

A THEORY OF NONPARAMETRIC COVARIANCE FUNCTION ESTIMATION FOR DISCRETELY OBSERVED DATA

BY YOSHIKAZU TERADA^{1,2,a} AND ATSUTOMO YARA¹

¹*Graduate School of Engineering Science, The University of Osaka, ^aterada.yoshikazu.es@osaka-u.ac.jp*

²*Center for Advanced Integrated Intelligence Research, RIKEN*

1. Introduction. Recent advances in measurement technology have led to the increasing availability of data in which each subject is observed over a domain, either continuously or at a finite set of locations, as commonly encountered in longitudinal and spatio-temporal studies. In functional data analysis (FDA), such data are viewed as realizations of random elements in a function space, or equivalently as sample paths of stochastic processes (Ramsey and Silverman, 2005; Horváth and Kokoszka, 2012; Hsing and Eubank, 2015; Wang, Chiou and Müller, 2016). Estimation of the mean and covariance functions is a fundamental problem in FDA. It plays a central role in many subsequent methods, most notably functional principal component analysis (Hall and Hosseini-Nasab, 2006; Hall, Müller and Wang, 2006; Zhou, Wei and Yao, 2025) and functional linear regression (Hall and Horowitz, 2007; Yuan and Cai, 2010; Zhou, Yao and Zhang, 2023), and is also essential for reconstructing latent trajectories from discretely observed data (Yao, Müller and Wang, 2005).

When functional data are observed continuously over the domain, covariance estimation is essentially no more difficult than in the multidimensional setting, as the empirical covariance function is directly available. By contrast, when measurements are available only at discrete locations, the covariance function estimation becomes a challenging nonparametric estimation problem. In this paper, we focus on this standard and practically important discretely observed setting.

Consider a centered, square-integrable stochastic process $X(\cdot)$ on a d -dimensional domain $\mathcal{T} \subset \mathbb{R}^d$ with the covariance function $K^\circ(s, t) = \mathbb{E}[X(s)X(t)]$ ($s, t \in \mathcal{T}$). We observe i.i.d. replicates $X_1, \dots, X_n$, but not the trajectories/surfaces themselves. Instead, for subject i we observe noisy measurements

$$Y_{ij} = X_i(T_{ij}) + \epsilon_{ij}, \quad j = 1, \dots, m_i,$$

where the observed points $T_{ij} \in \mathcal{T}$ are random, and ϵ_{ij} are mean-zero noises. For simplicity, the number of measurements is the same among the subjects (i.e., $m = m_1 = \dots = m_n$). Following Zhang and Wang (2016), we refer to the dense regime as the case where $m \rightarrow \infty$ as $n \rightarrow \infty$, and to the sparse regime as the case where m is uniformly bounded.

Existing estimators are typically based on smoothing splines (Cai and Yuan, 2010) or local linear smoothing (Yao, Müller and Wang, 2005; Hall, Müller and Wang, 2006; Li and Hsing, 2010; Zhang and Wang, 2016), and their statistical properties have been investigated in considerable detail. Local linear smoothing, for example, has the attractive feature that the uniform convergence can often be ensured (Li and Hsing, 2010; Zhang and Wang, 2016). Moreover, Cai and Yuan (2010) derives convergence rates for the smoothing spline estimators and proves their minimax optimality. Most of the existing literature focuses on functions defined on one-dimensional domains. In many modern applications, however, functions are defined on multidimensional domains, and the covariance function is therefore defined on the corresponding product domain, whose dimension is doubled, thereby exacerbating the curse of dimensionality.

Recent results in nonparametric regression suggest that adaptive deep learning estimators can, for suitably structured function classes, mitigate the curse of dimensionality (Schmidt-Hieber, 2019, 2020; Nakada and Imaizumi, 2020; Suzuki and Nitanda, 2021). This fact motivates the use of learning-based estimators for covariance function estimation. For instance, Yan, Yao and Zhou (2025) established theoretical guarantees for deep learning estimators of the mean function, and Sarkar and Panaretos (2022) proposed a deep learning approach to covariance function estimation.

In this paper, we study a general nonparametric least-squares (LS) approach to estimating K° . More precisely, we consider a learning-based estimator defined as a minimizer of the least-squares loss

$$(1) \quad \frac{1}{n m(m-1)} \sum_{i=1}^n \sum_{1 \leq j \neq k \leq m} \{Y_{ij}Y_{ik} - K(T_{ij}, T_{ik})\}^2.$$

This loss function is widely used (e.g., Yao, Müller and Wang, 2005; Hall, Müller and Wang, 2006; Li and Hsing, 2010; Cai and Yuan, 2010; Zhang and Wang, 2016).

For mean function estimation, Yan, Yao and Zhou (2025) adapted classical localization analysis (Bartlett, Bousquet and Mendelson, 2005; Koltchinskii, 2006; Blanchard, Bousquet and Massart, 2008) and thereby provided a general framework for analyzing learning-based estimators. For covariance function estimation, by contrast, an analogous framework is not yet available. Sarkar and Panaretos (2022) provided a theoretical analysis of deep learning estimators for discretely observed data on regular grids. However, their analysis requires the number of grid points in each coordinate direction to diverge for consistency. This condition appears overly restrictive. In addition, Paul and Peng (2009) derived theoretical properties of a restricted maximum likelihood estimator under very restrictive conditions, but their results do not attain optimal rates when the number of measurements m diverges. Moreover, their approach cannot accommodate both sparse and dense regimes in a unified manner.

At first sight, one might expect the classical localization analysis for nonparametric regression to extend directly to covariance function estimation, as in the case of mean function estimation. A naive application of the existing theory, however, yields loose bounds when the number of measurements m diverges. The main difficulty arises from the dependence structure induced by the least-squares loss (1), which is a sum of within-subject U -statistics. Unless this within-subject U -statistic structure is handled explicitly, sharp bounds cannot be obtained. We discuss this issue in detail in Section 2.2.

In this paper, we develop a framework for sharp localization analysis of covariance function estimation. Inspired by the prior work on empirical risk minimization of U -statistics (Cléménçon, Lugosi and Vayatis, 2008), we extend Talagrand’s inequality to the least-squares loss (1) for covariance function estimation. Combining this extension with localization analysis, we derive a sharp oracle inequality that treats sparse and dense regimes in a unified manner. As an application, we use this oracle inequality to establish, to the best of our knowledge, the first convergence-rate results for deep learning estimators over several classes of covariance functions. We observe a phase transition phenomenon similar to those reported in Cai and Yuan (2010, 2011) and Yan, Yao and Zhou (2025).

Our results suggest that, under highly anisotropic smoothness, deep learning estimators can mitigate the curse of dimensionality in covariance function estimation, as also demonstrated in other estimation problems (Suzuki and Nitanda, 2021; Yan, Yao and Zhou, 2025). In the sparse regime, the convergence rate of deep learning estimators matches, up to polylogarithmic factors, the minimax rate for nonparametric regression over $2d$ -dimensional domains, as established in Suzuki and Nitanda (2021).

Interestingly, under the dense regime in the one-dimensional setting, for the class of smooth functions considered by [Zhang and Wang \(2016\)](#), deep learning estimators are sub-optimal, whereas local linear smoothing attains a faster rate. In contrast, for the structured function classes studied by [Cai and Yuan \(2010\)](#), deep learning estimators attain the same minimax rate as smoothing spline estimators up to polylogarithmic factors, while local linear smoothing becomes suboptimal. Thus, these results reflect an inherent limitation of learning-based estimators for covariance function estimation. Such behavior appears to be uncommon in classical nonparametric estimation and may be specific to covariance function estimation.

2. Preliminaries. In this section, we describe the repeated measurement model and the least-squares estimator and fix notation. We then discuss the challenges in the theoretical analysis of covariance function estimation.

2.1. Model and least squares estimation. In this paper, we employ the repeated measurement model, which is commonly assumed in existing literature (e.g., [Zhang and Wang \(2016\)](#) and [Yan, Yao and Zhou \(2025\)](#)). Let $\mathcal{T} := [0, 1]^d$. Let $X(\cdot)$ be a square-integrable and centered stochastic process on $\mathcal{T}$ (i.e., $\forall t \in \mathcal{T}; \mathbb{E}[X(t)] = 0$), and assume that X is jointly measurable. The covariance function of X will be denoted by K° , that is,

$$K^\circ(t, t') := \mathbb{E}[X(t)X(t')], \quad t, t' \in \mathcal{T}.$$

Let $X_1, \dots, X_n$ be independent and identically distributed copies of X . The processes X_i ($i = 1, \dots, n$) are not observed directly. Instead, for the i th subject, noisy measurements are observed at m_i time points T_{ij} ($j = 1, \dots, m_i$). We assume that T_{ij} are i.i.d. from a distribution P_T on $\mathcal{T}$ and are independent of X_i ($i = 1, \dots, n$). That is, we assume the following repeated measurement model:

$$(2) \quad Y_{ij} = X_i(T_{ij}) + \epsilon_{ij} \quad (i = 1, \dots, n; j = 1, \dots, m_i)$$

where ϵ_{ij} are independent noises with $\mathbb{E}[\epsilon_{ij}] = 0$. Thus, we observe $\{(T_{ij}, Y_{ij}) : i = 1, \dots, n, j = 1, \dots, m_i\}$. For simplicity, we assume that the number of measurements is the same across subjects, that is, $m = m_1 = \dots = m_n$, and that P_T is the uniform distribution on $[0, 1]^d$.

To estimate the true covariance function K° , we consider the following general nonparametric least squares (LS) estimator:

$$(3) \quad \hat{K}_n := \arg \min_{K \in \mathcal{K}_{nm}} \frac{1}{nm(m-1)} \sum_{i=1}^n \sum_{1 \leq j \neq k \leq m} \{Y_{ij}Y_{ik} - K(T_{ij}, T_{ik})\}^2,$$

where $\mathcal{K}_{nm}$ is a given class of symmetric functions.

To evaluate the accuracy of the LS estimator $\hat{K}_n$, the following mean squared error is considered:

$$\mathbb{E} \left[\|\hat{K}_n - K^\circ\|_2^2 \right],$$

where $\|K\|_2^2 := \int_{\mathcal{T} \times \mathcal{T}} K^2(s, t) dP_T^{\otimes 2}(s, t)$.

2.2. Difficulty of covariance function estimation. We explain why classical localization analysis does not yield tight bounds for covariance function estimation, in contrast to mean function estimation.

In the theory of mean function estimation ([Yan, Yao and Zhou, 2025](#)), bounding the following quantity for $r > 0$ is essential:

$$\sup_{f \in \mathcal{F}} \left| \frac{1}{nm} \sum_{i=1}^n \sum_{j=1}^m \frac{\|f - f^\circ\|_2^2 - \{f(T_{ij}) - f^\circ(T_{ij})\}^2}{\|f - f^\circ\|_2^2 + r} \right|,$$

where f° is the true mean function of X , and $\mathcal{F}$ is a given class of mean functions. Since the T_{ij} are independent, each term in the sum is independent. Therefore, the classical Talagrand inequality (e.g., Theorem 3.3.16 of [Giné and Nickl \(2016\)](#)) can be applied to bound this quantity, and standard localization analysis yields a sharp oracle inequality.

In contrast, for covariance function estimation, bounding the following quantity for $r > 0$ is essential:

$$\sup_{K \in \mathcal{K}_{nm}} \left| \frac{1}{nm(m-1)} \sum_{i=1}^n \sum_{1 \leq j \neq k \leq m} \frac{\|K - K^\circ\|_2^2 - \{K(T_{ij}, T_{ik}) - K^\circ(T_{ij}, T_{ik})\}^2}{\|K - K^\circ\|_2^2 + r} \right|.$$

In the sparse regime, where the number of measurements m is bounded, the following within-subject U -statistic can be viewed as a single random variable Z_i :

$$Z_i := \frac{1}{m(m-1)} \sum_{1 \leq j \neq k \leq m} \frac{\|K - K^\circ\|_2^2 - \{K(T_{ij}, T_{ik}) - K^\circ(T_{ij}, T_{ik})\}^2}{\|K - K^\circ\|_2^2 + r}.$$

Thus, as in [Yan, Yao and Zhou \(2025\)](#), standard localization analysis combined with Talagrand's inequality yields an oracle inequality. However, in the dense regime, where the number of measurements m diverges, this naive approach yields a looser bound because it does not exploit the fact that the observations approach the continuous regime.

We address this gap by developing a unified nonasymptotic analysis of the learning-based estimator (3) that applies to both sparse and dense regimes.

3. A theory of nonparametric covariance function estimation. In this section, we establish an oracle inequality to derive the convergence rates of the mean squared error for a general LS estimator. We assume the following two general assumptions.

ASSUMPTION 1. (Boundedness) There exists a positive number $B_K > 0$ such that $\|K^\circ\|_\infty < B_K$ and $\forall K \in \mathcal{K}_{nm}; \|K\|_\infty < B_K$.

ASSUMPTION 2. (Sub-Gaussian) There exist positive numbers $B_1 > 0$ and $B_2 > 0$ such that

$$\forall \lambda \in \mathbb{R}; \mathbb{E} \left[e^{\lambda X_i(T_{ij})} \right] \leq e^{\lambda^2 B_1^2 / 2} \quad \text{and} \quad \mathbb{E} \left[e^{\lambda \epsilon_{ij}} \right] \leq e^{\lambda^2 B_2^2 / 2}.$$

To overcome the difficulty described in Section 2.2, we establish the following modified Talagrand inequality for covariance function estimation. The lemma builds on a key representation of U -statistics due to [Cl  men  on, Lugosi and Vayatis \(2008\)](#).

LEMMA 3.1. (Modified Talagrand inequality) Let $\{X_{ij}\}_{n \times m}$ be i.i.d. $\mathcal{S}$ -valued random variables. Let $\mathcal{F}$ be a countable class of measurable functions $f = (f_1, \dots, f_n) : \mathcal{S} \times \mathcal{S} \rightarrow \mathbb{R}^n$ such that $\|f_i\|_\infty \leq U < \infty$. Assume that $\mathbb{E}[f_i(X_{i1}, X_{i2})] = 0$ for any $f \in \mathcal{F}$ and any $i = 1, \dots, n$. Set

$$S_n := \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n \sum_{j \neq k} f_i(X_{ij}, X_{ik}) \right| \quad \text{and} \quad \tilde{S}_n := \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n \sum_{j=1}^{\lfloor m/2 \rfloor} f_i(X_{ij}, X_{i(\lfloor m/2 \rfloor + j)}) \right|.$$

Moreover, assume that

$$U^2 \geq \sigma^2 \geq \frac{1}{n} \frac{1}{\lfloor m/2 \rfloor} \sum_{i=1}^n \sum_{j=1}^{\lfloor m/2 \rfloor} \mathbb{E}[f_i^2(X_{ij}, X_{i(\lfloor m/2 \rfloor + j)})],$$

and define $v_n := 2U \mathbb{E}_X[\tilde{S}_n] + n \lfloor m/2 \rfloor \sigma^2$.

Then, for any $x \geq 0$ and any $\alpha > 0$, with probability at least $1 - e^{-x}$,

$$\frac{1}{nm(m-1)} S_n \leq (1 + \alpha) \mathbb{E} \left[\frac{\tilde{S}_n}{n \lfloor m/2 \rfloor} \right] + \sqrt{\frac{2\sigma^2}{n \lfloor m/2 \rfloor}} x + \left(\frac{3}{2} + \frac{1}{\alpha} \right) \frac{Ux}{n \lfloor m/2 \rfloor}.$$

Combined with a conditioning argument, this lemma enables the localization analysis by controlling the empirical process associated with a sum of U -statistics.

Now, we introduce several definitions for localization analysis (Bartlett, Bousquet and Mendelson, 2005; Koltchinskii, 2006).

DEFINITION 3.2 (Sub-root function). Let $r^* > 0$. A nondecreasing function $\phi : [0, \infty) \rightarrow [0, \infty)$ is called a sub-root function on $[r^*, \infty)$ if $r \mapsto \phi(r)/\sqrt{r}$ is nonincreasing for $r \geq r^*$.

DEFINITION 3.3 (Rademacher fixed point). Let $r_{nm}^* > 0$ and let $\phi_{nm} : [0, \infty) \rightarrow [0, \infty)$ be a sub-root function on $[r_{nm}^*, \infty)$. Let $\{\sigma_{ij}\}$ be i.i.d. Rademacher variables. Define

$$\mathcal{K}_{nm}(r) := \{K \in \mathcal{K}_{nm} \mid \|K - K^\circ\|_2^2 \leq r\}.$$

If $\phi_{nm}(r_{nm}^*) \leq r_{nm}^*$ and, for $r \geq r_{nm}^*$,

$$\phi_{nm}(r) \geq \mathbb{E} \left[\sup_{K \in \mathcal{K}_{nm}(r)} \left| \frac{1}{n \lfloor m/2 \rfloor} \sum_{i=1}^n \sum_{j=1}^{\lfloor m/2 \rfloor} \sigma_{ij} \left\{ K(T_{ij}, T_{i(\lfloor m/2 \rfloor + j)}) - K^\circ(T_{ij}, T_{i(\lfloor m/2 \rfloor + j)}) \right\} \right| \right],$$

then r_{nm}^* is called the Rademacher fixed point of $\mathcal{K}_{nm}$.

Localization analysis based on the modified Talagrand inequality yields the following oracle inequality.

THEOREM 3.4. Under Assumption 1 and Assumption 2, for $m \geq 2$, there exists a global constant $c > 0$ such that

$$\mathbb{E} \left[\|\hat{K}_n - K^\circ\|_2^2 \right] \leq c \left\{ \inf_{K \in \mathcal{K}_{nm}} \mathbb{E} [\|K - K^\circ\|_2^2] + \frac{1}{n} + \log^2(n \lfloor m/2 \rfloor) \left(\frac{1}{n \lfloor m/2 \rfloor} + r_{nm}^* \right) \right\},$$

where r_{nm}^* is the Rademacher fixed point of $\mathcal{K}_{nm}$.

This theorem can be seen as an extension of Theorem 2.1 in Yan, Yao and Zhou (2025), which establishes a sharp oracle inequality for mean function estimation, to covariance function estimation. It also extends the theory of Cai and Yuan (2010) to general least-squares (LS) estimators that may not admit an explicit form. Interestingly, our inequality matches the oracle inequality of Yan, Yao and Zhou (2025). For mean function estimation, a total of nm observations $\{Y_{ij}\}$ are used, whereas covariance estimation involves nm^2 pairwise products $\{Y_{ij}Y_{ik}\}$. However, these terms are highly redundant, as they are constructed from only nm underlying observations $\{Y_{ij}\}$. This point is also mentioned in Section 5 of Cai and Yuan (2010).

In contrast, when the estimator admits an explicit representation, such as local linear smoothing estimators (e.g., Zhang and Wang (2016)), a more refined analysis that exploits its structure may improve the $1/(nm)$ rate to $1/(nm^2)$. This improvement, however, requires imposing additional structural constraints on the estimator, thereby sacrificing adaptivity.

REMARK. In Theorem 3.4, Assumption 2 is used to control tail probabilities in the concentration arguments. Hence, it can be relaxed to a sub-Weibull assumption. Under this weaker condition, our results remain valid with the same polynomial rates; only the poly-logarithmic factors deteriorate (i.e., $\log^2(n\lfloor m/2\rfloor)$ is replaced by $\log^c(n\lfloor m/2\rfloor)$ for some larger $c > 2$).

Theorem 3.4 is not directly applicable in practice, since the fixed point r_{nm}^* depends on the population distribution, which is typically unknown. To address this issue, we extend Theorem 3.4 by replacing the Rademacher fixed point r_{nm}^* with a quantity depending only on the complexity of $\mathcal{K}_{nm}$.

COROLLARY 3.5. *Under Assumption 1 and Assumption 2, for $m \geq 2$, we have*

$$\mathbb{E}[\|\hat{K}_n - K^\circ\|_2^2] \leq c \left\{ \inf_{K \in \mathcal{K}_{nm}} \mathbb{E}[\|K - K^\circ\|_2^2] + \frac{1}{n} + \frac{\text{VCdim}(\mathcal{K}_{nm}) \log(n\lfloor m/2\rfloor)}{n\lfloor m/2\rfloor} \right\},$$

where c is a global constant and $\text{VCdim}(\mathcal{K}_{nm})$ denotes the Vapnik–Chervonenkis (VC) dimension of $\mathcal{K}_{nm}$.

4. Covariance function estimation via deep learning. In this section, we derive convergence rates for the learning-based estimators based on deep neural networks (DNNs). To illustrate the potential of deep learning estimators to mitigate the curse of dimensionality, we consider the anisotropic Besov space introduced by Suzuki and Nitanda (2021). This space forms a rich function class that includes a wide range of functions and is particularly suitable for capturing heterogeneous smoothness in real-world data. For example, in spatiotemporal functional data, the covariance function may be very smooth in the temporal direction but much less smooth in the spatial direction, and vice versa.

We define the anisotropic Besov space following the notation of Suzuki and Nitanda (2021). Let $\mathbb{R}_+ := \{x \in \mathbb{R} \mid x > 0\}$. For $\beta = (\beta_1, \dots, \beta_d) \in \mathbb{R}_+^d$, we define

$$\tilde{\beta} := \left(\sum_{j=1}^d \frac{1}{\beta_j} \right)^{-1}, \quad \bar{\beta} := \max_{j=1, \dots, d} \beta_j.$$

For a function $f : \mathbb{R}^d \rightarrow \mathbb{R}$, the r -th difference of f in the direction $h \in \mathbb{R}^d$ is recursively defined by

$$\Delta_h^r(f)(x) := \Delta_h^{r-1}(f)(x+h) - \Delta_h^{r-1}(f)(x), \quad \Delta_h^0(f)(x) := f(x),$$

for $x \in [0, 1]^d$ with $x + rh \in [0, 1]^d$; otherwise, we set $\Delta_h^r(f)(x) := 0$. Moreover, for $f \in L^p([0, 1]^d)$ with $p \in (0, \infty]$, the r -th modulus of smoothness of f is defined as

$$w_{r,p}(f, t) = \sup_{h \in \mathbb{R}^d: \|h_i\| \leq t_i} \|\Delta_h^r(f)\|_p,$$

where $t = (t_1, \dots, t_d) \in \mathbb{R}_+^d$. Using this modulus of smoothness, we define the anisotropic Besov space $B_{p,q}^\beta([0, 1]^d)$ for $\beta = (\beta_1, \dots, \beta_d) \in \mathbb{R}_+^d$.

DEFINITION 4.1 (Anisotropic Besov space). For $0 < p, q \leq \infty$, $\beta = (\beta_1, \dots, \beta_d) \in \mathbb{R}_+^d$, and $r := \max_i \lfloor \beta_i \rfloor + 1$, define the seminorm $|\cdot|_{B_{p,q}^\beta}$ by

$$|f|_{B_{p,q}^\beta} := \begin{cases} \left(\sum_{k=0}^{\infty} [2^k w_{r,p}(f, (2^{-k/\beta_1}, \dots, 2^{-k/\beta_d}))]^q \right)^{1/q}, & \text{if } q < \infty, \\ \sup_{k \geq 0} 2^k w_{r,p}(f, (2^{-k/\beta_1}, \dots, 2^{-k/\beta_d})), & \text{if } q = \infty. \end{cases}$$

The norm of the anisotropic Besov space $B_{p,q}^\beta([0, 1]^d)$ is defined by

$$\|f\|_{B_{p,q}^\beta} := \|f\|_p + |f|_{B_{p,q}^\beta},$$

and

$$B_{p,q}^\beta([0, 1]^d) = \left\{ f \in L^p([0, 1]^d) \mid \|f\|_{B_{p,q}^\beta} < \infty \right\}.$$

Intuitively, β characterizes the smoothness in each coordinate direction. If β_i is large, functions in $B_{p,q}^\beta$ are smooth along the i -th direction; otherwise, they may be less regular in that direction.

We consider the following class of covariance functions as a subset of the anisotropic Besov space:

$$\mathcal{K}_{p,q}^\beta := \left\{ K \in B_{p,q}^{(\beta,\beta)}(\mathcal{T}^2) \mid \|K\|_{B_{p,q}^{(\beta,\beta)}} \leq 1, \text{ and } K \text{ is symmetric and positive definite} \right\}.$$

Here, the choice of (β, β) for the smoothness parameters is natural, reflecting the symmetry of covariance functions.

We introduce a model for covariance functions based on deep feedforward neural networks. Let $\eta : \mathbb{R} \rightarrow \mathbb{R}$ denote the ReLU activation function, defined by $\eta(x) = \max\{x, 0\}$. For a vector x , $\eta(x)$ denotes the vector obtained by applying η to each component of x . The architecture of a feedforward neural network is specified by its depth L and a common width $W \in \mathbb{N}$. A deep neural network with depth L and width W is defined as a function of the form

$$(4) \quad F(x) = (W_L \eta(\cdot) + b_L) \circ (W_{L-1} \eta(\cdot) + b_{L-1}) \circ \cdots \circ (W_1 x + b_1), \quad x \in \mathbb{R}^{2d},$$

where $W_1 \in \mathbb{R}^{W \times 2d}$, $W_L \in \mathbb{R}^{1 \times W}$, $W_l \in \mathbb{R}^{W \times W}$ for $1 < l < L$, and $b_L \in \mathbb{R}$, $b_l \in \mathbb{R}^W$ ($1 \leq l < L$) are bias parameters. Since we consider estimating the covariance function, the input dimension is taken to be $2d$.

We now define the class of sparse networks by

$$\mathcal{F}(L, W, S) = \left\{ F \text{ of the form (4)} \mid \sum_{j=1}^L (\|W_j\|_0 + |b_j|_0) \leq S, \|F\|_\infty < B_K \right\}.$$

Here, $\|W_j\|_0$ and $|b_j|_0$ represent the numbers of nonzero entries of W_j and b_j , respectively. To ensure symmetry of the covariance function estimator, we consider the following class of covariance functions:

$$(5) \quad \mathcal{K}(L, W, S) := \left\{ K(\cdot, \star) = \frac{1}{2} \{h(\cdot, \star) + h(\star, \cdot)\} \mid h \in \mathcal{F}(L, W, S) \right\}.$$

This DNN model (5) is much simpler than the CovNet model (Sarkar and Panaretos, 2022), leading to a simpler convergence analysis. It may seem that the estimator should be positive definite, as in CovNet. However, post-estimation projection always ensures that

$$\|\tilde{K}_n - K^\circ\|_2 \leq \|\hat{K}_n - K^\circ\|_2,$$

where $\tilde{K}_n$ is the projection of $\hat{K}_n$ onto the class of positive definite functions. Thus, positive definiteness need not be imposed at the estimation stage, as is also the case in existing approaches.

The following theorem establishes convergence rates of the deep learning estimator over the class $\mathcal{K}_{p,q}^\beta$.

THEOREM 4.2. *Suppose $0 < p, q \leq \infty$ and $\tilde{\beta}/2 > (1/p - 1/2)_+$ for $\beta \in \mathbb{R}_+^d$. Consider the deep neural network model $\mathcal{K}(L, W, S)$ with*

- (i) $L \asymp \log(n \lfloor m/2 \rfloor)$, (ii) $W \asymp (n \lfloor m/2 \rfloor)^{1/(\tilde{\beta}+1)}$, and
- (iii) $S \asymp (n \lfloor m/2 \rfloor)^{1/(\tilde{\beta}+1)} \log^2(n \lfloor m/2 \rfloor)$.

If $K^\circ \in \mathcal{K}_{p,q}^\beta$ and $\|K^\circ\|_\infty \leq B_K$, under Assumption 2, we obtain

$$\mathbb{E}[\|\hat{K}_n - K^\circ\|_2^2] \leq c \left\{ \frac{1}{n} + \frac{1}{(n \lfloor m/2 \rfloor)^{\tilde{\beta}/(\tilde{\beta}+1)}} \log^3(n \lfloor m/2 \rfloor) \right\},$$

where c is a universal constant that does not depend on n and m .

In the sparse regime, the convergence rate of deep learning estimators coincides with that of deep nonparametric regression in $2d$ dimensions, as established in [Suzuki and Nitanda \(2021\)](#). The rates do not directly depend on the dimension d , and thus suggest that deep learning estimators can mitigate the curse of dimensionality. For example, set $\beta_1 = \alpha$ and $\beta_2 = \dots = \beta_d = (d-1)\alpha$ for some $\alpha > 0$. This represents an anisotropic setting in which the function is non-smooth in the first coordinate, while being considerably smoother in the other coordinates. Then $\tilde{\beta} = \alpha/2$, and the resulting rate does not depend on the dimension d .

The bound in Theorem 4.2 exhibits a phase transition with respect to the number of measurements m . When $m \ll n^{1/\tilde{\beta}}$, the second term dominates and the convergence rate is determined by $(nm)^{-\tilde{\beta}/(\tilde{\beta}+1)}$ up to polylogarithmic factors. In contrast, when $m \gg n^{1/(\tilde{\beta})}$, the first term n^{-1} dominates, and the estimator achieves the parametric rate $1/n$. The transition occurs at the critical scaling $m \asymp n^{1/\tilde{\beta}}$, where the two terms are of the same order.

REMARK. We can easily extend Theorem 4.2 to the affine composition model or the deep composition model in [Suzuki and Nitanda \(2021\)](#). For example, the affine composition model for covariance functions can be defined as follows:

$$\mathcal{K}_{\text{aff},p,q}^\beta := \left\{ H(As + b, At + b) \mid H \in \mathcal{K}_{p,q}^\beta, A \in \mathbb{R}^{\tilde{d} \times d}, b \in \mathbb{R}^{\tilde{d}} \right\}, \quad \beta \in \mathbb{R}_+^{\tilde{d}}.$$

Thus, the directions of smoothness need not be aligned with the coordinate directions and can be arbitrarily oriented through affine transformations.

5. Conclusion. In this paper, we studied nonparametric covariance function estimation for discretely observed functional data on a multidimensional domain. We established an oracle inequality for a broad class of learning-based estimators that applies to both sparse and dense observation regimes in a unified framework. As an important application, we derived convergence rates for deep learning estimators over several classes of covariance functions. Our results reveal a phase transition with respect to the number of measurements m . This transition highlights the interplay between sampling density and intrinsic smoothness in covariance function estimation.

We further demonstrated that the performance of estimators crucially depends on the underlying function class. In the one-dimensional setting, for a class of smooth functions considered by [Zhang and Wang \(2016\)](#), deep learning estimators are suboptimal, whereas local linear smoothing estimators achieve faster rates. In contrast, for a structured function class, such as the tensor product RKHS studied in [Cai and Yuan \(2010\)](#), deep learning estimators attain the minimax rate up to polylogarithmic factors, while local linear smoothing estimators are suboptimal.

These findings indicate that learning-based estimators are not always optimal, but can effectively exploit structural properties such as anisotropy. This leads to a distinctive adaptivity–variance trade-off that appears to be specific to covariance function estimation and is not commonly observed in classical nonparametric regression.

Funding. The first author was supported by JSPS KAKENHI Grants (JP24K14855, JP25K15032, JP25K21806). The second author was supported in part by JST BOOST (JP-MJBS2402).

REFERENCES

- BARTLETT, P. L., BOUSQUET, O. and MENDELSON, S. (2005). Local Rademacher complexities. *The Annals of Statistics* **33** 1497–1537.
- BLANCHARD, G., BOUSQUET, O. and MASSART, P. (2008). Statistical performance of support vector machines. *The Annals of Statistics* **36** 489–531.
- CAI, T. T. and YUAN, M. (2010). Nonparametric Covariance Function Estimation for Functional and Longitudinal Data Technical Report, University of Pennsylvania, Philadelphia, PA.
- 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.
- CLÉMENÇON, S., LUGOSI, G. and VAYATIS, N. (2008). Ranking and Empirical Minimization of U -Statistics. *The Annals of Statistics* **36** 844–874.
- GINÉ, E. and NICKL, R. (2016). *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge Series in Statistical and Probabilistic Mathematics **40**. Cambridge University Press, Cambridge, UK.
- HALL, P. and HOROWITZ, J. L. (2007). Methodology and Convergence Rates for Functional Linear Regression. *The Annals of Statistics* **35** 70–91.
- HALL, P. and HOSSEINI-NASAB, M. (2006). On properties of functional principal components analysis. *Journal of the Royal Statistical Society: Series B* **68** 109–126.
- HALL, P., MÜLLER, H.-G. and WANG, J.-L. (2006). Properties of Principal Component Methods for Functional and Longitudinal Data Analysis. *The Annals of Statistics* **34** 1493–1517.
- HORVÁTH, L. and KOKOSZKA, P. (2012). *Inference for Functional Data with Applications*. Springer, New York.
- HSING, T. and EUBANK, R. (2015). *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. Wiley.
- KOLTCHINSKII, V. (2006). Local Rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics* **34** 2593–2656.
- LI, Y. and HSING, T. (2010). Uniform Convergence Rates for Nonparametric Regression and Principal Component Analysis in Functional/Longitudinal Data. *The Annals of Statistics* **38** 3321–3351.
- NAKADA, R. and IMAIZUMI, M. (2020). Adaptive Approximation and Generalization of Deep Neural Network with Intrinsic Dimensionality. *Journal of Machine Learning Research* **21** 1–39.
- PAUL, D. and PENG, J. (2009). Consistency of restricted maximum likelihood estimators of principal components. *The Annals of Statistics* **37** 1229–1271.
- RAMSAY, J. O. and SILVERMAN, B. W. (2005). *Functional Data Analysis*. Springer, New York.
- SARKAR, S. and PANARETOS, V. M. (2022). CovNet: Covariance Networks for Functional Data on Multidimensional Domains. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **84** 1785–1820.
- SCHMIDT-HIEBER, J. (2019). Deep ReLU network approximation of functions on a manifold. *arXiv:1908.00695*.
- SCHMIDT-HIEBER, J. (2020). Nonparametric Regression Using Deep Neural Networks with ReLU Activation Function. *The Annals of Statistics* **48** 1875–1897.
- SUZUKI, T. and NITANDA, A. (2021). Deep learning is adaptive to intrinsic dimensionality of model smoothness in anisotropic Besov space. *Advances in Neural Information Processing Systems* **34** 3609–3621.
- WANG, J.-L., CHIOU, J.-M. and MÜLLER, H.-G. (2016). Functional Data Analysis. *Annual Review of Statistics and Its Application* **3** 257–295.
- YAN, S., YAO, F. and ZHOU, H. (2025). Deep Regression for Repeated Measurements. *Journal of the American Statistical Association* **120** 2461–2472.
- YAO, F., MÜLLER, H.-G. and WANG, J.-L. (2005). Functional Data Analysis for Sparse Longitudinal Data. *Journal of the American Statistical Association* **100** 577–590.
- YUAN, M. and CAI, T. T. (2010). A Reproducing Kernel Hilbert Space Approach to Functional Linear Regression. *The Annals of Statistics* **38** 3412–3444.

- ZHANG, X. and WANG, J.-L. (2016). From Sparse to Dense Functional Data and Beyond. *The Annals of Statistics* **44** 2281–2321.
- ZHOU, H., WEI, D. and YAO, F. (2025). Theory of Functional Principal Component Analysis for Discretely Observed Data. *The Annals of Statistics* **53** 2103–2127.
- ZHOU, H., YAO, F. and ZHANG, H. (2023). Functional Linear Regression for Discretely Observed Data: From Ideal to Reality. *Biometrika* **110** 381–393.