

Zhong, Q., Müller, J., and Wang, J.-L. (2022), "Deep Learning for the Partially Linear Cox Model," *The Annals of Statistics*, 50, 1348–1375. [2462]

Zou, C., Ke, Y., and Zhang, W. (2022), "Estimation of Low Rank High-Dimensional Multivariate Linear Models for Multi-Response Data," *Journal of the American Statistical Association*, 117, 693–703. [2470]