Nonparametric Estimation via Partial Derivatives

XIAOWU DAI

University of California, Los Angeles

To appear in Journal of the Royal Statistical Society: Series B.

Abstract

Traditional nonparametric estimation methods often lead to a slow convergence rate in large dimensions and require unrealistically enormous sizes of datasets for reliable conclusions. We develop an approach based on partial derivatives, either observed or estimated, to effectively estimate the function at near-parametric convergence rates. The novel approach and computational algorithm could lead to methods useful to practitioners in many areas of science and engineering. Our theoretical results reveal a behavior universal to this class of nonparametric estimation problems. We explore a general setting involving tensor product spaces and build upon the smoothing spline analysis of variance (SS-ANOVA) framework. For $d$-dimensional models under full interaction, the optimal rates with gradient information on $p$ covariates are identical to those for the $(d-p)$-interaction models without gradients and, therefore, the models are immune to the “curse of interaction.” For additive models, the optimal rates using gradient information are root-$n$, thus achieving the “parametric rate.” We demonstrate aspects of the theoretical results through synthetic and real data applications.

Key Words: Gradient; Interaction; Reproducing kernel Hilbert space; Smoothing spline ANOVA; Time series.

1 Address for correspondence: Xiaowu Dai, Department of Statistics and Data Science, UCLA, 8125 Math Sciences Bldg #951554, CA 90095, USA. E-mail: dai@stat.ucla.edu.
1 Introduction

Gradient information for complex systems arises in many areas of science and engineering. Economists estimate cost functions, where data on factor demands and costs are collected together. By Shephard’s Lemma, the demand functions are the first-order partial derivatives of the cost function (Hall and Yatchew 2007). In actuarial science, demography provides mortality force data, which, along with samples from the survival distribution, yield gradients for the survival distribution function (Frees and Valdez 1998). In stochastic simulation, gradient estimation has been studied for a large class of problems (Glasserman 2013). In discrete event simulation, the gradient can be estimated with a negligible computational burden compared to the effort for obtaining a new response (Chen et al. 2013). In meteorology, wind speed and direction are gradient functions of barometric pressure and are measured over broad geographic regions (Breckling 2012). In dynamical and time series applications, gradient information can be observed or estimated, as in biological and infectious disease modeling (Ramsay et al. 2007; Dai and Li 2022, 2024). In traffic engineering, real-time motion sensors can record velocity in addition to positions (Solak et al. 2002).

This paper focuses on nonparametric function estimation under smoothness constraints. Rates of convergence often limit the applications of traditional nonparametric estimation methods in high-dimensional settings, where the number of covariates is large (Stone 1980, 1982). A considerable amount of research effort has been devoted to countering this curse of dimensionality. The additive model is a popular choice (Stone 1985; Hastie and Tibshirani 1990). An additive model assumes the high-dimensional function to be a sum of one-dimensional functions and drops interactions among covariates in order to reduce the variability of an estimator. Stone (1985) showed that the optimal convergence rate for additive models is the same as that for univariate nonparametric estimation problems. Thus, the additive models effectively mitigate the curse of dimensionality. Additive models, however, could be too restrictive and lead to wrong conclusions in applications where interactions among the covariates may be present. As a more flexible alternative, smoothing spline
analysis of variance (SS-ANOVA) models, the analogs of parametric ANOVA models, have attracted lots of attention \cite{Wahba1995, Huang1998, Lin2006, Zhu2014}. In particular, SS-ANOVA models include additive models as special cases. \cite{Lin2000} proved that when the interactions among covariates are in tensor product spaces, the optimal rates of convergence for SS-ANOVA models exponentially depend on the order of interaction. Thus, when SS-ANOVA models are used in problems that involve high-order interactions, it leads to the requirement of unrealistically enormous dataset sizes for reliable conclusions. We call this phenomenon the \textit{curse of interaction}.

We develop a new approach based on partial derivatives to effectively compromise the curse of interaction. Let \( \{(t^{(0)}_i, y^{(0)}_i) : i = 1, \ldots, n\} \) be the function data that follow a regression model,

\[ Y^{(0)} = f_0(t^{(0)}) + \epsilon^{(0)}. \]  

(1)

Here \( \epsilon^{(0)} \in \mathbb{R} \) is a random error, \( f_0 : \mathcal{X}^d \mapsto \mathbb{R} \) is a function of \( d \) covariates \( t = (t_1, \ldots, t_d) \), and \( t^{(0)} \in \mathcal{X}^d \equiv [0,1]^d \) is the design point. Write \( \partial f_0(t)/\partial t_j \) as the \( j \)th partial derivative of \( f_0(t) \). Let \( \{(t^{(j)}_i, y^{(j)}_i) : i = 1, \ldots, n; j = 1, \ldots, p\} \) be the partial derivatives that follow a regression model,

\[ Y^{(j)} = \frac{\partial f_0(t^{(j)})}{\partial t_j} + \epsilon^{(j)}, \quad j = 1, \ldots, p. \]  

(2)

Here \( \epsilon^{(j)} \)s are random errors, and \( t^{(j)} \)s are the design points in \( \mathcal{X}^d \). We allow \( Y^{(j)} \) to be directly observable or estimated from function data. The \( p \in \{1, \ldots, d\} \) denotes the number of gradient types. Without loss of generality, we focus on the first \( p \) covariates in model (2). In particular, when \( p = d \), model (2) gives the \textit{full} gradient data. We allow for a relaxed error structure for both function and gradient data. Specifically, we assume the random errors \( \epsilon^{(0)} \) and \( \epsilon^{(j)} \)s in models (1) and (2) to satisfy,

\[ \mathbb{E}[\epsilon^{(j)}_i] = o(n^{-1/2}), \quad \text{Var}[\epsilon^{(j)}_i] = \sigma_j^2 < \infty, \]

\[ \text{Cov}[\epsilon^{(j)}_i, \epsilon^{(j')}_i] = O\left(|i - i'|^{-\gamma}\right) \quad \text{for some } \gamma > 1, \]  

(3)

where \( i \neq i' \) and \( j, j' = 0, 1, \ldots, p \). We assume the short-range correlation in (3) with some \( \gamma > 1 \). This assumption is generally valid in practice, as gradient data are often estimated by
using local function data through methods such as finite-difference techniques. We provide three concrete examples in Appendix to elaborate on the assumption \([3]\). Moreover, random errors in \([3]\) can be uncentered and correlated, which are typical for estimated gradients, and include the i.i.d. errors in Hall and Yatchew (2007) as a special case.

The SS-ANOVA model \([\text{Wahba et al.} 1995]\) amounts to the assumption that

\[
  f_0(t) = \text{constant} + \sum_{j=1}^{d} f_{0j}(t_j) + \cdots + \sum_{1 \leq j_1 < j_2 < \cdots < j_r \leq d} f_{0j_1 j_2 \cdots j_r}(t_{j_1}, t_{j_2}, \ldots, t_{j_r}),
\]

where the component functions include main effects \(f_{0j}\)s, two-way interactions \(f_{0j_1 j_2}\)s, and so on. Component functions are modeled nonparametrically, and we assume that they reside in certain reproducing kernel Hilbert spaces (RKHS, \text{Wahba} 1990). The series on the right-hand side of \([4]\) is truncated to some order \(r\) of interactions to enhance interpretability. We call \(f_0(t)\) as full or truncated interaction SS-ANOVA model if \(r = d\) or \(1 \leq r < d\), respectively. The SS-ANOVA model \([4]\) can be identified with space,

\[
  \mathcal{H} = \{1\} \oplus \sum_{j=1}^{d} \mathcal{H}_j^1 \oplus \cdots \oplus \sum_{1 \leq j_1 < j_2 < \cdots < j_r \leq d} \left[ \mathcal{H}_1^{j_1} \otimes \mathcal{H}_2^{j_2} \otimes \cdots \otimes \mathcal{H}_r^{j_r} \right].
\]

The components of the SS-ANOVA model in \([4]\) are in the mutually orthogonal subspaces of \(\mathcal{H}\) in \([5]\). The additive model can be viewed as a special case of the SS-ANOVA model \([4]\) by taking \(r = 1\). We assume that all component functions come from a common RKHS \((\mathcal{H}_1, \| \cdot \|_{\mathcal{H}_1})\) given by \(\mathcal{H}_j^1 \equiv \mathcal{H}_1\) for \(j = 1, \ldots, d\). Obviously the linear model is a special example of \([4]\) by taking \(r = 1\) and letting \(\mathcal{H}_1\) be the collection of all univariate linear functions defined over \(X\). Another canonical example of \(\{1\} \oplus \mathcal{H}_1\) is the \(m\)th order Sobolev space \(W_2^m(X)\); see, e.g., \text{Wahba} (1990) for further examples.

We study the possibility of near-parametric rates in the general setting of SS-ANOVA models. Suppose the eigenvalues of the kernel function decay polynomially, i.e., its \(\nu\)th largest eigenvalue is of the order \(\nu^{-2m}\). Our results show that the minimax optimal rates for estimating \(f_0\) under the full interaction (i.e., \(r = d\)) are

\[
  R(n, d, r, p) = \begin{cases} 
  \left[ n (\log n)^{1+p-d} \right]^{-2m/p+1}, & \text{if } 0 \leq p < d, \\
  n \left( \frac{2d}{(2m+1) \alpha - 2} \right)^{2d} \mathbb{1}_{d \geq 3} + n^{-1} (\log n)^{d-1} \mathbb{1}_{d < 3}, & \text{if } p = d.
  \end{cases}
\]
The rates in (6) present an interesting two-regime dichotomy between the scenarios of $0 \leq p < d$ and $p = d$. When $0 \leq p < d$, the rate given by (6) matches with the minimax optimal rate for estimating a $(d - p)$-interaction model without gradient information [Lin 2000]. For example, when $p = 0$ with no partial derivative data, the rate from (6) is $[n(\log n)^{1-d}]^{-2m/(2m+1)}$. This rate aligns with the known rate for estimating a $d$-interaction SS-ANOVA model [Lin 2000]. However, with a large $d$, this rate is heavily affected by the exponential term $(\log n)^{d-1}$, which makes the estimation challenging and leads to the curse of interaction. The inclusion of gradient data provides a substantial advantage in overcoming these challenges. For instance, when $p = d - 1$, the rate in (6) becomes $n^{-2m/(2m+1)}$, which is the same as the optimal rate for estimating additive models without gradient information and independent of $d$ [Stone 1985]. This indicates that SS-ANOVA models can be immune to the curse of interaction through the use of partial derivative data.

On the other hand, when $p = d \geq 3$, the rate in (6) becomes

$$R(n, d, r, p) = n^{-\frac{2md}{(2m+1)d-2}}.$$

This rate converges faster than the optimal rate for additive models $n^{-2m/(2m+1)}$. When $p = d = 2$, the rate in (6) is $R(n, d, r, p) = n^{-1} \log n$. If $p = d = 1$, the rate in (6) is the same as the parametric convergence rate, $R(n, d, r, p) = n^{-1}$. It is also worth noting that when $f_0$ has truncated interaction (i.e., $r < d$), the rates also improve by incorporating partial derivatives, which will be discussed in Section 3. In particular, the rate for additive models (i.e., $r = 1$) under $p = d$ matches with the parametric rate, $R(n, d, r, p) = n^{-1}$.

In the literature, various studies have outlined the construction of linear estimators for the linear functionals of $f_0$, with the difficulty of estimation characterized by a modulus of continuity [Donoho and Liu 1991; Donoho 1994; Klemelä and Tsybakov 2001; Cai and Low 2005]. These studies are relevant to our work in two ways: first, they demonstrate the feasibility of achieving a parametric rate in estimating a univariate function $f_0$ from noisy derivative data, which aligns with the rate in our paper as a special case in the univariate context. Second, they provide the optimal rate for estimating partial derivatives of $f_0$ from
observations of $f_0$, which differs from our target of estimating $f_0$ itself. Our methodology and new convergence rates bridge a gap in these studies by focusing on incorporating noisy gradient data for multivariate function estimation. A similar observation of accelerated rates has been noted earlier with higher-order derivatives (Hall and Yatchew, 2007, 2010). Our results suggest that such a phenomenon holds with first-order derivatives and applies to general SS-ANOVA models involving tensor product spaces. While our theoretical comparison primarily involves Hall and Yatchew (2007) due to its seminal importance and relevance to integrating noisy gradients in nonparametric regression, we recognize the continuous advancements in the field over the last decade. These developments include applications of joint models (1) and (2) in areas such as stochastic simulations and Gaussian process methodologies, where gradient data enhances estimation and prediction (see, e.g., Riihimäki and Vehtari, 2010; Chen et al., 2013; Fu and Qu, 2014; Wang and Berger, 2016; Zhang et al., 2023; Lim, 2024). Nonetheless, a comprehensive statistical theory explaining the benefit of incorporating noisy gradient data has been lacking. This paper develops a theoretical framework that shows how gradient data can mitigate the curse of interaction and significantly enhance the scalability of nonparametric modeling, especially for high-dimensional SS-ANOVA models.

1.1 Our contributions

We develop an approach and computational algorithm to incorporate partial derivatives and lead to methods useful to practitioners in many areas of science and engineering. We obtain a new theory that reveals a behavior universal to this class of nonparametric estimation problems. Our proposal and theoretical results considerably differ from the existing works in multiple ways, which are summarized as follows.

Firstly, our results broaden the i.i.d. error structure by allowing the random errors in function data and gradient data to be biased and correlated. This relaxed assumption is in line with applications when the gradient data are estimated (Chen et al., 2013).

Secondly, we develop a new approach and computational algorithm in RKHS that can easily incorporate gradient information. The proposed estimator also enjoys interpretability
by providing a direct description of interactions. We also find that partial derivatives can reduce interactions in terms of the minimax convergence rates.

Finally, we obtain a sharper theory on the estimation with partial derivatives. We show that when \( p = d - 1 \), the optimal rate for estimating \( d \)-dimensional SS-ANOVA models under full interaction is \( n^{-2m/(2m+1)} \), which is independent of the interaction order \( r \). Hence the SS-ANOVA models are immune to the curse of interaction via using gradients. In contrast, Hall and Yatchew (2007) showed that when \( p = d - 1 \), the convergence rate for estimating \( d \)-dimensional functions is \( n^{-2m/(2m+d-1)} \), which has the curse of dimensionality in \( d \). Therefore, our results show that partial derivatives are useful for the scalability of nonparametric estimation in high dimensions, particularly when using the SS-ANOVA models.

The rest sections are organized as follows. We first provide background in Section 2 and show main results in Section 3. Section 4 presents synthetic and real data examples. Section 5 discusses related works. We provide conclusion in Section 6. The results under other types of designs and their proofs, together with additional numerical examples, are relegated to the Appendix.

2 Background

We begin with a motivating example with partial derivatives. Then we briefly review basic facts about RKHS for the setting of our interest.

2.1 Motivating example

We study a stochastic simulation application to motivate models (1) and (2). Let \( h(t, \omega) \) be the response of a stochastic simulation, which has a design point \( t \in \mathcal{X}^d \) and a random variable \( \omega \). It is of interest to build fast and accurate estimation for \( f_0(t) = \mathbb{E}_\omega[h(t, \omega)] \) (Chen et al., 2013; Glasserman, 2013). At each replication \( k = 1, \ldots, q \), the stochastic simulation has a different random variable \( \omega_k \). A user can select design \( t^{(0)} \) and run the stochastic simulation to obtain a response \( Y_k(t^{(0)}) = h(t^{(0)}, \omega_k) = f_0(t^{(0)}) + \epsilon_k^{(0)} \), where \( \epsilon_k^{(0)} \) is...
Figure 1: Estimation error of our estimator incorporating different levels of gradient information, for the stochastic simulation example. The $y$-axis is in the log scale.

i.i.d. centered simulation noise. In practice, it is common to average responses to reduce the variance of simulation noises, i.e., let $Y^{(0)} = [Y_1(t^{(0)}) + Y_2(t^{(0)}) + \cdots + Y_q(t^{(0)})]/q$, where $q$ is the number of simulation replications and is at the order of hundreds or thousands. Then the response $Y^{(0)}$ follows model (1), where $\epsilon^{(0)}$ is the averaged simulation noise. Under regularity conditions ensuring the interchange of expectation and differentiation (L’Ecuyer, 1990), the infinitesimal perturbation analysis (IPA) gives the gradient estimator of $f_0(t)$ that follows model (2),

$$Y^{(j)} = \frac{\partial}{\partial t_j} h(t^{(j)}, \omega), \quad t^{(j)} \in X^d, \quad j = 1, \ldots, p, \quad 1 \leq p \leq d.$$  

Moreover, the IPA estimators are unbiased, $\mathbb{E}_\omega[Y^{(j)}] = \partial f_0/\partial t_j$ (Glasserman, 2013). We provide details of our stochastic simulation in Section 4.1. The results are reported in Figure 1, which shows mean-squared errors (MSEs) for varying sample size $n$, replication number $q$, and different methods. Those include stochastic kriging with function data (i.e., $p = 0$), our estimator with function and one type of gradient data (i.e., $p = 1$), two types of gradient data (i.e., $p = 2$), the full gradient data (i.e., $p = 3$). A significant decrease in MSEs is observed...
when incorporating partial derivatives. Moreover, the computational cost for obtaining the
gradient estimator is relatively low, as calculating the IPA estimator \( Y^{(j)} \) does not need
additional replication of the simulation. In contrast, getting a new function response \( Y^{(0)} \)
requires \( q \) new replications of the simulation, and each replication could incur a high cost.

2.2 Reproducing kernel for partial derivatives

We briefly review some basic facts about RKHS. Interested readers are referred to Aronszajn
(1950) and Wahba (1990) for further details. Let \( K \) be a Mercer kernel that is a symmetric
positive semi-definite and square-integrable function on \( \mathcal{X} \times \mathcal{X} \). It can be uniquely identified
with the Hilbert space \( \mathcal{H}_1 \) that is the completion of \( \{ \sum_{i=1}^{N} c_i K(t_i, \cdot) : t_i \in \mathcal{X}, c_i \in \mathbb{R}, i = 1, \ldots, N \} \) under the inner product \( \langle \sum_{i} c_i K(t_i, \cdot), \sum_{j} c_j K(t_j, \cdot) \rangle_{\mathcal{H}_1} = \sum_{i,j} c_i c_j K(t_i, t_j) \). Most
commonly used kernels are differentiable, which we shall assume in what follows. In partic-
ular, we assume that

\[
\frac{\partial^2}{\partial t_1 \partial t'_1} K(t, t') \in C(\mathcal{X} \times \mathcal{X}).
\]

where \( C(\cdot) \) is the space of continuous functions. Let the kernel \( K_d((t_1, \ldots, t_d)^\top, (t'_1, \ldots, t'_d)^\top) = K(t_1, t'_1) \cdots K(t_d, t'_d) \). Then \( K_d(\cdot, \cdot) \) is the kernel corresponding to the RKHS \( (\mathcal{H}, \| \cdot \|_\mathcal{H}) \) in
(5); see, e.g., Aronszajn (1950). The condition (7) together with the continuity of \( \langle \cdot, \cdot \rangle_\mathcal{H} \)
yield that for any \( g \in \mathcal{H}, \frac{\partial g(t)}{\partial t_j} = \frac{\partial (g, K_d(t, \cdot))_\mathcal{H}}{\partial t_j} = \langle g, \frac{\partial K_d(t, \cdot)}{\partial t_j} \rangle_\mathcal{H} \). Thus, the
gradient \( \frac{\partial g(t)}{\partial t_j} \) is a bounded linear functional in \( \mathcal{H} \) and has a representer \( \frac{\partial K_d(t, \cdot)}{\partial t_j} \).

By Mercer’s theorem (Riesz and Sz.-Nagy 1955), the kernel function \( K \) admits an eigenvalue
decomposition:

\[
K(t, t') = \sum_{\nu \geq 1} \lambda_\nu \psi_\nu(t) \psi_\nu(t'),
\]

where \( \lambda_1 \geq \lambda_2 \geq \cdots \geq 0 \) are eigenvalues and \( \{ \psi_\nu : \nu \geq 1 \} \) are the corresponding eigen-
functions. For example, \( \lambda_\nu \propto \nu^{-2m} \) for \( \mathcal{W}_2^m(\mathcal{X}) \) under the Lebesgue measure (Wahba 1990),
which will be also discussed in Appendix.
3 Main Results

In this section, we present a new approach for nonparametric estimation via partial derivatives and develop a fast algorithm. We also derive a new theory and show a convergence behavior universal to this class of estimation problems.

3.1 Estimation via partial derivatives

We introduce a method that merges function and derivative information for better estimation. When the function \( f_0 \) in \((4)\) is smooth in \( \mathcal{H} \), we add the empirical loss of partial derivatives as a penalty. Combining these information, we derive the function \( \hat{f}_n \) that meets the smoothness criteria and aligns closely with the observed data,

\[
\hat{f}_n = \arg\min_{\|f\|_{\mathcal{H}} \leq R_n} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left[ y_i^{(0)} - f(t_i^{(0)}) \right]^2 + \sum_{j=1}^{p} w_j \cdot \frac{1}{n} \sum_{i=1}^{n} \left[ y_i^{(j)} - \frac{\partial f}{\partial t_j}(t_i^{(j)}) \right]^2 \right\}. \tag{9}
\]

Here \( R_n \geq 0 \) is an appropriately chosen Hilbert radius, and \( w_j \geq 0 \) is a weight parameter, where a natural choice is \( w_j = \sigma_0^2 / \sigma_j^2 \). If \( \sigma_0^2 \) and \( \sigma_j^2 \) are unknown, we can replace them with consistent estimators for variances (Hall et al., 1990). The concept of derivative-based penalty has also been employed in the generalized profiling approach of Ramsay et al. (2007), which derives a penalty by comparing the derivative of the estimated function to a trajectory generated by ordinary differential equations (ODEs). However, the approach in (9) is different by directly comparing the derivative of the estimated function with either observed or estimated derivatives at discrete data points, which avoids the complexities associated with ODE computations. The following theorem gives a closed-form solution to (9).

**Theorem 1.** Assume that kernel \( K \) satisfies the differentiability condition \((7)\). Then, for any \( R_n \geq 0 \), there exists a minimizer \( \hat{f}_n(t) \) of (9) in a finite-dimensional space,

\[
\hat{f}_n(t) = \sum_{i=1}^{n} \alpha_{i0} K_d(t_i^{(0)}, t) + \sum_{j=1}^{p} \sum_{i=1}^{n} \alpha_{ij} \frac{\partial K_d}{\partial t_j}(t_i^{(j)}, t),
\]

where the coefficients \( \alpha_j = (\alpha_{1j}, \ldots, \alpha_{nj})^T \in \mathbb{R}^n \) for \( j = 0, 1, \ldots, p \).
This theorem is a generalization of the well-known representer lemma for smoothing splines (Wahba 1990). It in effect turns an infinity-dimensional optimization problem into an optimization problem over finite number of coefficients. We will devise a fast algorithm for this optimization in Section 3.2 and show its scalability for large data.

The estimator (9) is different from existing methods of incorporating gradients. For example, Morris et al. (1993) proposed a stationary Gaussian process to combine noiseless gradients, whereas the estimator (9) applies to noisy gradients. Hall and Yatchew (2007) studied a regression-kernel estimator to incorporate noisy derivatives and required special structures on the observed derivatives. However, the estimator (9) can incorporate all types of estimated or observed partial derivatives. Hall and Yatchew (2010) used a series-type estimator but could have a curse of dimensionality problem. In contrast, (9) can scale up to a large dimension $d$. Chen et al. (2013) considered a stochastic kriging method, where the correlation coefficients between gradients and function data are required to be estimated. Differently, it is unnecessary to estimate such correlations for implementing (9). Moreover, we will demonstrate that the estimator (9) outperforms competing alternatives through numerical examples in Section 4.

3.2 Computational algorithm

We now develop a fast algorithm for computing the minimizer $\hat{f}_n(t)$ in Theorem 1. Note that $\hat{f}_n(t)$ can be further written as, for any $t \in X^d$,

$$\hat{f}_n(t) = \tilde{\Psi}_d(t)^\top \tilde{c}_0 + \sum_{j=1}^p \frac{\partial \tilde{\Psi}_d(t)^\top \tilde{c}_j}{\partial t_j},$$

(10)

where $\tilde{\Psi}_d(t) = \begin{bmatrix} \Psi^{\otimes 1}(t_1)^\top, \ldots, \Psi^{\otimes 1}(t_d)^\top, \Psi^{\otimes 2}(t_1,t_2)^\top, \ldots, \Psi^{\otimes r}(t_{d-r+1},t_{d-r+2},\ldots,t_d)^\top \end{bmatrix}^\top$. The column vector $\tilde{\Psi}^{\otimes 1}(t)$ has the $\nu$th element equal to $\sqrt{\lambda_\nu} \psi_\nu(X)$ for $\nu \geq 1$. The vector $\Psi^{\otimes 2}(t_i,t_j) = \Psi^{\otimes 1}(t_i) \otimes \Psi^{\otimes 1}(t_j)$ is generated by the Kronecker product that combines two vectors $\Psi^{\otimes 1}(t_i)$ and $\Psi^{\otimes 1}(t_j)$ into a single vector, where for each element in the first vector $\Psi^{\otimes 1}(t_i)$, we multiply the entire second vector $\Psi^{\otimes 1}(t_j)$ by that element, and the resulting
vectors from each multiplication are then concatenated, forming a long vector that captures all pairwise interactions between the elements of \( \Psi^{\otimes_1}(t_i) \) and \( \Psi^{\otimes_1}(t_j) \). Similarly, 
\[
\Psi^{\otimes_r}(t_{d-r+1}, t_{d-r+2}, \ldots, t_d) = \Psi^{\otimes_1}(t_{d-r+1}) \otimes \Psi^{\otimes_1}(t_{d-r+2}) \otimes \cdots \otimes \Psi^{\otimes_1}(t_d)
\]
is the Kronecker product of the \( r \) corresponding vectors. Here \( \tilde{c}_j = [\tilde{\Psi}_d(t_j^{(j)}), \ldots, \tilde{\Psi}_d(t_n^{(j)})] \) \( \alpha_j \) is the infinite-dimensional coefficient vector, where \( j = 0, 1, \ldots, p \).

The key idea is to employ the random feature mapping [Rahimi and Recht 2007; Dai et al. 2023] to approximate the kernel function, which enables us to construct a projection operator between the RKHS and the original predictor space. Specifically, if the kernel functions that generate \( \mathcal{H}_1 \) are shift-invariant, i.e., \( K(t, t') = K(t - t') \), and integrate to one, i.e., \( \int_{\mathcal{X}} K(t - t')d(t - t') = 1 \), then the Bochner’s theorem [Bochner 1934] states that such kernel functions satisfy the Fourier expansion:

\[
K(t - t') = \int_{\mathbb{R}} p(w) \exp \left\{ \sqrt{-1}w(t - t') \right\} dw,
\]
where \( p(w) \) is a probability density defined by

\[
p(w) = \int_{\mathcal{X}} K(t)e^{-2\pi \sqrt{-1}wt}dt.
\]

We note that many kernel functions are shift-invariant and integrate to one. Examples include the Matérn kernel, \( K(t, t') = \tilde{\tau}_1(1 + |t - t'|/\tau_1 + |t - t'|^2/3\tau_1^2)e^{-|t-t'|/\tau_1} \), the Laplacian kernel, \( K(X, X') = \tilde{\tau}_2e^{-|X-X'|/\tau_2} \), the Gaussian kernel, \( K(X, X') = \tilde{\tau}_3e^{-\tau_3^2|X-X'|^2/2} \), and the Cauchy kernel, \( K(X, X') = \tilde{\tau}_4(1 + \tau_4^2|X - X'|^2)^{-1} \), where \( \tilde{\tau}_1, \tilde{\tau}_2, \tilde{\tau}_3, \tilde{\tau}_4 \) are the normalization constants, and \( \tau_1, \tau_2, \tau_3, \tau_4 \) are the scaling parameters. It is then shown that [Rahimi and Recht 2007] the minimizer in Theorem 1 can be approximated by,

\[
\tilde{f}_n(t) = \Psi_d(t)^\top c_0 + \sum_{j=1}^{p} \frac{\partial \Psi_d(t)^\top c_j}{\partial t_j},
\]
where \( \Psi_d(t) = [\Psi^{\otimes_1}(t_1)^\top, \ldots, \Psi^{\otimes_1}(t_d)^\top, \Psi^{\otimes_2}(t_1, t_2)^\top, \ldots, \Psi^{\otimes_r}(t_{d-r+1}, t_{d-r+2}, \ldots, t_d)^\top] ^\top \), and \( \Psi^{\otimes_1}(t_j) = [\tilde{\psi}_1(t_j), \ldots, \tilde{\psi}_s(t_j)] ^\top \subset \mathbb{R}^s \) is a vector of \( s \) Fourier bases with the frequencies drawn.
from the density \( p(w) \), i.e.,

\[
\omega_{j,\nu} \sim p(\omega), \quad b_{j,\nu} \sim \text{Uniform}[0, 2\pi],
\]

\[
\tilde{\psi}_\nu(t_j) = \sqrt{\frac{2}{s}} \cos(t_j \omega_{j,\nu} + b_{j,\nu}), \quad j = 1, \ldots, d, \ \nu = 1, \ldots, s,
\]

and \( \Psi^\otimes_2(t_i, t_j) = \Psi^\otimes_1(t_i) \otimes \Psi^\otimes_1(t_j) \in \mathbb{R}^s \), and so on. We write the augmented random feature vector as,

\[
\Psi_{(p+1)d}(t) = \left( \Psi_d(t)^\top, \frac{\partial \Psi_d(t)}{\partial t_1}, \ldots, \frac{\partial \Psi_d(t)}{\partial t_p} \right)^\top.
\]

(12)

Then the minimizer in Theorem 1 can be approximated by,

\[
\hat{f}_n(t) = \Psi_{(p+1)d}(t)^\top c_{(p+1)d}.
\]

(13)

We estimate the coefficient vector \( c_{(p+1)d} = (c_0^\top, c_1^\top, \ldots, c_p^\top)^\top \) by minimizing the following convex objective function,

\[
\frac{1}{n} \sum_{i=1}^n \left[ y_i^{(0)} - \hat{f}_n(t_i^{(0)}) \right]^2 + \sum_{j=1}^p w_j \cdot \frac{1}{n} \sum_{i=1}^n \left[ y_i^{(j)} - \frac{\partial \hat{f}_n(t_i^{(j)})}{\partial t_j} \right]^2 + \lambda \sum_{j=0}^p \| c_j \|_2^2,
\]

(14)

where \( \lambda \geq 0 \) is the penalty parameter. We remark that the penalty in (14) is different from the penalty in kernel ridge regression (Wainwright, 2019), which takes the form \( \| \Psi_{(p+1)d}(t)^\top c_{(p+1)d} \|_H^2 \). Since the random feature mapping generally cannot form an orthogonal basis, there is no closed-form representation of the RKHS norms \( \| \Psi_{(p+1)d}(t)^\top c_{(p+1)d} \|_H^2 \) in our setting. As a result, the kernel ridge regression penalty is difficult to implement, and instead we adopt the \( L_2 \) penalty in (14) that is easy for computing. We choose the smoothing parameter \( \lambda \) in (14) by generalized cross-validation (GCV) (Golub et al., 1979). Let \( A(\lambda) \) be the influence matrix as \( \tilde{y} = A(\lambda)y \), where \( y \) is the vector of function and gradient data \( y = (y_1^{(0)}, \ldots, y_n^{(0)}, \ldots, y_1^{(p)}, \ldots, y_n^{(p)})^\top \), and \( \tilde{y} \) is the estimate, \( \tilde{y} = (\hat{f}_n(t_1^{(0)}), \ldots, \hat{f}_n(t_n^{(0)}), \ldots, \partial \hat{f}_n/\partial t_1(t_1^{(p)}), \ldots, \partial \hat{f}_n/\partial t_p(t_n^{(p)}))^\top \). Then GCV selects \( \lambda \geq 0 \) by minimizing the following risk,

\[
\text{GCV}(\lambda) = \frac{\| \tilde{y} - y \|^2}{[n^{-1} \text{tr}(I - A(\lambda))]^2}.
\]
Algorithm 1 Estimation via partial derivatives.

1: **Input**: Function data \( \{(t_i^{(0)}, y_i^{(0)}) : i = 1, \ldots, n\} \), partial derivatives \( \{(t_i^{(j)}, y_i^{(j)}) : i = 1, \ldots, n; j = 1, \ldots, p\} \), weight parameters \( \{w_j : j = 1, \ldots, p\} \).
2: **Step 1**: Sample \( d \) of i.i.d. \( s \)-dimensional random features \( \{w_\nu, b_\nu\}_{\nu=1}^s \) by (11), and construct the augmented random feature vector \( \Psi_{(p+1)d}(t) \) by (12).
3: **Step 2**: Solve the coefficient vector \( c_{(p+1)d} \) by (14).
4: **Output**: Function estimate \( \widehat{f}_n(t) \) in (13).

The use of random feature mapping achieves potentially substantial dimension reduction. More specifically, the estimator in (13) only requires to learn the finite-dimensional coefficient \( c_{(p+1)d} \), compared to the estimator in (10) that involves an infinite-dimensional vector \( \tilde{c}_j \) for \( j = 0, 1, \ldots, p \). It is known that the random feature mapping obtains the optimal bias-variance tradeoff if \( s \) scales at a certain rate and \( s/n \to 0 \) when \( n \) grows (Rudi and Rosasco, 2017). We note that the random feature mapping also efficiently reduces the computational complexity. Given any \( (d, r, p) \), the computation complexity of the estimator in (13) is only \( O(ns^2) \), compared to the computation complexity of the kernel estimator in Theorem 1 that is \( O(n^3) \). The saving of the computation is substantial if \( s/n \to 0 \) as \( n \) grows.

We summarize the above estimation procedure in Algorithm 1.

### 3.3 Minimax optimality

We show that our proposed estimator achieves optimality. Suppose that design points \( t^{(0)} \) in (1) and \( t^{(j)} \)'s in (2) are independently drawn from \( \Pi^{(0)} \) and \( \Pi^{(j)} \)'s, respectively, where \( \Pi^{(0)} \) and \( \Pi^{(j)} \)'s have densities bounded away from zero and infinity. We first present a minimax lower bound in the presence of partial derivatives.

**Theorem 2.** Assume that \( \lambda_\nu \asymp \nu^{-2m} \) for some \( m > 3/2 \) and the kernel \( K \) admits the decomposition in (8). Under the regression models (1) and (2) where \( f_0 \) follows the SS-ANOVA model (4) and \( \|f\|_H \leq R_n \). Then under the error structure (3), there exists a
constant $c$ such that
\[
\liminf_{n \to \infty} \inf_f \sup_{f_0 \in H} \mathbb{P} \left\{ \int_{X^d} \left[ \tilde{f}(t) - f_0(t) \right]^2 dt \geq c \left( [n(\log n)^{1-(d-p)\wedge r}]^{-\frac{2m}{2m+1}} 1_{0 \leq p < d} \right. \\
+ \left. \left[ n^{-\frac{2mr}{(2m+1)r-2}} 1_{r \geq 3} + n^{-1}(\log n)^{r-1} 1_{r < 3} \right] 1_{p=d} \right) \right\} > 0,
\]
where the infimum of $\tilde{f}$ is taken over all measurable functions of the data.

This lower bound is new in the literature, and its proof is established via Fano’s lemma (Tsybakov 2009). Next, we show that the lower bound is attainable via our estimator.

**Theorem 3.** Assume that $\lambda_0 \asymp n^{-2m}$ for some $m > 3/2$ and the kernel $K$ admits the decomposition in (8). Under the regression models (1) and (2) where $f_0$ follows the SS-ANOVA model (4) and $\|f\|_H \leq R_n$. Then under the error structure (3) and with the number of random features in (11) set to $s = O(n \log n)$, the estimator $\hat{f}_n$ in (13) satisfies
\[
\lim_{C \to \infty} \limsup_{n \to \infty} \sup_{f_0 \in H} \mathbb{P} \left\{ \int_{X^d} \left[ \hat{f}_n(t) - f_0(t) \right]^2 dt \leq C \left( [n(\log n)^{1-(d-p)\wedge r}]^{-\frac{2m}{2m+1}} 1_{0 \leq p < d} \right. \\
+ \left. \left[ n^{-\frac{2mr}{(2m+1)r-2}} 1_{r \geq 3} + n^{-1}(\log n)^{r-1} 1_{r < 3} \right] 1_{p=d} \right) \right\} = 1.
\]

Here the tuning parameter $\lambda$ in (14) is chosen by $\lambda \asymp [n(\log n)^{1-(d-p)\wedge r}]^{-2m/(2m+1)}$ when $0 \leq p < d$, and $\lambda \asymp n^{-(2mr-2)/(2m+1)r-2}$ when $p = d, r \geq 3$, and $\lambda \asymp (n \log n)^{-(2m-1)/2m}$ when $p = d, r = 2$, and $\lambda \asymp n^{-(m-1)/m}$ when $p = d, r = 1$.

The proof of Theorem 3 relies on several techniques from empirical process and stochastic process theory, including the linearization method and operator gradients. In our analysis of SS-ANOVA models incorporating gradient information, unlike the approach by Lin (2000) which lacks such data, we have developed a method for the simultaneous diagonalization of two positive definite kernels: one including only function data, and the other incorporating both function and gradient data. We have obtained sharper results on the minimax rates of convergence than those in Lin (2000). Moreover, Theorem 3 demonstrates that the optimal rate in (15) can be achieved with the random feature estimator $\hat{f}_n(t)$, as defined in (13). This represents another contribution compared to Lin (2000).
Theorems 2 and 3 together immediately imply that the minimax optimal rate for estimating \( f_0 \in \mathcal{H} \) is

\[
[n(\log n)^{1-(d-p)\wedge r}]^{-\frac{2m}{2m+1}} \mathbb{1}_{0 \leq p < d} + \left[ n^{-\frac{2m}{(2m+1)\wedge r}} \mathbb{1}_{r \geq 3} + n^{-1}(\log n)^{r-1} \mathbb{1}_{r < 3} \right] \mathbb{1}_{p = d}.
\] (15)

This result connects with two strands of literature–estimating SS-ANOVA models without gradient information, and estimating nonparametric functions using derivatives.

Firstly, in the case of estimating SS-ANOVA models without gradient information, the result in (15) recovers the rate known in the literature (see, e.g., Lin (2000)),

\[
[n(\log n)^{1-r}]^{-\frac{2m}{2m+1}}.
\] (16)

For a high-order interaction \( r \), the exponential term \((\log n)^{r-1}\) in (16) introduces the curse of interaction and makes the SS-ANOVA models impractical. Surprisingly, the result in (15) shows that incorporating gradient data mitigates the curse of interaction. For example, when \( d - r \leq p \leq d - 1 \), the rate given by (15) becomes,

\[
[n(\log n)^{1-(d-p)}]^{-\frac{2m}{2m+1}}.
\] (17)

This rate is identical to the minimax optimal rate for estimating a \((d-p)\)-interaction model without gradient information (Lin 2000). When increasing \( p \) types of gradient data to \((p+1)\) types, the rate given by (17) accelerates at the order of \((\log n)^{-2m/(2m+1)}\), where \( p \geq d - r \) and \( p + 1 \leq d - 1 \). Moreover, when \( p = d - 1 \), the rate given by (17) is \( n^{-2m/(2m+1)} \), which coincides with the optimal rate for estimating additive models without gradient information (Stone 1985). The result in (15) indicates a phase transition from \( 0 \leq p < d \) to \( p = d \).

Specifically, the rate with full gradient \( p = d \) is further improved compared to that with \( p \leq d - 1 \). We also note that when the SS-ANOVA models have full interaction with \( r = d \), the result in (15) yields the special result in (6).

Secondly, in the case of estimating functions using derivatives, Hall and Yatchew (2007) pioneered the proposal of a regression-kernel method for incorporating derivative data under
random design and i.i.d. errors. Hall and Yatchew (2007) proved that with first-order partial derivatives, their estimator achieves the rate $n^{-2m/(2m+d-1)}$ for general Hölder spaces (e.g., their Theorem 3). This rate converges slower than the rate given by (15) when $d \geq 2$, and it suffers from the curse of dimensionality when $d$ is large. In contrast, our work, employing a reproducing-kernel approach within the function space of SS-ANOVA models, a subspace of Hölder spaces characterized by a tensor-product structure, achieves the improved convergence rate in (15). This new result shows the practical value of gradient information in enhancing the scalability of nonparametric modeling, especially in high-dimensional settings typical of SS-ANOVA models.

### 3.4 Extensions of the main results

We discuss various ways for extending the optimal rates established in Theorems 2 and 3. For instance, these rates can be extended to scenarios where the function values and partial derivatives have different sample sizes. Let $n_j$ denote the sample size for the dataset $\{(t_i^{(j)}, y_i^{(j)}) : i = 1, \ldots, n_j\}$, where $j = 0, 1, \ldots, p$. By applying the same arguments as in our proof, the rate in these theorems can be expressed as

$$
\min \left\{ \left[ n_0(\log n_0)^{1-r}\right]^{-\frac{2m}{2m+1}}, \left[ (\min_{j \geq 1} n_j) \left( \log \left( \min_{j \geq 1} n_j \right) \right)^{1-(d-p)\wedge r} \right]^{-\frac{2m}{2m+1}} \mathbb{1}_{0 \leq p < d}
+ \left[ (\min_{j \geq 1} n_j)^{-\frac{2m}{(2m+1)r-2}} \mathbb{1}_{r \geq 3} + \left( \min_{j \geq 1} n_j \right)^{-1} \left( \log \left( \min_{j \geq 1} n_j \right) \right)^{r-1} \mathbb{1}_{r < 3} \right] \mathbb{1}_{p = d} \right\}.
$$

This rate is essentially the minimum of two scenarios: the rate obtained by replacing (15) in terms of the value of $\min_{j \geq 1} n_j$ and the conventional rate (16) based solely on the function data with $n_0$ samples. If the sample size $n_0$ for noisy function values is significantly smaller than $\min_{j \geq 1} n_j$, the optimal rate in (15) still holds with $n = \min_{j \geq 1} n_j$. In this case, the noisy function values contribute to anchoring the absolute level of the function, making function estimation identifiable. Conversely, if the dataset of noisy function values alone is substantially large, i.e., $n_0$ is much greater than $\min_{j \geq 1} n_j$, the convergence rate by Theorems 2 and 3 aligns with the conventional rate (16) based solely on the noisy function values.
The optimal rates in Theorems 2 and 3 also apply under deterministic designs, where the design points $t^{(0)}$ in (1) and $t^{(j)}$s in (2) are equally spaced in $\mathcal{X}^d$, rather than independently drawn from distributions $\Pi^{(0)}$ and $\Pi^{(j)}$s, respectively. This adaptation demonstrates the robustness of our result to variations in design point selection. The results for deterministic designs are given in Appendix S1. Additionally, the optimal rates are valid under a more general error assumption than (3). Specifically, it holds when $\text{Var}(\epsilon_i^{(j)}) = \sigma_j^2 + o(n^{-1/2})$. A rigorous proof of Theorem 3 under this general error assumption follows a similar argument to that of the original proof.

Finally, we discuss additive models, which can be regarded as a special case of the SS-ANOVA model (4) by setting $r = 1$. In this scenario, with gradient data available where $p = d$, Theorems 2 and 3 suggest that the estimation of additive models can achieve the parametric rate of $n^{-1}$, which is a significant improvement over the traditional optimal rate of $n^{-2m/(2m+1)}$ typically achieved without gradient information (Stone, 1985). We provide intuition behind achieving the parametric rate in additive models to illustrate the benefits of incorporating gradient information in statistical estimations. Heuristically, for a univariate function $f_0$, the problem of estimating $f_0$ with noisy gradient data is analogous to settings where $f_0$ is observed with noise, but the integral of $f_0$ is the estimation target, which can achieve the parametric rate $n^{-1}$ through nonlocal averaging (Donoho and Liu, 1991; Donoho, 1994). This analogy suggests that the availability of gradient data eliminates the need for smoothing or local averaging, typically necessary in nonparametric estimation, thus allowing for a faster parametric rate. In the case of multivariate additive models, where $f_0 = f_{01} + \cdots + f_{0d}$, gradient data effectively provides observations on the derivatives of each component function, $df_{0j}(t_j)/dt_j$, enabling the estimation of each component at the parametric rate and, consequently, the entire function $f_0$. 

18
4 Applications

In this section, we demonstrate the aspects of our method and theory via various applications. We study a stochastic simulation example in Section 4.1 and an economics example in Section 4.2. We analyze a real data experiment of ion channel in Section 4.3.

4.1 Call option pricing with stochastic simulations

We discussed a motivating example of stochastic simulation in Section 2.1. Now we consider a detailed stochastic simulation of the call option pricing. The Black-Scholes stochastic differential equation is commonly used to model stock price $S_T$ at time $T$ through

$$dS_T = r_s S_T dT + \sigma_s S_T dW_T, \quad T \geq 0,$$

where $W_T$ is the Wiener process, $r_s$ is the risk-free rate, and $\sigma_s$ is the volatility of the stock price. The equation has a closed-form solution:

$$S_T = S_0 \exp\left\{\left(r_s - \frac{1}{2}\sigma_s^2\right)T + \sigma_s \sqrt{T} \omega\right\}$$

with the standard normal variable $\omega$ and initial stock price $S_0$. The European call option is the right to buy a stock at the prespecified time $T$ with a prespecified price $P_0$. The value of the European option is

$$h(t, \omega) = e^{-r_s T}(S_T - P_0)_+,$$

where $t = (S_0, r_s, \sigma_s)$. Our goal is to estimate the expected net present value of the option with fixed $T$ and $P_0$: $f_0(t) = \mathbb{E}_{\omega}[h(t, \omega)]$. It can be seen that $f_0(t)$ follows the SS-ANOVA model (4). In the experiment, we fix $T = 1$, $P_0 = 100$, and choose the design $t$ from equally spaced points from $S_0 \in [80, 120]$, $r_s \in [0.01, 0.05]$, and $\sigma_s \in [0.2, 1]$ with the sample size $n = 7^3, 14^3, 21^3$. The end points of each interval are always included. We set the number of random feature $s = n/10$ for constructing the random feature estimator in (13). To address the impact of stochastic simulation noise, we simulate $q = 1000, 2000, 5000$ i.i.d. replications of $S_T$ at each design point and then average the responses. Independent sampling is used across design points. It is known that IPA estimators for the gradient: $\partial f_0/\partial S_0, \partial f_0/\partial r_s,$
\( \frac{\partial f_0}{\partial \sigma_*} \) are given by Glasserman (2013),

\[
Y^{(1)} = e^{-r^* T} \frac{S_T}{S_0} \cdot 1\{S_T \geq P_0\},
\]

\[
Y^{(2)} = -TY^{(0)} + e^{-r^* T} TS_T \cdot 1\{S_T \geq P_0\},
\]

\[
Y^{(3)} = e^{-r^* T} \frac{1}{\sigma_*} \log \left( \frac{S_T}{S_0} \right) - \left( r_* + \frac{1}{2} \sigma_*^2 \right) T \right] S_T \cdot 1\{S_T \geq P_0\}. \tag{18}
\]

The IPA estimators (18) are unbiased, \( \mathbb{E}_{\omega}[Y^{(1)}] = \frac{\partial f_0}{\partial S_0} \), \( \mathbb{E}_{\omega}[Y^{(2)}] = \frac{\partial f_0}{\partial r_*} \), \( \mathbb{E}_{\omega}[Y^{(3)}] = \frac{\partial f_0}{\partial \sigma_*} \). We show in Appendix B that the error assumption 3 holds for IPA estimators in (18). In this example, obtaining function data at a new design point requires the generation of \( q \) new random numbers and the computation of \( S_T \) for each of these \( q \) simulation replications. In contrast, the gradient estimator given by (18) can be obtained at a negligible cost and without a new simulation.

**Comparison with existing method.** Stochastic kriging (Ankenman et al., 2010; Chen et al., 2013) is a popular method for the mean response estimation of a stochastic simulation. We compare the results of our estimator (13) incorporating gradient information and the stochastic kriging method without gradient. Consider the tensor product Matérn kernel,

\[
\prod_{j=1}^{3} \left( 1 + \frac{|t_j - t'_j|}{\tau_j} + \frac{|t_j - t'_j|^2}{3\tau_j^2} \right) \exp \left( -\frac{|t_j - t'_j|}{\tau_j} \right). \tag{19}
\]

This kernel satisfies the differentiability condition (7), where lengthscale parameters \( \tau_j \)'s are chosen by the five-fold cross-validation. We use the actual output as the reference given by \( f_0(S_0, r_*, \sigma_*) = S_0 \Phi(-d_1 + \sigma_*) - 100e^{-r_* \Phi(-d_1)} \) when \( T = 1, P_0 = 100 \), where \( d_1 = \sigma_*^{-1} [\log 100 - \log(S_0) - (r_* - \sigma_*^2/2)] \) and \( \Phi(\cdot) \) is the CDF of standard normal distribution. We estimate the MSE \( \mathbb{E}(\hat{f}_n - f_0)^2 \) by a Monte Carlo sample of \( 10^4 \) test points in \([80, 120] \times [0.01, 0.05] \times [0.2, 1]\).

Figure 1 reports the MSEs for different methods: stochastic kriging with only function data (i.e., \( p = 0 \)), our estimator with different types of gradient data. The results are averaged over 1000 simulations in each setting. It is seen that our estimator with gradient data gives a substantial improvement in estimation compared to stochastic kriging without
gradient. For example, the MSE of \( n = 7^3, q = 1000 \) with full gradient (i.e., \( p = 3 \)) is comparable to the MSE of \( n = 14^3, q = 1000 \) without gradient (i.e., \( p = 0 \)). Since it needs little additional cost to estimate gradients by (18), our estimator essentially saves the computational cost of sampling at new designs. It is also seen that a faster convergence rate is obtained when incorporating all gradient data (i.e., \( p = 3 \)) compared to \( p \leq 2 \). This confirms our theoretical finding in Section 3.3.

### Table 1: The ratios of MSE with two types of gradient data (i.e., \( p = 2 \)) relative to MSE with only function data (i.e., \( p = 0 \)), for the example in Section 4.1

| \( n \)         | \( q = 1000 \) | \( q = 2000 \) | \( q = 5000 \) |
|-----------------|---------------|---------------|---------------|
| \( 7^3 = 343 \) | 0.6818        | 0.6789        | 0.6612        |
| \( 14^3 = 2744 \) | 0.5850        | 0.5848        | 0.5835        |
| \( 21^3 = 9261 \) | 0.5484        | 0.5483        | 0.5294        |

Table 1 reports the ratios of the MSE of our estimator with two types of gradient data (i.e., \( p = 2 \)) relative to the MSE of stochastic kriging with only function data (i.e., \( p = 0 \)). It is seen that incorporating gradient data leads to a faster convergence rate, which also agrees with our finding in Section 3.3.

### 4.2 Cost estimation in economics

We consider an economic problem of the cost function estimation. Write the cost function \( f_0(t) = f_0(t_1, \ldots, t_d) \), where \( t_d \) denotes the level of output and \( (t_1, \ldots, t_{d-1}) \) represent the prices of \( d - 1 \) factor inputs. The Cobb-Douglas production function (Varian, 1992) yields that

\[
f_0(t_1, \ldots, t_d) = c_0 \left( \frac{c}{c_j} \right)^{c_j} \prod_{1 \leq j \leq d-1} \prod_{1 \leq j \leq d-1} t_j^{c_j} t_d^{c_j}.
\]

Here \( c_0 \) is the efficiency parameter, \( c_1, \ldots, c_{d-1} \) are elasticity parameters, and \( c = c_1 + \cdots + c_{d-1} \). Our goal is to estimate the cost function \( f_0(t) \). The function data of \( f_0(t) \) are observed at design \( t^{(0)} \in \mathcal{X}^d \). The gradient data of \( f_0(t) \) with respect to input prices are the quantities...
of factor inputs that are also observable \cite{Hall and Yatchew 2007},

\[ Y^{(j)} = \frac{\partial}{\partial t_j} f_0(t^{(j)}) + \epsilon^{(j)}, \quad t^{(j)} \in \mathcal{X}^d, \ j = 1, \ldots, d - 1. \]

Here \( t^{(j)} = t^{(0)} \in \mathcal{X}^d \) for \( 1 \leq j \leq d - 1 \) that typically follows a random design. Moreover, the observational errors are usually assumed to be i.i.d. \cite{Hall and Yatchew 2007} and hence satisfy the error structure (3). Since the gradient data about \( \partial f_0/\partial t_d \) is not usually observable, it motivates our modeling of \( p \in \{1, \ldots, d\} \) in model (2). Clearly, \( f_0(t) \) in this example follows the SS-ANOVA model (4). In the experiment, we consider \( d = 3 \) and fix \( t_3 = 1 \) since the cost function is homogeneous of degree one in \( (t_1, t_2, t_3) \), that is \( f_0(t_1, t_2, t_3, t_4) = t_3 f_0(t_1/t_3, t_2/t_3, 1, t_4) \). The data are generated through,

\[ Y^{(0)} = f_0(t_1, t_2, 1, t_4) + \epsilon^{(0)}, \quad Y^{(j)} = \frac{\partial f_0(t_1, t_2, 1, t_4)}{\partial t_j} + \epsilon^{(j)} \text{ for } j = 1, 2, \]

where \( c_0 = 1, c_1 = 0.8, c_2 = 0.7, c_3 = 0.6, \) and the designs \( t^{(j)}, j = 0, 1, 2 \) follow the i.i.d. uniform distribution in \([0.5, 1.5]^3\). Suppose that \( \epsilon^{(j)}, j = 0, 1, 2 \) are Gaussian with zero means, standard deviations 0.35, and correlation \( \rho \). We consider varying sample size \( n = 100, 200, 500, 1000 \), the correlation \( \rho = 0, 0.4, 0.9 \), and set the number of random feature \( s = n/10 \) for constructing the random feature estimator in (13).

Comparison with existing method. \cite{Hall and Yatchew 2007} proposed a regression-kernel method for incorporating gradient for cost function estimation. We compare the performance of our estimator (13) with that of Hall and Yatchew’s estimator. For the estimator in \cite{Hall and Yatchew 2007}, we follow Hall and Yatchew’s Example 3 to use the tensor product Matérn kernel (19) to average \( (t_1, t_4) \) and \( (t_2, t_4) \) directions locally, and then average the estimates. The MSE is estimated by a Monte Carlo sample of \( 10^4 \) test points in \([0.5, 1.5]^3\).

Table 2 reports the MSEs and standard errors for varying sample size \( n \), correlation \( \rho \), and different methods: our estimator with only function data (i.e., \( p = 0 \)), Hall and Yatchew’s estimator with function and gradient data (i.e., \( p = 2 \)), our estimator with function and
### Table 2: The comparison of average MSEs and standard errors of our estimator with those of Hall and Yatchew’s estimator, considering various gradient types, for the example in Section 4.2 with 1000 simulations. The table shows metrics: “average MSE (standard error),” in units of $10^{-4}$.

|                | Our Estimator (13) with only $Y^{(0)}$ | Hall and Yatchew (2007) with $Y^{(0)} + Y^{(1)} + Y^{(2)}$ | Our Estimator (13) with $Y^{(0)} + Y^{(1)} + Y^{(2)}$ |
|----------------|----------------------------------------|----------------------------------------------------------|--------------------------------------------------|
| $\rho = 0$    | 127.1471 (22.8495)                      | 61.4098 (17.4460)                                        | 47.4739 (13.5196)                               |
| $n = 100$     | 129.6300 (24.8577)                      | 64.5989 (19.8965)                                        | 51.9224 (13.6433)                               |
| $\rho = 0$    | 76.6199 (15.9333)                       | 33.3001 (11.5872)                                        | 24.1501 (8.2730)                                |
| $n = 200$     | 77.7602 (16.1079)                       | 35.0696 (11.7615)                                        | 25.5342 (8.3062)                                |
| $\rho = 0$    | 77.9138 (16.3593)                       | 36.2591 (11.9210)                                        | 27.0137 (8.6223)                                |
| $n = 500$     | 36.1925 (8.0550)                        | 16.3861 (5.5399)                                         | 9.3499 (2.5570)                                 |
| $\rho = 0$    | 38.0683 (8.2180)                        | 18.2355 (5.6164)                                         | 10.4708 (2.5619)                                |
| $n = 1000$    | 21.8570 (5.6051)                        | 9.2788 (2.2411)                                          | 4.5364 (1.6147)                                 |
| $\rho = 0$    | 22.4943 (5.6312)                        | 10.4801 (2.2433)                                         | 5.1468 (1.6561)                                 |
| $n = 1000$    | 22.9499 (5.6446)                        | 10.6193 (2.3386)                                         | 5.3288 (1.8550)                                 |

Gradient data (i.e., $p = 2$). The results are obtained over 1000 simulations in each setting. It is seen that MSEs and standard errors of incorporating gradient information are significantly smaller than that without gradient. Moreover, the performances of our estimator compare favorably with that of Hall and Yatchew’s estimator.

Table 3 reports the ratios of the MSE of our estimator incorporating two types of gradient data (i.e., $p = 2$) relative to the MSE of Hall and Yatchew’s estimator incorporating two types of gradient data (i.e., $p = 2$). It is seen that the ratio decreases with the sample size, which agrees with our theoretical finding in Section 3.3 since our estimator in this example converges at the rate $n^{-2m/(2m+1)}$ by Theorem 3 and Hall and Yatchew’s estimator converges at a slower rate $n^{-m/(m+1)}$.

Tables 2 and 3 also indicate that $s = n/10$ yields sufficient accuracy for the estimations by the random feature estimator in (13). Therefore, in practical applications, an $s$ significantly smaller than the theoretical minimum of $s = O(n \log n)$ in Theorem 3 might often suffice.
Table 3: The ratios of MSE of our estimator with two types of gradient data (i.e., $p = 2$) relative to MSE of Hall and Yatchew’s estimator with two types of gradient data (i.e., $p = 2$), for the example in Section 4.2.

| $n$    | $\rho = 0$ | $\rho = 0.4$ | $\rho = 0.9$ |
|--------|------------|--------------|--------------|
| 100    | 0.7731     | 0.7907       | 0.8038       |
| 200    | 0.7252     | 0.7281       | 0.7450       |
| 500    | 0.5706     | 0.5742       | 0.5887       |
| 1000   | 0.4889     | 0.4911       | 0.5018       |

4.3 Ion channel experiment

We consider a real data example from a single voltage clamp experiment. The experiment is on the sodium ion channel of the cardiac cell membranes. The experiment output $z_k$ measures the normalized current for maintaining a fixed membrane potential of $-35mV$ and the input $x_k$ is the logarithm of time. The sample size of the ion channel experiment is $N = 19$. Computer model has been used to study the ion channel experiment (Plumlee, 2017). Let $\eta(x, t)$ be the computer model that approximates the physical system for the ion channel experiment, where $x$ is the experiment input and $t \in \mathcal{X}^d$ is the calibration parameter whose value are unobservable. For analyzing the ion channel experiment, the computer model is given by $\eta(x, t) = e_1^\top \exp(\exp(x)A(t))e_4$, where $t = (t_1, t_2, t_3)^\top \in \mathcal{X}^d$, $d = 3$, $e_1 = (1, 0, 0, 0)^\top$, $e_4 = (0, 0, 1)^\top$, and

$$A(t) = \begin{pmatrix} -t_2 - t_3 & t_1 & 0 & 0 \\ t_2 & -t_1 - t_2 & t_1 & 0 \\ 0 & t_2 & -t_1 - t_2 & t_1 \\ 0 & 0 & t_2 & -t_1 \end{pmatrix}.$$ 

Our goal is to estimate the function, $f_0(t) = \mathbb{E}_{(x,z)}[z - \eta(x, t)]^2$, which is useful for visualization, calibration, and understanding how well the computer model approximates the physical system in this experiment (Kennedy and O’Hagan, 2001). The function data at design $t^{(0)} \in \mathcal{X}^3$ is generated by,

$$Y^{(0)} = \frac{1}{N} \sum_{k=1}^{N} [z_k - \eta(x_k, t^{(0)})]^2,$$

where $N = 19$. 

24
The gradient of computer model, i.e., \( \nabla_t \eta(x, t) \), can be obtained using the chain rule-based automatic differentiation. By the cheap gradient principle (Griewank and Walther, 2008), the cost for computing \( \nabla_t \eta(x, t) \) is at most four or five times the cost for function evaluation \( \eta(x, t) \) and hence, the gradient is cheap to obtain. Then the estimator for the gradient of \( f_0(t) \) is given by,

\[
Y^{(j)} = -\frac{2}{N} \sum_{k=1}^{N} \left[ z_k - \eta(x_k, t^{(j)}) \right] \frac{\partial}{\partial t_j} \eta(x_k, t^{(j)}), \quad t^{(j)} \in X^3, \ j = 1, 2, 3.
\]

In the experiment, we choose i.i.d. uniform designs for \( t^{(j)} \)s, \( j = 0, 1, 2, 3 \) from \( X^3 \) with the sample size \( n = 1000, 2000, 3000, 5000 \).

**Table 4:** The comparison of average MSEs and standard errors of our estimator with those of Morris et al.’s estimator, considering various gradient types, for the example in Section 4.3 with 1000 simulations. The table shows metrics: “average MSE (standard error),” in units of \( 10^{-6} \).

| \( n \)     | Our Estimator with only \( Y^{(0)} \) | Morris et al. 1993 with \( Y^{(0)} + \cdots + Y^{(d)} \) | Our Estimator with \( Y^{(0)} + \cdots + Y^{(3)} \) |
|------------|---------------------------------|---------------------------------|---------------------------------|
| \( n = 1000 \) | 10.6491 (4.9867) | 8.8956 (4.8729) | **7.7804** (3.6737) |
| \( n = 2000 \) | 8.5302 (4.3339) | 6.5494 (4.0728) | **5.1375** (2.4687) |
| \( n = 3000 \) | 6.4296 (3.9595) | 4.1940 (3.2242) | **3.1035** (1.7187) |
| \( n = 5000 \) | 5.4143 (3.2268) | 3.0910 (1.9073) | **2.1305** (0.9322) |

**Comparison with existing method.** Morris et al. (1993) proposed a stationary Gaussian process method to incorporate gradient data for estimation. We compare the performance of our estimator (13) with that of Morris et al.’s estimator. We use the Matérn kernel (19) for both our estimator and Morris et al.’s estimator, and estimate the MSE by a Monte Carlo sample of \( 10^4 \) test points in \( X^3 \). We set the number of random feature \( s = n/10 \) for constructing the estimator (13). Since the true function \( f_0(t) \) is unknown at each test point, we approximate it by using total \( N = 19 \) real ion channel samples at each test point. The function and gradient training data are generated using \( N' = 10 \) real ion channel samples, which are randomly chosen from the total \( N = 19 \) samples.
Table 4 reports the MSEs and standard errors for varying sample size $n$ and different methods: our estimator with only function data (i.e., $p = 0$), Morris et al.’s estimator with function and gradient data (i.e., $p = 3$), our estimator with function and gradient data (i.e., $p = 3$). The results are obtained over 1000 simulations in each setting. It is evident that the gradient data can significantly improve the estimation performance, and our estimator outperforms Morris et al.’s estimator.

5 Related Work

We review related work from multiple kinds of literature, including nonparametric regression, function interpolation, and dynamical systems.

There is growing literature on nonparametric regression with derivatives. Our work is related to the pioneering work of Hall and Yatchew (2007, 2010), which established the root-$n$ consistency for nonparametric estimation given mixed and sufficiently higher-order derivatives. However, it is difficult to obtain higher-order derivatives in practice, such as in economics and stochastic simulation. In contrast, we focus on gradient information that is first-order derivatives and are easier to obtain in practice. We show that the minimax optimal rates for estimating SS-ANOVA models are accelerated by using gradient data. In particular, we show that SS-ANOVA models are immune to the curse of interaction given gradient information.

The function interpolation with gradients has been widely studied. For exact data and one-dimensional functions, Karlin (1969) and Wahba (1971) showed that at certain deterministic design for data without gradients, incorporating gradient to the dataset provides no new information for function interpolation. This result, however, cannot be extended to the case of noisy data. Morris et al. (1993) incorporated noiseless derivatives for deterministic surface estimation in computer experiments. Unlike these works, we consider the noisy gradient information for nonparametric estimation.

Our work is also related to the literature on dynamical systems and stochastic simulation.
Solak et al. (2002) considered the identified linearization around an equilibrium point for estimating the derivatives in nonlinear dynamical systems. They used Gaussian processes for a combination of function and derivative observations. Chen et al. (2013) used stochastic kriging to incorporate gradient estimators and improve surface estimation, where stochastic kriging (Ankenman et al., 2010) is a metamodeling technique for representing the mean response surface implied by a stochastic simulation. However, the rates of convergence are not studied in Solak et al. (2002) and Chen et al. (2013). We quantify the improved rates of convergence in nonparametric estimation by using gradient data.

6 Conclusion

Statistical modeling of gradient information becomes an increasingly important problem in many areas of science and engineering. We develop an approach based on partial derivatives, either observed or estimated, to effectively estimate the nonparametric function. The proposed approach and computational algorithm could lead to methods useful to practitioners. Our theoretical results showed that SS-ANOVA models are immune to the curse of interaction using gradient information. Moreover, for the additive models, the rates using gradient information are root-$n$, thus achieving the parametric rate. As a working model, we assume that the eigenvalues decay at the same polynomial rate across component RKHS $\mathcal{H}_i$'s, which hold for Sobolev kernels, among other commonly used kernels. It is of interest to consider incorporating gradient information in more general settings, for example, when eigenvalues decay at different rates, or if the eigenvalues for some components decay even exponentially. It is conceivable that our analysis could be extended to deal with more general settings, which will be left for future studies.

Acknowledgement

The author thanks the Editor, Associate Editor, and three anonymous reviewers for their invaluable feedback, and thanks Prof. Grace Wahba for valuable advice on this work. The
author acknowledges the support of the California Center for Population Research as part of the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD) population research infrastructure grant P2C-HD041022.

References

Ankenman, B. E., Nelson, B. L., and Staum, J. (2010). Stochastic kriging for simulation metamodeling. *Operations Research*, 58(2):371–382.

Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404.

Bochner, S. (1934). A theorem on fourier-stieltjes integrals. *Bulletin of the American Mathematical Society*, 40(4):271–276.

Breckling, J. (2012). *The Analysis of Directional Time Series: Applications to Wind Speed and Direction*, volume 61. Springer Science & Business Media.

Cai, T. T. and Low, M. G. (2005). On adaptive estimation of linear functionals. *The Annals of Statistics*, 33(5):2311–2343.

Chen, X., Ankenman, B. E., and Nelson, B. L. (2013). Enhancing stochastic kriging metamodels with gradient estimators. *Operations Research*, 61(2):512–528.

Dai, X. and Li, L. (2022). Kernel ordinary differential equations. *Journal of the American Statistical Association*, 117(540):1711–1725.

Dai, X. and Li, L. (2024). Post-regularization confidence bands for ordinary differential equations. *Journal of Machine Learning Research*, 25(23):1–51.

Dai, X., Lyu, X., and Li, L. (2023). Kernel knockoffs selection for nonparametric additive models. *Journal of the American Statistical Association*, 118(543):2158–2170.
Donoho, D. L. (1994). Statistical estimation and optimal recovery. *The Annals of Statistics*, 22(1):238–270.

Donoho, D. L. and Liu, R. C. (1991). Geometrizing rates of convergence, iii. *The Annals of Statistics*, 19(2):668–701.

Donoho, D. L., Liu, R. C., and MacGibbon, B. (1990). Minimax risk over hyperrectangles, and implications. *The Annals of Statistics*, 18(3):1416–1437.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. New York: Chapman and Hall.

Frees, E. W. and Valdez, E. A. (1998). Understanding relationships using copulas. *North American Actuarial Journal*, 2(1):1–25.

Fu, M. C. and Qu, H. (2014). Regression models augmented with direct stochastic gradient estimators. *INFORMS Journal on Computing*, 26(3):484–499.

Gelfand, I. M. and Silverman, R. A. (2000). *Calculus of Variations*. Courier Corporation.

Glasserman, P. (2013). *Monte Carlo Methods in Financial Engineering*. New York: Springer Science & Business Media.

Golub, G. H., Heath, M., and Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223.

Griewank, A. and Walther, A. (2008). *Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation*. Philadelphia, PA: SIAM.

Hall, P. (1992a). Effect of bias estimation on coverage accuracy of bootstrap confidence intervals for a probability density. *The Annals of Statistics*, pages 675–694.

Hall, P. (1992b). On bootstrap confidence intervals in nonparametric regression. *The Annals of Statistics*, pages 695–711.
Hall, P., Kay, J. W., and Titterington, D. M. (1990). Asymptotically optimal difference-based estimation of variance in nonparametric regression. *Biometrika*, 77(3):521–528.

Hall, P. and Yatchew, A. (2007). Nonparametric estimation when data on derivatives are available. *The Annals of Statistics*, 35(1):300–323.

Hall, P. and Yatchew, A. (2010). Nonparametric least squares estimation in derivative families. *Journal of Econometrics*, 157(2):362–374.

Härdle, W. and Bowman, A. W. (1988). Bootstrapping in nonparametric regression: local adaptive smoothing and confidence bands. *Journal of the American Statistical Association*, 83(401):102–110.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*. London, UK: Chapman & Hall/CRC.

Huang, J. Z. (1998). Projection estimation in multiple regression with application to functional anova models. *The Annals of Statistics*, 26(1):242–272.

Jones, B. L. and Mereu, J. A. (2002). A critique of fractional age assumptions. *Insurance: Mathematics and Economics*, 30(3):363–370.

Karlin, S. (1969). The fundamental theorem of algebra for monosplines satisfying certain boundary conditions and applications to optimal quadrature formulas. *Approximations with Special Emphasis on Spline Functions*, pages 467–484.

Kennedy, M. C. and O’Hagan, A. (2001). Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(3):425–464.

Klemelä, J. and Tsybakov, A. B. (2001). Sharp adaptive estimation of linear functionals. *The Annals of Statistics*, 29(6):1567–1600.

L’Ecuyer, P. (1990). A unified view of the ipa, sf, and lr gradient estimation techniques. *Management Science*, 36(11):1293–1416.
Lim, E. (2024). Estimating a function and its derivatives under a smoothness condition. *Mathematics of Operations Research*.

Lin, Y. (1998). Tensor product space anova models in multivariate function estimation. *Thesis (Ph.D.)–University of Pennsylvania*.

Lin, Y. (2000). Tensor product space anova models. *The Annals of Statistics*, 28(3):734–755.

Lin, Y. and Zhang, H. H. (2006). Component selection and smoothing in multivariate nonparametric regression. *The Annals of Statistics*, 34(5):2272–2297.

Morris, M. D., Mitchell, T. J., and Ylvisaker, D. (1993). Bayesian design and analysis of computer experiments: Use of derivatives in surface prediction. *Technometrics*, 35(3):243–255.

Oden, J. T. and Reddy, J. N. (2012). *An Introduction to the Mathematical Theory of Finite Elements*. New York: John Wiley & Sons.

Plumlee, M. (2017). Bayesian calibration of inexact computer models. *Journal of the American Statistical Association*, 112(519):1274–1285.

Rahimi, A. and Recht, B. (2007). Random features for large-scale kernel machines. *Advances in Neural Information Processing systems*, 20:1177–1184.

Ramsay, J. O., Hooker, G., Campbell, D., and Cao, J. (2007). Parameter estimation for differential equations: a generalized smoothing approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(5):741–796.

Riesz, F. and Sz.-Nagy, B. (1955). *Functional Analysis*. New York: Dover Publications.

Riihimäki, J. and Vehtari, A. (2010). Gaussian processes with monotonicity information. In *International Conference on Artificial Intelligence and Statistics*, pages 645–652.
Rudi, A. and Rosasco, L. (2017). Generalization properties of learning with random features. *Advances in Neural Information Processing Systems (NeurIPS)*, 30.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric Regression*. New York: Cambridge University Press.

Schoenberg, I. J. (1964). Spline functions and the problem of graduation. *Proceedings of the National Academy of Sciences*, 52(4):947–950.

Solak, E., Murray-Smith, R., Leithead, W., Leith, D., and Rasmussen, C. (2002). Derivative observations in gaussian process models of dynamic systems. *Advances in Neural Information Processing Systems*, 15.

Stone, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *The Annals of Statistics*, 8(6):1348–1360.

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, 10(4):1040–1053.

Stone, C. J. (1985). Additive regression and other nonparametric models. *The Annals of Statistics*, 13(2):689–705.

Suri, R. and Leung, Y. T. (1987). Single run optimization of a siman model for closed loop flexible assembly systems. *Proceedings of the 19th Conference on Winter Simulation*, pages 738–748.

Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation*. New York: Springer.

van der Vaart, A. and Wellner, J. (1996). *Weak Convergence and Empirical Processes*. Springer, New York.

Varian, H. R. (1992). *Microeconomic Analysis*. New York: W. W. Norton & Company.
Wahba, G. (1971). On the regression design problem of sacks and ylvisaker. *The Annals of Mathematical Statistics*, pages 1035–1053.

Wahba, G. (1990). *Spline Models for Observational Data*. Philadelphia, PA: SIAM.

Wahba, G., Wang, Y., Gu, C., Klein, R., and Klein, B. (1995). Smoothing spline anova for exponential families, with application to the wisconsin epidemiological study of diabetic retinopathy. *The Annals of Statistics*, 23(6):1865–1895.

Wainwright, M. J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, volume 48. Cambridge University Press.

Wang, X. and Berger, J. O. (2016). Estimating shape constrained functions using gaussian processes. *SIAM/ASA Journal on Uncertainty Quantification*, 4(1):1–25.

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

Zhang, H., Zheng, Z., and Lavaei, J. (2023). Gradient-based algorithms for convex discrete optimization via simulation. *Operations Research*, 71(5):1815–1834.

Zhu, H., Yao, F., and Zhang, H. H. (2014). Structured functional additive regression in reproducing kernel hilbert spaces. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(3):581–603.
Supplementary Appendix for Nonparametric Estimation via Partial Derivatives

A Optimal Rates Under Deterministic Designs

We present the minimax optimal rates under deterministic designs. Specifically, we consider the regular lattice design, which is also called the tensor product design. A regular lattice of size $n = l_1 \times \cdots \times l_d$ on $\mathcal{X}^d$ is a collection of design points $\{t_1, \ldots, t_n\} = \{(t_{i_1,1}, t_{i_2,2}, \ldots, t_{i_d,d}) \mid i_j = 1, \ldots, l_j, j = 1, \ldots, d\}$, where $t_{i,j} = i/l_j$, $i = 1, \ldots, l_j$, $j = 1, \ldots, d$.

This design is widely used for SS-ANOVA models (Wahba et al., 1995; Lin, 2000). Under regular lattices, it is without loss of generality to assume that $f_0 : \mathcal{X}^d \rightarrow \mathbb{R}$ has a periodic boundary condition. This is because any finite sequence $\{f(t_1), \ldots, f(t_n)\}$ can be associated with a periodic sequence, $f_{\text{per}}(i_1/l_1, \cdots, i_d/l_d) = \sum_{q_1 = -\infty}^{\infty} \cdots \sum_{q_d = -\infty}^{\infty} f(i_1/l_1 - q_1, \cdots, i_d/l_d - q_d)$, $\forall (i_1, \ldots, i_d) \in \mathbb{Z}^d$, where $\mathbb{Z}$ is the set of integers, and let $f(\cdot) \equiv 0$ outside and on the unobserved boundaries of $\mathcal{X}^d$. On the other hand, any finite sequence $\{f(t_1), \ldots, f(t_n)\}$ can be recovered from periodic sequence $f_{\text{per}}(\cdot)$. We now present the main results under deterministic design by first stating a minimax lower bound.

**Theorem 4.** Assume that $\lambda_\nu \asymp \nu^{-2m}$ for some $m > 3/2$. Under the regression models (1) and (2) where $f_0$ follows the SS-ANOVA model (4) and the designs $t^{(0)}$ and $t^{(j)}$s are from the regular lattice. Then under the error structure (3), there exists a constant $c$ that does not depend on $n$ such that

$$\liminf_{n \to \infty} \inf_{f_0 \in \mathcal{H}} \sup_{f \in \mathcal{H}} \mathbb{E} \int_{\mathcal{X}^d} \left[ \tilde{f}(t) - f_0(t) \right]^2 \, dt \geq \begin{cases} c \left[ n(\log n)^{1-(d-p)/r} \right]^{2m/(2m+1)}, & \text{if } 0 \leq p < d, \\ c \left[ n^{-2m/(2m+1)} \mathbb{1}_{r \geq 3} + n^{-1}(\log n)^{-1} \mathbb{1}_{r < 3} \right], & \text{if } p = d, \end{cases}$$

where the infimum of $\tilde{f}$ is taken over all measurable functions of the data.
The lower bound is established via the analysis of a version of the hardest rectangular subproblem. See, e.g., [Donoho et al. (1990)]. We relegate its proof to Section E. Next, we show that the rates given in the lower bound in Theorem 4 is attainable by the estimator $\hat{b}_f n$ in (9). Hence $\hat{f}_n$ is also minimax rate optimal under deterministic design.

Theorem 5. Assume that $\lambda \asymp \nu^{-2m}$ for some $m > 3/2$. Under the regression models (1) and (2) where $f_0$ follows the SS-ANOVA model (4) and the designs $t^{(0)}$ and $t^{(j)}$s are from the regular lattice. Then under the error structure (3), there exists a constant $C$ that does not depend on $n$ such that the estimator $\hat{b}_f n$ defined by (9) satisfies

$$\limsup_{n \to \infty} \sup_{f_0 \in H} \mathbb{E} \int_{\mathbb{R}^d} \left[ \hat{f}_n(t) - f_0(t) \right]^2 dt \leq \begin{cases} C \left[ n (\log n)^{1-(d-p)\wedge r} \right]^{-2m/(2m+1)}, & \text{if } 0 \leq p < d, \\ C \left[ n^{-2mr/(2m+1)} \mathbb{1}_{r \geq 3} + n^{-1}(\log n)^{-1} \mathbb{1}_{r < 3} \right], & \text{if } p = d. \end{cases}$$

Here the tuning parameter $\lambda$ in (9) is chosen by $\lambda \asymp \left[ n (\log n)^{1-(d-p)\wedge r} \right]^{-2m/(2m+1)}$ when $0 \leq p < d$, and $\lambda \asymp n^{-(2mr-2)/(2m+1)r-2}$ when $p = d, r \geq 3$, and $\lambda \asymp (n \log n)^{-(2m-1)/2m}$ when $p = d, r = 2$, and $\lambda \asymp n^{-(m-1)/m}$ when $p = d, r = 1$.

The proof of Theorem 5 is also presented in Section E. Theorems 4 and 5 together imply that under deterministic design, the minimax optimal rate for estimating $f_0 \in H$ with partial derivatives is

$$\left[ n (\log n)^{1-(d-p)\wedge r} \right]^{-2m/(2m+1)} \mathbb{1}_{0 \leq p < d} + \left[ n^{-2mr/(2m+1)r-2} \mathbb{1}_{r \geq 3} + n^{-1}(\log n)^{-1} \mathbb{1}_{r < 3} \right] \mathbb{1}_{p = d}.$$ 

This result coincides with the rate given by (15) under random design. Different from ours, Hall and Yatchew (2010) proposed a series-type estimator for incorporating various derivative data under the regular lattice. Hall and Yatchew (2010) showed that their estimator achieves the $\sqrt{n}$-consistency when sufficiently high-order derivatives are available. However, it is difficult to obtain high-order derivative data in practice, such as in economics and stochastic simulation. In contrast, we focus on incorporating first-order partial derivatives that are easier to obtain in practice. Chen et al. (2013) studied a stochastic kriging method for
incorporating partial derivatives, and analyzed its estimation error under certain widely spread designs, where the spatial correlations of observational errors at distinct design points approximately vanish. However, rates of convergence are not studied in Chen et al. (2013). By contrast, we quantify the improved rates of convergence with partial derivatives, which result holds under the general error structure (3).

B Error structures of common gradient estimators

We give three examples to illustrate that the random error assumption in (3) holds for gradient estimators that are commonly used in real-world settings.

Example 1: Infinitesimal perturbation analysis (IPA). In Section 4.1, we studied the example of call option pricing with stochastic simulations, where the unbiased gradient estimators are derived using IPA. Generally, IPA estimators are obtained under the condition (see, Ankenman et al., 2010; Chen et al., 2013) that common random numbers are not used across design points. Then, correlation exists only within the error terms \((\epsilon_i^{(0)}, \epsilon_i^{(1)}, \ldots, \epsilon_i^{(p)})^\top\) for the same design point \(i\) and not between those of different design points, \(\text{Cov}[\epsilon_i^{(j)}, \epsilon_i^{(j')}] = 0\), where \(i \neq i'\) and \(j, j' = 0, 1, \ldots, p\). Therefore, the errors of IPA gradient estimators satisfy the error assumption (3).

Moreover, define the correlation between the simulation noise in the response and in the estimator of the \(r\)th gradient component as \(\rho_i^{(0,j)} = \text{Corr}[\epsilon_i^{(0)}, \epsilon_i^{(j)}, j = 1, \ldots, p\). Let the correlation between the simulation noise in the estimators of a pair of distinct gradient components be \(\rho_i^{(j_1,j_2)} = \text{Corr}[\epsilon_i^{(j_1)}, \epsilon_i^{(j_2)}, j_1, j_2 = 1, \ldots, p\) and \(j_1 \neq j_2\). Notably, our error assumption (3) accommodates the scenario where the correlations \(\rho_i^{(0,j)}\) and \(\rho_i^{(j_1,j_2)}\) at different design points are not necessarily equal. This characteristic is consistent with the properties of the IPA estimators as shown in Ankenman et al. (2010) and Chen et al. (2013).

Example 2: Observational gradients. In Section 4.2 we considered the example of cost estimation in economics, where the gradient data are directly observable. More specifically,
the partial derivatives of \( f_0(t) \) with respect to input prices correspond to observable quantities of factor inputs.

In such observational studies where derivative data are available, the errors are commonly assumed to be i.i.d. (Hall and Yatchew 2007). Then, \( \text{Cov}[\epsilon_i^{(j)}, \epsilon_i^{(j')}] = 0 \), where \( i \neq i' \) and \( j, j' = 0, 1, \ldots, p \). Therefore, the errors of observational gradients satisfy the error assumption (3).

**Example 3: Finite difference method.** We explore the finite difference method as an alternative approach to derivative estimation, as applied in the life table estimation example in Appendix C.2. Specifically, we consider the finite-difference gradient estimator at \( t_{i}^{(0)} \in \mathbb{R} \) for \( i = 1, \ldots, n - 1 \),

\[
\frac{df_0}{dt}(t_{i}^{(0)}) \equiv \frac{y_{i+1}^{(0)} - y_{i}^{(0)}}{t_{i+1}^{(0)} - t_{i}^{(0)}} = f(t_{i+1}^{(0)}) - f(t_{i}^{(0)}) + \frac{\epsilon_{i+1}^{(0)} - \epsilon_{i}^{(0)}}{t_{i+1}^{(0)} - t_{i}^{(0)}} \quad \text{term I} = f'(t_{i}^{(0)}) + \left( f(t_{i+1}^{(0)}) - f(t_{i}^{(0)}) \right) \left( \frac{1}{t_{i+1}^{(0)} - t_{i}^{(0)}} - f'(t_{i}^{(0)}) \right) \quad \text{term II}.
\]

By the Taylor expansion, we have

\[
\text{term I} = \frac{1}{2} f''(\tilde{t})(t_{i+1}^{(0)} - t_{i}^{(0)}),
\]

where \( \tilde{t} \) lies between \( t_{i}^{(0)} \) and \( t_{i+1}^{(0)} \). Assuming that the observation errors \( \epsilon_i^{(0)} \)'s of function data are i.i.d. and centered, and considering the continuity of the second-order derivative of \( f \) along with \( |t_{i+1}^{(0)} - t_{i}^{(0)}| = o(n^{-1/2}) \), the bias of the finite-difference gradient estimator satisfies,

\[
\mathbb{E}[\epsilon_i^{(1)}] = \mathbb{E}[\text{term I}] + \mathbb{E}[\text{term II}] = \frac{1}{2} f''(\tilde{t})(t_{i+1}^{(0)} - t_{i}^{(0)}) = o(n^{-1/2}).
\]

Note that the assumption \( |t_{i+1}^{(0)} - t_{i}^{(0)}| = o(n^{-1/2}) \) is mild and typically satisfied in practical settings, such as when \( t_i^{(0)} \)'s are equally spaced in \( X = [0, 1] \), where \( |t_{i+1}^{(0)} - t_{i}^{(0)}| = 1/n = o(n^{-1/2}) \). Moreover, for \( |i - i'| > 1 \), we have \( \text{Cov}[\epsilon_i^{(0)}, \epsilon_i^{(1)}] = 0 \) and \( \text{Cov}[\epsilon_i^{(1)}, \epsilon_i^{(1)}] = 0 \). Hence, the covariance of the finite-difference gradient estimator satisfies,

\[
\text{Cov}[\epsilon_i^{(j)}, \epsilon_i^{(j')}](|i - i'|)^{-2},
\]

37
where \( i \neq i' \) and \( j, j' = 0, 1 \). Therefore, the errors of finite-difference gradient estimators satisfy the error assumption (3).

## Additional Numerical Examples

In this section, we provide additional numerical examples. We study a manufacturing example in Section C.1, analyze a real dataset on an actuarial life table in Section C.2, and explore a statistical inference example on cost estimation in Section C.3.

### C.1 Flexible assembly systems in manufacturing

We study a stochastic simulation in manufacturing that generates partial derivatives. Closed-loop flexible assembly system (CLFAS) is a useful tool to lower production costs and increase flexibility in manufacturing (Suri and Leung, 1987; Chen et al., 2013).

Since building a CLFAS is expensive, it is important to provide a fast and accurate prediction to the CLFAS performance. We consider a CLFAS of six automatic workstations and a conveyor with six pallets shown in Figure 2. Note that our analysis can be extended to any number of workstations or pallets. In this CLFAS, unfinished parts are loaded and unloaded through workstation 1 and proceed on the pallets. The operation time at each workstation \( j, 1 \leq j \leq 6 \), is given by \( t_j + 1\{\text{jam at station } j\}R_j \), where \( t_j \) is the fixed machine time (in minutes) and \( R_j \) is the additional random time (in minutes) to clear the machine \( j \) if it jams. Let \( p_j \) be the probability of a part causing a jam at workstation \( j \). Since the operation time is random, queueing may occur in the system. Our goal is to estimate \( f_0(t_1, \ldots, t_6) \), which denotes the expected production time of the first 5000 parts completed by the CLFAS. Here \( f_0 \) can be approximated by a SS-ANOVA model in (4) because if there is no queue occurs, \( f_0 \) has an additive structure in the covariates \((t_1, \ldots, t_6)\). In the experiment, we fix \( p_j = 0.5\% \) and let \( R_j \) i.i.d. uniformly sample from [0, 1.1]. The design points of \((t_1, \ldots, t_6)\) are uniformly random in [3, 9]^6 with the sample size \( n = 100 \). To address the impact of stochastic simulation noise, we simulate 1000 stochastic simulations of CLFAS at
each design and then average the results.

**Figure 2**: Diagram of CLFAS for the example in Section C.1

Suri and Leung (1987) proposed an IPA derivative estimators for a CLFAS as follows.

**Step 1**: Let $A_{j_1,j_2}$ be accumulator variables. Initialize: $A_{j_1,j_2} = 0$ for $j_1, j_2 = 1, \ldots, 6$;

**Step 2**: At the end of an operation at station $j$, let $A_{j,j} \leftarrow A_{j,j} + 1$, $j = 1, \ldots, 6$;

**Step 3**: If a pallet leaving station $j_1$ going to station $j'_1$ terminates an idle period of station $j'_1$, let $A_{j'_1,j_2} \leftarrow A_{j_1,j_2}$, $j_2 = 1, \ldots, 6$;

**Step 4**: If a pallet leaving station $j_1$ going to station $j'_1$ terminates a blocked period of station $j_1$, let $A_{j_1,j_2} \leftarrow A_{j'_1,j_2}$, $j_2 = 1, \ldots, 6$;

**Step 5**: At the end of the simulation, let $P$ be the total number of parts completed and $L$ be the full length of simulation in minutes. Output the function data $Y^{(0)}(t) = L/P$ and the IPA derivative estimator $Y^{(j)}(t) = A_{6,j}/P$ for $j = 1, \ldots, 6$.

In the data generating process, the correlation only exists for function and derivative data at the same design, not data across different design points. Hence the random errors satisfy the error structure in (3). In this example, obtaining function data at a new design requires to conduct 1000 new simulation replications. However, it only needs to record a small matrix $\{A_{j_1,j_2}\}_{j_1,j_2=1}^6$ in the algorithm of Suri and Leung (1987) for obtaining the IPA derivative
estimators, whose computational cost is negligible compared to that of obtaining a new function data.

**Figure 3:** The box plots of MSEs of our estimator with derivative data and the stochastic kriging without derivative data, for the example in Section C.1

**Comparison to existing method.** We compare our estimator (13) and the stochastic kriging method (Ankenman et al., 2010). We use the 6-dimensional version of the tensor product Matérn kernel (19), and choose lengthscale parameters by the five-fold cross-validation. We estimate the MSE of estimation by a Monte Carlo sample of $10^4$ test points in $[3, 9]^6$. Since the true production time is unknown at each test point, we approximate it by replicating $10^6$ CLFAS experiments at each test point.

Figure 3 reports the MSEs for different methods: stochastic kriging with only function data (i.e., $p = 0$), and our estimator with derivative data (i.e., $p = 6$). The results are averaged over 1000 simulations. It is seen that incorporating partial derivatives leads to a significant improvement of estimation compared to without using the derivatives.
C.2 Life table estimation

We study a real data of U.S. 2015 period life table for the social security area (www.ssa.gov/OACT/STATS/table4c6.html#fn2), where the data separate the male and female population. The life table in actuarial science provides probabilities of survival and death at integer ages (Frees and Valdez 1998). To value payments that are not at integer ages, actuaries need to make a fractional age assumption of surviving at fractional ages. Our goal is to estimate the survival distribution function $f_0(t)$. Let $u(t)$ be the force of mortality function. It is known that (see, Frees and Valdez 1998),

$$f'_0(t) = -f_0(t)u(t).$$  \hspace{1cm} (20)

The function data $Y^{(0)}$ on $f_0(t)$ are generated using the death probability from life table. The force of mortality function $u(t)$ can be estimated using the number of people that survive at age $t$, where the detail is given as follows. Denote by $l(t)$ the number of people that survive at age $t$. Then a divided-difference estimator for $u(t)$ is (Jones and Mereu 2002),

$$u(0) = \frac{1}{2l(0)}[3l(0) - 4l(1) + l(2)], \quad u(t) = \frac{1}{2l(t)}[l(t - 1) - l(t + 1)] \text{ for } t > 0.$$ 

The function $Y^{(0)}$, together with the estimate of $u(t)$, yield the derivative $Y^{(1)}$ according to (20). We choose the design $t$ from equally spaced integers from $[0, 119]$ with the sample size $n = 5, 10, 15, 20$. The endpoints of $[0, 119]$ are included.

Table 5: The comparison of average MSEs and standard errors of our estimator with those of smoothing spline estimator, for the example in Section C.2 with 1000 simulations. The table shows metrics: “average MSE (standard error),” in units of $10^{-4}$.

|       | n = 5         | n = 10        | n = 15        | n = 20        |
|-------|---------------|---------------|---------------|---------------|
|       |               |               |               |               |
| M     | Smoothing spline estimator with $Y^{(0)}$ | 15.3674 (4.8815) | 6.7944 (2.2596) | 1.7687 (0.6676) | 0.1745 (0.0594) |
|       | Our estimator with $Y^{(0)} + Y^{(1)}$ | **7.4381 (2.5242)** | **1.6488 (0.5009)** | **0.3446 (0.1012)** | **0.0227 (0.0098)** |
| F     | Smoothing spline estimator with $Y^{(0)}$ | 23.0655 (7.1699) | 9.9948 (3.8025) | 2.2299 (0.8110) | 0.5925 (0.1569) |
|       | Our estimator with $Y^{(0)} + Y^{(1)}$ | **9.4745 (3.2385)** | **2.4790 (0.8654)** | **0.4091 (0.1015)** | **0.0755 (0.0152)** |
Comparison to existing method. Smoothing spline \cite{Wahba1990} is widely used for smoothing noisy data. We compare the results of our estimator \cite{13} using the estimated derivative and the smoothing spline without using the derivative. We use the Matérn kernel \cite{19} and estimate the MSE by using the full sample at \( t = 0, 1, \ldots, 119 \).

Table 5 reports the MSEs and standard errors for varying sample size \( n \), different population, and different methods: smoothing spline with only function data (i.e., \( p = 0 \)), and our estimator with function and derivative data (i.e., \( p = 1 \)). The results are obtained over 1000 simulations in each setting. It is seen that our estimator incorporating derivative data significantly improves the estimation results compared to the smoothing splines.

Table 6 reports the ratios of the MSE of our estimator incorporating derivative data (i.e., \( p = 1 \)) relative to the MSE of smoothing spline estimator with only function data (i.e., \( p = 0 \)). It is seen that the ratio decreases with the sample size, which agrees with our theory in Section A that incorporating derivative data accelerates the convergence rate.

\begin{table}[h]
\centering
\begin{tabular}{lcccc}
\hline
 & \( n = 5 \) & \( n = 10 \) & \( n = 15 \) & \( n = 20 \) \\
\hline
Male & 0.4840 & 0.2426 & 0.1948 & 0.1301 \\
Female & 0.4108 & 0.2480 & 0.1835 & 0.1274 \\
\hline
\end{tabular}
\caption{The ratios of MSE of our estimator with derivative data (i.e., \( p = 1 \)) relative to MSE of spline estimator with only function data (i.e., \( p = 0 \)), for the example in Section C.2}
\end{table}

C.3 Statistical inference for the cost estimation in economics

We consider the economic problem of the cost function estimation in Section 4.2. We employ the bootstrap method (see, e.g., \cite{Efron1993}) to quantify the uncertainty of our estimators \cite{13} for this example. The process for generating a bootstrap sample includes the following steps: (a) Produce \( B \) bootstrap samples by resampling centered residuals; (b) Re-estimate the functions to obtain \( B \) bootstrap estimates of \( f_0 \), denoted as \( \hat{f}_b^* \) for \( b = 1, \ldots, B \). From this, we can derive a bootstrap confidence interval for \( f_0 \) at any new input \( t_{new} \). Specifically, we determine the \( \alpha/2 \) and \( 1 - \alpha/2 \) sample quantiles from
\{\hat{f}_1^*(t_{\text{new}}), \ldots, \hat{f}_{B}^*(t_{\text{new}})\}, represented as \(z_{\alpha/2}^*\) and \(z_{1-\alpha/2}^*\), respectively. The confidence interval is thus \((z_{\alpha/2}^*, z_{1-\alpha/2}^*)\). Given that bias in non-parametric regression may affect the asymptotic coverage of bootstrap confidence intervals, two common correction strategies include undersmoothing and oversmoothing (see, e.g., Härdle and Bowman 1988; Hall 1992a,b). Undersmoothing is often preferred due to its simplicity and effectiveness (Hall 1992a). Our estimation procedure can be easily modified to incorporate undersmoothing by selecting a smaller smoothing parameter. Despite the potential for a modest gain in practical performance, these strategies require another ad hoc choice of the amount of undersmoothing or oversmoothing. Moreover, it is quite common to ignore this bias issue, essentially leading to the use of non-adjusted confidence intervals as suggested by Efron and Tibshirani (1993) and Ruppert et al. (2003). To keep the approach simple, we use the non-adjusted confidence intervals in this example with \(B = 2000\). We set the significance level at 95%. The empirical coverage probability is calculated as the percentage of instances in which the confidence interval covers \(f_0(t_{\text{new}})\) across 1000 repetitions, with \(t_{\text{new}}\) randomly drawn from \(X^d\) for each repetition.

Table 7 compares the coverage probability and interval length when incorporating various levels of gradients \((p = 0, 1, 2)\) using our method (13). The average length of the bootstrap confidence interval is computed across 1000 repetitions. We observe in Table 7 that the coverage probability of our estimator approximates 95% consistently across all gradient levels \((p = 0, 1, 2)\). However, intervals without gradient information have larger lengths compared to those incorporating gradients. This observation aligns with our theoretical finding in Section 3.3 that the inclusion of gradient data results in a faster decrease in the MSE of the estimator compared to excluding gradient data.

C.4 Additional comparisons with Hall and Yatchew’s estimator

We present two additional examples to compare our estimator with the regression-kernel estimator in Hall and Yatchew (2007).

The first example is the stochastic simulation on call option pricing in Section 4.1. We
Table 7: Coverage probability and length of 95% bootstrap confidence intervals, incorporating various levels of gradients \( (p = 0, 1, 2) \) using our method [13], for the example in Section C.3 with 1000 simulations.

| \( n \) = 100 | \( \rho = 0 \) | 95.9722 | 14.4226 | 95.9116 | 13.4315 | \( \rho = 0.4 \) | 94.4613 | 15.6566 | 96.1340 | 13.6916 | \( \rho = 0.9 \) | 94.1245 | 16.4833 | 94.1276 | 14.3109 |
| \( \rho = 0 \) | 96.8295 | 11.5146 | \( \rho = 0.4 \) | 96.1200 | 13.4637 | \( \rho = 0.9 \) | 96.9722 | 12.3477 | \( \rho = 0 \) | 97.3076 | 12.8906 | \( \rho = 0.4 \) | 96.9119 | 12.4307 | \( \rho = 0.9 \) | 97.0722 |
| \( \rho = 0 \) | 95.3200 | 8.6322 | \( \rho = 0.4 \) | 94.5275 | 9.4833 | \( \rho = 0.9 \) | 95.1417 | 9.4833 | \( \rho = 0 \) | 95.9650 | 8.6566 | \( \rho = 0.4 \) | 95.7476 | 12.1215 | \( \rho = 0.9 \) | 95.8447 |
| \( \rho = 0 \) | 95.8447 | 8.6322 | \( \rho = 0.4 \) | 95.9650 | 8.6566 | \( \rho = 0.9 \) | 95.1417 | 9.4833 | \( \rho = 0 \) | 95.9650 | 8.6566 | \( \rho = 0.4 \) | 95.7476 | 12.1215 | \( \rho = 0.9 \) | 95.8447 |
| \( \rho = 0 \) | 95.3200 | 8.6322 | \( \rho = 0.4 \) | 94.5275 | 9.4833 | \( \rho = 0.9 \) | 95.1417 | 9.4833 | \( \rho = 0 \) | 95.9650 | 8.6566 | \( \rho = 0.4 \) | 95.7476 | 12.1215 | \( \rho = 0.9 \) | 95.8447 |
| \( \rho = 0 \) | 95.9001 | 6.4732 | \( \rho = 0.4 \) | 95.3200 | 6.8322 | \( \rho = 0.9 \) | 95.0288 | 7.4983 | \( \rho = 0 \) | 95.9722 | 14.4226 | \( \rho = 0.4 \) | 94.4613 | 15.6566 | \( \rho = 0.9 \) | 94.1245 |
| \( \rho = 0 \) | 96.2507 | 5.2168 | \( \rho = 0.4 \) | 95.3559 | 5.7529 | \( \rho = 0.9 \) | 95.9213 | 5.9815 | \( \rho = 0 \) | 95.9116 | 13.4315 | \( \rho = 0.4 \) | 96.1340 | 13.6916 | \( \rho = 0.9 \) | 96.1276 |
| \( \rho = 0 \) | 96.1200 | 13.4637 | \( \rho = 0.4 \) | 96.2182 | 9.7177 | \( \rho = 0.9 \) | 96.9119 | 10.4494 | \( \rho = 0 \) | 97.3076 | 12.8906 | \( \rho = 0.4 \) | 96.9119 | 12.4307 | \( \rho = 0.9 \) | 96.9119 |

 adopting the same simulation setting, and use the actual output as the reference, which is given by \( f_0(S_0, r_*, \sigma_*) = S_0 \Phi(-d_1 + \sigma_*) - 100e^{-r_\cdot \Phi(-d_1)} \). Here \( d_1 = \sigma_*^{-1}[\log 100 - \log(S_0) - (r_* - \sigma_*)^2/2] \) and \( \Phi(\cdot) \) is the CDF of standard normal distribution. For the estimator in Hall and Yatchew (2007), we follow the approach in Hall and Yatchew’s Example 3 to average \((S_0, r_*)\) and \((S_0, \sigma_*)\) directions locally, and then average the estimates. The MSE \( = E((f_n - f_0)^2) \) is estimated using a Monte Carlo sample of \( 10^4 \) test points in \([80, 120] \times [0.01, 0.05] \times [0.2, 1] \).

Table 8 reports the MSEs and standard errors across varying sample size \( n \), replications of the simulation \( q \), and levels of gradient data. The results are summarized based on 1000 simulations for each scenario. It is seen that our estimator significantly enhances estimation accuracy compared to Hall and Yatchew’s estimator.

The second example is the single voltage clamp experiment in Section 4.3. We follow the same simulation setting. For the estimator in Hall and Yatchew (2007), we again follow the approach in Hall and Yatchew’s Example 3 to average \((t_1, t_2)\) and \((t_1, t_3)\) directions locally, and then average the estimates. The MSE \( = E((f_n - f_0)^2) \) is estimated using a Monte Carlo sample of \( 10^4 \) test points in \( \mathcal{X}^3 \). Since the true function \( f_0(t) \) is unknown at each test point,
Table 8: The average MSEs and standard errors of our estimator and those of Hall and Yatchew’s
estimator, considering various gradient types, for the example in Section 4.1 with 1000 simulations.
The table shows metrics: “average MSE (standard error),” in units of $10^{-2}$.

| $n$  | $q$  | Hall and Yatchew with $Y^{(0)} + Y^{(1)} + Y^{(2)}$ | Our Estimator with $Y^{(0)} + Y^{(1)} + Y^{(2)}$ | Hall and Yatchew with $Y^{(0)} + Y^{(1)} + Y^{(2)} + Y^{(3)}$ | Our Estimator with $Y^{(0)} + Y^{(1)} + Y^{(2)} + Y^{(3)}$ |
|------|------|--------------------------------------------------|---------------------------------|--------------------------------------------------|---------------------------------|
| 1000 | 12   | 12.1741 (3.8190)                                   | 8.5599 (3.8415)                  | 11.4690 (3.4460)                                  | 3.9507 (1.3516)                  |
| 7    | 2000 | 11.8920 (3.3524)                                   | 4.5767 (1.3534)                  | 10.8306 (3.1022)                                  | 2.2173 (0.6291)                  |
| 5000 | 10.9300 (2.8547)                               | 2.8012 (0.9527)                  | 10.1989 (2.6965)                                  | 1.8633 (0.5813)                  |
| 1000 | 7    | 7.6601 (2.9033)                                    | 2.2702 (0.8333)                  | 7.3001 (2.1872)                                   | 1.5684 (0.5730)                  |
| 14    | 2000 | 7.2160 (2.4019)                                   | 1.7510 (0.6079)                  | 7.0696 (2.0615)                                   | 1.2402 (0.5062)                  |
| 5000 | 6.9731 (2.3591)                               | 1.4351 (0.5593)                  | 6.2591 (1.9210)                                   | 1.1468 (0.4213)                  |
| 1000 | 6    | 6.1625 (2.0150)                                    | 1.3341 (0.5150)                  | 5.3861 (1.7399)                                   | 1.0912 (0.3570)                  |
| 21    | 2000 | 6.0483 (1.9180)                                   | 1.1994 (0.4180)                  | 5.0355 (1.6164)                                   | 0.8988 (0.2919)                  |
| 5000 | 5.7112 (1.8264)                               | 0.9541 (0.3654)                  | 4.7698 (1.4877)                                   | 0.7460 (0.2124)                  |

we approximate it by using total $N = 19$ real ion channel samples at each test point. The
function and gradient training data are generated using $N' = 10$ real ion channel samples,
which are randomly chosen from the total $N = 19$ samples. Table 9 reports the MSEs and
standard errors across varying sample size $n$, replications of the simulation $q$, and levels of
gradient data. The results are summarized based on 1000 simulations for each scenario.
Table 9 shows that our estimator outperforms Hall and Yatchew’s estimator in terms of
estimation accuracy.

Table 9: The average MSEs and standard errors of our estimator and those of Hall and Yatchew’s
estimator, considering various gradient types, for the example in Section 4.3 with 1000 simulations.
The table shows metrics: “average MSE (standard error),” in units of $10^{-6}$.

| $n$  | Hall and Yatchew with $Y^{(0)}$ | Our Estimator with $Y^{(0)}$ | Hall and Yatchew with $Y^{(0)} + Y^{(1)} + Y^{(2)}$ | Our Estimator with $Y^{(0)} + Y^{(1)} + Y^{(2)} + Y^{(3)}$ |
|------|---------------------------------|--------------------------------|--------------------------------------------------|--------------------------------------------------|
| 1000 | 11.0134 (5.6061)                | 10.6491 (4.9867)                | 8.6488 (4.4921)                                   | 7.7804 (3.6737)                                  |
| 2000 | 9.0626 (5.0207)                 | 8.5302 (4.3339)                 | 6.3674 (3.1476)                                   | 5.1375 (2.4687)                                  |
| 3000 | 7.0134 (4.2182)                 | 6.4296 (3.9595)                 | 5.0655 (2.4226)                                   | 3.1035 (1.7187)                                  |
| 5000 | 6.2315 (3.4613)                 | 5.4143 (3.2268)                 | 3.1745 (1.6182)                                   | 2.1305 (0.9322)                                  |
D Proofs of the Main Results

D.1 Proof of Theorem 1

We prove a more general result in the following lemma. Let

\[ l_n(f) \equiv \frac{1}{n} \sum_{i=1}^{n} \left[ y_i^{(0)} - f(t_i^{(0)}) \right]^2 + \sum_{j=1}^{p} w_j \cdot \frac{1}{n} \sum_{i=1}^{n} \left[ y_i^{(j)} - \frac{\partial f}{\partial t_j}(t_i^{(j)}) \right]^2. \]

Then the optimization problem (9) can be rewritten as,

\[ \min_{f \in \mathcal{H}} l_n(f) \text{ subject to } \|f\|_{\mathcal{H}} \leq R_n. \]

Lemma 1. Let \( f_{I,n} \) is the unique solution to the problem: \( \min_{f \in \mathcal{H}} \|f\|_{\mathcal{H}} \text{ subject to } l_n(f) = 0. \)

Then, for \( 0 \leq R_n < \|f_{I,n}\|_{\mathcal{H}}, \) there exists a unique minimizer \( \hat{f}_n(t) \) of (9) in a finite-dimensional space. Specifically, there exist coefficients \( \alpha_j = (\alpha_{1j}, \ldots, \alpha_{nj})^T \in \mathbb{R}^n \) for \( j = 0, 1, \ldots, p \) such that,

\[ \hat{f}_n(t) = \sum_{i=1}^{n} \alpha_{i0} K_d(t_i^{(0)}, t) + \sum_{j=1}^{p} \sum_{i=1}^{n} \alpha_{ij} \frac{\partial K_d}{\partial t_j}(t_i^{(j)}, t), \]

(21)

and \( \|\hat{f}_n\|_{\mathcal{H}} = R_n. \) For \( R_n \geq \|f_{I,n}\|_{\mathcal{H}}, \) \( \hat{f}_n(t) \) in (21) is one of the minimizers of (9).

Proof. Following the proof of Lemma 1 and Proposition 3 of Lim (2024), there exists a unique solution to the problem:

\[ \min_{f \in \mathcal{H}} \|f\|_{\mathcal{H}} \text{ subject to } l_n(f) = 0, \]

which is denoted by \( f_{I,n}. \) Additionally, if \( 1 \leq R_n < J(f_I) \), there exists a unique minimizer \( \hat{f}_n(t) \) of (9) that satisfies \( \|\hat{f}_n\|_{\mathcal{H}} = R_n. \) A similar result can be found in Theorem 3 of Schoenberg (1964).

Since the optimization problem of (9) is convex, by Lagrangian duality, it can be reformulated as

\[ \hat{f}_n = \arg \min_{f \in \mathcal{H}} \{ l_n(f) + \lambda \|f\|_{\mathcal{H}}^2 \}. \]
Here, for a fixed set of function data and partial derivatives, the smoothing parameter $\lambda \geq 0$ is a function of the radius $R_n \geq 0$. Under the condition (7), the derivative $\partial f / \partial t_j$ is a bounded linear functional in $\mathcal{H}$. Following a similar argument to that of Theorem 1.3.1 in Wahba (1990), $\hat{f}_n(t)$ takes the form in (21). For $R_n \geq \|f_{I,n}\|_{\mathcal{H}}$, following the proof of Proposition 6 of Lim (2024), $\hat{f}_n(t)$ in (21) is one of the minimizers of (9). This completes the proof of Lemma 1.

Next, by Lemma 1, we know that for any $R_n \geq 0$, $\hat{f}_n(t)$ in (21) is a minimizer of (9) and it is in a finite-dimensional space spanned by $\{K_d(t^{(0)}_i, \cdot), \partial K_d / \partial t_j(t^{(j)}_i, t); 1 \leq i \leq n, 1 \leq j \leq p\}$. This completes the proof of Theorem 1.

D.2 Proof of Theorem 2

We establish the lower bound under random design via Fano’s lemma (Tsybakov, 2009). It suffices to consider a particular case where the random errors $\epsilon^{(0)}$ and $\epsilon^{(j)}$s are independent Gaussian with zero mean and unit variance, and $\Pi^{(0)}$ and $\Pi^{(j)}$s are uniform distributions, and $\mathcal{H}_1$ is generated by periodic kernels. The lower bound established for this case is at least for the general cases (Tsybakov, 2009).

Let $N$ be a natural number whose value will be clear later. We first derive the eigenvalue decay rate for the kernel $K_d$, which generates the RKHS $\mathcal{H}$. We introduce some additional notation. Define a family of the multi-index $\vec{\nu}$ by

$$V = \{\vec{\nu} = (\nu_1, \ldots, \nu_d)^\top \in \mathbb{N}^d, \text{ where at most } r \geq 1 \text{ of } \nu_k \text{s are not 1}\}. \quad (22)$$

For a given $\tau > 0$, the number of multi-indices $\vec{\nu} = (\nu_1, \ldots, \nu_r) \in \mathbb{N}^r$ satisfying

$$\nu_1^{-2m} \cdots \nu_r^{-2m} \geq \tau$$

is the same as the number of multi-indices such that $\nu_1 \cdots \nu_r \leq \tau^{-1/(2m)}$, which amounts to

$$\sum_{\nu_2 \cdots \nu_r \leq \tau^{-1/(2m)}} \tau^{-1/(2m)} / (\nu_2 \cdots \nu_r) = \tau^{-1/(2m)} \left( \sum_{\nu \leq \tau^{-1/(2m)}} 1/\nu \right)^{r-1} \approx \tau^{-1/(2m)} \left( \log \frac{1}{\tau} \right)^{r-1}. \quad (23)$$
Denote by $\lambda_N(K_d)$ the $N$th eigenvalues of $K_d$. By inverting (23), we obtain
\[
\lambda_N(K_d) \asymp \left[N(\log N)^{1-r}\right]^{-2m}.
\]
Hence, the multi-indices $\vec{\nu} = (\nu_1, \ldots, \nu_r) \in \mathbb{N}^r$ satisfying $\nu_1 \cdots \nu_r \leq N$ correspond to the first
\[
c_0 N(\log N)^{r-1}
\]
eigenvalues of $K_d$, for some constant $c_0$. Let $b$ be a length-$\{c_0 N(\log N)^{r-1}\}$ binary sequence,
\[
b = \{b_{\vec{\nu}} : \nu_1 \cdots \nu_r \leq N\} \in \{0, 1\}^{c_0 N(\log N)^{r-1}}.
\]
Let $\{\tilde{\lambda}_{\vec{\nu}} : \nu_1 \cdots \nu_r \leq N\}$ be the first $c_0 N(\log N)^{r-1}$ eigenvalues of $K_d$. Denote by
\[
\{\tilde{\lambda}_{\vec{\nu} + c_0 N(\log N)^{r-1}} : \nu_1 \cdots \nu_r \leq N\}
\]
the $\{c_0 N(\log N)^{r-1} + 1\}$th, $\{c_0 N(\log N)^{r-1} + 2\}$th, $\ldots$, $\{2c_0 N(\log N)^{r-1}\}$th eigenvalues of $K_d$.

For brevity, we only prove for the case $p = d$ and $r \geq 3$. The other cases $p = d$, $r \leq 2$ and $0 \leq p < d$ can be showed similarly. Write
\[
f_b(t_1, \ldots, t_r) = N^{-\frac{1}{2} + \frac{1}{r}} \sum_{\nu_1 \cdots \nu_r \leq N} b_{\vec{\nu}} (1 + \nu_1^2 + \cdots + \nu_r^2)^{-\frac{1}{2}}
\]
\[
\times \tilde{\lambda}_{\vec{\nu} + c_0 N(\log N)^{r-1}}^{\frac{1}{2}} \psi_{\vec{\nu} + c_0 N(\log N)^{r-1}}(t_1, \ldots, t_r),
\]
where $\psi_{\vec{\nu} + c_0 N(\log N)^{r-1}}(t_1, \ldots, t_r)$ are the corresponding eigenfunctions of $\tilde{\lambda}_{\vec{\nu} + c_0 N(\log N)^{r-1}}$ of $K_d$. Note that
\[
\|f_b\|^2_H = N^{-1 + \frac{2}{r}} \sum_{\nu_1 \cdots \nu_r \leq N} b_{\vec{\nu}}^2 (1 + \nu_1^2 + \cdots + \nu_r^2)^{-1}
\]
\[
\leq N^{-1 + \frac{2}{r}} \sum_{\nu_1 \cdots \nu_r \leq N} (1 + \nu_1^2 + \cdots + \nu_r^2)^{-1} \asymp 1,
\]
where the last step by Lemma 6, and this implies $f_b(\cdot) \in \mathcal{H}$. 48
By the Varshamov-Gilbert bound, e.g., Tsybakov (2009), there exists a collection of binary sequences \( \{ b^{(1)}, \ldots, b^{(M)} \} \subset \{0, 1\}^{c_0 N (\log N)^{r-1}} \) such that

\[
M \geq 2^{c_0 N (\log N)^{r-1/8}},
\]

and

\[
H(b^{(l)}, b^{(q)}) \geq c_0 N (\log N)^{r-1/8}, \quad \forall 1 \leq l < q \leq M.
\]

Here \( H(\cdot, \cdot) \) denotes the Hamming distance. Then, for \( b^{(l)}, b^{(q)} \in \{0, 1\}^{c_0 N (\log N)^{r-1}} \),

\[
\| f_{b^{(l)}} - f_{b^{(q)}} \|^2_{L_2} \geq N^{-1+2/r}(2N)^{-2m} \sum_{\nu_1 \cdots \nu_r \leq N} (1 + \nu_1^2 + \cdots + \nu_r^2)^{-1} [ \| f_{\bar{b}^{(l)}} - f_{\bar{b}^{(q)}} \|_2^2 ]
\]

\[
\geq N^{-1+2/r}(2N)^{-2m} \sum_{c_1 \nu_1 / \nu_r \leq N} (1 + \nu_1^2 + \cdots + \nu_r^2)^{-1}
\]

\[
= c_2 N^{-2m}
\]

for some constants \( c_1 \) and \( c_2 \), where the last step is by Lemma 6.

On the other hand, for any \( b^{(l)} \in \{ b^{(1)}, \ldots, b^{(M)} \} \), again by Lemma 6

\[
\| f_{b^{(l)}} \|^2_{L_2} \leq \sum_{j=1}^p \| \partial f_{b^{(l)}} / \partial t_j \|^2_{L_2} \leq N^{-1+2/r} \sum_{\nu_1 \cdots \nu_r \leq N} \nu_1^{-2m} \cdots \nu_r^{-2m} [ \| f_{\bar{b}^{(l)}} \|_2^2 ]
\]

\[
\leq N^{-1+2/r} \sum_{\nu_1 \cdots \nu_r \leq N} \nu_1^{-2m} \cdots \nu_r^{-2m} = c_3 N^{-2m+2/r}(\log N)^{r-1}
\]

for some constant \( c_3 \).

A standard argument gives that the lower bound can be reduced to the error probability in a multi-way hypothesis test (Tsybakov, 2009). Specifically, let \( \Theta \) be a random variable uniformly distributed on \( \{1, \ldots, M\} \). Note that

\[
\inf_{\bar{f}} \sup_{f_0 \in \mathcal{H}} \mathbb{P} \left\{ \| \bar{f} - f_0 \|^2_{L_2} \geq \frac{1}{4} \min_{b^{(l)} \neq b^{(q)}} \| f_{b^{(l)}} - f_{b^{(q)}} \|^2_{L_2} \right\} \geq \inf_{\hat{\Theta}} \mathbb{P} \{ \hat{\Theta} \neq \Theta \}. \quad (24)
\]

The infimum on the right-hand side is taken over all decision rules that are measurable functions of the data. By Fano’s lemma,

\[
\mathbb{P} \left\{ \hat{\Theta} \neq \Theta | t_1^{(0)}, \ldots, t_n^{(0)}; \ldots; t_1^{(p)}, \ldots, t_n^{(p)} \right\}
\]

\[
\geq 1 - \frac{1}{\log M} \times \left[ \prod_{i=1}^n \prod_{j=1}^p (y_1^{(0)}, \ldots, y_n^{(0)}, \ldots, y_1^{(p)}, \ldots, y_n^{(p)}; \Theta) + \log 2 \right], \quad (25)
\]

49
where
\[
\mathbb{I}_{t_1^{(0)}, \ldots, t_n^{(0)}; \ldots; t_1^{(p)}, \ldots, t_n^{(p)}} (y_1^{(0)}, \ldots, y_n^{(0)}, \ldots, y_1^{(p)}, \ldots, y_n^{(p)})
\]
is the mutual information between \( \Theta \) and \( \{y_1^{(0)}, \ldots, y_n^{(0)}, \ldots, y_1^{(p)}, \ldots, y_n^{(p)}\} \), and we fix the design points \( \{t_1^{(0)}, \ldots, t_n^{(0)}; \ldots; t_1^{(p)}, \ldots, t_n^{(p)}\} \). Thus,
\[
\begin{align*}
\mathbb{E}_{t_1^{(0)}, \ldots, t_n^{(0)}; \ldots; t_1^{(p)}, \ldots, t_n^{(p)}} \left[ \mathbb{I}_{t_1^{(0)}, \ldots, t_n^{(0)}; \ldots; t_1^{(p)}, \ldots, t_n^{(p)}} (y_1^{(0)}, \ldots, y_n^{(0)}, \ldots, y_1^{(p)}, \ldots, y_n^{(p)}; \Theta) \right] \\
\leq \left( \frac{M}{2} \right)^{-1} \sum_{b^{(l)} \neq b^{(q)}} \mathbb{E}_{t_1^{(0)}, \ldots, t_n^{(0)}; \ldots; t_1^{(p)}, \ldots, t_n^{(p)}} \mathcal{K}(P_{f_l^{(l)}}, P_{f_q^{(q)}}) \\
\leq \frac{n(p+1)}{2} \left( \frac{M}{2} \right)^{-1} \sum_{b^{(l)} \neq b^{(q)}} \mathbb{E}_{t_1^{(0)}, \ldots, t_n^{(0)}; \ldots; t_1^{(p)}, \ldots, t_n^{(p)}} \|f_{b^{(l)}} - f_{b^{(q)}}\|_{\ast n}^2 .
\end{align*}
\]

Here \( \mathcal{K}(\cdot, \cdot) \) is the Kullback-Leibler distance, \( P_f \) is conditional distribution of \( y_i^{(0)} \) and \( y_i^{(j)} \)'s given \( \{t_1^{(0)}, \ldots, t_n^{(0)}; \ldots; t_1^{(p)}, \ldots, t_n^{(p)}\} \), and the norm \( \| \cdot \|_{\ast} \) is defined as follows,
\[
\|f\|_{\ast n}^2 = \frac{1}{n(p+1)} \sum_{i=1}^{n} \left\{ [f(t_i^{(0)})]^2 + \sum_{j=1}^{p} \left| \frac{\partial f(t_i^{(j)})}{\partial t_j} \right|^2 \right\} , \quad \forall f : \mathcal{X} \mapsto \mathbb{R}.
\]

Thus,
\[
\begin{align*}
\mathbb{E}_{t_1^{(0)}, \ldots, t_n^{(0)}; \ldots; t_1^{(p)}, \ldots, t_n^{(p)}} \left[ \mathbb{I}_{t_1^{(0)}, \ldots, t_n^{(0)}; \ldots; t_1^{(p)}, \ldots, t_n^{(p)}} (y_1^{(0)}, \ldots, y_n^{(0)}, \ldots, y_1^{(p)}, \ldots, y_n^{(p)}; \Theta) \right] \\
\leq \frac{n(p+1)}{2} \left( \frac{M}{2} \right)^{-1} \sum_{b^{(l)} \neq b^{(q)}} \left\{ \|f_{b^{(l)}} - f_{b^{(q)}}\|_{L_2}^2 + \sum_{j=1}^{p} \|\partial f_{b^{(l)}}/\partial t_j - \partial f_{b^{(q)}}/\partial t_j\|_{L_2}^2 \right\} \\
\leq \frac{n(p+1)}{2} \max_{b^{(l)} \neq b^{(q)}} \left\{ \|f_{b^{(l)}} - f_{b^{(q)}}\|_{L_2}^2 + \sum_{j=1}^{p} \|\partial f_{b^{(l)}}/\partial t_j - \partial f_{b^{(q)}}/\partial t_j\|_{L_2}^2 \right\} \\
\leq 2n(p+1) \max_{b^{(l)} \in \{b^{(1)}, \ldots, b^{(M)}\}} \left\{ \|f_{b^{(l)}}\|_{L_2}^2 + \sum_{j=1}^{p} \|\partial f_{b^{(l)}}/\partial t_j\|_{L_2}^2 \right\} \\
\leq 2c_3 n(p+1) N^{-2m+\frac{2}{7}} (\log N)^{r-1}.
\end{align*}
\]
Now, (25) yields that
\[
\inf_{\tilde{f}} \sup_{f_0 \in \mathcal{H}} P \left\{ \| \tilde{f} - f_0 \|_{L^2}^2 \geq \frac{1}{4} c_2 N^{-2m} \right\} \\
\geq \inf_{\hat{\Theta}} P \{ \hat{\Theta} \neq \Theta \} \\
\geq 1 - \frac{1}{\log M} \left[ \mathbb{E} \left[ \Pi_{t_1^{(0)}, \ldots, t_{m}^{(0)}, \ldots, t_{m}^{(p)}, \ldots, t_{m}^{(p)}; y_{1}^{(0)}, \ldots, y_{n}^{(0)}, \ldots, y_{1}^{(p)}, \ldots, y_{n}^{(p)} ; \Theta \right] + \log 2 \right] \\
\geq 1 - \frac{2c_3 n(p + 1)N^{-2m+\frac{1}{2}}(\log N)^{r-1} + \log 2}{c_0 (\log 2) N (\log N)^{r-1/8}}.
\]
Taking \( N = c_4 n^{r/(2mr+r-2)} \) with an appropriate choice of \( c_4 \), we have
\[
\limsup_{n \to \infty} \inf_{f_0 \in \mathcal{H}} P \left\{ \| \tilde{f} - f_0 \|_{L^2}^2 \geq c n^{-2m+1} \right\} > 0,
\]
where \( c \) does not depend on \( n \). In addition, \( \| \tilde{f} - f_0 \|_{L^2}^2 \geq \min_{\| f \|_{\mathcal{H}} \leq R_n} \| f - f_0 \|_{L^2}^2 \). This completes the proof of this theorem.

### D.3 Proof of Theorem 3

**Preliminaries.** We consider a general quadratic penalty \( J(\cdot) \) for the proposed method (9), where \( J(\cdot) \) is any squared semi-norm on the RKHS \( \mathcal{H} \). For example, when \( \mathcal{H}_1 = W_m^2(\mathcal{X}) \), it is common to choose \( J(\cdot) \) for penalizing only the smooth component of a function. In this case, an explicit form of \( J(\cdot) \) is presented in Wahba (1990). The following analysis holds for replacing \( J(\cdot) \) with the squared norm \( \| \cdot \|_{\mathcal{H}}^2 \).

We define a new norm for any \( f \in \mathcal{H} \),
\[
\| f \|_R^2 = \frac{1}{p + 1} \left[ \frac{1}{\sigma_0^2} \int f^2(t) d\Pi^{(0)}(t) + \sum_{j=1}^{p} \frac{1}{\sigma_j^2} \int \left\{ \partial f(t)/\partial t_j \right\}^2 d\Pi^{(j)}(t) \right] + J(f). \tag{28}
\]
Note that \( \| \cdot \|_R \) is a norm since it is a quadratic form and is equal to zero if and only if \( f = 0 \). Let \( \langle \cdot, \cdot \rangle_R \) be the inner product associated with \( \| \cdot \|_R \). Then by Lemma 7 the norm \( \| \cdot \|_R \) is equivalent to the norm \( \| \cdot \|_{\mathcal{H}} \) in RKHS \( \mathcal{H} \). In particular, \( \| f \|_R < \infty \) if and only if \( \| f \|_{\mathcal{H}} < \infty \).

We introduce another norm \( \| \cdot \|_0 \) given by
\[
\| f \|_0^2 = \frac{1}{p + 1} \left[ \frac{1}{\sigma_0^2} \int f^2(t) d\Pi^{(0)}(t) + \sum_{j=1}^{p} \frac{1}{\sigma_j^2} \int \left\{ \partial f(t)/\partial t_j \right\}^2 d\Pi^{(j)}(t) \right]. \tag{29}
\]

51
Let a function space $F_0$ be the direct sum of some set of the orthogonal subspaces in the decomposition of $\bigotimes_{j=1}^d L_2(\mathcal{X})$ as in (5) and equipped with the norm $\| \cdot \|_0$. Write $\langle \cdot, \cdot \rangle_0$ as the inner product associated with $\| \cdot \|_0$ in $F_0$.

Finally, we define the following norm. For $f \in H$,

$$
\|f\|_{L^2(a)}^2 = \sum_{\nu \in \mathcal{V}} \left( 1 + \frac{\rho_\nu}{\|\phi_\nu\|_{L^2}^2} \right)^a f_\nu^2 \|\phi_\nu\|_{L^2}^2,
$$

for $0 \leq a \leq 1$, \hspace{1cm} (30)

where $f_\nu = \langle f, \phi_\nu \rangle_0$. By direct calculations, when $a = 0$ this norm coincides with $\| \cdot \|_{L^2}$ on $F_0$, and when $a = 1$ this norm is equivalent to $\| \cdot \|_R$ on $H$.

Denote the loss function in (9) by $l_n(f)$, that is,

$$
l_n(f) = \frac{1}{n(p+1)} \left[ \frac{1}{\sigma^2_0} \sum_{i=1}^n (f(t_i^{(0)}) - y_i^{(0)})^2 + \sum_{j=1}^p \frac{1}{\sigma^2_j} \sum_{i=1}^n \left( \frac{\partial f(t_i^{(j)})}{\partial t_j} - y_i^{(j)} \right)^2 \right],
$$

and write $l_{n\lambda}(f) = l_n(f) + \lambda J(f)$. Then the estimator $\hat{f}_n = \arg \min_{f \in H} l_{n\lambda}(f)$. Denote the expected loss by $l_\infty(f) = \mathbb{E}l_n(f)$ and write $l_{\infty\lambda}(f) = l_\infty(f) + \lambda J(f)$. Since $l_{\infty\lambda}(f)$ a positive quadratic form in $f \in H$, it has a unique minimizer in $H$ given by

$$
\bar{f}_{\infty\lambda} = \arg \min_{f \in H} l_{\infty\lambda}(f).
$$

Let $f^* = \arg \min_{J(f) \leq R^*_n} \| f - f_0 \|_{L^2}^2$. Thus, we decompose

$$
\hat{f}_n - f_0 = (\hat{f}_n - \bar{f}_{\infty\lambda}) + (\bar{f}_{\infty\lambda} - f^*) + (f^* - f_0), \hspace{1cm} (31)
$$

where $(\hat{f}_n - \bar{f}_{\infty\lambda})$ is referred to stochastic error, $(\bar{f}_{\infty\lambda} - f^*)$ is referred to deterministic error, and $(f^* - f_0)$ is referred to approximation error; see, e.g., van der Vaart and Wellner (1996).

We omit the subscripts of $\bar{f}_{\infty\lambda}$ and $\hat{f}_n$ hereafter if no confusion occurs.

**Outline of the proof.** Since the distributions $\Pi^{(0)}$ and $\Pi^{(j)}$s are known, it suffices to consider the uniform distributions by the inverse transform sampling in Lemma\hspace{1cm}8. Moreover, since $f_0$ is a functional ANOVA model with component function spaces supported in a compact domain $\mathcal{X}^d \equiv [0, 1]^d$, one can smoothly extend $f_0$ to a larger compactly supported
domain \([0, 1 + \delta]^d\) and achieve periodicity on the new boundary. This is proven in Lemma \[2\] which also shows that the eigenvalue decay rate for the RKHS associated with the extended periodic function remains the same as that for the RKHS associated with the original function. Moreover, the probability of selecting \(t\) within the range \([[0, 1 + \delta]^d \setminus X^d\]) is \(O(\delta)\), which is negligible for a sufficiently small \(\delta\). Lemma \[10\] shows that the estimation error of \(f_0\) can be upper bounded by that of the extended periodic function. Hence, the upper bound of the estimation for the periodic function also applies to the original function \(f_0\). Therefore, we consider \(f_0\) has a periodic boundary in \(X^d\) in the proof. A similar technique has been used in literature; e.g., Hall and Yatchew (2010).

Write the trigonometrical basis on \(L_2(X)\) as \(\psi_1(t) = 1\), \(\psi_{2\nu}(t) = \sqrt{2}\cos 2\pi \nu t\) and \(\psi_{2\nu+1}(t) = \sqrt{2}\sin 2\pi \nu t\) for \(\nu \geq 1\). Let

\[
\phi_\nu(t_1, \ldots, t_d) = \frac{\psi_{\nu_1}(t_1) \cdots \psi_{\nu_d}(t_d)}{\|\psi_{\nu_1}(t_1) \cdots \psi_{\nu_d}(t_d)\|_0}.
\]

(32)

Since \(f_0\) has a periodic boundary in \(X^d\) and \(\pi^{(j)} \equiv 1\), \(\{\phi_\mu(t) : \nu \in \mathbb{V}\}\), where \(\mathbb{V}\) in (22) forms an orthogonal basis for \(\mathcal{H}\) in \(\langle \cdot, \cdot \rangle_R\); an orthogonal system for \(L_2(X^d)\); and an orthonormal basis for \(F_0\) in \(\langle \cdot, \cdot \rangle_0\), that is \(\langle \phi_\mu(t), \phi_\mu(t) \rangle_0 = \delta_{\nu\mu}\), where \(\delta_{\nu\mu}\) is Kronecker’s delta; see, e.g., Chapter 2 in Wahba (1990). The concept of simultaneous orthogonality of a basis in multiple inner product spaces has been explored in other RKHS settings; see, e.g., Section 3 in Yuan and Cai (2010). Hence, any \(f \in \mathcal{H}\) has the decomposition

\[
f(t_1, \ldots, t_d) = \sum_{\nu \in \mathbb{V}} f_\nu \phi_\nu(t_1, \ldots, t_d), \quad \text{where } f_\nu = \langle f(t), \phi_\nu(t) \rangle_0.
\]

(33)

We denote a positive scalar series \(\{\rho_\nu\}_{\nu \in \mathbb{V}}\) such that \(\langle \phi_\nu, \phi_\mu \rangle_R = (1 + \rho_\nu)\delta_{\nu\mu}\). Then,

\[
J(f) = \langle f, f \rangle_R - \langle f, f \rangle_0 = \sum_{\nu \in \mathbb{V}} \rho_\nu f_\nu^2.
\]

(34)

First, we analyze the deterministic error \((\bar{f} - f^\dagger)\). By (33), write \(f^\dagger(t) = \sum_{\nu \in \mathbb{V}} f^\dagger_\nu \phi_\nu(t)\) and \(\bar{f}(t) = \sum_{\nu \in \mathbb{V}} \bar{f}_\nu \phi_\nu(t)\). Note the bias satisfies \(\mathbb{E}[\epsilon_i^{(j)}] = o(n^{-1/2})\), we have

\[
l_\infty(f) = \sum_{\nu \in \mathbb{V}} (f_\nu - f^\dagger_\nu)^2 + o(n^{-1/2}) \sqrt{\sum_{\nu \in \mathbb{V}} (f_\nu - f^\dagger_\nu)^2} + 1,
\]

53
and
\[ \bar{f}_\nu = \frac{f^\dagger(1 + \kappa \bar{\nu})}{1 + \kappa \bar{\nu} + \lambda \rho \bar{\nu}}, \] where \( \kappa \bar{\nu} = o(1), \ \forall \bar{\nu} \in \mathbb{V}. \)

(35)

An upper bound of the deterministic error will be given in Lemma 2.

Second, we analyze the stochastic error \((\hat{f}_n - \bar{f})\). The existence of the following Fréchet derivatives is guaranteed by Lemma 3.

\[
Dl_n(f)g = \frac{2}{n(p + 1)} \left[ \frac{1}{\sigma^2} \sum_{i=1}^{n} \{ f(t_i^{(0)}) - y_i^{(0)} \} g(t_i^{(0)}) \right.
+ \sum_{j=1}^{p} \frac{1}{\sigma_j^2} \sum_{i=1}^{n} \left\{ \frac{\partial f(t_i^{(j)})}{\partial t_j} - y_i^{(j)} \right\} \frac{\partial g(t_i^{(j)})}{\partial t_j} \right],
\]

(36)

\[
Dl_\infty(f)g = \frac{2}{p + 1} \left[ \frac{1}{\sigma^2} \int \left\{ f(t) - f_0(t) + o(n^{-1/2}) \right\} g(t) d\Pi^{(0)}(t) \right.
+ \sum_{j=1}^{p} \frac{1}{\sigma_j^2} \int \left\{ \frac{\partial f(t)}{\partial t_j} - \frac{\partial f_0(t)}{\partial t_j} + o(n^{-1/2}) \right\} \frac{\partial g(t)}{\partial t_j} d\Pi^{(j)}(t) \right] \right],
\]

(37)

\[
D^2l_n(f)gh = \frac{2}{n(p + 1)} \left[ \frac{1}{\sigma^2} \sum_{i=1}^{n} g(t_i^{(0)}) h(t_i^{(0)}) \right.
+ \sum_{j=1}^{p} \frac{1}{\sigma_j^2} \sum_{i=1}^{n} \frac{\partial g(t_i^{(j)})}{\partial t_j} \frac{\partial h(t_i^{(j)})}{\partial t_j} \right],
\]

(38)

\[
D^2l_\infty(f)gh = \frac{2}{p + 1} \left[ \frac{1}{\sigma^2} \int g(t) h(t) d\Pi^{(0)}(t) \right.
+ \sum_{j=1}^{p} \frac{1}{\sigma_j^2} \int \frac{\partial g(t)}{\partial t_j} \frac{\partial h(t)}{\partial t_j} d\Pi^{(j)}(t) \right] = 2\langle g, h \rangle_0,
\]

(39)

where \(Dl_n(f), Dl_\infty(f), D^2l_n(f)g,\) and \(D^2l_\infty(f)g\) are bounded linear operators on \(\mathcal{H}\). By Riesz representation theorem, with a slight abuse of notation, write

\[ Dl_n(f)g = \langle Dl_n(f), g \rangle_R, \quad Dl_\infty(f)g = \langle Dl_\infty(f), g \rangle_R, \]

\[ D^2l_n(f)gh = \langle D^2l_n(f), gh \rangle_R, \quad D^2l_\infty(f)gh = \langle D^2l_\infty(f), gh \rangle_R. \]

From Oden and Reddy (2012), there exists a bounded linear operator \(U : F_0 \mapsto \mathcal{H}\) such that \(U\phi_\nu = (1 + \rho \nu)^{-1} \phi_\nu\) and \(\langle f, Ug \rangle_R = \langle f, g \rangle_0\) for any \(f \in \mathcal{H}\) and \(g \in F_0\), and the restriction of
U to $\mathcal{H}$ is self-adjoint and positive definite. By (39), we further derive

$$D^2l_{\infty\lambda}(f)\phi_\vartheta(t) = 2(U + \lambda(I - U))\phi_\vartheta(t) = 2(1 + \rho_\vartheta)^{-1}(1 + \lambda\rho_\vartheta)\phi_\vartheta(t).$$

Define that $G_\lambda\phi_\vartheta = \frac{1}{2}D^2l_{\infty\lambda}(\bar{f})\phi_\vartheta$. By the Lax-Milgram theorem, $G_\lambda : \mathcal{H} \mapsto \mathcal{H}$ has a bounded inverse $G_\lambda^{-1}$ on $\mathcal{H}$, and

$$G_\lambda^{-1}\phi_\vartheta = (1 + \rho_\vartheta)(1 + \lambda\rho_\vartheta)^{-1}\phi_\vartheta. \quad (40)$$

Define

$$\tilde{f}^* = \bar{f} - \frac{1}{2}G_\lambda^{-1}Dl_{n\lambda}(\bar{f}).$$

Then the stochastic error can be decomposed as

$$\tilde{f}_n - \bar{f} = (\tilde{f}^* - \bar{f}) + (\tilde{f}_n - \tilde{f}^*).$$

The two terms on the right-hand side will be studied separately, and their upper bounds will be given in Lemma 4 and Lemma 5 respectively.

**Main proof.** Now, we give the details by following the above outline. First, we present an upper bound of the deterministic error $(\bar{f} - f^\dagger)$ in (31).

**Lemma 2.** For any $0 \leq a \leq 1$, the deterministic error in (31) satisfies

$$\|\bar{f} - f^\dagger\|_{L^2(a)}^2 = \begin{cases} O\{\lambda^{1-a}R_n^2\} & \text{when } 0 \leq p < d, \\ O\{\lambda^{\frac{(1-a)m}{m+1}}R_n^2\} & \text{when } p = d. \end{cases}$$

**Proof.** We first introduce some notations. For two positive sequences $a_n$ and $b_n$, we write $a_n \lesssim b_n$ (or $a_n \gtrsim b_n$) means that there exists a constant $c > 0$ (or $c' > 0$) such that $a_n \leq cb_n$ (or $a_n \geq c'b_n$) for all $n$. We write $a_n \asymp b_n$ if $a_n/b_n$ is bounded away from both zero and infinity as $n \to \infty$. 
For any \(0 \leq a \leq 1\), by (34) and (35), we have
\[
\|\bar{f} - f^\dagger\|^2_{L_2(a)} = \sum_{\nu \in \mathcal{V}} \left(1 + \frac{\rho_\nu}{\|\phi_\nu\|^2_{L_2}}\right)^a \left(1 + \frac{\lambda \rho_\nu}{1 + \kappa_\nu + \lambda \rho_\nu}\right)^2 \|\phi_\nu\|^2_{L_2} (f^\dagger_\nu)^2 \sum_{\nu \in \mathcal{V}} \rho_\nu (f^\dagger_\nu)^2
\]
\[
\lesssim \lambda^2 \sup_{\nu \in \mathcal{V}} \frac{(1 + \rho_\nu/\|\phi_\nu\|^2_{L_2})^a \rho_\nu \|\phi_\nu\|^2_{L_2}}{(1 + \lambda \rho_\nu)^2} \sum_{\nu \in \mathcal{V}} \rho_\nu (f^\dagger_\nu)^2 \tag{41}
\]
\[
\lesssim \lambda^2 R^2 \sup_{\nu \in \mathcal{V}} \frac{(\prod_{k=1}^d \nu_k^{2m})^{1+a}}{(1 + \sum_{j=1}^p \nu_j^2 + \lambda \prod_{k=1}^d \nu_k^{2m})^2}, \quad \nu \in \mathcal{V}.
\]
Write
\[
B_\lambda(\nu) = \frac{(\prod_{k=1}^d \nu_k^{2m})^{1+a}}{(1 + \sum_{j=1}^p \nu_j^2 + \lambda \prod_{k=1}^d \nu_k^{2m})^2}, \quad \nu \in \mathcal{V}.
\]
We discuss \(B_\lambda(\nu)\) for \(0 \leq p \leq d - 1\) and \(p = d\) separately.

For \(0 \leq p \leq d - 1\), since \(\nu \in \mathcal{V}\), there are at most \(r\) of \(\nu_1, \ldots, \nu_d\) not equal to 1. Suppose for any \(x = \prod_{k=1}^d \nu_k^{2m} > 0\) fixed. Then \(B_\lambda(\nu)\) is maximized by letting \(\sum_{j=1}^p \nu_j^2\) be as small as possible, which implies \(\nu_1 = \nu_2 = \cdots = \nu_p = 1\). Then,
\[
\sup_{\nu \in \mathcal{V}} B_\lambda(\nu) \lesssim \sup_{(\nu_{p+1}, \ldots, \nu_{(p+r)\wedge d})^\top \in \mathbb{N}^{r \wedge (d-p)}} \frac{\prod_{k=p+1}^{(p+r)\wedge d} \nu_k^{2m(1+a)}}{(1 + \lambda \prod_{k=p+1}^{(p+r)\wedge d} \nu_k^{2m})^2} \tag{42}
\]
\[
\lesssim \sup_{x > 0} \frac{x^{-(1+a)}}{(1 + \lambda x^{-1})^2} \lesssim \lambda^{-(a+1)},
\]
where the last step is achieved when \(x \asymp \lambda\).

For \(p = d\), since \(\nu \in \mathcal{V}\) and by the symmetry of coordinates \(\nu_1, \ldots, \nu_d\), assume that all indices except \(\nu_1, \ldots, \nu_r\) being 1. Letting \(z = \prod_{j=1}^r \nu_j^{2m} > 0\), we have
\[
\sup_{\nu \in \mathcal{V}} B_\lambda(\nu) \asymp \sup_{z > 0} \frac{z^{-(1+a)}}{(z^{1/mr} + \lambda z^{-1})^2} \asymp \lambda^{2-\frac{1+a}{mr-1}}, \tag{43}
\]
where the last step is achieved when \(z \asymp \lambda^{mr/(mr-1)}\). Combining (41), (42) and (43) we complete the proof.

Before we establish an upper bound for the stochastic error, we present the Fréchet derivative of the operator that will be used in the proof. Let \(X\) and \(Y\) be the normed linear
spaces. The Fréchet derivative of an operator $F : X \mapsto Y$ is a bounded linear operator $DF(a) : X \mapsto Y$ with
\[
\lim_{h \to 0, h \in X} \frac{\|F(a + h) - F(a) - DF(a)h\|_Y}{\|h\|_X} = 0.
\]
For example, if $F(a + h) - F(a) = Lh + R(a, h)$ with a linear operator $L$ and
\[
\frac{\|R(a, h)\|_Y}{\|h\|_X} \to 0, \quad \text{as } h \to 0,
\]
by definition then $L = DF(a)$ is the Fréchet derivative of $F(\cdot)$. The reader is referred to Gelfand and Silverman (2000) for a thorough investigation of the Fréchet derivative. We give the Fréchet derivative of the operator in our setting.

**Lemma 3.** Denote the loss function in (9) by $l_n(f)$. With the norm $\| \cdot \|_R$ in (28), the first-order Fréchet derivative of the functional $l_n(\cdot)$ for any $f, g \in H$ is
\[
Dl_n(f)g = \frac{2}{n(p + 1)} \left[ \frac{1}{\sigma_0^2} \sum_{i=1}^{n} \{ f(t_i^{(0)}) - y_i^{(0)} \} g(t_i^{(0)}) 
+ \sum_{j=1}^{p} \frac{1}{\sigma_j^2} \sum_{i=1}^{n} \left\{ \frac{\partial f(t_i^{(j)})}{\partial t_j} - y_i^{(j)} \right\} \frac{\partial g(t_i^{(j)})}{\partial t_j} \right].
\]
The second-order Fréchet derivative of $l_n(\cdot)$ for any $f, g, h \in H$ is
\[
D^2l_n(f)gh = \frac{2}{n(p + 1)} \left[ \frac{1}{\sigma_0^2} \sum_{i=1}^{n} g(t_i^{(0)})h(t_i^{(0)}) 
+ \sum_{j=1}^{p} \frac{1}{\sigma_j^2} \sum_{i=1}^{n} \frac{\partial g(t_i^{(j)})}{\partial t_j} \frac{\partial h(t_i^{(j)})}{\partial t_j} \right].
\]

**Proof.** By direct calculations, we have
\[
l_n(f + g) - l_n(f) = \frac{2}{n(p + 1)} \left[ \frac{1}{\sigma_0^2} \sum_{i=1}^{n} \{ f(t_i^{(0)}) - y_i^{(0)} \} g(t_i^{(0)}) 
+ \sum_{j=1}^{p} \frac{1}{\sigma_j^2} \sum_{i=1}^{n} \left\{ \frac{\partial f(t_i^{(j)})}{\partial t_j} - y_i^{(j)} \right\} \frac{\partial g(t_i^{(j)})}{\partial t_j} \right] + R_n(f, g),
\]
where
\[
\mathcal{R}_n(f, g) = \frac{1}{n(p+1)} \left[ \frac{1}{\sigma^2_0} \sum_{i=1}^n g_i^2(t_i^{(0)}) + \sum_{j=1}^p \frac{1}{\sigma^2_j} \sum_{i=1}^n \left( \frac{\partial g(t_i^{(j)})}{\partial t_j} \right)^2 \right]^{2/3}
= \|g\|_0^2 + \mathcal{O}(n^{-1/2}),
\]
and the \(\| \cdot \|_0\) norm is given in (29). Note that \(|\mathcal{R}_n(f, g)|/\|g\|_R \to 0\) as \(\|g\|_R \to 0\) and \(n^{1/2}\|g\|_R \to \infty\). This proves the first part of the lemma. For the second-order Fréchet derivative, note that
\[
Dl_n(f + h)g - Dl_n(f)g = \frac{2}{n(p+1)} \left[ \frac{1}{\sigma^2_0} \sum_{i=1}^n g(t_i^{(0)}) h(t_i^{(0)}) + \sum_{j=1}^p \frac{1}{\sigma^2_j} \sum_{i=1}^n \partial g(t_i^{(j)}) \partial h(t_i^{(j)}) \right],
\]
which is linear in \(h\). By definition, the \(D^2l_n(f)gh\) in the lemma is the valid second-order Fréchet derivative of \(l_n(\cdot)\). 

By following a similar derivation for Lemma 3, it is easy to obtain the first and the second-order Fréchet derivatives of the functional \(l_\infty(\cdot)\) in (37) and (39), respectively.

We now establish an upper bound for the term \((\tilde{\mathcal{F}}^* - \bar{f})\), which is a part of the stochastic error.

**Lemma 4.** When \(0 \leq p < d\), we have for any \(0 \leq a < 1 - 1/2m\),
\[
\|\tilde{\mathcal{F}}^* - \bar{f}\|_{L_2(a)}^2 = \mathbb{O}_P \left\{ n^{-1} \lambda^{-1/,2m} \right\}.
\]
When \(p = d\), we have for any \(0 \leq a \leq 1\),
\[
\|\tilde{\mathcal{F}}^* - \bar{f}\|_{L_2(a)}^2 = \begin{cases} 
\mathbb{O}_P \left\{ n^{-1} R_n^2 \lambda^{-1/,2m} \right\}, & \text{if } r \geq 3; \\
\mathbb{O}_P \left\{ n^{-1} R_n^2 \log(1/\lambda) \right\}, & \text{if } r = 2, a = 0; \quad \mathbb{O}_P \left\{ n^{-1} R_n^2 \right\}, & \text{if } r = 2, 0 < a \leq 1; \\
\mathbb{O}_P \left\{ n^{-1} R_n^2 \right\}, & \text{if } r = 1, a < \frac{1}{2m}; \quad \mathbb{O}_P \left\{ n^{-1} \log(1/\lambda) R_n^2 \right\}, & \text{if } r = 1, a = \frac{1}{2m}; \\
\mathbb{O}_P \left\{ n^{-1} \lambda^{1/,2m} R_n^2 \right\}, & \text{if } r = 1, a > \frac{1}{2m}.
\end{cases}
\]
Proof. Notice that $Dl_{n,\lambda}(\bar{f}) = Dl_{n,\lambda}(\bar{f}) - Dl_{\infty,\lambda}(\bar{f}) = Dl_{n}(\bar{f}) - Dl_{\infty}(\bar{f})$. Hence, for any $g \in \mathcal{H}$,

$$
\mathbb{E} \left[ \frac{1}{2} Dl_{n,\lambda}(\bar{f}) g \right]^2 = \mathbb{E} \left[ \frac{1}{2} Dl_n(\bar{f}) g - \frac{1}{2} Dl_{\infty}(\bar{f}) g \right]^2 \\
\leq \frac{1}{n(p+1)^2} \sum_{j=0}^p \text{Var} \left[ \frac{1}{\sigma_j^2} \left\{ \frac{\partial \bar{f}(t^{(j)})}{\partial t_j} - Y^{(j)} \right\} \frac{\partial g(t^{(j)})}{\partial t_j} \right] \\
+ \sum_{j=0}^p \frac{\sigma_j^{-4}}{n^2(p+1)^2} \sum_{i \neq i'} \text{Cov} \left[ \left( \frac{\partial \bar{f}(t_i^{(j)})}{\partial t_j} - y_i^{(j)} \right) \frac{\partial g(t_i^{(j)})}{\partial t_j} , \left( \frac{\partial \bar{f}(t_{i'}^{(j)})}{\partial t_j} - y_{i'}^{(j)} \right) \frac{\partial g(t_{i'}^{(j)})}{\partial t_j} \right] \\
+ \sum_{j=0}^p \frac{\sigma_j^{-2}\sigma_k^{-2}}{n^2(p+1)^2} \sum_{i \neq i', k} \text{Cov} \left[ \left( \frac{\partial \bar{f}(t_i^{(j)})}{\partial t_j} - y_i^{(j)} \right) \frac{\partial g(t_i^{(j)})}{\partial t_j} , \left( \frac{\partial \bar{f}(t_{i'}^{(k)})}{\partial t_k} - y_{i'}^{(k)} \right) \frac{\partial g(t_{i'}^{(k)})}{\partial t_k} \right],
$$

where the second step is due to $\sum_{i \neq i'} \text{Cov} [\epsilon_i^{(j)}, \epsilon_{i'}^{(k)}] = \sum_{i \neq i'} o(|i - i'|^{-\gamma}) = o(n)$. Note that (44) can be further bounded up to some constant by,

$$
\mathbb{E} \left[ \frac{1}{2} Dl_{n,\lambda}(\bar{f}) g \right]^2 = \frac{1}{n(p+1)} \mathbb{E} \left\{ \frac{1}{\sigma_0^2} \right\} \left\{ \bar{f}(t^{(0)}) - f_0(t^{(0)}) \right\}^2 \left\{ g(t^{(0)}) \right\}^2 + \frac{1}{\sigma_0^2} \mathbb{E} \left\{ g(t^{(0)}) \right\}^2 \\
+ \sum_{j=1}^p \frac{1}{\sigma_j^2} \mathbb{E} \left\{ \frac{\partial \bar{f}(t^{(j)})}{\partial t_j} - \frac{\partial f_0(t^{(j)})}{\partial t_j} \right\}^2 \left\{ \frac{\partial g(t^{(j)})}{\partial t_j} \right\}^2 + \sum_{j=1}^p \frac{1}{\sigma_j^2} \mathbb{E} \left\{ \frac{\partial g(t^{(j)})}{\partial t_j} \right\}^2 \\
+ o(n^{-1}) \frac{1}{(p+1)^2} \sum_{j,k=0}^p \mathbb{E} \left[ \frac{\partial g(t^{(j)})}{\partial t_j} \right] \mathbb{E} \left[ \frac{\partial g(t^{(k)})}{\partial t_k} \right].
$$

By Lemma 7, Lemma 11 and Cauchy-Schwarz inequality, we have that (45) is bounded up to some constant by

$$
\frac{1}{n(p+1)} \mathbb{E} \left\{ \frac{1}{\sigma_0^2} \right\} \left\{ \bar{f}(t^{(0)}) - f_0(t^{(0)}) \right\}^2 \left\{ g(t^{(0)}) \right\}^2 + \frac{1}{\sigma_0^2} \mathbb{E} \left\{ g(t^{(0)}) \right\}^2 \\
+ \sum_{j=1}^p \frac{1}{\sigma_j^2} \mathbb{E} \left\{ \frac{\partial \bar{f}(t^{(j)})}{\partial t_j} - \frac{\partial f_0(t^{(j)})}{\partial t_j} \right\}^2 \left\{ \frac{\partial g(t^{(j)})}{\partial t_j} \right\}^2 + \sum_{j=1}^p \frac{1}{\sigma_j^2} \mathbb{E} \left\{ \frac{\partial g(t^{(j)})}{\partial t_j} \right\}^2 \\
\leq n^{-1} R_n^2 \|g\|_0^2.
$$

where the last step above is by Lemma 2 and the definition of the norm $\| \cdot \|_0$. From the definition of $G^{-1}_\lambda$ in (40), we have that $\forall g \in \mathcal{H}$,

$$
\| G^{-1}_\lambda g \|_{L^2(a)}^2 = \sum_{\nu \in \mathcal{Y}} \left( 1 + \frac{\rho^2}{\| \phi_{\nu} \|_{L^2}^2} \right)^a \left( 1 + \lambda \rho^2 \right) \| \phi_{\nu} \|_{L^2}^2 \langle g, \phi_{\nu} \rangle_R^2.
$$
Then by the definition of $\tilde{f}^*$,
\[
\mathbb{E}\|\tilde{f}^* - \tilde{f}\|_{L_2(a)}^2 = \mathbb{E}\left\| \frac{1}{2} G_\lambda^{-1} Dl_{n\lambda}(\tilde{f}) \right\|_{L_2(a)}^2
= \frac{1}{4} \mathbb{E} \left[ \sum_{\tilde{v} \in \mathcal{V}} \left( 1 + \frac{\rho_{\tilde{v}}}{\|\phi_{\tilde{v}}\|_{L_2}^2} \right)^a (1 + \lambda \rho_{\tilde{v}})^{-2} \|\phi_{\tilde{v}}\|_{L_2}^2 \langle Dl_{n\lambda}(\tilde{f}), \phi_{\tilde{v}} \rangle_R \right]^2
\leq \sum_{\tilde{v} \in \mathcal{V}} \left( 1 + \frac{\rho_{\tilde{v}}}{\|\phi_{\tilde{v}}\|_{L_2}^2} \right)^a (1 + \lambda \rho_{\tilde{v}})^{-2} \|\phi_{\tilde{v}}\|_{L_2}^2 \mathbb{E} \left[ \frac{1}{2} Dl_{n\lambda}(\tilde{f}) \phi_{\tilde{v}} \right]^2
\lesssim n^{-1} R_n^2 \sum_{\tilde{v} \in \mathcal{V}} \left( 1 + \frac{\rho_{\tilde{v}}}{\|\phi_{\tilde{v}}\|_{L_2}^2} \right)^a (1 + \lambda \rho_{\tilde{v}})^{-2} \|\phi_{\tilde{v}}\|_{L_2}^2 \|\phi_{\tilde{v}}\|_0^2
\lesssim n^{-1} R_n^2 N_a(\lambda),
\]
where the fourth step is by \[46\] and the last step is because of $\|\phi_{\tilde{v}}\|_0 = 1$, $\|\phi_{\tilde{v}}\|_{L_2}^2 \asymp (1 + \sum_{j=1}^p \nu_j^2)^{-1}$, $\rho_{\tilde{v}} \asymp (1 + \sum_{j=1}^p \nu_j^2)^{-1} \prod_{k=1}^d \nu_k^{2m}$, and $N_a(\lambda)$ is defined in Lemma \[12\] Hence, by Lemma \[12\] we complete the proof.

We now give an upper bound of $(\hat{f}_n - \tilde{f}^*)$, which is another part of the stochastic error. Since $l_{n\lambda}(f)$ is a quadratic form of $f$, the Taylor expansion of $Dl_{n\lambda}(\hat{f}_n) = 0$ at $\bar{f}$ gives
\[
Dl_{n\lambda}(\bar{f}) + D^2l_{n\lambda}(\bar{f})(\hat{f}_n - \bar{f}) = 0,
\]
and by the definition of $\tilde{f}^*$ and $G_\lambda$, we have
\[
Dl_{n\lambda}(\bar{f}) + D^2l_{\infty\lambda}(\bar{f})(\tilde{f}^* - \bar{f}) = 0.
\]
Thus, $G_\lambda(\hat{f}_n - \tilde{f}^*) = \frac{1}{2} D^2l_{\infty}(\bar{f})(\hat{f}_n - \bar{f}) - \frac{1}{2} D^2l_n(\bar{f})(\hat{f}_n - \bar{f})$, and
\[
\hat{f}_n - \tilde{f}^* = G_\lambda^{-1} \left[ \frac{1}{2} D^2l_{\infty}(\bar{f})(\hat{f}_n - \bar{f}) - \frac{1}{2} D^2l_n(\bar{f})(\hat{f}_n - \bar{f}) \right]. \tag{47}
\]

**Lemma 5.** If $n^{-1} \lambda^{-(2a+3/2m)[\log(1/\lambda)]^{-r-1}} \to 0$ and $1/2m < a < (2m - 3)/4m$, we have for any $0 \leq c \leq a + 1/m$,
\[
\|\hat{f}_n - \tilde{f}^*\|_{L_2(c)}^2 \leq o_p \left\{ \|\tilde{f}^* - \bar{f}\|_{L_2(c)}^2 \right\}.
\]
Proof. A sufficient condition for this lemma is that for any $1/(2m) < a < (2m-3)/(4m)$ and $0 \leq c \leq a+1/m$,

$$\|\hat{f}_n - \tilde{f}^*\|_{L^2(c)}^2 = \begin{cases} O_p \{ n^{-1} \lambda^{-(c+a+1/2m)} [\log(1/\lambda)]^{r_n(d-p)-1} \}, & \text{if } 0 \leq p < d, \\
O_p \{ n^{-1} \lambda^{a+c+\frac{r_n}{2m^2}} \|\hat{f}_n - \tilde{f}\|_{L^2(a+1/m)}^2 \}, & \text{if } p = d, r \geq 3, \\
O_p \{ n^{-1} \|\hat{f}_n - \tilde{f}\|_{L^2(a+1/m)} \}, & \text{if } p = d, r = 2, \\
O_p \{ n^{-1} \lambda^{1-2m(a+1)/2m^2} \|\hat{f}_n - \tilde{f}\|_{L^2(a+1/m)} \}, & \text{if } p = d, r = 1. \tag{48} \end{cases}$$

This is because once (48) established, by letting $c = a+1/m$ and under the assumption that $n^{-1} \lambda^{-(2a+3/2m)} [\log(1/\lambda)]^{-1} \to 0$, we have

$$\|\hat{f}_n - \tilde{f}^*\|_{L^2(a+1/m)}^2 = O_p(1) \|\hat{f}_n - \tilde{f}\|_{L^2(a+1/m)}^2.$$ 

By the triangle inequality, we have $\|\hat{f}_n - \tilde{f}\|_{L^2(a+1/m)}^2 \geq \|\hat{f}_n - \tilde{f}\|_{L^2(a+1/m)}^2 - \|\hat{f}_n - \tilde{f}^*\|_{L^2(a+1/m)}^2 = [1 - O_p(1)] \|\hat{f}_n - \tilde{f}\|_{L^2(a+1/m)}^2$, which implies $\|\hat{f}_n - \tilde{f}\|_{L^2(a+1/m)}^2 = O_p(\|\hat{f}_n - \tilde{f}\|_{L^2(a+1/m)}^2)$. Thus, by (48) and Lemma 4, we complete the proof.

We now are in the position to prove (48). For any $0 \leq c \leq a+1/m$, by (47), we have

$$\|\hat{f}_n - \tilde{f}^*\|_{L^2(c)}^2 \leq \sum_{\nu \in \mathbb{V}} \left( 1 + \frac{\rho_{\nu}}{\|\phi_{\nu}\|_{L^2}} \right)^c (1 + \lambda \rho_{\nu})^{-2} \|\phi_{\nu}\|_{L^2}^2 \cdot \frac{1}{p+1}. \tag{49}$$

Let $g_j(t) = \frac{1}{\sigma_j} \frac{\partial (\hat{f}_n - \tilde{f})}{\partial t_j} \frac{\partial \phi_{\mu}}{\partial t_j}$ and $g_0(t) = \frac{1}{\sigma_0} (\hat{f}_n - \tilde{f}) \phi_{\mu}$. Hence, we can do the expansion on the basis $\{\phi_{\mu}\}_{\mu \in \mathbb{N}^d}$,

$$g_j(t) = \sum_{\mu \in \mathbb{N}^d} Q^j_{\mu} \phi_{\mu}(t), \quad \text{where } Q^j_{\mu} = \langle g_j(t), \phi_{\mu}(t) \rangle_0. \tag{50}$$
 Unlike (33) with the multi-index $\mathbf{\nu} \in \mathcal{V}$, we require $\mathbf{\mu} \in \mathbb{N}^d$ in (50) since now $g_j(t)$ is a product function. By Cauchy-Schwarz inequality,

$$
\left[ \frac{1}{n\sigma_j^2} \sum_{i=1}^{n} \left( \frac{1}{n} \sum_{i=1}^{n} \phi_{\mathbf{\mu}}(t_i^{(j)}) - \int \phi_{\mathbf{\mu}}(t) \right) \right]^2
$$

$$
\leq \left[ \sum_{\mathbf{\mu} \in \mathbb{N}^d} (Q_j^j)^2 \left( 1 + \frac{\rho_{\mathbf{\mu}}}{\|\phi_{\mathbf{\mu}}\|_{L_2}^2} \right) \|\phi_{\mathbf{\mu}}\|_{L_2}^2 \right]
\cdot \left[ \sum_{\mathbf{\mu} \in \mathbb{N}^d} (1 + \frac{\rho_{\mathbf{\mu}}}{\|\phi_{\mathbf{\mu}}\|_{L_2}^2})^{-a} \|\phi_{\mathbf{\mu}}\|_{L_2}^{-2} \left( \frac{1}{n} \sum_{i=1}^{n} \phi_{\mathbf{\mu}}(t_i^{(j)}) - \int \phi_{\mathbf{\mu}}(t) \right)^2 \right].
$$

By Lemma 13 if $a > 1/2m$, then the sum of the first part in the right-hand side of (51) over $j = 0, 1, \ldots, p$ is bounded by

$$
\sum_{j=0}^{p} \sum_{\mathbf{\mu} \in \mathbb{N}^d} \left( 1 + \frac{\rho_{\mathbf{\mu}}}{\|\phi_{\mathbf{\mu}}\|_{L_2}^2} \right)^a \|\phi_{\mathbf{\mu}}\|_{L_2}^2 \left\langle \left( \frac{\partial(f_n - \bar{f})}{\partial t_j} \right), \phi_{\mathbf{\mu}} \right\rangle_0^2
$$

$$
\lesssim \|f_n - \bar{f}\|_{L_2(a+1/m)}^2 \sum_{j=0}^{p} \sum_{\mathbf{\mu} \in \mathbb{N}^d} \left( 1 + \frac{\rho_{\mathbf{\mu}}}{\|\phi_{\mathbf{\mu}}\|_{L_2}^2} \right)^a \|\phi_{\mathbf{\mu}}\|_{L_2}^2 \left\langle \left( \frac{\partial \phi_{\mathbf{\nu}}}{\partial t_j} \right), \phi_{\mathbf{\mu}} \right\rangle_0^2
$$

$$
\lesssim \|f_n - \bar{f}\|_{L_2(a+1/m)}^2 \left( 1 + \frac{\rho_{\mathbf{\nu}}}{\|\phi_{\mathbf{\nu}}\|_{L_2}^2} \right)^a \|\phi_{\mathbf{\nu}}\|_{L_2}^2 \left( 1 + \sum_{j=1}^{p} \nu_j^2 \right)
$$

$$
\lesssim \|f_n - \bar{f}\|_{L_2(a+1/m)}^2 \left( 1 + \frac{\rho_{\mathbf{\nu}}}{\|\phi_{\mathbf{\nu}}\|_{L_2}^2} \right)^a.
$$
The second part on the right-hand side of (51) can be bounded by
\[
\mathbb{E} \left[ \sum_{\mu \in \mathbb{N}^d} \left( 1 + \frac{\rho \bar{\mu}}{\|\bar{\mu}\|_{L_2}^2} \right)^{-a} \|\bar{\mu}\|_{L_2}^{-2} \left( \frac{1}{n} \sum_{i=1}^{n} \phi_{\bar{\mu}}(t_i^{(j)}) - \int \phi_{\bar{\mu}}(t) \right)^2 \right] \leq n^{-1} \sum_{\mu \in \mathbb{N}^d} \left( 1 + \frac{\rho \bar{\mu}}{\|\bar{\mu}\|_{L_2}^2} \right)^{-a} \|\bar{\mu}\|_{L_2}^{-2} \int \phi_{\bar{\mu}}^2(t) \lesssim n^{-1} \sum_{\mu \in \mathbb{N}^d} \mu_1^{-2ma} \cdots \mu_d^{-2ma} \leq n^{-1} \left( \sum_{\mu_1=1}^{\infty} \mu_1^{-2ma} \right)^d \approx n^{-1},
\]
where the third step uses \( \rho \bar{\mu}/\|\bar{\mu}\|_{L_2}^2 \approx \mu_1^{2m} \cdots \mu_d^{2m} \), and the fourth step holds for \( a > 1/2m \).

Combing (52) and (53), we have that for \( a > 1/2m \),
\[
\sum_{j=0}^{p} \mathbb{E} \left[ \sum_{\mu \in \mathbb{N}^d} Q_{\bar{\mu}} \left( \frac{1}{n} \sum_{i=1}^{n} \phi_{\bar{\mu}}(t_i^{(j)}) - \int \phi_{\bar{\mu}}(t) \right) \right]^2 \lesssim n^{-1} \|\bar{f}_n - \bar{f}\|_{L_2(a+1/m)}^2 \left( \sum_{\mu \in \mathbb{N}^d} \frac{\rho \bar{\mu}}{\|\bar{\mu}\|_{L_2}^2} \right)^a.
\]

**Putting all together.** Therefore, if \( 1/2m < a < (2m - 3)/4m \) and \( 0 \leq c \leq a + 1/m \), (49) and (54) imply that
\[
\mathbb{E} \|\hat{f}_n - \hat{f}^*\|_{L_2(c)}^2 \lesssim n^{-1} \|\hat{f}_n - \hat{f}\|_{L_2(a+1/m)}^2 N_{a+c}(\lambda).
\]
By Lemma 12 we complete the proof for (48) and this lemma.

Finally, we combine (31) and Lemmas 2 and 5 to obtain the following proposition.

**Proposition 1.** Under the conditions of Theorem 3 and assuming the distributions \( \Pi^{(i)} \) and \( \Pi^{(j)} \) are known. If \( 1/2m < a < (2m - 3)/4m \), and \( n^{-1} \lambda^{-(2a+3)/2m}[\log(1/\lambda)]^{r-1} \to 0 \), then

\[
|\hat{f}_n - f_0|_{L_2(c)}^2 = O \left\{ \min_{j(f) \leq R_n} |f - f_0|_{L_2(c)}^2 + \lambda^{1-c} R_n^2 \right\} + O_{\mathbb{P}} \left\{ n^{-1} R_n^2 \lambda^{-(c+1/2m)}[\log(1/\lambda)]^{r+(d-p)-1} \right\},
\]

63
and when \( p = d \),

\[
\| \hat{f}_n - f_0 \|_{L^2(c)}^2 = \begin{cases}
O \left\{ \min_{f(f) \leq R_n} \| f - f_0 \|_{L^2(c)}^2 + \frac{(1-c)m}{m} R_n^2 \right\} + O_p \left\{ n^{-1} R_n^2 \lambda \frac{m}{c+\frac{r-2}{2m}} \right\} & \text{if } r \geq 3, \\
O \left\{ \min_{f(f) \leq R_n} \| f - f_0 \|_{L^2(c)}^2 + \frac{2m}{r-2} R_n^2 \right\} + O_p \left\{ n^{-1} R_n^2 \log(1/\lambda) \right\} & \text{if } r = 2, c = 0, \\
O \left\{ \min_{f(f) \leq R_n} \| f - f_0 \|_{L^2(c)}^2 + \frac{(1-c)m}{m} R_n^2 \right\} + O_p \left\{ n^{-1} R_n^2 \lambda \frac{2mc}{2m} \right\} & \text{if } r = 2, c > 0, \\
O \left\{ \min_{f(f) \leq R_n} \| f - f_0 \|_{L^2(c)}^2 + \frac{1-c}{m} R_n^2 \right\} + O_p \left\{ n^{-1} R_n^2 \lambda \frac{1-2mc}{2m} \right\} & \text{if } r = 1, c > \frac{1}{2m},
\end{cases}
\]

By Proposition 1 we can derive the convergence rates by the estimator \( \hat{f}_n \) defined by (9). In fact, for \( p = d \) and \( r \geq 3 \), by letting \( \lambda \propto n^{-\frac{2mr-2}{(2m+1)r^2}} \), \( a = 1/2m + \epsilon \) for some \( \epsilon > 0 \) and \( c = 0 \), we have that \( n^{-1} \lambda^{-\frac{2a+3}{2m}} \| \log(1/\lambda) \|^{r-1} \to 0 \) is equivalent to

\[ -1 + \frac{5(mr - 1)}{2m^2r + mr - 2m} < 0. \]

Thus, the conditions for Proposition 1 are satisfied. Similarly, we can verify that when \( p = d \) and \( r = 2 \), \( \lambda \propto [n(\log n)]^{-\frac{2m-1}{2m} \frac{1}{2m}} \) satisfies the conditions for Proposition 1. When \( p = d \) and \( r = 1 \), \( \lambda \propto n^{-\frac{m-1}{m}} \) satisfies the conditions for the above proposition. When \( 0 \leq p \leq d - r \), \( \lambda \propto [n(\log n)^{1-r}]^{-2m/(2m+1)} \) satisfies the conditions for the above Proposition, as well as when \( d - r < p < d \) by letting \( \lambda \propto [n(\log n)^{1+p-d}]^{-2m/(2m+1)} \). This observation leads to the following theorem for \( \hat{f}_n \) in (9).

**Theorem 6.** Assume that \( \lambda_\nu \propto \nu^{-2m} \) for some \( m > 3/2 \). Under the regression models (1) and (2) where \( f_0 \) follows the SS-ANOVA model (4) and \( \| f \|_H \leq R_n \). Then under the general error structure (3), the estimator \( \hat{f}_n \) defined by (9) satisfies

\[
\lim_{C \to \infty} \limsup_{n \to \infty} \sup_{f_0 \in H} \mathbb{P} \left\{ \int_{X^d} \left[ \hat{f}_n(t) - f_0(t) \right]^2 dt \leq C \left( [n(\log n)^{1-(d-p)^{r}}]^{-\frac{2m}{r+1}} 1_{0 \leq p < d} + n^{-\frac{2mr}{(2m+1)r^2}} 1_{l \geq 3} + n^{-1} (\log n)^{r-1} 1_{r < 3} \right) 1_{p = d} \right\} = 1.
\]

64
Here the tuning parameter $\lambda$ in (14) is chosen by 
$\lambda \asymp \left( n \log n \right)^{1-(d-p)/r} \left[ \frac{1}{2}m/(2m+1) \right] \quad \text{when} \quad 0 \leq p < d, \quad \text{and} \quad \lambda \asymp n^{-2(2mr-2)/(2m+1)r-2} \quad \text{when} \quad p = d, r \geq 3, \quad \text{and} \quad \lambda \asymp (n \log n)^{-(2m-1)/2m} \quad \text{when} \quad p = d, r = 2, \quad \text{and} \quad \lambda \asymp n^{-(m-1)/m} \quad \text{when} \quad p = d, r = 1.$

Finally, we approximate the estimator $\hat{f}_n$ in (9) with the random feature estimator defined by (13),

$$\hat{f}_n^{\mathrm{RF}} = \Psi_{(p+1)d}(t)^\top c_{(p+1)d}. $$

We have the following decomposition,

$$\hat{f}_n^{\mathrm{RF}} - \hat{f}_n = \left( S_s \tilde{C}_{s,\lambda}^{-1} \tilde{S}_s^* y - L_s L_{s,\lambda}^{-1} y \right) + \left( L_s L_{s,\lambda}^{-1} y - LL_{s,\lambda}^{-1} y \right). $$

(55)

Here the notations are similar to those in the Definition 2 of Rudi and Rosasco (2017). Specifically, let $y$ be the vector of data, $y = (y_1^{(0)}, \ldots, y_n^{(0)}, \ldots, y_1^{(p)}, \ldots, y_n^{(p)})^\top$. Moreover,

- The approximated kernel $K_s = \Psi_{(p+1)d}(t)^\top (t) \Psi_{(p+1)d}(t')$.
- $S_s$: $(S_s \beta)(\cdot) = \Psi_{(p+1)d}(\cdot)^\top \beta$.
- $S_s^*$: $S_s^* g = \frac{1}{\sqrt{s}} \int \Psi_{(p+1)d}(t) g(t) dt$.
- $L_s$: $(L_s g)(\cdot) = \int K_s(\cdot, t) g(t) dt$.
- $C_s$: $C_s = \int \Psi_{(p+1)d}(t) \Psi_{(p+1)d}(t)^\top \int dt$.
- $\tilde{C}_s$: $\tilde{C}_s = \frac{1}{n} \sum_{i=1}^n \Psi_{(p+1)d}(t_i) \Psi_{(p+1)d}(t_i)^\top$.
- The random feature mapping estimator $\hat{f}_n^{\mathrm{RF}} = S_s \tilde{C}_{s,\lambda}^{-1} \tilde{S}_s^* y$.

We analyze the two error terms in (55) separately. For the Error I, note that, $L_s L_{s,\lambda}^{-1} = S_s C_{s,\lambda}^{-1} S_s^*$. Then,

Error I $= S_s \tilde{C}_{s,\lambda}^{-1} \tilde{S}_s^* y - L_s L_{s,\lambda}^{-1} y$

$= S_s \tilde{C}_{s,\lambda}^{-1} (\tilde{S}_s^* - S_s^*) y + S_s (\tilde{C}_{s,\lambda}^{-1} - C_{s,\lambda}^{-1}) S_s^* y$

$= S_s \tilde{C}_{s,\lambda}^{-1} (\tilde{S}_s^* - S_s^*) y + S_s \tilde{C}_{s,\lambda}^{-1} (C_{s,\lambda} - \tilde{C}_{s,\lambda}) C_{s,\lambda}^{-1} S_s^* y$

$= S_s \tilde{C}_{s,\lambda}^{-1} (\tilde{S}_s^* - S_s^*) y + (S_s \tilde{C}_{s,\lambda}^{-1} C_{s,\lambda}^{1/2}) [C_{s,\lambda}^{-1/2} (C_{s,\lambda} - \tilde{C}_{s,\lambda})] C_{s,\lambda}^{-1} S_s^* y$.
By Lemma 7 in [Rudi and Rosasco (2017)], we obtain that,

\[ \|\text{Error I}\|_{L^2} \leq O\left(\lambda^{-1/2}n^{-1} + n^{-1/2}\lambda^{-1/4m}\right). \]

By Lemma 4, both the term \(\lambda^{-1/2}n^{-1}\) and the term \(n^{-1/2}\lambda^{-1/4m}\) are dominated by \(\|\hat{f}^* - \bar{f}\|_{L^2}\).

By Proposition 1,

\[ \|\text{Error I}\|_{L^2} = O(\|\hat{f}_n - f_0\|_{L^2}^2). \] (56)

For the Error II, by Lemma 8 and Equation (14) in [Rudi and Rosasco (2017)], we have

\[ \|\text{Error II}\|_{L^2} = O\left(\frac{\log(1/\lambda)}{s}\right), \]

By Proposition 1 and letting \(s = O(n \log n)\), we have \(\|\text{Error II}\|_{L^2} = O(n^{-1/2})\). Hence

\[ \|\text{Error II}\|_{L^2} = O(\|\hat{f}_n - f_0\|_{L^2}^2). \] (57)

By combining (56) and (57), we have \(\|\hat{f}_{RF}^n - f_n\|_{L^2} = O(\|\hat{f}_n - f_0\|_{L^2})\). By triangle inequality,

\[ \|\hat{f}_{RF}^n - f_0\|_{L^2} = \|\{\hat{f}_{RF}^n - \hat{f}_n\} + \{\hat{f}_n - f_0\}\|_{L^2} \]
\[ \leq \|\hat{f}_{RF}^n - \hat{f}_n\|_{L^2} + \|\hat{f}_n - f_0\|_{L^2} \]
\[ = O(\|\hat{f}_n - f_0\|_{L^2}). \] (58)

Using Theorem 6 and (58), we complete the proof of Theorem 3.

D.4 Auxiliary Lemmas for Theorems 2 and 3

Lemma 6. Suppose that \(\beta \geq 0\) and \(0 < \alpha \leq 2\). Then, as \(\Xi \to \infty\),

\[ \int_{x_1 \ldots x_r \leq \Xi, x_k \geq 1} \prod_{k=1}^{r} x_k^\beta (x_1^\alpha + x_2^\alpha + \ldots + x_r^\alpha)^{-1} \, dx_1 \ldots dx_r \]
\[ \asymp \begin{cases} 
\Xi^{\beta + 1 - \alpha/r}, & \text{if } r \geq 3; \\
\log(\Xi), & \text{if } r = 2, \beta = \alpha/2 - 1; \\
\Xi^{\beta + 1 - \alpha/2} & \text{if } r = 2, \beta > \alpha/2 - 1; \\
1, & \text{if } r = 1, \beta < \alpha - 1; \\
\log(\Xi) & \text{if } r = 1, \beta = \alpha - 1; \\
\Xi^{\beta - \alpha + 1} & \text{if } r = 1, \beta > \alpha - 1.
\end{cases} \]
Proof. By the symmetry of covariates,
\[
\int_{x_1 \cdots x_r \leq \Xi, x_k \geq 1} \prod_{k=1}^r x_k^\beta (x_1^\alpha + x_2^\alpha + \cdots + x_r^\alpha)^{-1} dx_1 \cdots dx_r
\]
\[
\asymp \int_{x_1 \cdots x_r \leq \Xi, x_1 \geq x_2 \geq \cdots \geq x_r \geq 1} \prod_{k=1}^r x_k^\beta (x_1^\alpha + x_2^\alpha + \cdots + x_r^\alpha)^{-1} dx_r \cdots dx_1
\]
\[
:= \mathcal{E}.
\]

First, we prove when \( r \geq 3 \), as \( \Xi \to \infty \), we have
\[
\mathcal{E} \lesssim \Xi^{\beta + 1 - \alpha/r}.
\]
(59)

For this, define the set \( \mathcal{K} = \left\{ 0 \leq k \leq r - 2 : \left( \frac{\Xi}{x_1 \cdots x_{r-k-1}} \right)^{1/(k+1)} \leq x_{r-k} \right\} \). If \( \mathcal{K} \) is not empty, we denote the smallest element in \( \mathcal{K} \) by \( k^* \). Then \( 0 \leq k^* \leq r - 2 \). For any \( (x_1, \ldots, x_r) \in \{ (x_1, \ldots, x_r) : x_1 \cdots x_r \leq \Xi, x_1 \geq x_2 \geq \cdots \geq x_r \geq 1, x_r \leq x_{r-1} \leq \frac{\Xi}{x_1 \cdots x_{r-k}} \} \), we have
\[
\begin{cases}
1 \leq x_{r-k} \leq x_{r-k-1} & \text{for } 0 \leq k \leq k^* - 1, \\
1 \leq x_{r-k^*} \leq \left( \frac{\Xi}{x_1 \cdots x_{r-k^*}} \right)^{1/(k^*+1)} & \text{for } k = k^*, \\
x_{r-k} \geq \left( \frac{\Xi}{x_1 \cdots x_{r-k-1}} \right)^{1/(k+1)} & \text{for } k^* + 1 \leq k \leq r - 2, \\
x_1 \geq \Xi^{1/r} & \text{for } k = r - 1.
\end{cases}
\]
(60)
Thus, as $\Xi \to \infty$,

$$
\mathcal{E} \leq \int_{x_1 \ldots x_r \leq \Xi, x_1 \geq x_2 \geq \cdots \geq x_r \geq 1} \left\{ (x_1)^{\beta-\alpha/(r-1)} \cdots (x_{r-k^*+1})^{\beta-\alpha/(r-1)} \right\} \cdot \left\{ (x_{r-k^*+1})^{\beta-\alpha/(r-1)} \cdots (x_r)^{\beta-\alpha/(r-1)} \right\} \, dx
$$

$$
\leq \int_{x_1 \ldots x_r \leq \Xi, x_1 \geq x_2 \geq \cdots \geq x_r \geq 1} \left\{ (x_1)^{\beta-\alpha/(r-1)} \cdots (x_{r-k^*+1})^{\beta-\alpha/(r-1)} \right\} \cdot (x_{r-k^*})^{[\beta+1-\alpha/(r-1)]k^* + \beta} \, dx_{r-k^*} dx_{r-k^*-1} \cdots dx_1
$$

$$
\leq \Xi^{\beta+1-\alpha/r},
$$

where the first step uses $x_{r-k^*} \geq 1$ and Lemma 14, the second step uses $x_{r-k} \leq x_{r-k-1}$ for all $k \leq k^* - 1$ in (60), the third step uses the upper bound on $x_{r-k^*}$ in (60), the fourth step uses the lowers bounds on $x_{r-k}$ for all $k^* + 1 \leq k \leq r - 2$ in (60). If $\mathcal{K}$ is empty, then for any $(x_1, \ldots, x_r) \in \{(x_1, \ldots, x_r) : x_1 \cdots x_r \leq \Xi, x_1 \geq x_2 \geq \cdots \geq x_r \geq 1, x_r \leq x_r \leq \Xi/(x_1 \cdots x_r - 1)\}$, it satisfies

$$
1 \leq x_k \leq x_{k-1} \text{ for any } 2 \leq k \leq r, \quad \text{and} \quad 1 \leq x_1 \leq \Xi^{1/r}.
$$

Thus, as $\Xi \to \infty$,

$$
\mathcal{E} = \int_{1}^{\Xi^{1/r}} \cdots \int_{1}^{\Xi^{1/r}} \int_{1}^{\Xi^{1/r}} \prod_{k=1}^{r} x_k^{\beta}(x_1^\alpha + x_2^\alpha + \cdots + x_{r-1}^\alpha + x_r^\alpha)^{-1} \, dx_r \, dx_{r-1} \cdots \, dx_1
$$

$$
\leq \int_{1}^{\Xi^{1/r}} \cdots \int_{1}^{\Xi^{1/r}} \int_{1}^{\Xi^{1/r}} x_1^{\beta-\alpha/r} \cdots x_{r-1}^{\beta-\alpha/r} x_r^{\beta-\alpha/r} \, dx_r \, dx_{r-1} \cdots \, dx_1 \leq \Xi^{\beta+1-\alpha/r}.
$$

Combining (61) and (62) completes the proof for (59).
On the other hand, when $r \geq 3$ and as $\Xi \to \infty$,

$$
\mathcal{E} \geq \int_1^{\Xi^{1/r}} \cdots \int_1^{x_{r-2}} \int_1^{x_{r-1}} \prod_{k=1}^r x_k^\beta (x_1^\alpha + \cdots + x_1^\alpha + x_1^\alpha - 1) dx_1 \cdots dx_1 \geq \int_1^{\Xi^{1/r}} \cdots \int_1^{x_{r-2}} \int_1^{x_{r-1}} \prod_{k=1}^r x_k^\beta \cdot \frac{1}{x_1^\alpha} \cdots dx_1 \cdots dx_1 \Xi^{\beta+1-\alpha/r}. $$

(63)

Therefore, combining (59) and (63) completes the proof of the lemma for $r \geq 3$.

Then we consider for $r = 2$. For $0 < \alpha \leq 2$,

$$
\mathcal{E} \leq 2 \int_1^{\sqrt{\Xi}} \int_1^{x_1} x_1^{\beta-\alpha} x_2^\beta dx_2 dx_1 + 2 \int_1^{\Xi} \int_1^{\Xi/x_1} x_1^{\beta-\alpha} x_2^\beta dx_2 dx_1
\times \begin{cases} 
\log(\Xi) & \text{when } 2\beta + 2 - \alpha = 0 \\
\Xi^{\beta+1-\alpha/2} & \text{when } 2\beta + 2 - \alpha > 0 
\end{cases}
$$

as $\Xi \to \infty$. (64)

On the other hand, we have

$$
\mathcal{E} \geq \int_1^{\sqrt{\Xi}} \int_1^{x_1} x_1^{\beta} (x_1^\alpha + x_2^\alpha - 1) dx_2 dx_1
\geq 2^{-1} \int_1^{\sqrt{\Xi}} \int_1^{x_1} x_1^{\beta} x_2^\beta dx_2 dx_1
\times \begin{cases} 
\log(\Xi) & \text{when } 2\beta + 2 - \alpha = 0 \\
\Xi^m & \text{when } 2\beta + 2 - \alpha > 0 
\end{cases}
\text{ as } \Xi \to \infty.
$$

(65)

Combining (64) and (65) completes the proof of the lemma for $r = 2$.

Finally, we consider for $r = 1$. Note that $\int_1^{\Xi} x_1^{\beta} x_1^{-\alpha} dx_1 \asymp 1$ when $0 \leq \beta < \alpha - 1$, and $\int_1^{\Xi} x_1^{\beta} x_1^{-\alpha} dx_1 \asymp \log(\Xi)$ when $\beta = \alpha - 1$, and $\int_1^{\Xi} x_1^{\beta} x_1^{-\alpha} dx_1 \asymp \Xi^{\alpha-1}$ when $\beta > \alpha - 1$. This completes the proof.

Lemma 7. The norm $\| \cdot \|_R$ is equivalent to $\| \cdot \|_H$ in $\mathcal{H}$.

Proof. Observe that for any $g \in \mathcal{H}$, by the assumption that $\Pi^{(0)}$ and $\Pi^{(j)}$s are bounded away
from 0 and infinity, we have
\[
\frac{1}{p+1} \left[ \frac{1}{\sigma_0^2} \int g^2(t) \Pi(0)(t) + \sum_{j=1}^{p} \frac{1}{\sigma_j^2} \int \left\{ \frac{\partial g(t)}{\partial t_j} \right\}^2 \Pi(j)(t) \right] 
\leq c_1 \left[ \int g^2(t) + \sum_{j=1}^{p} \int \left\{ \frac{\partial g(t)}{\partial t_j} \right\}^2 \right] 
\leq c_2 \cdot c_2^d \|g\|_H^2,
\]
for some constant \(c_1\) and \(c_2\), where the last step is by Lemma \([\text{II}]\). Hence
\[
\|g\|_R^2 \leq (c_2 c_2^d + 1) \|g\|_H^2. \tag{66}
\]

On the other hand, for any \(g \in \mathcal{H}\) we can do the orthogonal decomposition \(g = g^0 + g^1\) where \(\langle g^0, g^1 \rangle_{\mathcal{H}} = 0\), \(g^0\) is in the null space of \(J(\cdot)\) and \(g^1\) is in the orthogonal space of the null space of \(J(\cdot)\) in \(\mathcal{H}\). Since the null space of \(J(\cdot)\) has a finite basis that forms a positive definite kernel matrix, we assume the minimal eigenvalue of the kernel matrix is \(\mu'_{\min} > 0\). Then there exists a constant \(c_3 > 0\) such that
\[
\|g^0\|_R^2 \geq c_3 \|g^0\|_L^2 \geq c_3 \mu'_{\min} \|g^0\|_H^2. \tag{67}
\]
For \(g^1\), we have \(\|g^1\|_R^2 \geq J(g^1) = \|g^1\|_H^2\). Thus, for any \(g \in \mathcal{H}\),
\[
\|g\|_R^2 \geq c_3 \left( \|g^0\|_L^2 + 1 \right) + \|g^1\|_H^2 
\geq c_3 \left\{ \|g^0\|_L^2 + \frac{1 + c_3}{c_3} \|g^1\|_L^2 - 2 \|g^0\|_L \|g^1\|_L \right\} 
\geq \frac{c_3}{1 + c_3} \|g^0\|_L^2,
\]
where the second inequality is by \(\|g^1\|_H^2 \geq \|g^1\|_L^2\). By \(67\), we obtain \(\|g\|_R^2 \geq 1 + c_3^{-1} c_3 \mu'_{\min} \|g^0\|_H^2\). Together with \(\|g\|_R^2 \geq J(g^1) = \|g^1\|_H^2\), we have
\[
\|g\|_R^2 \geq \left( 1 + \frac{1 + c_3}{c_3 \mu'_{\min}} \right)^{-1} \|g\|_H^2. \tag{68}
\]
Combining \(66\) and \(68\) completes the proof. \hfill \blacksquare

**Lemma 8 (Inverse transformation).** Suppose that designs \(t^{(j)}, j = 0, \ldots, p\) are independently drawn from known distributions \(\Pi^{(j)}\) supported in \(\mathcal{X}^d\). Then, there exists a linear transformation to data \((t^{(j)}, Y^{(j)})\) such that transformed design points \(x^{(j)}\)s are independently uniformly distributed on \(\mathcal{X}^d\).
Proof. First, we consider function and derivative data sharing a common design, i.e., \( t_i^{(j)} = t_i^{(k)}, \forall 1 \leq i \leq n, 0 \leq j < k \leq p \). Write \( t^{(j)} = (t_1^{(j)}, \ldots, t_d^{(j)}) \in \mathcal{X}^d \). We allow covariates of \( t^{(j)} \) can be correlated; that is, the density of \( t^{(j)} \) is decomposed as:

\[
d\Pi^{(j)}(t_1, \ldots, t_d) = d\Pi_d^{(j)}(t_d)d\Pi_{d-1}^{(j)}(t_{d-1}|t_d) \cdots d\Pi_1^{(j)}(t_1|t_d, t_{d-1}, \ldots, t_2).
\]

Now let

\[
x_d^{(j)} = \Pi_d^{(j)}(t_d^{(j)}), \quad x_{d-1}^{(j)} = \Pi_{d-1}^{(j)}(t_{d-1}^{(j)}|t_d^{(j)}), \ldots, \quad x_1^{(j)} = \Pi_1^{(j)}(t_1^{(j)}|t_d^{(j)}, t_{d-1}^{(j)}, \ldots, t_2^{(j)}).
\]

Then, \( \mathbf{x}^{(j)} = (x_1^{(j)}, x_2^{(j)}, \ldots, x_d^{(j)}) \) is uniformly distributed on \( \mathcal{X}^d \). Define that

\[
h(x_1, x_2, \ldots, x_d)
= f(\{\Pi_1^{(j)}\}^{-1}(x_1|x_d, \ldots, x_2), \{\Pi_2^{(j)}\}^{-1}(x_2|x_d, \ldots, x_3), \ldots, \{\Pi_d^{(j)}\}^{-1}(x_d)).
\]

Thus,

\[
\frac{\partial h(\mathbf{x})}{\partial x_j} = \sum_{k=1}^{j} \frac{\partial f(\mathbf{t})}{\partial t_k} \cdot \frac{\partial t_k}{\partial x_j} = \sum_{k=1}^{j-1} \frac{\partial f}{\partial t_k} \cdot \frac{\partial t_k}{\partial x_j} + \frac{\partial f}{\partial t_j} \cdot \frac{1}{d\Pi_j^{(j)}(t_j|t_d, \ldots, t_{j+1})}.
\]

With the design \( \mathbf{x}^{(j)} \) defined, we transform the responses \( Y^{(j)} \)s to \( Z^{(j)} \)s by letting \( Z^{(0)} = Y^{(0)} \) and for any \( j = 1, \ldots, p \),

\[
Z^{(j)} = \sum_{k=1}^{j-1} Y^{(k)} \frac{\partial \Pi_k^{(j)}(x_d^{(j)}, x_{d-1}^{(j)}, \ldots, x_k^{(j)})}{\partial x_j} + \frac{Y^{(j)}}{d\Pi_j^{(j)}(t_j^{(j)}|t_d^{(j)}, \ldots, t_{j+1}^{(j)})}.
\]

Write

\[
\tilde{\sigma}_j^2 = \sum_{k=1}^{j-1} \sigma_k^2 \left[ \frac{\partial t_k^{(j)}}{\partial x_j} (x_d^{(j)}, x_{d-1}^{(j)}, \ldots, x_k^{(j)}) \right]^2 + \frac{\sigma_j^2}{[d\Pi_j^{(j)}(t_j^{(j)}|t_d^{(j)}, \ldots, t_{j+1}^{(j)})]^2}.
\]

Then it is clear that \( Z^{(j)} = \partial h/\partial x_j(\mathbf{x}^{(j)}) + \tilde{\epsilon}^{(j)} \), where the errors \( \tilde{\epsilon}^{(j)} \)s are independent and centered noises with variance \( \tilde{\sigma}_j^2 \)s.

Second, we consider that not all types of function observations and partial derivatives data share a common design, i.e., \( \exists 0 \leq j \neq k \leq p \) and \( 1 \leq i \leq n \) such that \( t_i^{(j)} \neq t_i^{(k)} \). We require the covariates of each \( t^{(j)} \) are independent; that is, the density of \( t^{(j)} \) can be decomposed as:

\[
d\Pi^{(j)}(t_1, \ldots, t_d) = d\Pi_1^{(j)}(t_1)d\Pi_2^{(j)}(t_2) \cdots d\Pi_d^{(j)}(t_d)
\]
Now let
\[ x_1^{(j)} = \Pi_1^{(j)}(t_1^{(j)}), \quad x_2^{(j)} = \Pi_2^{(j)}(t_2^{(j)}), \quad \ldots, \quad x_d^{(j)} = \Pi_d^{(j)}(t_d^{(j)}). \]
Then \( x^{(j)} = (x_1^{(j)}, x_2^{(j)}, \ldots, x_d^{(j)}) \) is uniformly distributed on \( X^d \). Define the function
\[ h(x_1, \ldots, x_d) = f(\{\Pi_1^{(j)}\}^{-1}(x_1), \{\Pi_2^{(j)}\}^{-1}(x_2), \ldots, \{\Pi_d^{(j)}\}^{-1}(x_d)). \]
Thus, we have
\[ \frac{\partial h(x)}{\partial x_j} = \frac{\partial f(t)}{\partial t_j} \cdot \frac{\partial t_j(x_j)}{\partial x_j} = \frac{\partial f(t)}{\partial t_j} \cdot \frac{1}{d\Pi_j^{(j)}(t_j)}. \]
Correspondingly, the responses \( Y^{(j)} \) is transformed to \( Z^{(j)} \), \( 0 \leq j \leq p \), by letting \( Z^{(0)} = Y^{(0)} \) and \( Z^{(j)} = Y^{(j)}/d\Pi_j^{(j)}(t_j^{(j)}) \) for \( 1 \leq j \leq d \), and write the transformed variance \( \tilde{\sigma}_j^2 = \sigma_j^2/[d\Pi_j^{(j)}(t_j^{(j)})]^2 \). 

\section*{Lemma 9}
Suppose that \( f_0 \) follows the SS-ANOVA model in (4), defined on \( X^d \equiv [0, 1]^d \).
Then, there exists a periodic function \( \tilde{f}_0 \) on the expanded domain \([0, 1 + \delta]^d\) for any \( \delta > 0 \) such that \( \tilde{f}_0(t) \equiv f_0(t) \) for \( t \in X^d \), and \( \tilde{f}_0 \) maintains the same order of smoothness as \( f_0 \), in the sense that \( \tilde{f}_0 \) follows the same RKHS in (5), defined on \([0, 1 + \delta]^d\).

\begin{proof}
The construction of the periodic function consists of four main steps.

\textit{Step 1:} We show that when \( \lambda_\nu \simeq \nu^{-2m} \), the \( m \)-th order Sobolev space on \( X \) can be embedded into the RKHS \( H_1 \). Specifically, let \( \mathcal{W}_2^m(X) \) denote the Sobolev space of order \( m \), consisting of functions whose derivatives up to order \( m - 1 \) are absolutely continuous and whose \( m \)-th derivative is square-integrable:
\[ \mathcal{W}_2^m(X) = \left\{ g : X \to \mathbb{R}, dg/dt, \ldots, d^{m-1}g/dt^{m-1} \text{ are absolutely continuous, and } d^mg/dt^m \in L_2 \right\}. \]
There are many possible norms that can be quipped with \( \mathcal{W}_2^m \) to make it a RKHS. For example, it can be endowed with the norm,
\[ \|g\|_{\mathcal{W}_2^m}^2 = \sum_{q=0}^{m-1} \left( \int g^{(q)} \right)^2 + \int (g^{(m)})^2. \]
Following the results in Chapter 2 of [Wahba (1990)], the eigenvalues of the associated kernel decay at a rate of $\lambda_\nu \asymp \nu^{-2m}$ for $\nu \geq 1$.

**Step 2:** For any $f_{0j} \in \mathcal{H}_1$ on $\mathcal{X}$, $j = 1, \ldots, d$, we construct the function $g_j$ as,

$$g_j(t_j) = \sum_{k=0}^{2m+1} c_{jk} t_j^k, \text{ for } t_j \in [1, 1 + \delta], \quad (69)$$

where the coefficients $\{c_{jk}\}_{k=0}^{2m+1}$ are computed by satisfying the linear system:

$$g_j^{(q)}(1) = f_{0j}^{(q)}(1) \quad \text{and} \quad g_j^{(q)}(1 + \delta) = f_{0j}^{(q)}(0), \quad \forall q = 0, 1, \ldots, m. \quad (70)$$

Since the linear system (70) has $2m + 2$ equations and the function $g_j$ in (69) has $2m + 2$ free coefficients $\{c_{jk}\}_{k=0}^{2m+1}$, there is a unique solution. We define the extended function as,

$$\tilde{f}_{0j}(t_j) = \begin{cases} f_{0j}(t_j), & t_j \in [0, 1], \\ g_j(t_j), & t_j \in [1, 1 + \delta], \end{cases}$$

where $g_j$ is the $(2m + 1)$-th order polynomial defined in (69). Since $g_j$ is continuous and has $m - 1$ absolutely continuous derivatives, together with the property that the $m$-th derivative of $g_j$ is in $L_2$, we know that $\tilde{f}_{0j}(t_j) \in \mathcal{W}_2^m([0, 1 + \delta])$. By the result in Step 1, the $m$-th order Sobolev space $\mathcal{W}_2^m(\mathcal{X})$ can be embedded to the RKHS $\mathcal{H}_1$. Hence $\tilde{f}_{0j}(t_j)$ follows the same RKHS as $f_{0j}(t_j)$ with the expanded domain on $[0, 1 + \delta]$.

**Step 3:** For any $f_{0j_1j_2\ldots j_r} \in \mathcal{H}_1 \otimes \mathcal{H}_1 \otimes \cdots \otimes \mathcal{H}_1$, $1 \leq j_1 < j_2 < \cdots < j_r \leq d$ and $1 \leq r \leq d$, there exists a finite integer $s$ and functions $f_{0j_1\nu}, f_{0j_2\nu}, \ldots, f_{0j_r\nu} \in \mathcal{H}_1$ for $\nu = 1, \ldots, s$, such that

$$f_{0j_1j_2\ldots j_r}(t_{j_1}, t_{j_2}, \ldots, t_{j_r}) = \sum_{\nu=1}^{s} f_{0j_1\nu}(t_{j_1}) f_{0j_2\nu}(t_{j_2}) \cdots f_{0j_r\nu}(t_{j_r}).$$

By the construction in Step 2, we can find $g_{j_\nu}(t_{j_1}) = \sum_{k=0}^{2m+1} c_{j_\nu k} t_{j_1}^k$ for $t_{j_1} \in [1, 1 + \delta]$ and $j = j_1, j_2, \ldots, j_r$, such that,

$$g_{j_\nu}^{(q)}(1) = f_{0j_\nu}^{(q)}(1) \quad \text{and} \quad g_{j_\nu}^{(q)}(1 + \delta) = f_{0j_\nu}^{(q)}(0), \quad \forall q = 0, 1, \ldots, m. \quad (71)$$

Since the linear system (71) has $(2m + 2)$ equations and the function $g_{j_\nu}(t_{j_1}) = \sum_{k=0}^{2m+1} c_{j_\nu k} t_{j_1}^k$ has $(2m + 2)$ free coefficients $\{c_{j_\nu k}\}_{k=0}^{2m+1}$, there is a unique solution. We define the extended
function as,

$$\tilde{f}_{0:j_1:j_2:...:j_r}(t_{j_1}, t_{j_2}, \ldots, t_{j_r}) = \sum_{\nu=1}^{s} h_{j_1 \nu}(t_{j_1}) h_{j_2 \nu}(t_{j_2}) \cdots h_{j_r \nu}(t_{j_r}),$$

for any \((t_{j_1}, t_{j_2}, \ldots, t_{j_r}) \in [0, 1 + \delta]^r\), where for any \(j = j_1, j_2, \ldots, j_r\), the function \(h_{j \nu}\) is defined as,

$$h_{j \nu}(t_j) = \begin{cases} f_{0:j \nu}(t_j) & t_j \in [0, 1], \\ g_{j \nu}(t_j) & t_j \in [1, 1 + \delta], \end{cases}$$

and \(g_{j \nu}(t_j) = \sum_{k=0}^{2m+1} c_{j \nu k} t_j^k\) is the \((2m + 1)\)-th order polynomial. Since \(g_{j \nu}\) is continuous and has \((m - 1)\) absolutely continuous derivatives, together with the property that the \(m\)-th derivative of \(g_{j \nu}\) is in \(L_2\), we know that \(\tilde{f}_{0:j_1:j_2:...:j_r}(t_{j_1}, t_{j_2}, \ldots, t_{j_r}) \in \mathcal{W}_2^m([0, 1 + \delta]) \otimes \cdots \otimes \mathcal{W}_2^m([0, 1 + \delta])\). By the result in Step 1, the \(m\)-th order Sobolev space \(\mathcal{W}_2^m(\mathcal{X})\) can be embedded to the RKHS \(\mathcal{H}_1\). Hence \(\tilde{f}_{0:j_1:j_2:...:j_r}(t_{j_1}, t_{j_2}, \ldots, t_{j_r})\) follows the same RKHS as \(f_{0:j_1:j_2:...:j_r}\) with the expanded domain on \([0, 1 + \delta]^r\).

**Step 4:** For \(f_0\) follows the SS-ANOVA model (4) on \(\mathcal{X}^d\), we can define the function \(\tilde{f}_0(t)\) that extends \(f_0\) from \(\mathcal{X}^d\) to \([0, 1 + \delta]^d\) for any \(\delta > 0\). Specifically, let

$$\tilde{f}_0(t) = \text{constant} + \sum_{j=1}^{d} \tilde{f}_0(t_j) + \cdots + \sum_{1 \leq j_1 < j_2 < \cdots < j_r \leq d} \tilde{f}_{0:j_1:j_2:...:j_r}(t_{j_1}, t_{j_2}, \ldots, t_{j_r}).$$

By the construction in Steps 2 and 3, we have that \(\tilde{f}_0(t) = f_0(t)\) for \(t \in \mathcal{X}^d\), which implies that \(\tilde{f}_0(t)\) coincides with \(f_0(t)\) on the original domain \(\mathcal{X}^d\). Moreover, \(\tilde{f}_0(t)\) the same order of smoothness as \(f_0(t)\) in the sense that \(\tilde{f}_0(t)\) follows the same RKHS in (5) defined on \([0, 1 + \delta]^d\). Hence, the eigenvalue decay rate of the RKHS for \(\tilde{f}_0(t)\) is the same as that of the RKHS for \(f_0(t)\). Finally, by (70) and (71), we have that for any \(j = 1, \ldots, d\) and \((t_1, \ldots, t_{j-1}, t_{j+1}, \ldots, t_d) \in [0, 1 + \delta]^{d-1},\)

$$\tilde{f}_0(t_1, \ldots, t_{j-1}, 0, t_{j+1}, \ldots, t_d) = \tilde{f}_0(t_1, \ldots, t_{j-1}, 1 + \delta, t_{j+1}, \ldots, t_d),$$

which shows that the extended function \(\tilde{f}_0\) has a periodic boundary on the expanded domain \([0, 1 + \delta]^d\) for any \(\delta > 0\). This completes the proof. 

74
Lemma 10. Suppose that $f_0$ follows the SS-ANOVA model in (4), defined on $\mathcal{X}^d \equiv [0,1]^d$, and the periodic function $\tilde{f}_0$ is constructed in Lemma 9, defined on $[0,1+\delta]^d$. Then, if $\delta > 0$ and for any estimator $\hat{f}_n$ on $[0,1+\delta]^d$, we have that $\int_{\mathcal{X}^d} [\hat{f}_n(t) - f_0(t)]^2 dt \leq \int_{[0,1+\delta]^d} [\hat{f}_n(t) - \tilde{f}_0(t)]^2 dt$.

Proof. We decompose the estimation error of $\tilde{f}_0$ as follows:

$$\int_{[0,1+\delta]^d} [\hat{f}_n(t) - \tilde{f}_0(t)]^2 dt = \int_{\mathcal{X}^d} [\hat{f}_n(t) - f_0(t)]^2 dt + \int_{[0,1+\delta]^d \setminus \mathcal{X}^d} [\hat{f}_n(t) - \tilde{f}_0(t)]^2 dt$$

$$\geq \int_{\mathcal{X}^d} [\hat{f}_n(t) - f_0(t)]^2 dt,$$

where the second step uses the property that $\tilde{f}_0(t) \equiv f_0(t)$ for $t \in \mathcal{X}^d$.

Lemma 11. For any $g \in \mathcal{H}$, there exists a constant $c_K$ which is independent of $g$ such that

$$\sup_{t \in \mathcal{X}^d} |g(t)| \leq c_K^d \|g\|_{\mathcal{H}},$$

and

$$\sup_{t \in \mathcal{X}^d} \left| \frac{\partial g(t)}{\partial t_j} \right| \leq c_K^d \|g\|_{\mathcal{H}}, \quad \forall 1 \leq j \leq d.$$

Proof. Since we assume that $K$ is continuous in the compact domain $\mathcal{X}$ and satisfies (7), there exists some constant $c_K$ such that

$$\sup_{t \in \mathcal{X}} |K(t,t)| \leq c_K$$

and

$$\sup_{t \in \mathcal{X}^d} \left| \frac{\partial^2 K(t,t)}{\partial t_j \partial t_j'} \right| \leq c_K.$$

This implies for any $t \in \mathcal{X}^d$,

$$\left\| \frac{\partial K_d(t,\cdot)}{\partial t_j} \right\|_{\mathcal{H}}^2 = \left| \frac{\partial^2 K(t_j,t_j)}{\partial t_j \partial t_j'} \right| \prod_{i \neq j} |K(t_i,t_i)| \leq c_K^d.$$

Thus, for any $g \in \mathcal{H}$, by the Cauchy-Schwarz inequality,

$$\sup_{t \in \mathcal{X}^d} \left| \frac{\partial g(t)}{\partial t_j} \right| \leq \sup_{t \in \mathcal{X}^d} \left\| \frac{\partial K_d(t,\cdot)}{\partial t_j} \right\|_{\mathcal{H}} \|g\|_{\mathcal{H}} \leq c_K^d \|g\|_{\mathcal{H}}, \quad \forall 1 \leq j \leq d.$
Similarly, we can show that $\sup_t |g(t)| \leq c_k^d \|g\|_H$.

**Lemma 12.** Recall that $\mathcal{V}$ as a family of multi-index $\vec{\nu}$ is defined in (22). We let

$$N_a(\lambda) = \sum_{\vec{\nu} \in \mathcal{V}} \left( \prod_{k=1}^d \nu_k^{2m} \right)^a \left( 1 + \sum_{j=1}^p \nu_j^2 \right)^{\frac{a}{2}} \left( 1 + \sum_{j=1}^p \nu_j^2 + \lambda \prod_{k=1}^d \nu_k^{2m} \right)^2. \tag{72}$$

Then, when $0 \leq p < d$, we have for any $0 \leq a < 1 - 1/2m$,

$$N_a(\lambda) = O \left\{ \lambda^{-a-1/2m} (\log(1/\lambda))^{(d-p)r-1} \right\},$$

and when $p = d$, we have for any $0 \leq a \leq 1$,

$$N_a(\lambda) = \begin{cases} O \left\{ \lambda^{1 - \frac{1}{2m} \cdot \frac{r}{2}} \right\}, & \text{if } r \geq 3; \\ O \{\log(1/\lambda)\}, & \text{if } r = 2, a = 0; \\ O \{1\}, & \text{if } r = 2, 0 < a \leq 1; \\ O \{1\}, & \text{if } r = 1, a < \frac{1}{2m}; \\ O \{\log(1/\lambda)\}, & \text{if } r = 1, a = \frac{1}{2m}; \\ O \left\{ \lambda^{1 - \frac{2ma}{2}} \right\}, & \text{if } r = 1, a > \frac{1}{2m}. \end{cases}$$

**Proof.** We will discuss three separate cases for $0 \leq p \leq d - r$, $d - r < p < d$ and $p = d$.

First, consider $0 \leq p \leq d - r$. Since $\vec{\nu} \in \mathcal{V}$, there are at most $r$ of $\nu_1, \ldots, \nu_d$ not equal to 1, which implies that the number of combinations of non-1 indices being summed in (72) is no greater than $C_d^1 + C_d^2 + \cdots + C_d^r < \infty$. Due to the appearance of $(1 + \sum_{j=1}^p \nu_j^2)$ in the denominator of (72), the largest terms of the summation (72) over $\vec{\nu} \in \mathcal{V}$ correspond to the combinations of $r$ indices whereas few $\nu_1, \ldots, \nu_p$ being summed as possible, which is the indices $\vec{\nu} = (\nu_{k_1}, \nu_{k_2}, \ldots, \nu_{k_r})^\top \in \mathbb{N}^r$ with $k_1, k_2, \ldots, k_r > p$. Thus, by the integral approximation,

$$N_a(\lambda) \asymp \sum_{\nu_{p+1} = 1}^\infty \cdots \sum_{\nu_{p+r-1} = 1}^\infty \sum_{\nu_{p+r} = 1}^\infty \left( \prod_{k=p+1}^{p+r} \nu_k^{2ma} \right)^a \left( 1 + \lambda \prod_{k=p+1}^{p+r} \nu_k^{2m} \right)^2 \left( 1 + \sum_{j=1}^p \nu_j^2 \right)^{\frac{a}{2}}$$

$$\asymp \int_1^\infty \int_1^\infty \cdots \int_1^\infty (1 + \lambda x_{p+1}^\delta \cdots x_{p+r-1}^\delta x_{p+r}^\delta)^{-2} dx_{p+1} \cdots dx_{p+r-1} dx_{p+r},$$
where \( b = 2m/(2ma + 1) \). Let \( z_k = x_{p+1}x_{p+2} \cdots x_k \) for \( k = p + 1, \ldots, p + r \). By using the change of variables to replace \((x_{p+1}, \ldots, x_{p+r})\) by \((z_{p+1}, \ldots, z_{p+r})\) and \( z_{p+r} \) by \( x = \lambda^{1/b} z_{p+r} \),

\[
N_a(\lambda) \asymp \int_{1}^{\infty} \cdots \int_{1}^{\infty} \left( 1 + \lambda z_{p+r}^{b} \right)^{-2} \left( z_{p+1}^{-1} \cdots z_{p+r-1}^{-1} d z_{p+1} \cdots d z_{p+r-1} d z_{p+r} \right)
\]

\[
\asymp \int_{1}^{\infty} \left( 1 + \lambda z_{p+r}^{b} \right)^{-2} (\log z_{p+r})^{-1} d z_{p+r}
\]

\[
\asymp \lambda^{-1/b} \int_{\lambda^{1/b}}^{\infty} (1 + x^{b})^{-2} (\log x - b^{-1} \log \lambda)^{-1} d x \asymp \lambda^{-a-1/2m} [\log(1/\lambda)]^{-1},
\]

where the last step follows from the fact that \( 2b > 1 \) for any \( 0 \leq a < (2m - 1)/(2m) \).

Second, we consider \( d - r < p < d \). As discussed in the previous case, the number of combinations of non-1 indices being summed is finite, and the largest terms of the summation \[72\] over \( \mathbf{\nu} \in \mathbb{V} \) correspond to the indices \( \mathbf{\nu} = (\nu_{k_1}, \ldots, \nu_{k_{r+d-p}}, \nu_{p+1}, \ldots, \nu_d)^{\top} \in \mathbb{N}^r \), where the indices \( k_1, \ldots, k_{r+d-p} \leq p \). Thus, by the integral approximation,

\[
N_a(\lambda) \asymp \sum_{\nu_{d-r+1} = 1}^{\infty} \cdots \sum_{\nu_d = 1}^{\infty} \frac{\prod_{k=d-r+1}^{d} \nu_k^{2ma}}{\left( 1 + \sum_{k=d-r+1}^{p} \nu_k^{2} + \lambda \prod_{k=d-r+1}^{d} \nu_k^{2m} \right)^{2}}
\]

\[
\asymp \int_{1}^{\infty} \cdots \int_{1}^{\infty} \left( 1 + x_{d-r+1}^{b/m} + \cdots + x_{p}^{b/m} + \lambda x_{d-r+1}^{b} \cdots x_{d}^{b} \right)^{-2} \frac{d x_{d-r+1} \cdots d x_{d}},
\]

where \( b = 2m/(2ma + 1) \). Set \( z_k = x_{p+1}x_{p+2} \cdots x_k \) for \( k = p + 1, \ldots, d \). By using the change the variables to replace \((x_{p+1}, \ldots, x_d)\) by \((z_{p+1}, \ldots, z_d)\), and \( z_d \) by \( x = \lambda^{1/b} z_d \), and \( x \) by \( u = x_{d-r+1} \cdots x_p \cdot x \). We have

\[
N_a(\lambda) \asymp \int_{1}^{\infty} \cdots \int_{1}^{\infty} \left[ \int_{1}^{\infty} \cdots \int_{1}^{\infty} \left( 1 + x_{d-r+1}^{b/m} + \cdots + x_{p}^{b/m} + \lambda x_{d-r+1}^{b} \cdots x_{d}^{b} \right)^{-2} \right]
\]

\[
\cdot \left( x_{p+1}^{-1} \cdots x_{d-r-1}^{-1} d z_{p+1} \cdots d z_{d-r-1} d z_d \right) d x_{d-r+1} \cdots d x_{p}
\]

\[
\asymp \lambda^{-1/b} \int_{\lambda^{1/b}}^{\infty} \left[ \int_{1}^{\infty} \left( 1 + x_{d-r+1}^{b/m} + \cdots + x_{p}^{b/m} + u \right)^{-2} \right]
\]

\[
\cdot (\log x - b^{-1} \log \lambda)^{-1} d x \asymp \lambda^{-a-1/(2m)} [\log(1/\lambda)]^{-1},
\]

\[
77
\]
By Lemma 14 then for any $0 < \tau < 1$,
\[
\left(1 + x_{d-r+1}^{b/m} + x_{d-r+2}^{b/m} + \cdots + x_p^{b/m} + u^b\right)^{-2} \\
\lesssim \left(1 + x_{d-r+2}^{b/m} + \cdots + x_p^{b/m} + u^b\right)^{-1+\tau} \left(x_{d-r+1}^{b/m}\right)^{-(1+\tau)}.
\]
Together with the fact $\int_t^\infty \frac{1}{t^{1+\tau} (\log t)^k} dt < \infty$ for any $k < \infty$, we have
\[
N_a(\lambda) \lesssim \lambda^{-1/b} \int_{\lambda^{1/b}}^\infty \left(1 + u^b\right)^{(1-\tau)p-d+r} \left(\log u - b^{-1} \log \lambda\right)^{d-p-1} du.
\]
Continuing this procedure gives
\[
N_a(\lambda) \lesssim \lambda^{-1/b} \int_{\lambda^{1/b}}^\infty \left(1 + u^b\right)^{(1-\tau)p-d+r} \left(\log u - b^{-1} \log \lambda\right)^{d-p-1} du.
\]
Since for any $\epsilon > 0$ and $d - r < p < d$, we know if $\tau < \epsilon/d$,
\[
(1 - \tau)^{p-d+r} \geq 1 - \tau(p - d + r) \geq 1 - \tau(d - 1) \geq 1 - \epsilon.
\]
Hence, for any $0 < a < (2m - 1)/(2m)$, there exists $\tau$ such that $(1 - \tau)^{p-d+r} > a + 1/(2m) = 1/b$. Therefore,
\[
N_a(\lambda) \lesssim \lambda^{-1/b} [\log(1/\lambda)]^{d-p-1} = \lambda^{-a-1/2m} [\log(1/\lambda)]^{d-p-1}.
\]
Finally, we consider $p = d$. As argued in the previous two cases, the number of combinations of non-1 indices being summed is finite. Now since $p = d$, by the symmetry of indices, the largest terms of the summation \([72]\) over $\vec{\nu} \in \mathbb{V}$ correspond to any combinations of $r$ non-1 indices, for example, the first $r$ indices. Thus, by the integral approximation,
\[
N_a(\lambda) \asymp \sum_{\nu_1=1}^{\infty} \cdots \sum_{\nu_r=1}^{\infty} \sum_{\nu_i=1}^{\infty} \frac{\prod_{k=1}^{r} \nu_k^{2ma} (1 + \sum_{k=1}^{r} \nu_k^2)}{(1 + \sum_{k=1}^{r} \nu_k^2 + \lambda \prod_{k=1}^{r} \nu_k^{2m})^2} dx_1 \cdots dx_r \int_1^\infty \cdots \int_1^\infty \frac{1 + x_1^{b/m} + \cdots + x_r^{b/m}}{\left(1 + x_1^{b/m} + \cdots + x_r^{b/m} + \lambda x_1^{b/m} \cdots x_r^{b/m}\right)^2} dx_1 \cdots dx_r \cdots dx_r, \quad \text{where } b = 2m/(2ma + 1).
\]
Observe that if $x_1 \cdots x_{r-1} x_r \lesssim \lambda^{mr/[b(1-mr)]}$, then
\[
\lambda x_1^{b/m} \cdots x_{r-1}^{b/m} x_r^{b/m} \lesssim x_1^{b/m} + \cdots + x_r^{b/m} + x_{r-1}^{b/m} + x_r^{b/m}.
\]
By Lemma 6 with $\beta = 0$ and $\alpha = b/m \leq 2$, we have

$$N_a(\lambda) \geq \int_{x_1 \cdots x_r \cdots x_r \leq \lambda^{m/a}/(b(1-mr))} \left(1 + x_1^{b/m} + \cdots + x_r^{b/m} + \lambda x_1^{b/m} \cdots x_r^{b/m} \right)^{-1} dx_1 \cdots dx_r$$

$$\approx \begin{cases} \lambda^{m/a}(a + \frac{r-2}{2m}), & \text{if } r \geq 3; \\ \log(1/\lambda), & \text{if } r = 2, a = 0; \\ \lambda^{2m/a}, & \text{if } r = 2, 0 < a \leq 1; \\ 1, & \text{if } r = 1, a < \frac{1}{2m}; \\ \lambda^{1-2m/a}, & \text{if } r = 1, a > \frac{1}{2m}. \end{cases}$$ (73)

On the other hand, if $\lambda^{m/a}/(b(1-mr)) (x_1 \cdots x_{r-1} x_r)^{-1} = o(1)$, then without loss of generality we assume $x_r = \min\{x_1, \cdots, x_r\}$. Let $z = \lambda^{1/b} x_1 \cdots x_{r-1} x_r$. By changing $x_r$ to $z$, we have

$$N_a(\lambda) \approx \int_{\lambda^{m/a}/(b(1-mr)) (x_1 \cdots x_{r-1} x_r)^{-1} = o(1)} \left(1 + x_1^{b/m} + \cdots + x_r^{b/m} + \lambda x_1^{b/m} \cdots x_r^{b/m} \right)^{-1} dx_1 \cdots dx_r$$

$$\approx \lambda^{-1/b} \int_{\lambda^{m/a}/(b(1-mr)) (z-1)^{-1} = o(1)} \left(1 + x_1^{b/m} + \cdots + x_r^{b/m} + z^{b} \right)^{-1} x_1^{-1} \cdots x_r^{-1} dx_1 \cdots dx_r dz$$

$$\approx \lambda^{-1/b} \int_{\lambda^{m/a}/(b(1-mr)) (z-1)^{-1} = o(1)} \left[ \int_{x_1^{b/m} + \cdots + x_r^{b/m} = (r-1)/(br)}} \left(1 + x_1^{b/m} + \cdots + x_r^{b/m} + z^{b} \right)^{-1} x_1^{-1} \cdots x_r^{-1} dx_1 \cdots dx_r \right] z^{b(1+\tau)} dz$$

$$\approx \lambda^{-1/b} \int_{\lambda^{m/a}/(b(1-mr)) (z-1)^{-1} = o(1)} \lambda^{\tau/(mr)} z^{-\tau b/(mr)} \cdot z^{b(1+\tau)} dz = o \left[ \lambda^{m/a}(a + \frac{r-2}{2m}) \right],$$ (74)

where the third step follows from the Lemma 15 for $\beta = -1$ and $\alpha = \tau b/m$. Combining (73) and (74), we complete the proof for $p = d$ and this lemma.

**Lemma 13** (Bounding the norm of the product of functions). For any $f, g \in \otimes^d \mathcal{H}_1$, $a > 1/2m$, and $1 \leq p \leq d$, we have that

$$\sum_{\varphi \in \mathcal{N}} \left(1 + \frac{\rho \varphi}{\|\varphi\|_2^2} \right)^a \|\varphi\|_2^2 \left\langle \frac{\partial f(t)}{\partial t_j}, \frac{\partial g(t)}{\partial t_j}, \varphi(t) \right\rangle_0^2 \leq \|f\|_2^2 (a+1/m) \left[ \sum_{\varphi \in \mathcal{N}} \left(1 + \frac{\rho \varphi}{\|\varphi\|_2^2} \right)^a \|\varphi\|_2^2 \left\langle \frac{\partial g(t)}{\partial t_j}, \varphi(t) \right\rangle_0^2 \right].$$

79
Proof. Recall that \( \{ \psi_\nu(t) \}_{\nu \geq 1} \) is the trigonometrical basis on \( L_2(\mathcal{X}) \) and \( \phi_\nu(\cdot) \) is defined in [32]. Write \( \psi_\nu = \psi_{\nu_1}(t_1)\psi_{\nu_2}(t_2) \cdots \psi_{\nu_d}(t_d) \). Note that

\[
\sum_{\nu \in \mathbb{N}^d} \left( 1 + \frac{\rho \nu}{\| \phi_\nu \|^2_{L_2}} \right) \| \phi_\nu \|^2_{L_2} \left( f, \phi_\nu \right)_0^2 = \sum_{\nu \in \mathbb{N}^d} \left( 1 + \frac{\rho \nu}{\| \phi_\nu \|^2_{L_2}} \right) \left( \int_{\mathcal{X}^d} f \psi_\nu \right)^2.
\]

By Theorem A.2.2 and Corollary A.2.1 in Lin [1998], if \( a > 1/2m \), then for any \( f, g \in \mathcal{O}_d \mathcal{H}_1 \),

\[
\sum_{\nu \in \mathbb{N}^d} \left( 1 + \frac{\rho \nu}{\| \phi_\nu \|^2_{L_2}} \right) \left( \int_{\mathcal{X}^d} f g \psi_\nu \right)^2 \leq \left[ \sum_{\nu \in \mathbb{N}^d} \left( 1 + \frac{\rho \nu}{\| \phi_\nu \|^2_{L_2}} \right) \left( \int_{\mathcal{X}^d} f \psi_\nu \right)^2 \right] \left[ \sum_{\nu \in \mathbb{N}^d} \left( 1 + \frac{\rho \nu}{\| \phi_\nu \|^2_{L_2}} \right) \left( \int_{\mathcal{X}^d} g \psi_\nu \right)^2 \right].
\]

Thus,

\[
\sum_{\nu \in \mathbb{N}^d} \left( 1 + \frac{\rho \nu}{\| \phi_\nu \|^2_{L_2}} \right) \| \phi_\nu \|^2_{L_2} \left( \partial f(t) \frac{\partial g(t)}{\partial t_j}, \phi_\nu(t) \right)_0^2 = \sum_{\nu \in \mathbb{N}^d} \left( 1 + \frac{\rho \nu}{\| \phi_\nu \|^2_{L_2}} \right) \left( \int_{\mathcal{X}^d} \frac{\partial f(t)}{\partial t_j} \frac{\partial g(t)}{\partial t_j} \psi_\nu(t) \right)^2 \leq \left[ \sum_{\nu \in \mathbb{N}^d} \left( 1 + \frac{\rho \nu}{\| \phi_\nu \|^2_{L_2}} \right) \left( \int_{\mathcal{X}^d} f(t) \psi_\nu(t) \right)^2 \right] \left[ \sum_{\nu \in \mathbb{N}^d} \left( 1 + \frac{\rho \nu}{\| \phi_\nu \|^2_{L_2}} \right) \left( \int_{\mathcal{X}^d} \frac{\partial g(t)}{\partial t_j} \psi_\nu(t) \right)^2 \right].
\]

This completes the proof.

Lemma 14 (A variant of Young’s inequality). For any \( a, b \geq 0 \) and \( 0 < \tau < 1 \), we have

\[
(a + b)^{-2} \leq \frac{(1 - \tau)^{1-\tau}(1 + \tau)^{1+\tau}}{4} a^{-(1+\tau)} b^{-(1-\tau)}.
\]

When \( \tau \) is small, the coefficient \( (1 - \tau)^{1-\tau}(1 + \tau)^{1+\tau} / 4 \) is close to \( 1/4 \).

Proof. To prove (75), it is sufficient to show

\[
a + b \geq 2(1 - \tau)^{-(1-\tau)/2}(1 + \tau)^{-(1+\tau)/2} a^{(1+\tau)/2} b^{(1-\tau)/2}.
\]

80
Letting \( p = 2/(1 + \tau) \), \( a^\prime = a^{1/p} \), \( b^\prime = [b/(p - 1)]^{(p-1)/p} \), the above formula is equivalent to
\[
\frac{a^\prime}{p} + \frac{(b^\prime)^{p/(p-1)}}{p/(p-1)} \geq a^\prime b^\prime,
\]
which holds by Young’s inequality. This completes the proof.

**Lemma 15.** Suppose that \( \beta \leq -1 \) and \( \alpha > 0 \). Then, as \( \Xi \to \infty \),
\[
\int_{x_1 \cdots x_r \geq \Xi} x_1^\beta (x_1^\alpha + x_2^\alpha + \cdots + x_r^\alpha)^{-1} dx_1 \cdots dx_r \approx \Xi^{\beta + 1 - \alpha/r}.
\]

**Proof.** The proof is similar to the proof for Lemma 6. We omit the details here. 

### E  Proofs for Section A

For brevity, we consider the regular lattice \( l_1 = \cdots = l_d = l \) and \( n = l^d \). Other regular lattices can be shown similarly. Write
\[
\psi_1(t) = 1, \quad \psi_{2\nu}(t) = \sqrt{2} \cos 2\pi \nu t, \quad \psi_{2\nu+1}(t) = \sqrt{2} \sin 2\pi \nu t,
\]
for \( \nu \geq 1 \). As discussed in Section A, it is without loss of generality to assume that \( f_0 : \mathcal{X}^d \to \mathbb{R} \) has a periodic boundaries on \( \mathcal{X}^d \). Hence \( \{\psi_\nu(t)\}_{\nu \geq 1} \) forms an orthonormal system in \( L_2(\mathcal{X}) \) and an eigenfunction system for \( K \). For a \( d \)-dimensional vector \( \vec{\nu} = (\nu_1, \ldots, \nu_d) \in \mathbb{N}^d \), write
\[
\psi_{\vec{\nu}}(t) = \psi_{\nu_1}(t_1) \cdots \psi_{\nu_d}(t_d) \quad \text{and} \quad \lambda_{\vec{\nu}} = \lambda_{\nu_1} \lambda_{\nu_2} \cdots \lambda_{\nu_d},
\]
where \( \lambda_{\nu_j} \)'s and \( \psi_{\nu_j}(t_j) \)'s are defined according to the spectral theorem, \( j = 1, \ldots, d \). Then, any function \( f(\cdot) \) in \( \mathcal{H} \) admits the Fourier expansion \( f(t) = \sum_{\vec{\nu} \in \mathbb{N}^d} \theta_{\vec{\nu}} \psi_{\vec{\nu}}(t) \), where \( \theta_{\vec{\nu}} = \langle f(t), \psi_{\vec{\nu}}(t) \rangle_{L_2} \), and \( J(f) = \sum_{\vec{\nu} \in \mathbb{N}^d} \lambda_{\vec{\nu}}^{-1} \theta_{\vec{\nu}}^2 \). We also write \( f_0(t) = \sum_{\vec{\nu} \in \mathbb{N}^d} \theta_{\vec{\nu}}^0 \psi_{\vec{\nu}}(t) \).

By Page 23 of [Wahba 1990], it is known that
\[
l^{-1} \sum_{i=1}^l \psi_{\mu}(i/l) \psi_{\nu}(i/l) = \begin{cases} 
1, & \text{if } \mu = \nu = 1, \ldots, l, \\
0, & \text{if } \mu \neq \nu, \mu, \nu = 1, \ldots, l.
\end{cases}
\]
Define
\[
\vec{\psi}_{\vec{\nu}} = (\psi_{\vec{\nu}}(t_1), \ldots, \psi_{\vec{\nu}}(t_n))^\top,
\]
for
\[
\bar{\psi}_{\vec{\nu}} = (\psi_{\vec{\nu}}(t_1), \ldots, \psi_{\vec{\nu}}(t_n))^\top.
\]
where \( \{t_1, \ldots, t_n\} \) are the regular lattice design points. Thus, we have

\[
\langle \tilde{\psi}_\theta, \tilde{\psi}_\mu \rangle_n = \begin{cases} 1, & \text{if } \nu_j = \mu_j = 1, \ldots, l; j = 1, \ldots, d, \\ 0, & \text{if there exists some } j \text{ such that } \nu_j \neq \mu_j, \end{cases}
\]

where \( \langle \cdot, \cdot \rangle_n \) is the empirical inner product in \( \mathbb{R}^n \). This implies that \( \{\tilde{\psi}_\nu \mid \nu_j = 1, \ldots, l; j = 1, \ldots, d\} \) form an orthogonal basis in \( \mathbb{R}^n \) with respect to the empirical norm \( \| \cdot \|_n \). Denote the observed data vectors by \( y^{(0)} = (y_1^{(0)}, \ldots, y_n^{(0)})^T \) and \( y^{(j)} = (y_1^{(j)}, \ldots, y_n^{(j)})^T \), and write

\[
\begin{aligned}
\begin{cases}
z_{\nu}^{(0)} &= \langle y^{(0)}, \tilde{\psi}_\nu \rangle_n, \\
z_{\nu_1,\ldots,2\nu_j-1,\ldots,2\nu_j-1,\ldots,\nu_d}^{(j)} &= (2\pi)^{-1} \langle y^{(j)}, \tilde{\psi}_{\nu_1,\ldots,2\nu_j-1,\ldots,\nu_d} \rangle_n, \\
z_{\nu_1,\ldots,2\nu_j-1,\ldots,\nu_d}^{(j)} &= - (2\pi)^{-1} \langle y^{(j)}, \tilde{\psi}_{\nu_1,\ldots,2\nu_j-1,\ldots,\nu_d} \rangle_n,
\end{cases}
\end{aligned}
\tag{78}
\]

for \( \nu_j = 1, \ldots, l \) and \( j = 1, \ldots, d \). Then, \( z_{\nu}^{(0)} = \tilde{\theta}_{\nu}^0 + \delta_{\nu}^{(0)} \) and \( z_{\nu}^{(j)} = \nu_j \tilde{\theta}_{\nu}^0 + \delta_{\nu}^{(j)} \), where \( \tilde{\theta}_{\nu}^0 = \theta_{\nu}^0 + \sum_{\mu_j \geq t+1, j=1,\ldots,d} \theta_{\mu_j}^0 \langle \tilde{\psi}_{\nu}, \tilde{\psi}_{\mu_j} \rangle_n \). The errors \( \delta_{\nu}^{(0)} \) satisfy

\[
\mathbb{E}[\delta_{\nu}^{(0)}] = \frac{1}{n} \sum_{i=1}^n \mathbb{E}[\epsilon_i^{(0)}] \tilde{\psi}_{\nu}(i) \leq \frac{1}{n} \sqrt{\sum_{i=1}^n \mathbb{E}[\epsilon_i^{(0)}]^2} \sqrt{\sum_{i=1}^n \tilde{\psi}_{\nu}^2(i)} = o(n^{-1/2}),
\]

\[
\text{Var}[\delta_{\nu}^{(0)}] = \frac{1}{n^2} \sum_{i=1}^n \text{Var}[\epsilon_i^{(0)}] \tilde{\psi}_{\nu}^2(i) + \frac{1}{n^2} \sum_{i \neq i'} \text{Cov}[\epsilon_i^{(0)}, \epsilon_i^{(0)}] \tilde{\psi}_{\nu}(i) \tilde{\psi}_{\nu}(i') \leq \frac{\sigma^2}{n} \cdot \frac{1}{n} \sum_{i=1}^n \tilde{\psi}_{\nu}^2(i) + \frac{2}{n^2} \sum_{i \neq i'} \text{Cov}[\epsilon_i^{(0)}, \epsilon_i^{(0)}] \]

\[
= O(n^{-1}) + \frac{2}{n^2} \sum_{i \neq i'} o(|i - i'|^{-\gamma}) = O(n^{-1}) + o(n^{-1}) = O(n^{-1}).
\]

Similarly for any \( j \), \( \delta_{\nu}^{(j)} \)'s have mean \( o(n^{-1/2}) \) and covariances \( O(n^{-1}) \).

### E.1 Proof of Theorem 4

We now prove the lower bound under the regular lattices. By the data transformation (78), it suffices to show the optimal rate in a special case

\[
\begin{cases}
z_{\nu}^{(0)} = \theta_{\nu}^0 + \delta_{\nu}^{(0)}, \\
z_{\nu}^{(j)} = \nu_j \theta_{\nu}^0 + \delta_{\nu}^{(j)},
\end{cases}
\tag{79}
\]

for \( 1 \leq j \leq p \).
where \( \sigma^{(j)} \sim \mathcal{N}(0, \sigma^2_j/n) \) are independent. For any \( \tilde{\nu} \in \mathbb{N}^d \), if we have the prior that \( |\tilde{\theta}_0^T| \leq \pi_{\tilde{\nu}} \), then the minimax linear estimator is

\[
\tilde{\theta}_0^T = \frac{\sigma_{0}^{-2}z_{(0)} + \sum_{j=1}^{p} \sigma_{j}^{-2}z_{(j)}^{(j)}}{n^{-1}{\pi_{\tilde{\nu}}^{-2}} + \sigma_{0}^{-2} + \sum_{j=1}^{p} \sigma_{j}^{-2}{\nu}^{2}_j},
\]

and the minimax linear risk is

\[
n^{-1} \left[ n^{-1}{\pi_{\tilde{\nu}}^{-2}} + \sigma_{0}^{-2} + \sum_{j=1}^{p} \sigma_{j}^{-2}{\nu}^{2}_j \right]^{-1}.
\]

By Lemma 6 and Theorem 7 in Donoho et al. (1990), if \( \sigma^2_j \)s are known, the minimax risk of estimating \( \theta_0^T \) under the model (79) is larger than 80% of the minimax linear risk of the hardest rectangle subproblem, and the latter linear risk is

\[
R^L = n^{-1} \max_{\sum_{\tilde{\nu} \in V}(1 + \lambda_{\tilde{\nu}})\pi_{\tilde{\nu}}^2 = 1} \sum_{\tilde{\nu} \in V} \left[ n^{-1}{\pi_{\tilde{\nu}}^{-2}} + \sigma_{0}^{-2} + \sum_{j=1}^{p} \sigma_{j}^{-2}{\nu}^{2}_j \right]^{-1}, \tag{80}
\]

where \( \lambda_{\tilde{\nu}} \) is the product of eigenvalues in (77) and recall that the set \( V \) is defined in (22).

We use the Lagrange multiplier method to find \( \pi_{\tilde{\nu}}^2 \) for solving (80). Let \( a \) be the scalar multiplier and define

\[
L(\pi_{\tilde{\nu}}^2, a) = \sum_{\tilde{\nu} \in V} \left[ n^{-1}{\pi_{\tilde{\nu}}^{-2}} + \sigma_{0}^{-2} + \sum_{j=1}^{p} \sigma_{j}^{-2}{\nu}^{2}_j \right]^{-1} - a(1 + \lambda_{\tilde{\nu}})\pi_{\tilde{\nu}}^2.
\]

Taking partial derivative with respect to \( \pi_{\tilde{\nu}}^2 \) gives

\[
\frac{\partial L}{\partial \pi_{\tilde{\nu}}^2} = n^{-1} \left[ n^{-1} + \left( \sigma_{0}^{-2} + \sum_{j=1}^{p} \sigma_{j}^{-2}{\nu}^{2}_j \right) \pi_{\tilde{\nu}}^2 \right]^{-2} - a(1 + \lambda_{\tilde{\nu}}) = 0.
\]

This implies

\[
\hat{\pi}_{\tilde{\nu}}^2 = \left( \sigma_{0}^{-2} + \sum_{j=1}^{p} \sigma_{j}^{-2}{\nu}^{2}_j \right)^{-1} \left[ b(1 + \lambda_{\tilde{\nu}})^{-1/2} - n^{-1} \right]_+,
\]

where \( b = (na)^{-1/2} \). On one hand, plugging the above formula into the constraint \( \sum_{\tilde{\nu} \in V}(1 + \lambda_{\tilde{\nu}})\pi_{\tilde{\nu}}^2 = 1 \) gives,

\[
\sum_{\tilde{\nu} \in V} \prod_{k=1}^{d} \nu_{k}^{2m} \left( \sigma_{0}^{-2} + \sum_{j=1}^{p} \sigma_{j}^{-2}{\nu}^{2}_j \right)^{-1} \left[ b \prod_{k=1}^{d} \nu_{k}^{-m} - n^{-1} \right]_+ \approx 1.
\]
By restricting \( \prod_{j=1}^{d} \nu_j \leq (nb)^{1/m} \), this becomes

\[
\sum_{\tilde{\nu} \in \mathbb{V}, \prod_{k=1}^{d} \nu_k \leq (nb)^{1/m}} \left( \sigma_0^{-2} + \sum_{j=1}^{p} \sigma_j^{-2} \nu_j^2 \right)^{-1} \left( b \prod_{k=1}^{d} \nu_k^m - n^{-1} \prod_{k=1}^{d} \nu_k^{2m} \right) \asymp 1. \tag{81}
\]

On the other hand, the linear risk in (80) can be written as

\[
R^L \asymp n^{-1} \sum_{\tilde{\nu} \in \mathbb{V}, \prod_{k=1}^{d} \nu_k \leq (nb)^{1/m}} \left( 1 - \frac{1}{nb} \prod_{k=1}^{d} \nu_k^m \right) \times \left( \sigma_0^{-2} + \sum_{j=1}^{p} \sigma_j^{-2} \nu_j^2 \right)^{-1}. \tag{82}
\]

We discuss for \( R^L \) in the above (82) under the condition (81) for three cases with \( 0 \leq p \leq d - r, \ d - r < p < d \) and \( p = d \).

If \( 0 \leq p \leq d - r \), since \( \tilde{\nu} \in \mathbb{V} \), there are at most \( r \) of \( \nu_1, \ldots, \nu_d \) not equal to 1, which implies that the number of combinations of non-1 indices being summed in (81) is no greater than \( C_d^1 + C_d^2 + \cdots + C_d^r < \infty \). Due to the term \( (\sigma_0^{-2} + \sum_{j=1}^{p} \sigma_j^{-2} \nu_j^2)^{-1} \), the largest terms of the summation (81) over \( \tilde{\nu} \in \mathbb{V} \) correspond to the combinations of indices whereas fewer \( \nu_1, \ldots, \nu_p \) being summed as possible, for example, \( \nu_k \equiv 1 \) for \( k \leq p \) and \( k > p + r \), and \( (\nu_{p+1}, \ldots, \nu_{p+r}) \in \mathbb{N}^r \) are non-1. Thus, (81) is equivalent to

\[
\sum_{\prod_{k=1}^{d} \nu_{p+k} \leq (nb)^{1/m}} \left( b \prod_{k=1}^{r} \nu_{p+k}^m - n^{-1} \prod_{k=1}^{r} \nu_{p+k}^{2m} \right) \asymp 1.
\]

Using the integral approximation, we have

\[
\int_{\prod_{k=1}^{d} x_{p+k} \leq (nb)^{1/m}, x_{p+k} \geq 1} \left( b \prod_{k=1}^{r} x_{p+k}^m - n \prod_{k=1}^{r} x_{p+k}^{2m} \right) dx_{p+1} \cdots dx_{p+r} \asymp 1.
\]

By letting \( z_j = \prod_{1 \leq k \leq j} x_{p+k}, \ j = 1, 2, \ldots, r \), we have

\[
\int_{(nb)^{1/m}} \left[ \int_{1}^{z_{2r}} \cdots \int_{1}^{z_{2}} \left( b z_{r}^{m} - \frac{1}{n} z_{r}^{2m} \right) z_{1-1}^{1} \cdots z_{r-1}^{1} dz_1 \cdots dz_{r-1} \right] dz_r \asymp 1,
\]

where the left-hand side term is the order of \( n^{(m+1)/mb(2m+1)/m[\log(nb)]^r} \) and hence

\[
b \asymp n^{-(m+1)/(2m+1)(\log n)^{-m(r-1)/(2m+1)}}. \tag{83}
\]
The linear risk in (82) becomes

\[ R^L \asymp n^{-1} \int \prod_{k=1}^{r} x_{p+k}^{\nu_{d-r+k}} \left( 1 - \frac{1}{nb} \prod_{k=1}^{r} x_{p+k}^{m} \right) \]

\[ \asymp [\log(nb)]^{-1} n^{-1+1/m} b^{1/m} \asymp [n(\log n)^{1-r}]^{-2m/(2m+1)}, \]

where the last step is by (83).

If \( d - r < p < d \), as discussed in the previous case, the number of combinations of non-1 indices being summed is finite, and the largest terms of the summation (81) over \( \vec{v} \in \mathbb{V} \) correspond to the combinations of indices whereas fewer than \( \nu_1, \ldots, \nu_p \) being summed as possible, for example, \( \nu_k \equiv 1 \) for \( k \leq d - r \), and \( (\nu_{d-r+1}, \ldots, \nu_d) \in \mathbb{N}^r \) are non-1. Thus, (81) is equivalent to

\[ \sum_{\prod_{k=1}^{r} x_{d-r+k} \leq (nb)^{1/m}} \left( b^{r} \prod_{k=1}^{r} \nu_{d-r+k} = n^{-1} \prod_{k=1}^{r} \nu_{d-r+k}^{2m} \right) \left( 1 + \sum_{j=d-r+1}^{p} \nu_{j}^{2} \right)^{-1} \asymp 1. \]

Using the integral approximation, we have

\[ \int \prod_{k=1}^{r} x_{d-r+k}^{\nu_{d-r+k} \leq (nb)^{1/m}} \left( b^{r} \prod_{k=1}^{r} x_{d-r+k}^{m} = n^{-1} \prod_{k=1}^{r} x_{d-r+k}^{2m} \right) \]

\[ \times \left( 1 + \sum_{j=d-r+1}^{p} x_{j}^{2} \right)^{-1} dx_{d-r+1} \cdots dx_{d} \asymp 1. \]

By letting \( z_j = x_{p+1} x_{p+2} \cdots x_j, j = p + 1, \ldots, d \), we get

\[ 1 \asymp \int_{x_{d-r+1} \cdots x_p z_d \leq (nb)^{1/m}} \left[ \int_{1}^{z_d} \cdots \int_{1}^{z_{p+2}} \left( b x_{d-r+1}^{m} \cdots x_p^{m} z_d^{m} - \frac{1}{n} x_{d-r+1}^{2m} \cdots x_p^{2m} z_d^{2m} \right) \right] \]

\[ \times \left( 1 + x_{d-r+1}^{2} + \cdots + x_{p}^{2} \right)^{-1} dz_{p+1} \cdots dz_{d-1} \]

\[ \times dx_{d-r+1} \cdots dx_p dz_d \]

\[ = \int_{x_{d-r+1} \cdots x_p z_d \leq (nb)^{1/m}} b x_{d-r+1}^{m} \cdots x_p^{m} z_d^{m} \left( 1 - \frac{1}{nb} x_{d-r+1}^{m} \cdots x_p^{m} z_d^{m} \right) \]

\[ \times (\log z_d)^{d-p-1} \left( 1 + x_{d-r+1}^{2} + \cdots + x_{p}^{2} \right)^{-1} dx_{d-r+1} \cdots dx_p dz_d \]

\[ \asymp [\log(nb)]^{d-p-1} n^{1+1/m} b^{2+1/m}. \]
The last step is by Lemma 16. Hence,
\[ b \asymp n^{-(m+1)/(2m+1)}(\log n)^{-m(d-p-1)/(2m+1)}. \] (84)

The linear risk in (82) becomes
\[
R^L \asymp n^{-1} \int_{\prod_{k=d-r+1}^{d} x_k \leq (nb)^{1/m}, x_k \geq 1} \left( 1 - \frac{1}{nb} x_{d-r+1}^m \cdots x_d^m \right) \cdot (1 + x_{d-r+1}^2 + \cdots + x_p^2)^{-1} dx_{d-r+1} \cdots dx_d \\
\asymp n^{-1} \int_{x_{d-r+1} \cdots x_p, z_d \leq (nb)^{1/m}} \left( 1 - \frac{1}{nb} x_{d-r+1}^m \cdots x_p^m z_d^m \right) (\log z_d)^{d-p-1} \\
\quad \cdot (1 + x_{d-r+1}^2 + \cdots + x_p^2)^{-1} dx_{d-r+1} \cdots dx_p dz_d \\
\asymp [\log(nb)]^{d-p-1} n^{-1+1/m} b^{1/m},
\]
where the second step uses the same change of variables by letting \( z_j = x_{p+1} x_{p+2} \cdots x_j, \ j = p+1, \ldots, d, \) and the last step is by Lemma 16. By (84), we have
\[
R^L \asymp [n(\log n)^{1+p-d}]^{-2m/(2m+1)}.
\]

If \( p = d, \) as discussed in the previous two cases, the number of combinations of non-1 indices being summed is finite, and the largest terms of the summation (81) over \( \vec{\nu} \in \mathbb{V} \) correspond to any combinations of \( r \) non-1 indices, for example, \( \nu_k \equiv 1 \) for \( k \geq r + 1, \) and \( (\nu_1, \ldots, \nu_r) \in \mathbb{N}^r. \) Thus, (81) is equivalent to
\[
\sum_{\prod_{k=1}^r \nu_k \leq (nb)^{1/m}} \left( b \prod_{k=1}^r \nu_k^m - n^{-1} \prod_{k=1}^r \nu_k^{2m} \right) \left( 1 + \sum_{k=1}^r \nu_k^2 \right)^{-1} \asymp 1.
\]

Using the integral approximation, we have
\[
1 \asymp \int_{\prod_{k=1}^r x_k \leq (nb)^{1/m}, x_k \geq 1} \left( b \prod_{k=1}^r x_k^m - n^{-1} \prod_{k=1}^r x_k^{2m} \right) \left( 1 + \sum_{k=1}^r x_k^2 \right)^{-1} dx_1 \cdots dx_r \\
\asymp \int_{\prod_{k=1}^r x_k \leq (nb)^{1/m}, x_k \geq 1} b \prod_{k=1}^r x_k^m \left( 1 + \sum_{k=1}^r x_k^2 \right)^{-1} dx_1 \cdots dx_r
\]
By letting \( \beta = m > 1 \) and \( \alpha = 2 \) in Lemma 6, we have for any \( r \geq 1, \)
\[
b \asymp n^{-(mr+r-2)/(2mr+r-2)}.
\] (85)
The linear risk in (82) becomes
\[ R^L \propto n^{-1} \int \prod_{k=1}^{r} x_k (nb)^{1/m} x_k \geq 1 \left( 1 - \frac{1}{n b x_1^m \ldots x_r^m} \right) (1 + x_1^2 + \ldots + x_r^2)^{-1} dx_1 \ldots dx_r \]
\[ \propto n^{-1} \int \prod_{k=1}^{r} x_k (nb)^{1/m} x_k \geq 1 (1 + x_1^2 + \ldots + x_r^2)^{-1} dx_1 \ldots dx_r \]
\[ \propto \left[ n^{-1} (nb)^{(r-2)/(mr)} 1_{r \geq 3} + \left[ n^{-1} \log(nb) \right] 1_{r=2} + (n^{-1}) 1_{r=1} \right] \]
where the last step uses Lemma 6 with \( \beta = 0 \) and \( \alpha = 2 \). By (83), we have
\[ R^L \propto \left[ n^{-2mr}/((2m+1)r-2) \right] 1_{r \geq 3} + \left[ n^{-1} \log(n) \right] 1_{r=2} + n^{-1} 1_{r=1} \]
where the constant factor does not depend on \( n \). This completes the proof.

### E.2 Proof of Theorem 5

We now prove the theorem for only \( r = d \) and \( p = d - 1 \). Other settings can be shown similarly. Using the discrete transformed data (78), the estimator \( \hat{f}_n \) in (9) can be obtained through
\[
\hat{\theta}_\varphi = \arg\min_{\hat{\theta}_\varphi \in \mathbb{R}} \left\{ \frac{1}{n(p+1)} \left[ \frac{1}{\sigma_\varphi^2} \sum_{\varphi \in V ||\varphi||_{min} \leq t} \left( z_\varphi^{(0)} - \theta_\varphi \right)^2 \right. \\
+ \sum_{j=1}^{p} \frac{1}{\sigma_j^2} \sum_{\varphi_j \in V ||\varphi||_{min} \leq t} \left( z_\varphi^{(j)} - \nu_j \theta_\varphi \right)^2 \right. \\
+ \lambda \sum_{\varphi_j \in V ||\varphi||_{min} \leq t} \lambda \theta_\varphi^2 \right\}
\]
and \( \hat{f}_n(t) = \sum_{\varphi \in V ||\varphi||_{min} \leq t} \hat{\theta}_\varphi \varphi(t) \), where \( V \) is defined in (22). Direct calculations give
\[
\hat{\theta}_\varphi = \frac{\sigma_\varphi^{-2} z_\varphi^{(0)} + \sum_{j=1}^{p} \sigma_j^{-2} \nu_j z_\varphi^{(j)}}{\sigma_\varphi^{-2} + \sum_{j=1}^{p} \sigma_j^{-2} \nu_j^2 + \lambda \sigma_\varphi^{-1}}.
\]

The deterministic error of \( \hat{f}_n \) can be analyzed in two parts. On one hand, since \( f_0 \in \mathcal{H} \) and \( \lambda_\nu \asymp \nu^{-2m} \), we know \( \sum_{\varphi \in V ||\varphi||_{min} \geq t+1} (\theta_\varphi^0)^2 \asymp n^{-2m} \). This is the truncation error due to \( \hat{\theta}_\varphi = 0 \) for \( \nu_k \geq l+1, 1 \leq k \leq d \). On the other hand, note that \( \langle \varphi, \nu \rangle_n \leq 1 \) and then
\[
\left( \sum_{\varphi \in V ||\varphi||_{min} \geq t+1} \theta_\varphi^0 \varphi, \nu \right)_n^2 \leq \sum_{\varphi \in V ||\varphi||_{min} \geq t+1} (\theta_\varphi^0)^2 \asymp n^{-2m}.
\]
Thus,
\[
\sum_{\tilde{\nu} \in \mathcal{V}, \|\tilde{\nu}\|_{\min} \leq l} \left( \mathbb{E}\tilde{\theta}_{\tilde{\nu}} - \theta^0_{\tilde{\nu}} \right)^2
\]
\[
\leq \lambda^2 \sup_{\tilde{\nu} \in \mathcal{V}, \|\tilde{\nu}\|_{\min} \leq l} \lambda_{\tilde{\nu}}^{-1} \left( \frac{\sigma_0^{-2} + \sum p_j \sigma_j^{-2} \nu_j^2 + \lambda \lambda_{\tilde{\nu}}^{-1}}{\lambda_{\tilde{\nu}}^{-1}} \right) \leq \sum_{\tilde{\nu} \in \mathcal{V}, \|\tilde{\nu}\|_{\min} \leq l} \lambda_{\tilde{\nu}}^{-1} (\theta^0_{\tilde{\nu}})^2
\]
\[+ o(n^{-1}) \sum_{\tilde{\nu} \in \mathcal{V}, \|\tilde{\nu}\|_{\min} \leq l} \frac{1 + \sum j \nu_j^2}{(1 + \sum j \nu_j^2 + \lambda \nu_1^2 \cdots \nu_d^2)^2} + n^{-2m+1} \]
\[
\leq \lambda^2 J(f_0) \sup_{\tilde{\nu} \in \mathcal{V}} \frac{\nu_1^{2m} \cdots \nu_d^{2m}}{(1 + \sum j \nu_j^2 + \lambda \nu_1^2 \cdots \nu_d^2)^2} + o(n^{-1} \lambda^{-1/2m}) + n^{-2m+1},
\]
where the last step uses Lemma 12 with \( a = 0 \) and \( p = d - 1 \). Define that
\[
B_\lambda(\tilde{\nu}) = \frac{\nu_1^{2m} \cdots \nu_d^{2m}}{(1 + \sum j \nu_j^2 + \lambda \nu_1^2 \cdots \nu_d^2)^2}.
\]
For the \( \sup_{\tilde{\nu} \in \mathcal{V}} B_\lambda(\tilde{\nu}) \) term above, suppose that \( \prod_{j=1}^d \nu_j^{2m} > 0 \) is fixed and denoted by \( x^{-1} \), then \( B_\lambda(\tilde{\nu}) \) is maximized by letting \( \sum j \nu_j^2 \) be as small as possible, where \( p = d - 1 \). This suggests \( \nu_1 = \nu_2 = \cdots = \nu_p = 1 \), and
\[
\sup_{\tilde{\nu} \in \mathcal{V}} B_\lambda(\tilde{\nu}) \asymp \sup_{x > 0} \frac{x^{-1}}{(1 + \lambda x^{-1})^2} \asymp \lambda^{-1},
\]
where the last step is achieved when \( x \asymp \lambda \). Combining all parts of bias gives
\[
\sum_{\tilde{\nu} \in \mathcal{V}} \left( \mathbb{E}\tilde{\theta}_{\tilde{\nu}} - \theta^0_{\tilde{\nu}} \right)^2 = O \{ \lambda J(f_0) + n^{-2m+1} \} + o(n^{-1} \lambda^{-1/2m}). \tag{86}
\]
The constant factor on the upper bound does not depend on \( n \).

The stochastic error is bounded as follows:
\[
\sum_{\tilde{\nu} \in \mathcal{V}} \mathbb{E} \left( \tilde{\theta}_{\tilde{\nu}} - \mathbb{E}\tilde{\theta}_{\tilde{\nu}} \right)^2 \leq \sum_{\tilde{\nu} \in \mathcal{V}, \|\tilde{\nu}\|_{\min} \leq l} \frac{n^{-1}(\sigma_0^{-2} + \sum j \sigma_j^{-2} \nu_j^2)}{(\sigma_0^{-2} + \sum j \sigma_j^{-2} \nu_j^2 + \lambda \lambda_{\tilde{\nu}}^{-1})^2}
\]
\[
\leq \sum_{\tilde{\nu} \in \mathcal{V}, \|\tilde{\nu}\|_{\min} \leq l} \frac{1 + \sum j \nu_j^2}{n(1 + \sum j \nu_j^2 + \lambda \nu_1^2 \cdots \nu_d^2)^2}.
\]
Using Lemma 12 with $a = 0$ and $p = d - 1$, we have

$$\sum_{\tilde{\nu} \in \mathcal{V}} \mathbb{E} \left( \tilde{\theta}_{\tilde{\nu}} - \mathbb{E} \tilde{\theta}_{\tilde{\nu}} \right)^2 = O \left\{ n^{-1} \lambda^{-1/2m} \right\}. \tag{87}$$

Combining (86) and (87) and letting $\lambda \asymp n^{-2m/(2m+1)}$ completes the proof.

### E.3 Auxiliary Lemmas for Theorems 4 and 5

**Lemma 16.** Suppose that $s \geq 1$, $\beta \geq 0$ and $\beta \neq 1$, and $r \geq 1$. Then as $\Xi \to \infty$,

$$\int_{x_1 \cdots x_r \cdot z \leq \Xi, x_k \geq 1, z \geq 1} x_1^\beta \cdots x_r^\beta (log z)^s (x_1^2 + \cdots + x_r^2)^{-1} dx_1 \cdots dx_r dz \asymp \Xi^{\beta+1} (log \Xi)^s.$$

**Proof.** For any $\tau \geq 1$, we have

$$\{1 \leq z \leq \Xi r^{-r}, 1 \leq x_k \leq \tau, k = 1, \ldots, r\} \subset \{x_1 \cdots x_r \cdot z \leq \Xi, z \geq 1, x_k \geq 1, k = 1, \ldots, r\}.$$

Thus, if $\Xi \to \infty$,

$$\int_{x_1 \cdots x_r \cdot z \leq \Xi, x_k \geq 1, z \geq 1} x_1^\beta \cdots x_r^\beta (log z)^s (x_1^2 + \cdots + x_r^2)^{-1} dx_1 \cdots dx_r dz$$

$$\geq \int_{1}^{\Xi r^{-r}} \int_{1}^{\tau} \cdots \int_{1}^{\tau} z^\beta (log z)^s x_1^{\beta-2} \cdots x_r^{\beta-2} dx_1 \cdots dx_r dz$$

$$\asymp \Xi^{\beta+1} r^{-r(\beta+1)} (log \Xi - r \log \tau)^s r^{\beta-1}.$$

Let $\tau \to 1$, we have $\int_{x_1 \cdots x_r \cdot z \leq \Xi, x_k \geq 1, z \geq 1} (log z)^s (x_1^2 + \cdots + x_r^2)^{-1} dx_1 \cdots dx_r dz \gtrsim \Xi^{\beta+1} (log \Xi)^s$.

On the other hand, define $u = x_1 \cdots x_r \cdot z$ and change the variable $z$ to $u$. We have that as $\Xi \to \infty$,

$$\int_{x_1 \cdots x_r \cdot z \leq \Xi, x_k \geq 1, z \geq 1} x_1^\beta \cdots x_r^\beta (log z)^s (x_1^2 + \cdots + x_r^2)^{-1} dx_1 \cdots dx_r dz$$

$$= \int_{1}^{\Xi} \int_{1}^{u/x_r} \cdots \int_{1}^{u/(x_r \cdots x_2)} u^\beta (log u - log x_r - \cdots - log x_1)^s$$

$$\cdot (x_1^2 + \cdots + x_r^2 + x_r^2 - 1) x_1^{1-2/r} \cdots x_r^{1-2/r} dx_1 \cdots dx_r dx_r du$$

$$\lesssim \int_{1}^{\Xi} \int_{1}^{u/x_r} \cdots \int_{1}^{u/(x_r \cdots x_2)} u^\beta (log u - log x_r - \cdots - log x_1)^s$$

$$\cdot x_1^{1-2/r} \cdots x_r^{1-2/r} dx_1 \cdots dx_r dx_r du$$

$$\lesssim \int_{1}^{\Xi} u^\beta (log u)^s du \asymp \Xi^{\beta+1} (log \Xi)^s.$$