Electronic Journal of Statistics
Vol. 17 (2023) 3660–3727
ISSN: 1935-7524
https://doi.org/10.1214/23-EJS2188

Regression in tensor product spaces by the method of sieves

Tianyu Zhang and Noah Simon

Department of Biostatistics, University of Washington, Seattle, USA
e-mail: zty@uw.edu; nrsimon@uw.edu

Abstract: Estimation of a conditional mean (linking a set of features to an outcome of interest) is a fundamental statistical task. While there is an appeal to flexible nonparametric procedures, effective estimation in many classical nonparametric function spaces, e.g., multivariate Sobolev spaces, can be prohibitively difficult – both statistically and computationally – especially when the number of features is large. In this paper, we present some sieve estimators for regression in multivariate product spaces. We take Sobolev-type smoothness spaces as an example, though our general framework can be applied to many reproducing kernel Hilbert spaces. These spaces are more amenable to multivariate regression, and allow us to, in-part, avoid the curse of dimensionality. Our estimator can be easily applied to multivariate nonparametric problems and has appealing statistical and computational properties. Moreover, it can effectively leverage additional structure such as feature sparsity.

MSC2020 subject classifications: Primary 62G05, 62G08.
Keywords and phrases: Multivariate regression, sparse nonparametric models, orthonormal basis.

Received January 2023.

1. Introduction

Understanding the relationship between an outcome of interest and a set of predictive features is an important topic across domains of scientific research. To this end, one often needs to estimate an underlying predictive function, e.g., the conditional mean function, that best relates the features and the outcome using available noisy observations. During the past two decades, there has been extensive research focusing on nonparametric learning methods that only require the outcome to vary smoothly with the features.

One challenge of applying nonparametric methods in multivariate problems is the "curse of dimensionality" [34]. Briefly, as the number of features grows linearly, we need an exponentially growing number of samples to achieve a specified threshold of predictive performance. In real-world applications, although the total number of candidate features may be large, it is very likely that only a small proportion are conditionally associated with the outcome. This smaller number, D, of *active features* should be the primary driver of the difficulty of the

arXiv: 2206.02994

problem, in a minimax sense. Sparse estimation [21, 48] is a vast field addressing such data science problems and developing effective estimation procedures, which is especially interesting when the total number of features, d, is much larger than D.

In this paper, we consider nonparametric procedures that can simultaneously select important features and estimate the conditional mean function (using only those selected features). For this procedure, the estimation error scales favorably with total dimension (proportional to $\log(d)$). Moreover, engaging with a tensor product space additionally means that our active dimension, D, only shows up multiplicatively in a $\log^D(n)$ term (as compared to modifying the rate of convergence in n in classical multivariate Sobolev/Holder spaces). Finally, our proposed framework is also seen to be empirically effective in our data example comparisons in Section 8.

The proposed method considers (penalized) sieve estimation in multivariate tensor product spaces. Sieve estimation, also known as projection estimation [49] or estimation using orthogonal series [59], is a classical estimation strategy that has been shown to be very effective in univariate regression problems. Although sieve-type estimation has a long history and, in many cases, deep theoretical exploration, the literature does not offer effective guidance on how to apply this method for moderate dimensional problems in practice. Instead, classical sieve estimation is not generally considered fruitful in more than a few (maybe 1-2) dimensions, as statistical convergence gets horrible without extremely high order smoothness assumptions (e.g., in the classical smoothness class we discuss further in Section 3.1). A key, and novel, aspect of this work is our focus on tensor product spaces: We show that this allows sieve estimation to be a tractable option even if only the existence of first order mixed partial derivatives is assumed. For example, if we aim to estimate nuisance parameters for causal effects, e.g. when using AIPW (augmented inverse propensity weighted estimation) [15, 25], tensor product spaces are more relevant than classical models. In particular, if we only assume existence of first order partials, the classical framework [43] can accommodate dimension up to 2, whereas the work in this manuscript can accommodate arbitrary fixed dimension.

In addition to the above statistical difficulty, there is also computational difficulty: The most popular multivariate extension of sieve estimators, which is presented in (8), implicates the use of $n^{\alpha d}$ basis functions with $\alpha \in (0, 1/2)$ (depending on the smoothness of the problem). In contrast, our proposal uses, essentially, on the order of $C(D)d^D n^{\alpha}(\log n)^D$ basis functions (Here $C(D)$ is a constant that depends on D but not sample size). This is much more computationally attractive. This work can be seen as an attempt to extend the method of sieves toward multivariate models that scale more efficiently, both statistically and computationally, with dimension.

When engaging with multivariate sieve estimators it is critically important to identify an ordering of multivariate basis elements from "most" to "least" important. In addition, one must identify how many basis functions to include to get an estimator with suitably low misspecification bias (this depends on the smoothness of the space). In univariate problems, there is usually a nat-

ural ordering, based on, e.g., frequency or polynomial degree. Extending this to multivariate settings is not as simple: We studied the spectrum of certain "covariance operators" to identify the appropriate strategy for ordering multivariate basis functions in tensor product models. This is critical for both method implementation and theoretical understanding: For multivariate problems, the "ordering question" has received little discussion in the literature.

The main contributions of this paper can be summarized as follows:

- We propose a rigorous extension of sieve estimators to multivariate problems. To the best of our knowledge, this is the first methodological treatment other than the (naive) direct extension repeatedly appearing in the literature. The direct extension is not computationally feasible even with moderate feature dimension. In contrast, the proposal in this manuscript is much more computationally tractable, and we provide theoretical predictive performance guarantees that scale favorably with the feature dimension.
- After identifying the proper ordering of multivariate basis functions, we give a direct and explicit implementation of the sieve estimator. In addition, we give a transparent result for computational expense even in the "large d, small n" genuine multivariate case. We demonstrate the effectiveness of (penalized) sieve estimation here with both theoretical guarantees and extensive simulation studies.
- We engage a relatively basic result in number theory ("the average order of divisor functions") to reduce a multivariate "Sobolev ellipsoid" to a (formally) univariate ellipsoid. This technique can also be applied to quantify the asymptotic eigenvalues of multivariate reproducing kernels. We believe it may be of independent interest and can be widely applied in reducing multivariate nonparametric problems to (formally) univariate ones.

Notation. In this paper, we will use bold letters to emphasize a Euclidean vector $\mathbf{x} \in \mathbb{R}^d$ when its dimension d is strictly greater than 1. The notation $\mathbf{x}^k \in \mathbb{R}$ is the k-th entry of $\mathbf{x} \in \mathbb{R}^d$ (rather the k-th power of it). We use $\mathbb{N}$ to represent the non-negative integer set $\{0, 1, 2, \ldots\}$, and use $\mathbb{N}^+$ for strictly positive integers $\{1, 2, 3, \ldots\}$. The $(\mathbb{N}^+)^d$ is the set of positive d-tuple grids: for example $(\mathbb{N}^+)^2 = \{(1,1), (1,2), (2,1), (3,1), (2,2), \ldots\}$.

2. Univariate nonparametric problems with sieve estimation

One can frame the goal of regression as estimating the function f that minimizes the population mean-squared error (MSE): $E[(Y - f(\mathbf{X}))^2]$, where Y is our outcome of interest, and $\mathbf{X}$ are our predictive features. We denote the distribution of $\mathbf{X}$ as ρ_X. The minimizer is the well-known condition mean function $f^0(\mathbf{X}) = E[Y|\mathbf{X}]$. In nonparametric regression, we assume f^0 belongs to some regular function space. An informative univariate model space that we will

engage with is the 1^{st}-order Sobolev space $W_1([0,1])$:

$$f^0 \in W_1\big([0,1]\big) = \big\{ f \in L_2\big([0,1]\big) \;\mid\; f' \text{ exists and } f' \in L_2\big([0,1]\big) \big\}. \tag{1}$$

Here f' can be understood as the weak derivative of f. In this framing, the set of piece-wise linear functions is a subset of $W_1([0,1])$. Without loss of generality, we will assume feature $\mathbf{X}$ belongs to the d-dimensional unit cube $[0,1]^d$. Sieve estimation for f^0 in the W_1 space is built upon the following basic fact: It is possible to express f^0 as an infinite linear combination of some basis functions $\{\phi_j\}$. Among many possibilities, we choose the following function system as a concrete example:

$$\phi_1(x) = 1, \phi_j(x) = \sqrt{2}\cos\big((j-1)\pi x\big). \tag{2}$$

The aforementioned "infinite linear combination" can be expressed as: $f^0 = \sum_{j=1}^{\infty} \beta_j^0 \phi_j$. Moreover, it is also known that the (generalized) Fourier coefficients β_j^0 decay at a rate faster than $j^{-1.5}$ for $f^0 \in W_1([0,1])$. Therefore, it is plausible to truncate the infinite series at a certain finite level J_n: Using only the first more important J_n basis vectors, one can construct an estimator of f^0 with relatively small bias. Formally, a sieve estimator f_n takes the form that $\hat{f}_n = \sum_{j=1}^{J_n} \hat{\beta}_j \phi_j$ where the coefficients are determined using the available training data $\{(\mathbf{X}_i, Y_i), i = 1, \ldots, n\}$. The coefficients can be determined by solving least-square problems [49] or using stochastic approximation methods [64], both strategies would lead to rate-optimal generalization error (in a minimax-rate sense).

Remark 2.1. The cosine functions ϕ_j presented above are not periodic over our domain themselves, and thus do not impose a periodic assumption on f^0. This is in contrast to periodic sine/cosine systems that are more commonly engaged with, and would imply a periodic assumption on f^0 [56]. One can add polynomials to the periodic systems to fit non-period functions [10]. For simplicity of exposition and to provide our readers a basis that is easy to implement, we choose to proceed with paper primarily using this cosine basis. For readers more familiar with the topic, the above rate statement on β_j^0 ("faster than $j^{-1.5}$") can be more precisely stated as Sobolev ellipsoid conditions. For more discussion, see Appendix B.

3. Multivariate nonparametric models

3.1. Additive models and classical smoothness classes

In most real-world problems, we have more than one feature under consideration. In addition it is not always apriori clear which model space to use. The nonparametric additive model [13] has been seen as one of the most direct models for multivariate nonparametric learning problems. There, we assume features

do not interact, or more formally that the regression function takes the following additive form:

$$f^0(\mathbf{x}) = \sum_{k=1}^{d} f_k^0\big(\mathbf{x}^k\big), \quad f_k^0 \in W_1\big([0,1]\big).$$

There are also some more flexible models widely discussed in the literature, such as Sobolev-type smooth function spaces. Formally, let $\mathbf{a} = (\mathbf{a}^1, \ldots, \mathbf{a}^d) \in (\mathbb{N})^d$, we define the (weak) partial derivative function $D^{\mathbf{a}} f$ of f as:

$$D^{\mathbf{a}} f = \frac{\partial^{\|\mathbf{a}\|_1}}{\partial x_1^{\mathbf{a}^1} \cdots \partial x_d^{\mathbf{a}^d}} f, \text{ where } \|\mathbf{a}\|_1 = \sum_{k=1}^{d} \mathbf{a}^k.$$

In this notation, people may assume that f^0 satisfies the following smoothness conditions:

$$f^0 \in W_s\big([0,1]^d\big) = \big\{ f \in L_2\big([0,1]^d\big) \mid D^{\mathbf{a}} f \in L_2\big([0,1]^d\big) \text{ for all } \|\mathbf{a}\|_1 \le s \big\}. \quad (3)$$

These types of smooth classes do not explicitly assume any specific form such as additivity, but as a cost, suffer substantially more from the curse of dimensionality. Specifically, the minimax rate (in MSE) of estimation in $W_s([0,1]^d)$ is of order $n^{-2s/(2s+d)}$ [43]. Although less likely to be miss-specified, this type of model is sometimes thought to be too large to explain the success of many machine learning methods, or be directly applied.

In the literature it is typical to put a more strict regularity requirement to cancel out the influence of dimension, that is, only considering smoother models in higher dimensions. Formally, this can be easily done by increasing the parameter s to ask for regular higher-order derivatives. For many statistical procedures that need to estimate conditional mean as a nuisance, e.g. semiparametric inference [25] and independence structure inference [61], we typically have to require the smoothness parameter s to be at least $d/2$. The resulting model space $W_{d/2}([0,1]^d)$ is sufficiently tame to allow estimators of f^0 that can satisfy certain (minimax rate) benchmark conditions. However, for d as small as 4, such a requirement already prevents f^0 from being a piece-wise linear function.

3.2. Tensor product models

Additive models (mentioned earlier) are an attractive approach for extending univariate smooth functions to multivariate regression. If the true regression function is nearly additive, then with a relatively small number of samples, one can fit a strong additive estimate. However, in some applications there may be important interactions between features to consider. One natural extension to the additive model is to include product-terms of basis functions between individual features. For example, we may consider:

$$f^0(\mathbf{x}) = \sum_{k=1}^{d} f_k^0\big(\mathbf{x}^k\big) + a\big(\mathbf{x}^1\big)b\big(\mathbf{x}^2\big) + c\big(\mathbf{x}^1\big)d\big(\mathbf{x}^3\big) + e\big(\mathbf{x}^1\big)f\big(\mathbf{x}^2\big)g\big(\mathbf{x}^3\big) + \cdots, \quad (4)$$

where all the univariate functions above belong to a smooth function class such as $W_1([0,1])$. This type of model has been studied in the literature as a *Tensor Product Space models* [29]. In a more compact notation:

$$f^0 \in \left\{ f = \sum_{m=1}^{N} \prod_{k=1}^{d} f_{mk}(\mathbf{x}^k) \text{ with finite } N, \text{ and } f_{mk} \in W_1\big([0,1]\big) \right\}. \tag{5}$$

Although we defined the tensor product space (5) by addition and multiplication of univariate regular functions (algebraic manipulations), there is an almost[1] equivalent characterization of it in the language of partial derivatives:

$$\begin{aligned} f^0 &\in S_1\big([0,1]^d\big) \\ &:= \big\{ f \in L_2\big([0,1]^d\big) \;\mid\; D^{\mathbf{a}} f \in L_2\big([0,1]^d\big) \text{ for all } \|\mathbf{a}\|_\infty \le 1 \big\}. \end{aligned} \tag{6}$$

Function spaces similar to (6) are called Sobolev spaces with dominating mixed derivatives. They are also characterized as the tensor product spaces of univariate Sobolev spaces $W_1([0,1])$. Compared with the (isotropic) Sobolev spaces defined in (3), tensor product spaces may appear to be formally similar, but have different (and favorable) properties related to statistical estimation. For function space $W_1([0,1]^d)$, we required regular partial derivatives for any index $\mathbf{a}$ satisfying $\|\mathbf{a}\|_1 \le 1$. But for tensor product space $S_1([0,1]^d)$, we require partial derivatives for those indices satisfying $\|\mathbf{a}\|_\infty \le 1$. The latter requirement is strictly stronger and as the dimension d increases, the difference between these two requirements becomes more meaningful. At the same time, the $S_1([0,1]^d)$ space requires less regularity than the d-th order isotropic Sobolev space $W_d([0,1]^d)$. In particular, assuming $f^0 \in W_d([0,1]^d)$ means that $\frac{\partial^d}{\partial^d \mathbf{x}^k} f^0$ exists and is square-integrable for any $k = 1, 2, \ldots, d$, however functions in $S_1([0,1])$ space do not need to have second partial derivatives $\frac{\partial^2}{\partial^2 \mathbf{x}^k} f$ for any k (so piece-wise linear functions can be elements of $S_1([0,1]^d)$). More formally, we have the following inclusion relationship:

$$W_d\big([0,1]^d\big) \subsetneq S_1\big([0,1]^d\big) \subsetneq W_1\big([0,1]^d\big). \tag{7}$$

The space S_1 can be generalized to function spaces with stronger smoothness restrictions by replacing the restriction $\|\mathbf{a}\|_\infty \le 1$ by $\|\mathbf{a}\|_\infty \le s$ for some $s > 1$. However, we choose not to pursue this generalization in this paper, as it would make the exposition and notation unnecessarily more complicated. We refer the interested reader to Section 9 for more discussion.

4. Literature review

In this section, we will provide a quick overview of the literature on tensor product models in statistical learning and nonparametric sieve estimators.

[1] Space $S_1([0,1]^d)$ contains finite linear combination of functions as in (5) as well as their limits with respect to a certain norm ($N = \infty$). This is in line with a reproducing kernel Hilbert space contains both the finite and infinite linear combination of the kernel functions.

In [29], the author presents regression estimators in *tensor product models* by the method of smoothing spline/kernel ridge regression. The estimators achieve the nonparametric minimax rate but typically have a high computational expense when directly implemented. Compared with the proposal in this work, it has a limited ability to perform variable selection and is shown to be adaptive to the active dimension D. Other work in this line of research includes Wahba et al. [57], Lin et al. [28] and Gao et al. [14].

In addition to using product reproducing kernels, other types of product bases are also used to construct multivariate regression estimators. For example, there are multivariate adaptive regression spline [12] and the highly adaptive lasso [2]. This class of methods select a collection of adaptive basis functions that center on the training data points. The set of basis functions, unlike the sieve estimator basis, are usually not orthogonal to each other under any natural measures. More comprehensive discussion can be found in the monograph [17].

A lot of work has been done over the last decade to adapt the tensor product model to ultra-high dimensional settings. This line of research typically assumes that the features must have a main effect on the outcome in order to have second-order interaction effects (formalized as some heredity assumptions). These methods target application cases when the feature dimension is very large and computational resources are restricted (For example, assuming d^2 derived features would not fit into the memory). See Haris et al. [20], Tan [44], and the references therein for a more detailed description of these computationally efficient methods.

In contrast to the kernel or spline-based methods, in this paper, we will discuss how to apply sieve estimators in tensor product models. In [49], the author presents the classical least-square sieve estimator (termed as a projection estimator) with theoretical discussion (many parts in our exposition will be of that flavor). In [4], the author provides an extensive review of commonly used/theoretically interesting sieve basis. Efromovich [9] provides an extensive review of the method of sieves in density estimation. See also Section 7.5 of Efromovich [8] for a discussion of sieve estimation for multivariate analytic functions. In [22], the authors discuss estimation with orthogonal series under additive models. However, there is no existing work that formally engages with tractable sieve estimation procedures under tensor product models to the best knowledge of the authors. In contrast, it has been repeatedly discussed in the literature to directly generalize univariate sieve estimators to multivariate settings with estimators of the form (e.g. here we take the dimension $d = 3$)

$$\hat{f}_n\big(\mathbf{x}^1, \mathbf{x}^2, \mathbf{x}^3\big) = \sum_{i=1}^{J_n} \sum_{j=1}^{J_n} \sum_{k=1}^{J_n} \theta_{ijk} \phi_i\big(\mathbf{x}^1\big) \phi_j\big(\mathbf{x}^2\big) \phi_k\big(\mathbf{x}^3\big). \tag{8}$$

This kind of direct extension does not lead to rate-optimal estimators in commonly discussed function classes and is not computationally scalable to even moderate dimension d in practice (the number of basis functions increases exponentially with respect to d).

The study of properties of (Sobolev) spaces with mixed derivatives [55, 36, 26, 31, 45] and related numerical problems [3] is an active field of mathematics. Applicable numerical methods in these fields are usually called "sparse grids" [39] or hyperbolic cross [38, 7, 46]. The work in this manuscript connects to those ideas but also engages statistical and computational questions.

5. Least-square sieve estimators

In this section, we will discuss ordinary least-square sieve estimators that is applicable to moderate-dimensional problems. Discussion of this kind of preliminary estimators may be of interest itself and will pave our road to the more practical proposal presented in Section 7.

Sieve estimation leverages the fact that smooth functions can be written as an infinite linear combination of some basis functions whose coefficients decay quickly. To construct estimates, we can use a truncated series to balance the approximation and estimation errors. Since functions in $S_1([0,1]^d)$ can be approximately written as the addition and multiplication of a set of univariate functions in $W_1([0,1])$, we may expect a function $f \in S_1([0,1]^d)$ to have the expansion

$$f^0(\mathbf{x}) = \sum_{\mathbf{j}\in(\mathbb{N}^+)^d} \beta^0_{\mathbf{j}} \psi_{\mathbf{j}}(\mathbf{x}), \text{ for some } \beta^0_{\mathbf{j}} \in \mathbb{R},$$

where $\mathbf{j} = (\mathbf{j}^1, \mathbf{j}^2, \ldots, \mathbf{j}^d) \in (\mathbb{N}^+)^d$, and $\psi_{\mathbf{j}}$ is a product of the univariate cosine basis $\psi_{\mathbf{j}}(\mathbf{x}) = \prod_{k=1}^d \phi_{\mathbf{j}^k}(\mathbf{x}^k)$ described in (2).

In contrast to the univariate case, there is no single obvious natural ordering of the basis functions $\psi_{\mathbf{j}}$ since they are indexed by some d-tuples $\mathbf{j}$. Recall that in the univariate case, the basis functions are naturally ordered by their trigonometric frequency. To apply sieve estimation in tensor product spaces (or for any multivariate nonparametric models), we need to establish an order on $\{\psi_{\mathbf{j}}\}$ and determine which basis functions should be used first. In other words, we need to unravel the *set* $\{\psi_{\mathbf{j}}, \mathbf{j} \in (\mathbb{N}^+)^d\}$ to a *sequence* of functions $\{\psi_j, j \in \mathbb{N}^+\}$. They contain the same set of functions but the latter is an ordered sequence.

Let (ψ_j) be the sequence of functions unravelled from $\{\psi_{\mathbf{j}}\}$ (we postpone the details of the rearrangement rule to Section 6). In the new notation, any $f^0 \in S_1([0,1]^d)$ has the expansion $f^0(\mathbf{x}) = \sum_{j=1}^\infty \beta^0_j \psi_j(\mathbf{x})$, $\beta^0_j \in \mathbb{R}$. To perform sieve estimation in $S_1([0,1]^d)$, we also truncate the series at a proper level J_n. The least-square sieve estimator f_n^{OLS} is $f_n^{OLS}(\mathbf{x}) = \sum_{j=1}^{J_n} \beta^{OLS}_{jn} \psi_j(\mathbf{x})$, whose coefficients are the minimizers of the following empirical least-squares problem:

$$\left(\beta^{OLS}_{1n}, \ldots, \beta^{OLS}_{J_n n}\right) = \underset{(\beta_1,\ldots,\beta_{J_n})\in\mathbb{R}^{J_n}}{\operatorname{argmin}} \sum_{i=1}^n \left\{ Y_i - \sum_{j=1}^{J_n} \beta_j \psi_j(\mathbf{X}_i) \right\}^2. \tag{9}$$

Using analysis tools from empirical process theory, it is possible to derive some theoretical guarantees regarding the performance of f_n^{OLS}.

Theorem 5.1. *Suppose* $\{(\mathbf{X}_i, Y_i) \in [0,1]^d \times \mathbb{R}, i = 1, 2, \ldots, n\}$ *is an independent and identically distributed (i.i.d.) training sample and the true regression function* $f^0 \in S_1([0,1]^d)$*, formally*

$$\sum_{\|\mathbf{a}\|_\infty \le 1} \left\| D^{\mathbf{a}} f^0 \right\|^2_{L_2([0,1]^d)} \le Q^2.$$

Let $\epsilon_i = Y_i - f^0(\mathbf{X}_i)$ *be sub-Gaussian, mean-zero random variables. We further assume that the distribution of* $\mathbf{X}$*,* ρ_X*, is continuous with an upper-bounded density function.*

Then, for the least-square sieve estimator f_n^{OLS}*, constructed with product of cosine basis functions* (2)*, we have:*

$$\left\| f_n^{OLS} - f^0 \right\|^2_{2,\rho_X} = O_P\left(\left(\frac{\log^{d-1}(n)}{n} \right)^{2/3} \log(n) \right), \tag{10}$$

when $J_n = \Theta(n^{1/3} \log^{2(d-1)/3}(n))$*. The ordering of the multivariate product basis is described in detail in Section 6.*

We present the proof of Theorem 5.1 in Appendix D. The overall proof structure for the least-square estimator is similar to that of Theorem 1 in [65]. However, to determine the proper truncation level J_n and approximation error, we need the new technical results presented in Lemma C.7. The above theoretical guarantee is almost minimax-optimal [29], up to a logarithm term. Specifically, when d is a given fixed number, the minimax-rate of estimation in S_1 is $(n^{-1} \log^{d-1}(n))^{2/3}$.

The generalization MSE of this least-squares sieve estimator only differs from $n^{-2/3}$—the rate for univariate Sobolev space $W_1([0,1])$—by a polylog term (with the dimension d in the exponent). This is much improved as compared with estimation in spaces such as $W_s([0,1]^d)$. For that classical space, the minimax rate is of order $n^{-2s/(2s+d)}$. The dimension d shows up in the exponent of n rather than $\log n$. That horrible dependence on the dimension is one manifestation of the curse of dimensionality. It is much alleviated but still exists, in tensor product spaces. Many semiparametric procedures require convergence of intermediate components at a rate of at least $n^{-1/2}$ [25]. Classical Sobolev models must assume $s \ge d/2$ to give such a guarantee. This requirement may be too strong for many applications: specifically, it already rules out all the piece-wise linear truths when $d \ge 4$.

Remark 5.2 (Metric entropy comparison)**.** From a learning theory perspective, regression problems are of different degrees of difficulty is due to the difference in the metric entropy of the hypothesis spaces. Let $N(\delta, \mathcal{F})$ denote the δ-covering numbers of function space $\mathcal{F}$. Then for the unit balls in W_1, S_1, W_d spaces, we have

$$\begin{aligned} \log N\big(\epsilon, W_1([0,1]^d)\big) &\asymp \epsilon^{-d}, \\ \log N\big(\epsilon, S_1([0,1]^d)\big) &\asymp \epsilon^{-1} \log^{d-1}(1/\epsilon), \\ \log N\big(\epsilon, W_d([0,1]^d)\big) &\asymp \epsilon^{-1}. \end{aligned} \tag{11}$$

(To clarify, the δ-covering is defined with respect to $\|\cdot\|_{2,\rho_X}$-norm and the "unit balls" aforementioned are defined using their corresponding Sobolev-norms.) This shows that the S_1 space is a slightly richer function class than W_d. Note that W_d is also the largest classical Sobolev space that is a strict subset of S_1. The metric entropy results for W spaces are known in the literature (e.g. Proposition 6, page 15, [5], Example 5.12 of [58]). We derive the metric entropy of the S_1 space in Proposition C.8.

Remark 5.3 (Minimal number of basis functions)**.** The least-square estimator f_n^{OLS} constructed with $J_n = n^{1/3}\log^{2(d-1)/3}(n)$ basis functions uses the minimal basis number among all the "linear" estimators that essentially achieve the minimax rate. That is, there are no other sets of pre-specified-functions $\{\zeta_j\}$ and estimators of form $f_n^{not} = \sum_{j=1}^{J_n^{not}} \beta_j \zeta_j$ that can achieve the minimax-rate with $J_n^{not} \ll J_n$. One way to see this is by examining the metric entropy $\log N(\cdot)$ of S_1 space. The magnitude of metric entropy $\log N(\epsilon, S_1)$ characterizes the minimal number of digits required to specify every function in (a ball in) S_1 up to ϵ-accurately. Let ϵ be the root-MSE minimax-rate, $n^{-1/3}\log^{(d-1)/3} n$: plugging it into the entropy magnitude (11), we know that there are at least $n^{1/3}\log^{2(d-1)/3} n$ digits required to specify every function in S_1 to this accuracy. The least-square estimator f_n^{OLS} records the coefficients of J_n-many basis functions. Assuming we use a constant number of bits for each coefficient (32 or 64), the total number of bits of this estimator is the same order as the minimal requirement (but f_n^{OLS} achieves a slightly worse convergence rate than the minimax limit). If there indeed were some f_n^{not} that used significantly fewer pre-specified basis functions, it would lead to a contradiction with the metric entropy limit.

6. Important technical details: unravelling

In this section, we are going to talk about how to rearrange a set of functions $\{\psi_{\mathbf{j}}\}$ indexed by d-tuples to a sequence of functions (ψ_j). For ease of discussion, we will term this kind of rearrangement process as *unravelling*. Now we present our proposed unravelling rule for tensor product models.

In Fig. 1, we present how to rearrange 2-tuples $(\mathbb{N}^+)^2$ into a sequence (this corresponds to the statistical question when $d = 2$). We first assign a number $c_{\mathbf{j}}$ to each grid element $\mathbf{j} \in (\mathbb{N}^+)^2$ that equals to the elemental product $c_{\mathbf{j}} = \mathbf{j}^1 \cdot \mathbf{j}^2$. We then rearrange the 2-tuples on the left based on the product value $c_{\mathbf{j}}$ in increasing order. In the right panel of Fig. 1, we can see the tuples assigned with smaller $c_{\mathbf{j}}$ values get a more prioritized position in the sequence indexed by $j \in \mathbb{N}^+$. For example, $(1,1)$ is mapped to the first element on the right because it has the smallest product. In contrast, $(2,2)$ gets the 7-th position because there are 6 tuples having product less or equal to it. For tuples with the same $c_{\mathbf{j}}$ values (such as $(1,2)$ and $(2,1)$), their relative order can be defined arbitrarily. We put $(2,1)$ in front of $(1,2)$ because it has a larger value in the first dimension. In many parts of the analysis, we are interested in how fast the

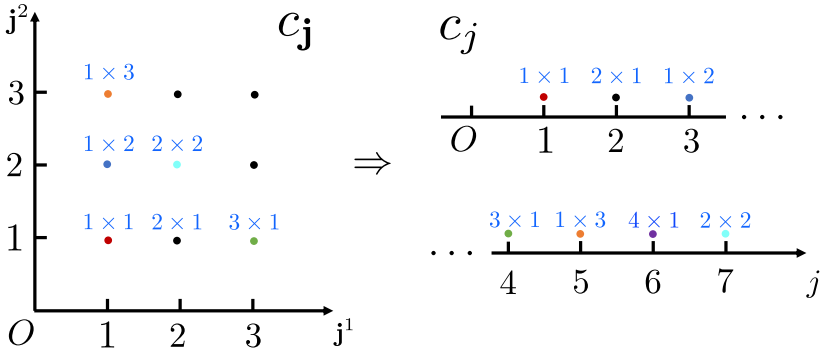

FIG 1. *Illustration of unravelling. The unravelling rule function is* $c_{\mathbf{j}}^{CPR} = \prod_{k=1}^{d} \mathbf{j}^k$.

unravelled sequence (c_j) diverges with respect to j (the series presented in the right panel of Fig. 1).

The unravelling rule described above can also be to applied to rearrange other objects originally indexed by 2-tuples (for example, the basis functions $\{\psi_{\mathbf{j}}\}$). Using the unravelling rule presented in Fig. 1, the first several basis functions in the unravelled basis sequence are,

$$\begin{aligned}
\psi_1(\mathbf{x}) &= \psi_{(1,1)}(\mathbf{x}) = \phi_1(\mathbf{x}^1)\phi_1(\mathbf{x}^2) = 1\\
\psi_2(\mathbf{x}) &= \psi_{(2,1)}(\mathbf{x}) = \phi_2(\mathbf{x}^1)\phi_1(\mathbf{x}^2) = \sqrt{2}\cos(\pi\mathbf{x}^1)\\
\psi_3(\mathbf{x}) &= \psi_{(1,2)}(\mathbf{x}) = \phi_1(\mathbf{x}^1)\phi_2(\mathbf{x}^2) = \sqrt{2}\cos(\pi\mathbf{x}^2)\\
\psi_7(\mathbf{x}) &= \psi_{(2,2)}(\mathbf{x}) = \phi_2(\mathbf{x}^1)\phi_2(\mathbf{x}^2) = 2\cos(\pi\mathbf{x}^1)\cos(\pi\mathbf{x}^2).
\end{aligned}$$

These are exactly the basis functions we used in constructing least-square sieve estimators in (9). We now give a formal definition of the unravelling rules:

Definition 6.1. Given a function $c : (\mathbb{N}^+)^d \to \mathbb{R}^+$ defined on the d-tuple grid-points, we define $\mathcal{U}(\mathbf{m}) = \mathcal{U}_c(\mathbf{m}) : (\mathbb{N}^+)^d \to \mathbb{N}^+$ to be the unique surjective mapping satisfying the following conditions:

1. $\mathcal{U}(\mathbf{m}) \le \mathcal{U}(\mathbf{n})$ if and only if $c_{\mathbf{m}} \le c_{\mathbf{n}}$;
2. (tie-breaker) For $\mathbf{m}, \mathbf{n} \in (\mathbb{N}^+)^d$ with the same c values: $c_{\mathbf{m}} = c_{\mathbf{n}}$, we set $\mathcal{U}(\mathbf{m}) < \mathcal{U}(\mathbf{n})$ if and only if the following conditions holds: There exists a value $k \in \{1, 2, \ldots, d\}$ such that, $\mathbf{m}^l = \mathbf{n}^l$ for all $l \le k$, but $\mathbf{m}^k > \mathbf{n}^k$.

We call such a mapping, $\mathcal{U}$, the c-unravelling rule.

The unravelling mapping is in nature a way to sort d-tuples into a sequence. Condition 1 in Definition 6.1 is essential: tuples with smaller $c_{\mathbf{j}}$ values get a more prioritized position in the unravelled sequence. Condition 2 is an arbitrary tie-breaking rule and can be modified.

Definition 6.2. Let $c : (\mathbb{N}^+)^d \to \mathbb{R}^+$ be a function and $\mathcal{U}(\mathbf{m}) = \mathcal{U}_c(\mathbf{m}) : (\mathbb{N}^+)^d \to \mathbb{N}^+$ be its corresponding unravelling mapping. When $\{\beta_{\mathbf{j}}\}$ denotes a

set of numbers indexed by d-tuples, the c-unravelling sequence of $\{\beta_{\mathbf{j}}\}$ means a sequence of numbers (β_j), such that $\beta_j = \beta_{\mathcal{U}^{-1}(j)}$ (note that $\mathcal{U}^{-1}(j)$ is an element in $(\mathbb{N}^+)^d$). Similarly, when $\{\psi_{\mathbf{j}}\}$ denote a set of functions indexed by d-tuples, the c-unravelling sequence of $\{\psi_{\mathbf{j}}\}$ means a sequence of functions (ψ_j) such that $\psi_j = \psi_{\mathcal{U}^{-1}(j)}$.

For each function c defined on $(\mathbb{N}^+)^d$, there is a uniquely defined unravelling rule $\mathcal{U} = \mathcal{U}_c$, which gives one way to rearrange a set of basis functions into a sequence. For tensor product models such as $S_1([0,1]^d)$, we propose using what we will henceforth refer to as the *Canonical Product unravelling Rule*:

$$c_{\mathbf{j}}^{CPR} = c^{CPR}(\mathbf{j}) = \prod_{k=1}^{d} \mathbf{j}^k,$$

which leads to computationally more feasible and statistically near-optimal estimators. Using this notation, the (c_j) sequence in Fig. 1 can be mathematically described as the c^{CPR}-unravelling sequence of $c_{\mathbf{j}}^{CPR}$.

The results of Theorem 5.1 imply the CPR ordering strategy is not arbitrary. Instead, it properly balanced the estimation error and the approximation error under the tensor product model. Other ways to increase the basis function sets that are essentially different from our proposal, for example the one in (8), would not lead to (near-)optimal estimators.

When analyzing the magnitude of (c_j) in Fig. 1, we will use a key expression for the asymptotic average order of some "divisor functions" (Lemma F.3). Briefly speaking, it states that, on average, there are $\log^{d-1}(T)/(d-1)!$ ways to factor a natural number less than T into a product of d positive integers. This can be translated into the magnitude of (c_j) and explains the presence of terms like $\log^{d-1}(n)$ in Theorem 5.1. Number theory fact can also give us an estimate of the eigenvalues of relevant general product reproducing kernels. We formally state this fact as follows:

Corollary 6.3. *Let $K(x,z)$ be a Mercer kernel defined over $\mathcal{X}\times\mathcal{X} \subset \mathbb{R}\times\mathbb{R}$. Let ρ_X be a measure defined over $\mathcal{X}$. Assume K has the following Mercer expansion*

$$K(x,z) = \sum_{j=1}^{\infty} j^{-2s}\phi_j(x)\phi_j(z),$$

for some $s > 1/2$ and $\{\phi_j\}$ is a uniformly bounded orthonormal basis of $L_2(\rho_X)$. Then we know its product kernel $K^{prod} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$

$$K^{prod}(\mathbf{x},\mathbf{z}) = \prod_{k=1}^{d} K\left(\mathbf{x}^k, \mathbf{z}^k\right)$$

has the following Mercer expansion

$$K^{prod}(\mathbf{x},\mathbf{z}) = \sum_{j=1}^{\infty} \lambda_j\psi_j(\mathbf{x})\psi_j(\mathbf{z}),$$

where the non-increasing sequence $\lambda_j \asymp (j^{-1}\log^{d-1}(j))^{2s}$ *(as* $j \to \infty$*). And the functions* (ψ_j) *is the* c^{CPR}*-unravelling sequence of* $\{\psi_{\mathbf{j}}(\mathbf{x})\} = \{\prod_{k=1}^{d}\phi_{\mathbf{j}^k}(\mathbf{x}^k)\}$.

The above characterization of the order of λ_j allows us to easily transport most of the results of this paper to general multivariate RKHS problems (by unravelling multivariate RKHS problems into well characterized univariate problems). It also implies that for any RKHS with known feature mappings ϕ_j, we can perform sieve-type estimation instead of kernel ridge regression to potentially gain some feature sparsity (Section 7) or better computational efficiency (Section 8.2). For the proof of Corollary 6.3 and more discussion, see Appendix B.

7. Penalized sieve estimators in sparse models

In this section, we will discuss how to apply l_1-penalized sieve estimators for nonparametric sparse models. The difference between this section and the previous is analogous to the difference between sparse additive models [32] and additive models (discussed in Section 3), though the technical tools employed differ.

Although there may be a substantial number of features collected, it is common that only a small active subset of those features are needed to build the optimal predictive model. We will show that, similar to many other sparse methods, our proposed method is relatively robust to the ambient dimension d. It is the active dimension of the problem that has a significant impact. We now formalize our nonparametric sparse model:

Condition 7.1. There exists a D-variate function $f^* : [0,1]^D \to \mathbb{R}$, and a set of indices $\{k_1, \ldots, k_D\} \subset \{1, 2, \ldots, d\}$ such that for any $\mathbf{u} \in [0,1]^d$: we have $f^0(\mathbf{u}) = f^*(\mathbf{u}^{k_1}, \mathbf{u}^{k_2}, \ldots, \mathbf{u}^{k_D})$. Moreover, we assume

$$f^* \in S_1\big([0,1]^D\big).$$

The first half of Condition 7.1 formally states that there are D features that have dominating association with the outcome; The later half is a smoothness assumption, which can potentially be replaced by other nonparametric model assumptions. Here, we take the S_1 space as an example for presenting our ideas, for general discussion and theory, see Condition C.10 in Appendix C.4.

There are relaxations of Condition 7.1 that may be considered more interesting in practice. For example, we can consider truth that can be decomposed as a *finite* sum of M feature-sparse functions:

$$f^0(\mathbf{u}) = \sum_{m=1}^{M} f_m^*\big(\mathbf{u}^{k_{m,1}}, \mathbf{u}^{k_{m,2}}, \ldots, \mathbf{u}^{k_{m,D}}\big), \tag{12}$$

and each component satisfies the smoothness condition $f_m^* \in S_1([0,1]^D)$. The index sets $\{k_{m,1}, \ldots, k_{m,D}\}$ for different m are allowed to be different or overlapping. This relaxed condition will not neither significantly affect the implementation of our (upcoming) proposed methods nor their convergence *rate* guarantees.

In this manuscript we focus on the essential case stated in Condition 7.1 for simpler presentation.

In Sections 5 and 6, we discussed the need to order the multivariate basis functions; we additionally showed that using the unravelling rule $c_{\mathbf{j}}^{CPR} = \prod_{k=1}^{d} \mathbf{j}^k$ would lead to nearly rate-optimal least-square estimators (up to polylog). In the sparse model setting, the unravelling rule is very similar except that we allow ourselves to remove some higher-order interaction terms for computational ease. In particular, we begin with a conservative guess D' for the active dimension D. We then remove any interactions of order $> D'$. So long as $D \leq D'$ this will not affect the theoretical performance of our estimator. Formally, our new unravelling rule is:

Condition 7.2. Let $\{\phi_j\}$ be the univariate cosine basis: $\phi_1(x) = 1$, $\phi_j(x) = \sqrt{2}\cos((j-1)\pi x)$. Consider their natural d-dimensional product extension $\psi_{\mathbf{j}}(\mathbf{x}) = \prod_{k=1}^{d} \phi_{\mathbf{j}^k}(\mathbf{x}^k)$, denote (ψ_j) as the c-unravelling sequence of $\{\psi_{\mathbf{j}}\}$. The unravelling rule $c(\cdot)$ is defined as

$$c_{\mathbf{j}} = \begin{cases} c^{CPR}(\mathbf{j}), & \text{if at most } D' \text{ entries of } \mathbf{j} \text{ are greater than 1} \\ \infty, & \text{otherwise} \end{cases} \tag{13}$$

Suppose $d = 3$ and we choose the working dimension $D' = 2$. Then $\psi_{(1,1,1)}$ will get the first place when unravelling $\{\psi_{\mathbf{j}}\}$ to the (ψ_j) sequence. Similarly, $\psi_{(2,1,1)}$ gets the second position and $\psi_{(1,2,1)}$ gets the third. However, basis functions that vary in more than $D' = 2$ dimensions will not be used for our estimate. For example, $\psi_{(2,2,2)}(\mathbf{x}) = 2^{3/2}\prod_{k=1}^{3}\cos(\pi \mathbf{x}^k)$ is excluded since it varies in all three dimensions. We formalize this using an infinite value for the index in our rule (13).

For problems with higher feature dimension d and limited samples, the empirical least-squares problem (9) is likely to be under-determined (one have more basis functions than samples), and thus regularization is required for numerical stability. In addition, basis functions from non-active features should have 0 coefficient which further motivates introducing a regularization term. Toward this end, we add a simple l_1 sparsity-inducing penalty to the original loss. Formally, we need to solve the following optimization problem:

$$\left(\beta_1^{PLS}, \ldots, \beta_{J_n}^{PLS}\right) = \underset{(\beta_1,\ldots,\beta_{J_n})\in\mathbb{R}^{J_n}}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n}\left\{Y_i - \sum_{j=1}^{J_n}\beta_j \cdot \psi_j(\mathbf{X}_i)\right\}^2 + \lambda_n \sum_{j=1}^{J_n}|\beta_j|, \tag{14}$$

and our estimate is given by $f_n^{PLS}(\mathbf{x}) = \sum_{j=1}^{J_n}\beta_j^{PLS}\psi_j(\mathbf{x})$. In Appendix A.2 we include more details on the implementation of the above method. Specifically, in Algorithm 1 we show how to effectively use computational resource to generate the index set for the basis functions in (14), which would have been quite a burden if not well taken care of. We have the following theoretical guarantee for this estimator's generalization error:

Theorem 7.3. *Suppose* $\{(\mathbf{X}_i, Y_i) \in [0,1]^d \times \mathbb{R}, i = 1, 2, \ldots, n\}$ *is an i.i.d. training sample and the true regression function* f^0 *satisfies Condition 7.1. Let*

$\epsilon_i = Y_i - f^0(\mathbf{X}_i)$ be sub-Gaussian, mean-zero random variables. We further assume that the distribution of $\mathbf{X}$, ρ_X, is continuous with a bounded density function (from above and away from zero), and the working dimension D' in Condition 7.2 is no smaller than the active dimension D in Condition 7.1.

Then, for the l_1-penalized sieve estimator f_n^{PLS}, constructed with basis functions described in Condition 7.2, we have:

$$\left\|f_n^{PLS} - f^0\right\|_{2,\rho_X}^2 = O_p\left(\log(d)\log(n)\left(\frac{\log^{D-1}(n)}{n}\right)^{2/3}\right), \tag{15}$$

when $J_n = C(D)d^{D'}n^{1/3}(\log n)^{D'-1}$ and $\lambda_n = (\log(J_n)/n)^{1/2}$. Here $C(D)$ is a constant that only depends on D.

This convergence rate for f_n^{PLS} looks similar to the rate obtained for the unpenalized estimator f_n^{OLS} with two substantial differences: 1) The $\log^{d-1}(n)$ has been replaced by $\log^{D-1}(n)$ which now only involves the active dimension; and 2) The ambient dimension d is only included through a $\log(d)$ term (as is common in sparse regression).

The l_1-penalized optimization problem in (14) can be solved directly using standard lasso solvers such as `glmnet` [40]. The overall task of fitting the nonparametric estimator f_n^{PLS} can be done with R package `Sieve`. Asymptotically, the time complexity for constructing the above l_1-penalized sieve estimator is of order $O(nJ_n) = O(d^{D'}n^{4/3}\log^{D'-1}n)$. In contrast, standard applications of reproducing kernel ridge regression require $\Theta(n^3)$ computation and give no adaptivity guarantees under feature sparsity. Computationally, the proposed sieve estimator is more suitable for large data sets as its dependency on sample size is almost linear. Other theoretically guaranteed methods, such as highly adaptive lasso [2], require solving optimization problems that scale as $2^d n$, which is substantially more resource intensive than the proposed method.

The order of J_n presented in Theorem 7.3 serves as a theoretical guidance of the number of basis required to achieve the presented accuracy. The proposed procedure is computationally more tractable in the sense that: the ambient dimension d does not show up in the exponent, the exponent of the polynomial term of n does not have D' in the exponent. In practice, generic model selection procedures such as cross-validation can be applied to tune the data adaptive hyper-parameters J_n and λ_n. Since f_n^{PLS} is a nonparametric estimator, cross-validation can consistently select the best hyper-parameter combination specified by the user [62].

8. Numerical examples

So far in this manuscript, we have introduced and discussed the (dense and sparse) tensor product models and the theoretical performance guarantees of sieve estimators. In this section, we will demonstrate the finite-sample performance of the proposed methods and their applicability in practice via simulated and real data sets. The methods discussed in this manuscript, penalized

TABLE 1
Functional form and highest interaction order for simulated data.

Example 1 f^0	Example 2 f^0
$\sum_{k=1}^{D-1} Leg(2(\mathbf{x}^k - 0.5), 3) + Leg(2(\mathbf{x}^k - 0.5), 2) \cdot Leg(2(\mathbf{x}^{k+1} - 0.5), 2)$	$\sum_{\mathbf{j} \in (\mathbb{N}^+)^d : c_{\mathbf{j}}^{CPR} \leq 8} \prod_{k=1}^{D} \cos((\mathbf{j}^k - 1)\pi \mathbf{x}^k)$
Highest interaction: second order	Highest interaction: third order

and least-square sieve estimators, are implemented in the R package `Sieve`. Currently, the package is available on the Comprehensive R Archive Network (CRAN).

8.1. Performance comparison with simulated data

We first present some numerical results based on simulated data sets. In this section we will consider two types of true regression functions. In Table 1 we present the detailed functional forms of the true regression functions. The $Leg(x, j)$ function in the table is the j-th Legendre polynomial: $Leg(x, 2) = x$, $Leg(x, 3) = (3x^2 - 1)/2$. We give an additional example in Appendix A.1 where the true conditional means only contain interaction terms without main effects. In this setting the proposed methods perform much better than tree-based methods.

In the simulation study, we considered active dimension $D \in \{2, 4\}$ and ambient dimension $d \in \{4, 8, 16\}$. We used signal-noise-ratio (SNR) $= 3$ and 30 with normally distributed noise random variables. Here SNR is defined as the ratio between the squared 2-norm of f^0 and the variance of the noise variables. This means the oracle (best possible) testing R^2 should be 0.75 (SNR $= 3$) and 0.97 (SNR $= 30$). We choose sample size $n \in \{400, 800\}$. The feature vectors $\mathbf{X}$ we consider are uniformly distributed over the $[0, 1]^d$ cube. We performed 100 simulations for each setting. We use oracle hyperparameters for each method (number of basis functions, regularization parameter, number of trees, etc.), which is determined based on an independent $n = 2000$ testing data set.

The regression estimators we considered in the simulation study are: sieve estimators proposed in this work (least-square and penalized), random forest (RF, R package randomForest), gradient boosting (GBM, R package gbm), Gaussian kernel ridge regression (also known as radial kernel support vector machine), highly adaptive lasso (HAL, R package hal9001, only applied for the lower dimension case $d = 4$ due to the exponential memory requirement) and sparse additive models. We also include some oracle estimators that know which D dimensions are truly associated with the outcome Y in order to demonstrate the dimension adaptivity of the other methods. The univariate basis ϕ_j we used for sieve estimators are: $\phi_1(x) = 1$, $\phi_j(x) = \sin((j + 1/2)\pi x)$ (sine basis, for the $f^0_{\cos}$ settings) and $\phi_j(x) = \cos((j-1)\pi x)$ (cosine basis, for all the other truth f^0). The oracle kernel ridge regression method, denoted as oKRR in Fig. 2 and Fig. 3, uses the reproducing kernel of $S_1([0, 1]^D)$, see Appendix B. In Fig. 2, we present the results under high SNR settings and we evaluate the performance of each method using (absolute) testing MSE. In Fig. 3, the larger noise settings, model

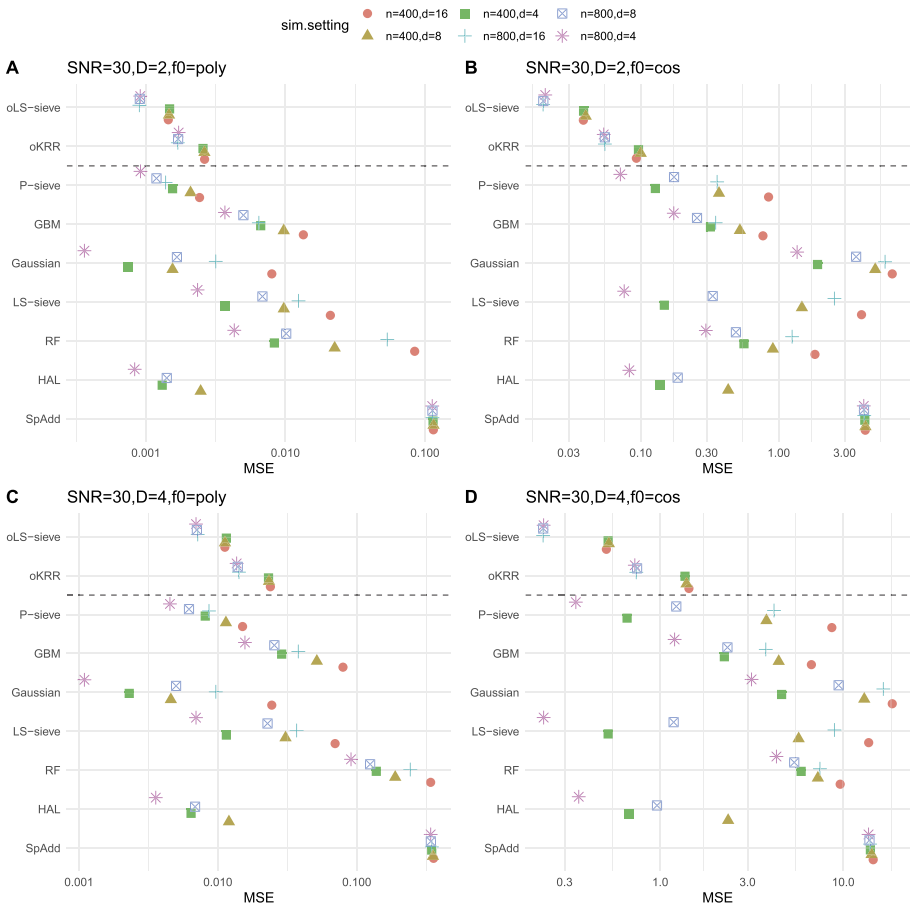

FIG 2. *Simulation study results. Low noise settings, SNR* $= 30$.

performance is evaluated via testing R^2. Sometimes R^2 is more interpretable in practice than absolute MSE, but we chose to present absolute MSE in Fig. 2 simply because it can differentiate methods better (all methods have high R^2 values in some settings).

8.2. CPU time benchmark with larger scale data

In Section 8.1, we chose several moderate sample size n and feature dimension d simulation settings. We restrict ourselves to scenarios where we can compare the predictive performance of a large library of estimators — many of the comparison estimators are computationally inefficient, so we had to limit n and d. In this section, we present some extra experiments with larger sample sizes to benchmark the computational expense of our proposed method. We use $d = 10$

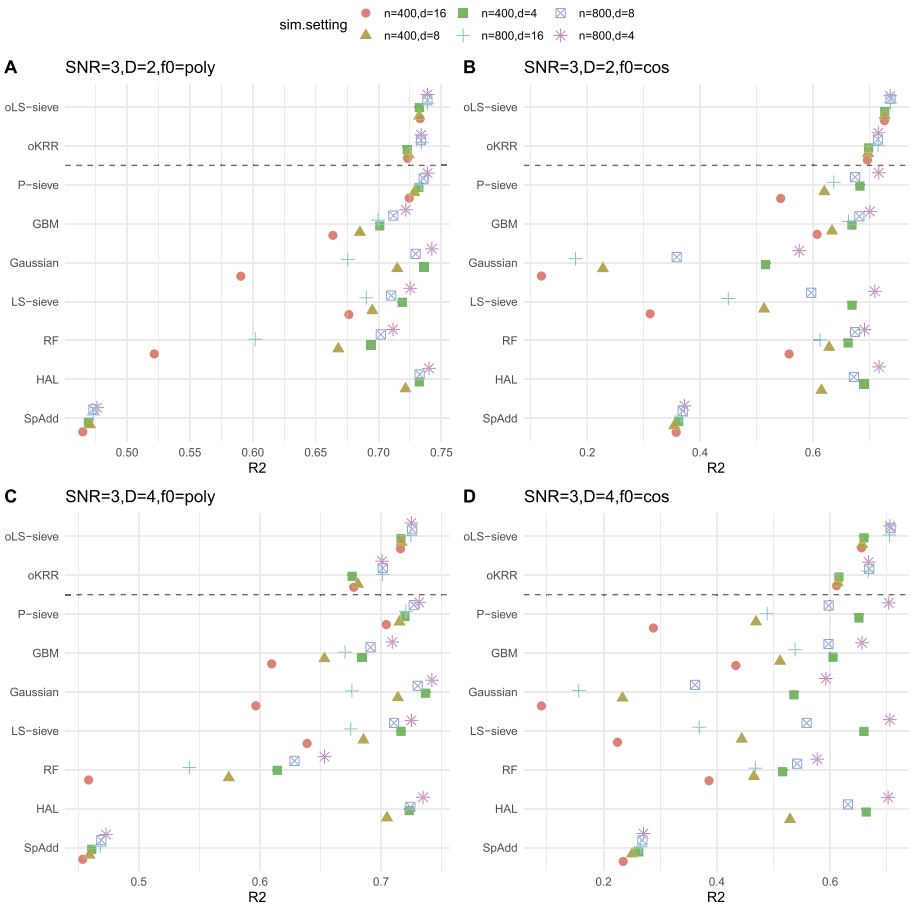

FIG 3. *Simulation study results. High noise settings, SNR* $= 3$.

and sample size $n \in [250, 5000]$. The truth f^0 takes a nonparametric additive form. Formally, we generate data from the following scheme:

$$\begin{aligned}
&\mathbf{X}_i \sim \text{Unif}[0,1]^d \\
&\epsilon_i \sim \text{Normal}(0,1) \\
&f^0(\mathbf{x}) = \sum_{j \text{ is odd}} 0.5 - |\mathbf{x}^j - 0.5| + \sum_{j \text{ is even}} \exp(-\mathbf{x}^j) \\
&Y_i = f^0(\mathbf{X}_i) + \epsilon_i.
\end{aligned}$$

The CPU time and MSE ($= \|f_n^{PLS} - f^0\|_{2,\rho_X}^2$) are shown in Fig. 4. We consider l_1-penalized estimators with varied numbers of basis functions (we label them as "less", "moderate" and "more"), though D' is fixed to be 3 for all examples and a cosine basis is always used. We use canonical unravelling rule

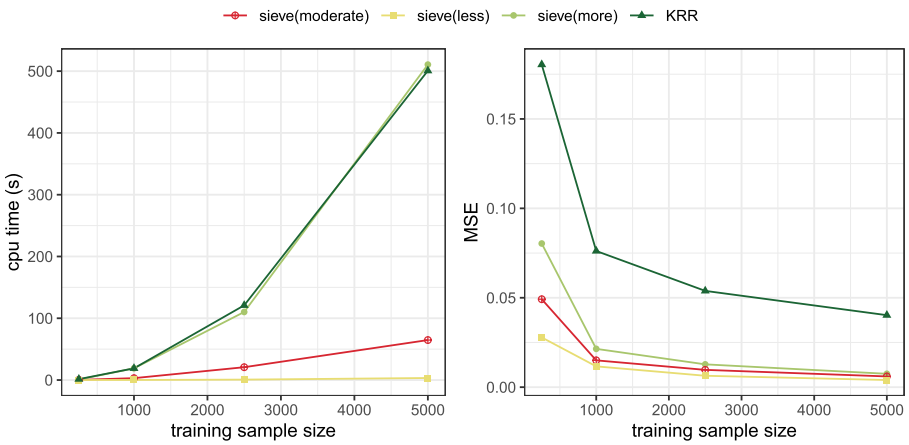

FIG 4. *CPU timing benchmark with larger sample sizes. Sieve(moderate), l_1-penalized sieve estimator, number of basis function J_n is equal to sample size n; Sieve(less), $J_n = n/10$; Sieve(more), $J_n = 10n$; KRR, kernel ridge regression. Left panel, running time of each method, including both perform model fitting and hyperparameter tuning. We use an independent validation data set, which has the same sample size as the training data, to select the hyperparameters. Right, MSE $\|\hat{f}_n - f^0\|^2_{2,\rho_X}$ of the selected estimators. The most parsimonious sieve estimator achieves the best performance among the four.*

c^{CPR} to reorder the multivariate basis functions. Instead of using exactly the theoretical number of basis functions listed in Theorem 7.3, which is in nature an asymptotic upper bound, we chose number of basis function to lie in a "convenient range" that users may consider in practice. We also include the running time and performance of kernel ridge regression for comparison. We are also performing penalty parameter tuning with some validation data: for sieve estimator, 100 candidate λ-values are considered where as for KRR only 10 are considered (due to computational expense). As we can see, the most parsimonious estimator sieve(less) is the fastest and the one with smallest MSE. This is consistent with the theory of univariate sieve-methods where a small number of basis functions appropriately balances estimation and approximation error. We did not use estimators with fewer basis functions since they cannot be distinguished in the CPU plot from sieve(less). The sieve estimators may additionally have leveraged some adaptivity to the additive structure of the true regression function, which could explain their improved MSE as compared to KRR estimators.

Both the KRR estimators and sieve-estimators, were written by us: They are coded in `C++` and called from our `R` package `sieve`. While we attempted to write efficient code, we naturally imagine that others might be able to write more performant code (for both KRR and our proposed method), thus the above results should mainly be used to give a general idea of timing comparison. The experiments in this section are run on a AMD Opteron 6300 Processor, 2.8 GHz.

TABLE 2
Basic information for public data sets used in performance comparison.

Name	Sample size (n)	Feature type	References
gdp	616	6	Liu and Stengos [30]
fev	654	4	Rosner [35]
fev50	654	54	–
bio	779	9	Grisoni et al. [16]
aba	4177	8	Waugh [60]
supc	21263	81	Hamidieh [18]

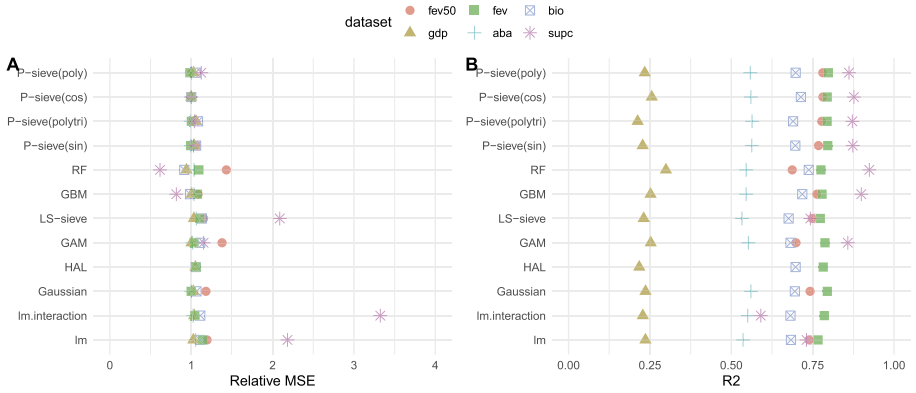

FIG 5. *Relative MSE and R^2 on real data sets. The MSE values are normalized to that of penalized sieve estimator with cosine functions. Methods requiring significantly more computational resource are not reported.*

8.3. Performance comparison with real data

We also compare the predictive performance of these methods on 5 publicly available data sets. Some basic information for the data sets is reported in Table 2. In Fig. 5, we present the relative testing MSE and (absolute) R^2 of each method. We saved 30% of the samples as the test set and the hyperparameters of each method are determined using a 5-fold cross-validation on the training set (more details presented in Table 3 of the supplement). The `fev50` data set combines the true outcome and features from `fev`, with 50 artificially constructed non-informative features (independent, Unif$[0, 1]$). We use this data set as a moderately high-dimensional, sparse feature example. One of the data sets, `supc`, has been used as an example to demonstrate the effectiveness of tree-based methods [18], so we also include it for a more comprehensive comparison. We only applied highly adaptive lasso to 3 data sets and Gaussian kernel ridge regression to 5 data sets due to their high computational resource requirement: These would not efficiently run on a machine with 1 Intel Core m3 processor, 1.2 GHz, with 8 GB of RAM. The linear model with all interaction terms is not applicable to `fev50` because the empirical problem is not well-posed without further modification (number of coefficients is larger than the sample size).

We compared sieve estimators based on different univariate bases ϕ_j, including polynomial, cosine basis and sine basis (the basis defined earlier in this section), as well as a combination of polynomial and trigonometric functions [10]. The performance of penalized sieve methods using different basis functions is quite similar. The random forest estimator is more sensitive to the extra dimensions of `fev50` than penalized sieve and GBM. For more information on the data sets, see Table 2 in the supplementary material.

9. Discussion

In this paper, we discussed sieve-type (basis-expansion) methods for multivariate nonparametric regression problems. Under certain tensor-product space assumptions, least-squares and penalized estimators were shown to have favorable theoretical guarantees. Specifically, they have a moderate dependence on the dimension of features and are adaptive when a small subset of the features are of primary importance for determining the outcome. We now give a bit more discussion and contextualization of our work as well as noting possible future directions.

Rate-optimality of our guarantees The minimax rate of estimation under the setting in Theorem 5.1 is known to be $n^{-2/3}(\log n)^{2(d-1)/3}$ (proved in Lin [29]). Both the proposed least square estimators and the l_1-penalized estimators can achieve this rate up to a $\log n$ term (Recall that in this setting we are treating the total dimension d as "fixed", not increasing with n).

For the "high-dimensional" estimation setting in Theorem 7.3, we conjecture [63] the minimax rate to be

$$\frac{D\log(d/D)}{n} + \left(\frac{\log^{D-1}(n)}{n}\right)^{2/3}.$$

This rate is formally analogous to the more well-known sparse additive models' minimax rate. Specifically, the $\log d$ term should not multiply the nonparametric rate and instead go into another additive term. In contrast, our theoretical guarantee is

$$\log d \log n \left(\frac{\log^{D-1}(n)}{n}\right)^{2/3},$$

where the $\log d$ term multiplies the nonparametric term.

When the ambient dimension d increases with n polynomially fast $d = d_n = n^\gamma$, the conjectured minimax rate is of order

$$\frac{\log n}{n} + \underbrace{\left(\frac{\log^{D-1}(n)}{n}\right)^{2/3}}_{\text{main term}}.$$

And our theoretical guarantee for the penalized estimator is

$$\gamma \log^2 n \left(\frac{\log^{D-1}(n)}{n} \right)^{2/3}, \tag{16}$$

which is $\log^2 n$ slower than the conjectured minimax rate. The multiplicative $\log d$ in (15) term, resulting in a multiplicative $\log n$ in (16), may be due to artifacts in our proof technique, but may also be an intrinsic limitation of a our simple $\|\cdot\|_1$-penalty.

More general models and diverging active dimension The theoretical guarantee of Theorem 7.3 tells us $d = d_n$ is allowed to increase at a polynomial rate n^γ and we would still obtain tractable estimators (having a $\log d_n = \gamma \log n$ term). However, the active dimension D has to increase very slowly if people are interested in such a regime. For example, if we plug $D = D_n = a \log n, a > 0$ into the minimax rate of $S_1([0,1]^d)$ space: $n^{-2/3}(\log n)^{2(D-1)/3}$, it would become

$$n^{-2/3}(\log n)^{2(a \log n - 1)/3} \to \infty, \text{ as } n \to \infty.$$

(Here we used the fact that for any $\gamma > 0$

$$\log n^{a \log n} = \big(\log n \exp(\gamma/a) \exp(-\gamma/a)\big)^{a \log n} = n^\gamma \big(\log n \exp(-\gamma/a)\big)^{a \log n},$$

which means $\log n^{a \log n}$ diverges faster than any polynomial in n as $n \to \infty$. Specifically, it diverges faster than $n^{2/3}$.)

The above calculation implies such a statistical problem is too hard even for $D \sim \log n$. It seems like $D = \log \log n$ would give more interesting models to perform estimation within. This is a curious demonstration that working under S_1 can only partially avoid the curse of dimentionality (as we claimed in the manuscript).

Alternatively, one could assume the true regression function is a linear combination of M_n component functions $\{f_i^0, i = 1, \ldots, M_n\}$, where each f_i^0 depends on only D features $\mathbf{x}$ (but those features could be different across f_i^0) and we assume each f_i^0 lies in a proper $S_1([0,1]^D)$ space. In this case, we expect the l_1-penalized estimator to converge at a rate no slower than

$$M_n^2 \log d \log n \left(\frac{\log^{D-1}(n)}{n} \right)^{2/3},$$

which may allow M_n to be on the order of $\log n$.

Tensor product spaces with higher order smoothness The main focus of this paper is the impact of the overall feature dimension d on non-parametric estimation quality. Our methods and discussion can also be extended to more general tensor product spaces. Formally, one may investigate the setting where f^0 belongs to space $S_s([0,1]^d)$:

$$S_s\big([0,1]^d\big) = \big\{ f \in L_2\big([0,1]^d\big) \mid D^{\mathbf{a}} f \in L_2\big([0,1]^d\big) \text{ for all } \|\mathbf{a}\|_\infty \le s \big\},$$

for some integer $s \geq 1$. One difficulty here is that the most appropriate sieve basis functions (orthogonal with respect to both the L_2 and Sobolev inner products, but not necessarily periodic) no longer take a simple closed-form, unlike in the S_1 case. However, our theoretical analysis can be directly applied here (e.g. Theorem C.5): We consider a true regression function that lies in some multivariate Sobolev ellipsoid with general smoothness parameter s.

Appendix A: More numerical examples and method implementation discussion

A.1. Supplementary numerical results

In the main text, we present selected results from our simulation study. In this section we will provide more details together with another data generation setting that only has interaction terms.

In the simulation study, we have been using the oracle hyperparameters for each method under comparison, that is, those parameters that lead to minimal testing error. In Table 3 we present the hyper-parameters that are tuned for each method.

In Figs. 6 and 7, we present the simulation results under the same setting as in the main text. The performance is evaluated using multiple metrics as in Figs. 2 and 3.

We also present the simulation results from another data generating mechanism that does not have an additive component (results are in Fig. 8 and 9). The data generating mechanism is defined as:

$$f^0_{interaction} = \sum_{k=1}^{D-1} Leg\big(2\big(\mathbf{x}^k - 0.5\big), 2\big) \cdot Leg\big(2\big(\mathbf{x}^{k+1} - 0.5\big), 3\big) \tag{17}$$

where the $Leg(x, j)$ function is the j-th Legendre polynomial

$$Leg(x, 2) = x, \quad Leg(x, 3) = \big(3x^2 - 1\big)/2 \tag{18}$$

This conditional mean has no main effects, meaning that

$$E\big[f^0_{interaction}(\mathbf{x}) \mid \mathbf{x}^k\big] = 0$$

TABLE 3
Hyperparameters for each method.

Method	Hyper-parameter
l_1-penalized sieve estimator (P-sieve)	number of basis functions, penalty parameter
gradient boosting machine (GBM)	number of iteration, tree depth
Gaussian kernel ridge regression	bandwidth, penalty parameter
least-square sieve estimator (LS-sieve)	number of basis functions
random forest (RF)	number of sampled features, tree depth
highly adaptive lasso (HAL)	penalty parameter
sparse additive model (SpAdd)	number of basis functions, penalty parameter

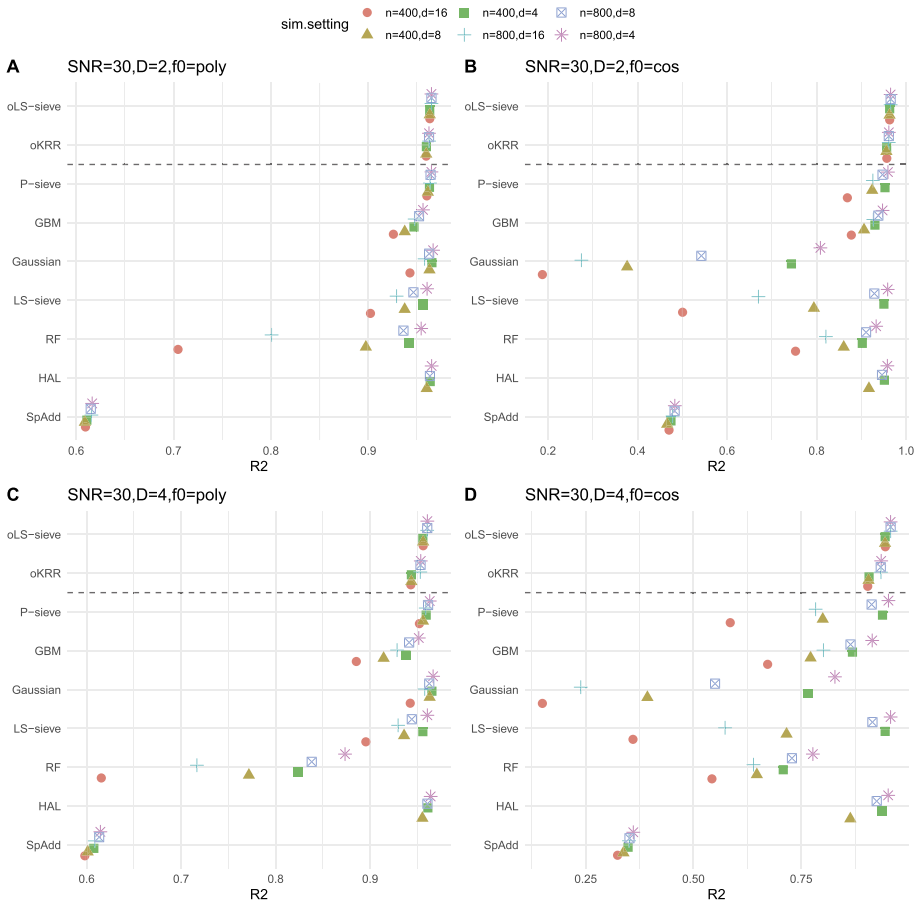

FIG 6. *Simulation study results.* $SNR = 30$.

for any $1 \leq k \leq d$. We can verify this by direct calculation (recall that $\mathbf{x} \sim Uniform([0,1]^d)$). Although $f^0_{interaction}$ is a simple polynomial with nice smoothness properties, the lack of main effects (or additive components) messes up the performance of many methods. The almost zero testing R^2 of additive models demonstrates that in this setting they are no better than taking an unconditional mean of the outcome. Tree-based methods (gradient boosting and random forest) have more difficulties in this setting, especially when compared with their outstanding performance when the main effect components do exist. Tree-based methods cannot readily decide at which point to divide the feature space. For any binary cut only engaged with one feature, the mean of the outcome on one side of the division would be very similar to that of the other side under this specific setting.

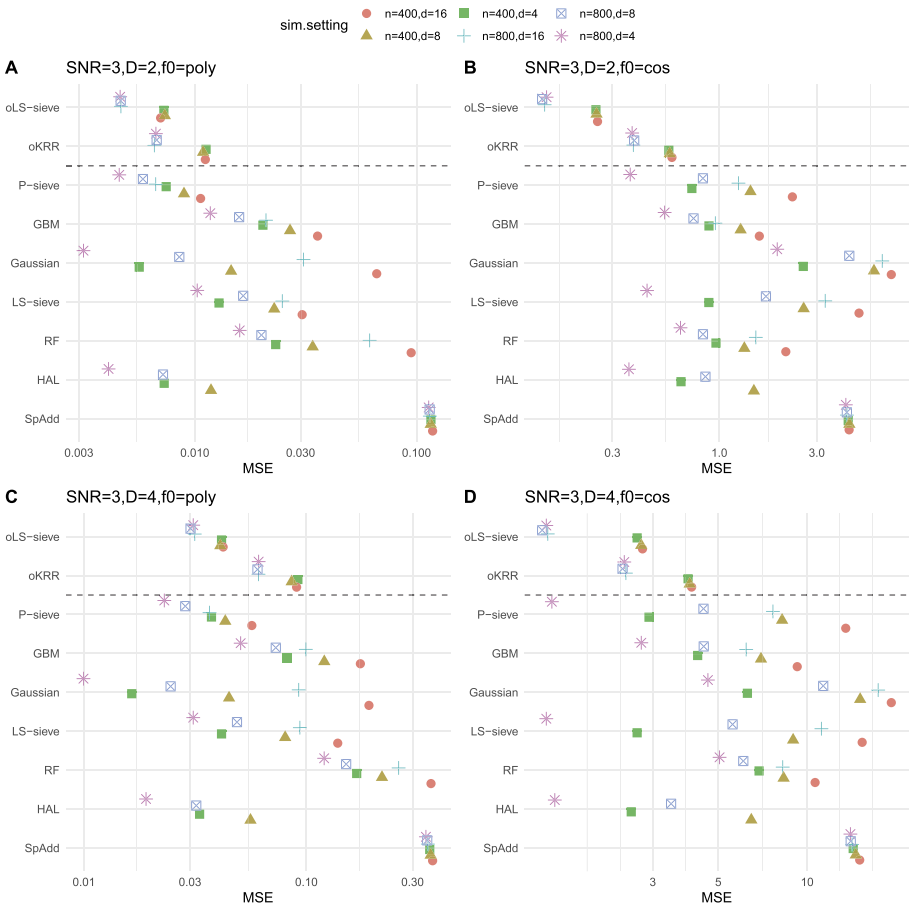

FIG 7. *Simulation study results. SNR* = 3.

A.2. Generating the design matrices

In this section we present more details on efficiently constructing the design matrix for multivariate sieve estimators. In the main text, we mention that the numerical implementation of sieve estimators is reduced to solving a least-square problem or a l_1-penalized optimization problem. In both cases we need to construct a design matrix $\hat{\Psi}$ whose elements are $\hat{\Psi}_{ij} = \psi_j(\mathbf{x}_i)$.

Given a set of multivariate product basis functions $\psi_{\mathbf{j}}(\mathbf{x}) = \prod_{k=1}^{d} \phi_{\mathbf{j}^k}(\mathbf{x}^k)$ indexed by $\mathbf{j} \in (\mathbb{N}^+)^d$, the unravelling rule $c_{\mathbf{j}} = \prod_{k=1}^{d} \mathbf{j}^k$ tells us how to sequentially use them to construct estimators. However, we only have a nonconstructive description of the elements in the unravelled sequence (ψ_j). To construct the design matrix $\hat{\Psi}$, we need to know the explicit form of each ψ_j. In practice, we need to first create an index matrix from which the algorithm identifies

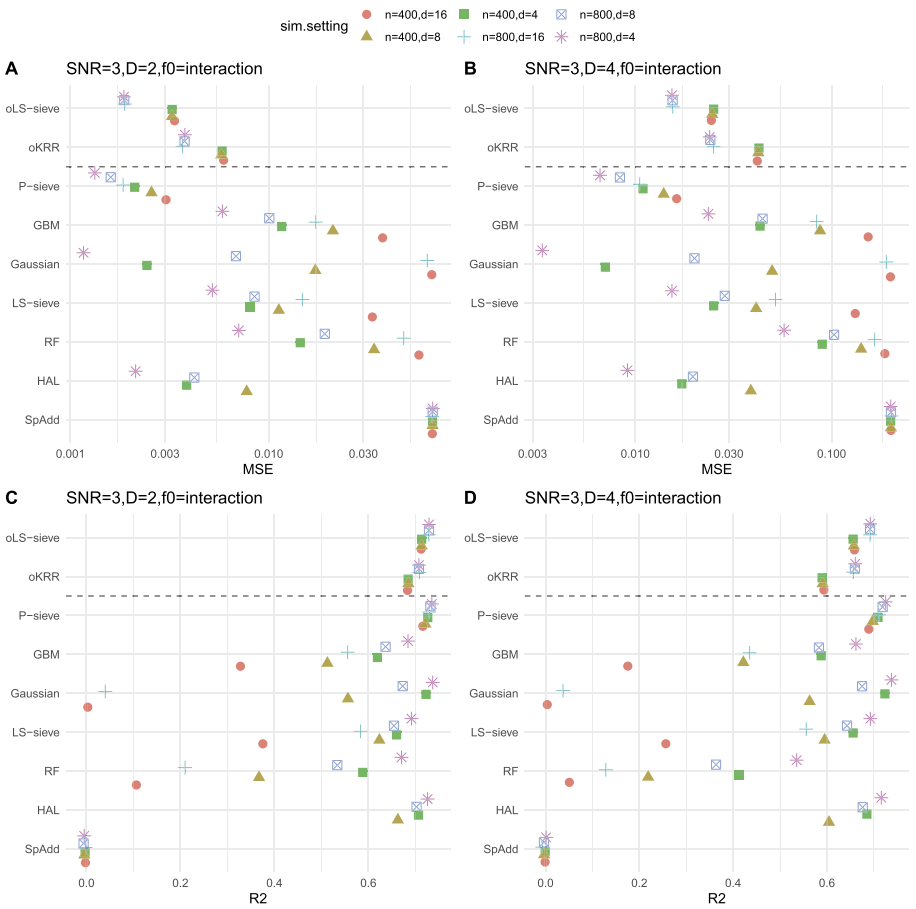

FIG 8. *Additional settings, true regression function does not have main effect components. SNR* = 3.

the analytical form of (ψ_j). For example, in the case $d = D' = 3$, we should construct an index matrix M of three columns (corresponding to the three dimensions). The first row has elements: $M_{11} = M_{12} = M_{13} = 1$, corresponding to the constant function $\psi_{(1,1,1)}$. And the following six rows are all 1 except for $M_{21} = M_{32} = M_{43} = 2$, and $M_{51} = M_{62} = M_{73} = 3$. They correspond to the second through seventh basis functions $\psi_{(2,1,1)}, \psi_{(1,2,1)}, \psi_{(1,1,2)}, \psi_{(3,1,1)}$, etc. By reading through this matrix, the algorithm can directly figure out the analytical form of the basis functions.

There are multiple ways to construct such an index matrix. When $D' = d$ (dense setting), one straightforward strategy is: 1) the user specifies the maximum index product C; 2) identify all the indices $\mathbf{j} \in (\mathbb{N}^+)^d$ whose maximum entry is smaller or equal to C; 3) sort the indices increasingly according to the index product $c_{\mathbf{j}}^{CPR}$; 4) keep only the earlier indices whose product is less

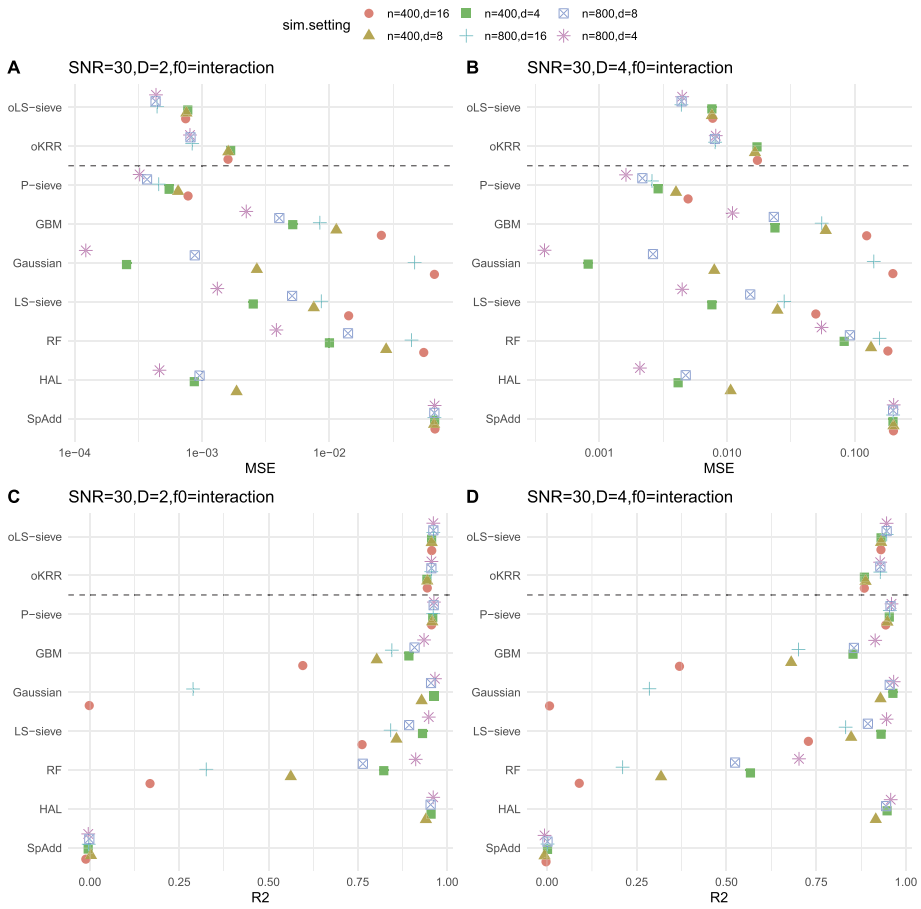

FIG 9. *Additional settings, true regression function does not have main effect components.* $SNR = 30$.

than or equal to C. This algorithm is simple but is computationally wasteful. In step 2), C^d indices must be stored (this is memory intensive, even for moderate C and d). According to our theoretical results, we only need a subset of size $C \log^d(C)$. This issue is further exacerbated in the sparse case when $D' \ll d$. Therefore, we seek an alternative, computationally more efficient strategy, which includes some integer factorization, for generating the index matrix.

In Algorithm 1, we provide the details of the procedure. By factoring each positive integer as a product of D' numbers sequentially, we can fill out the matrix M. In the case when $d = D' = 3$, there is one row with row product equal to 1, two rows having row product equals to 2, and six rows having a product equal to 4. When D' is much smaller than d, as for the sparse sieve estimators, there should be many fewer rows corresponding to the same row product. The algorithm is presented below, followed by an example to explain

some of the steps.

Set the maximum row product as `ProdMax`, feature dimension `d`, working dimension `D'`.
Define $C_m^d = d!/\{m!(d-m)!\}$, the combination number of "choosing m out of d".
`M` ← An all 1 matrix of size $1 \times$ `d`.
FOR `Prod` = 2 TO `Prod` = `ProdMax`
  Find all $\tau_{\mathtt{D}'}$(`Prod`) ways to factorize `Prod` as a product of `D'` numbers. *
  Omit all values of "1" in the products and combine identical factorizations.
  `GreaterThanOne` ← A list. Each element corresponds to one of the factorizations.
  FOR `i` = 1 TO `i` = list length of `GreaterThanOne`
    `Gi` ← The `i`-th element in `GreaterThanOne`.
    `m` ← The length of the array `Gi`.
    `Position` ← A matrix of size $C_{\mathtt{m}}^{\mathtt{d}} \times$ `m`.
      Each row corresponds to a unique way of choosing `m` elements from $\{1, \ldots, \mathtt{d}\}$.
    `NewIndexMatrix` ← A matrix of size $C_{\mathtt{m}}^{\mathtt{d}} \times$ `d`. All elements are 1.
    FOR `j` = 1 TO `j` = row number of `Position`
      `NewIndexMatrix[j, Position[j,]]` ← `Gi` **
    ENDFOR
    `M` ← Stack `M` above `NewIndexMatrix` to form a longer matrix.
  ENDFOR
ENDFOR
RETURN `M`.

Algorithm 1: Algorithm for generating the index matrix. For the definition of the $\tau_{D'}$ function mentioned in step *, see Definition C.1. In ** step we use the notation from the `R` programming language to express our matrix update.

We present some examples to better explain the compactly written algorithm above. Let's assume `d` = 3, `D'` = 2. Suppose we are currently at `Prod` = 6 in the first layer of FOR loops. The $\tau_2(6) = 4$ ways to factorize 6 are:

$$6 = 6 \times 1 = 1 \times 6 = 2 \times 3 = 3 \times 2. \tag{19}$$

After the "Omit all values of 1 in the products and combine identical factorizations" step, we have three ways to factor 6 (the first two above are combined). Therefore, the `GreaterThanOne` list is

$$\mathtt{GreaterThanOne} = \mathrm{list}\big([6], [2,3], [3,2]\big). \tag{20}$$

The arrays in `GreaterThanOne` are of different lengths. Suppose we are at `i` = 2 in the second layer of the FOR loop. Then `Gi` = $[2,3]$, `m` = 2. The `Position` matrix we constructed is

$$\mathtt{Position} = \begin{bmatrix} 1 & 2 \\ 1 & 3 \\ 2 & 3 \end{bmatrix}. \tag{21}$$

This matrix specifies at which positions we are going to insert `Gi`. In the inner most FOR loop, we are going to update the all 1 matrix `NewIndexMatrix` using the information of `Position` and `Gi`: `Position`. In particular, `Gi`: `Position`

specifies where to update, and `Gi` specifies what the elements are updated to. When $\mathtt{i} = 2, \mathtt{j} = 1$, we update the 1^{st} and 2^{nd} columns in the 1^{st} row of `NewIndexMatrix` to be $[2, 3]$, that is

$$\texttt{NewIndexMatrix}: \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \xrightarrow{\text{Update}} \begin{bmatrix} 2 & 3 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \tag{22}$$

When $\mathtt{i} = 2, \mathtt{j} = 2$, we update the 1^{st} and 3^{rd} columns in the 2^{nd} row of `NewIndexMatrix` to be $[2, 3]$:

$$\texttt{NewIndexMatrix}: \begin{bmatrix} 2 & 3 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \xrightarrow{\text{Update}} \begin{bmatrix} 2 & 3 & 1 \\ 2 & 1 & 3 \\ 1 & 1 & 1 \end{bmatrix} \tag{23}$$

After looping through all the `j`, `i` and `Prod`, we have our desired index matrix `M`. Its first several rows are:

$$\text{row 1 to 10:} \begin{bmatrix} 1 & 1 & 1 \\ 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \\ 3 & 1 & 1 \\ 1 & 3 & 1 \\ 1 & 1 & 3 \\ 4 & 1 & 1 \\ 1 & 4 & 1 \\ 1 & 1 & 4 \end{bmatrix} \quad \text{row 11 to 20:} \begin{bmatrix} 2 & 2 & 1 \\ 2 & 1 & 2 \\ 1 & 2 & 2 \\ 5 & 1 & 2 \\ 1 & 5 & 1 \\ 1 & 1 & 5 \\ 6 & 1 & 1 \\ 1 & 6 & 1 \\ 1 & 1 & 6 \\ 2 & 3 & 1 \end{bmatrix} \quad \text{row 21 to 30:} \begin{bmatrix} 2 & 1 & 3 \\ 1 & 2 & 3 \\ 3 & 2 & 1 \\ 3 & 1 & 2 \\ 1 & 3 & 2 \\ 7 & 1 & 1 \\ 1 & 7 & 1 \\ 1 & 1 & 7 \\ 8 & 1 & 1 \\ 1 & 8 & 1 \end{bmatrix} \tag{24}$$

So we can read $\psi_1 = \psi_{(1,1,1)}$, $\psi_{11} = \psi_{(2,2,1)}$ and $\psi_{21} = \psi_{(2,1,3)}$, etc.

Appendix B: Product kernels and tensor product spaces

B.1. Univariate RKHS and Sobolev ellipsoids

In Appendix B, we will review the concept of Mercer kernels and reproducing kernel Hilbert spaces (RKHS). We will first engage with univariate RKHSs and their Sobolev ellipsoid representation in Appendix B.1. By considering the tensor product kernel, we can extend our discussion to multivariate tensor product models (Appendix B.2). Later in this section, we will arrive at some multivariate Sobolev ellipsoid models. These models can be seen as abstractions of the example function spaces (such as $S_1([0,1]^d)$) discussed in the main text.

There is a vast literature on univariate nonparametric regression problem. We list a few of them here: Sobolev space and smoothing spline estimators [56]; reproducing kernel Hilbert space and kernel ridge regression estimators [41]; Sobolev ellipsoid and sieve-type projection estimators [49]. These function spaces are closely related to each other: Sobolev spaces can sometimes be treat

as a special case of RKHS and there is usually an equivalence between a ball in an RKHS and a Sobolev ellipsoid. We will try to give a brief review of this part of nonparametric learning through some examples.

First we are going to present the concept of Mercer-kernels and their related reproducing kernel Hilbert spaces (on the real line).

Definition B.1. A symmetric bivariate function $k : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is positive semi-definite (PSD) if for any $n \geq 1$ and $(x_i)_{i=1}^n \subset \mathbb{R}$, the $n \times n$ matrix $\mathbb{K}$ whose elements are $\mathbb{K}_{ij} = k(x_i, x_j)$ is always a PSD matrix.

A continuous, bounded, PSD kernel function k is called a *Mercer kernel*.

The following theorem [5] states the existence and uniqueness of a reproducing Hilbert space with respect to a Mercer kernel. The domain $\mathbb{R}$ in the following theorem can replaced by a subset such as $[0, 1]$ or $[0, +\infty)$.

Theorem B.2. *For a Mercer Kernel $k : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, there exists an unique Hilbert Space $(\mathcal{H}_k, \langle \cdot, \cdot \rangle_k)$ of functions on $\mathbb{R}$ satisfying the following conditions. Let $k_x : z \mapsto k(x, z)$:*

1. *For all $x \in \mathbb{R}$, $k_x \in \mathcal{H}_k$.*
2. *The linear span of $\{k_x \mid x \in \mathbb{R}\}$ is dense (w.r.t $\| \cdot \|_k$) in $\mathcal{H}_k$.*
3. *For all $f \in \mathcal{H}_k, x \in \mathbb{R}$, $f(x) = \langle f, k_x \rangle_k$ (reproducing property).*

We call this Hilbert space the Reproducing kernel Hilbert space (RKHS) *associated with kernel k.*

Example. The space $W_1([0, 1])$ is a RKHS with kernel

$$k(s, t) = \frac{\cosh(\min(s, t)) \cosh(1 - \max(s, t))}{\sinh(1)} \tag{25}$$

For the proof, see Appendix A of Fasshauer and McCourt [11] or Akgül et al. [1]. The RKHS inner product for this kernel is defined as

$$\langle f, g \rangle_{W_1([0,1])} = \int_0^1 f(\tau) g(\tau) d\tau + \int_0^1 f'(\tau) g'(\tau) d\tau \tag{26}$$

The reproducing property reads as: for any $x \in [0, 1]$ and any $f \in W_1([0, 1])$

$$\begin{aligned} f(x) &= \langle f, k_x \rangle_{W_1([0,1])} \\ &= \int_0^1 f(\tau) k(x, \tau) d\tau + \int_0^1 f'(\tau) \frac{\partial}{\partial \tau} k(x, \tau) d\tau. \end{aligned} \tag{27}$$

Under mild conditions [42], a Mercer kernel has the following Mercer expansion.

$$k(s, t) = \sum_{j \in \mathcal{J}} \lambda_j \phi_j(s) \phi_j(t), \tag{28}$$

where $\mathcal{J}$ is an at most countably infinite index set. The eigenvalues λ_j are real numbers. The eigenfunctions (basis functions) $\{\phi_j\}$ can also be a complete basis of some L_2 space or the RKHS.

Although the majority of estimation procedures under RKHS models leverage the reproducing property, the method considered in this paper uses the feature maps directly (which is of a sieve nature). There have been studies showing that considering the problem from this perspective can give substantial computational advantage over standard kernel methods [65, 64]. In this manuscript we will also show how sieve estimators can be more easily adapted to employ variable selection and can additionally be adaptive to dimension. Now, we present the important connection between a RKHS and a Sobolev ellipsoid established in the literature (e.g., p. 37, Theorem 4 in Cucker and Smale [5]).

Theorem B.3. *Under mild conditions, the Hilbert space $\mathcal{H}_k$ of the kernel k (defined in Theorem B.2) is identical – same function class with the same inner product – to the following Hilbert space $\mathbb{H}_k$.*

$$\mathbb{H}_k = \left\{ f \mid f = \sum_{j=1}^{\infty} a_j \phi_j \quad \textit{with} \quad \sum_{j=1}^{\infty} a_j^2 \lambda_j^{-1} < \infty \right\} \tag{29}$$

The RKHS inner product can be explicitly written as:

$$\langle f, g \rangle_k = \sum_{j=1}^{\infty} \lambda_j^{-1} a_j b_j \tag{30}$$

for $f = \sum_j a_j \phi_j$, and $g = \sum_j b_j \phi_j$. The functions ϕ_j and real numbers λ_j are the eigen-system in the Mercer expansion (28) (assuming $\mathcal{J} = \mathbb{N}^+$).

Example. The reproducing kernel for $W_1([0,1])$ has the following Mercer expansion:

$$k(s,t) = \sum_{j=1}^{\infty} \lambda_j \phi_j(s) \phi_j(t), \tag{31}$$

with

$$\begin{aligned} &\lambda_1 = 1, \quad \phi_1(x) = 1, \\ &\lambda_j = \frac{1}{1 + ((n-1)\pi)^2}, \quad \phi_j(x) = \sqrt{2} \cos\big((n-1)\pi x\big) \text{ for } j \geq 2. \end{aligned} \tag{32}$$

Therefore, we also have the following characterization of a ball in $W_1([0,1])$:

$$\big\{ f \in W_1\big([0,1]\big) \mid \|f\|_{W_1}^2 \leq Q^2 \big\} = \left\{ f = \sum_{j=1}^{\infty} a_j \phi_j \quad \text{with} \quad \sum_{j=1}^{\infty} a_j^2 \lambda_j^{-1} \leq Q^2 \right\} \tag{33}$$

To summarize, a ball in a RKHS is a Sobolev ellipsoid.

B.2. Multivariate RKHS and Sobolev ellipsoids

Given a univariate RKHS, one of the most naturally related multivariate RKHS is the one corresponding to the product kernel. This also happens to correspond to one of the most commonly used multivariate kernels in practice: The multivariate Gaussian kernel which is a product of univariate Gaussian kernels.

Definition B.4. Given a univariate Mercer kernel $k : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, we define its (natural, d-dimensional) product kernel $k^d : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ to be:

$$k^d(\mathbf{s}, \mathbf{t}) = \prod_{j=1}^{d} k\big(\mathbf{s}^j, \mathbf{t}^j\big). \tag{34}$$

We can also define the RKHS of k^d using the fact that k^d is also a Mercer kernel (Proposition 12.31 of Wainwright [58]). Typical elements in this multivariate RKHS take the following form:

$$f(\mathbf{x}) = \sum_{l=1}^{m} \prod_{k=1}^{d} f_{kl}\big(\mathbf{x}^k\big), \text{ with } f_{kl} \in \mathcal{H}_k. \tag{35}$$

There are multiple ways to engage with an element in $\mathcal{H}_{k^d}$ and its inner product. One way, as presented above, is using the property that $\mathcal{H}_{k^d}$ is a tensor product Hilbert space of d univariate Hilbert spaces. This would lead to the following characterization of its inner product.

Proposition B.5. *The RKHS for k^d, $\mathcal{H}_{k^d}$, is equipped with the inner product:*

$$\langle h, g \rangle_{k^d} = \sum_{i=1}^{n} \sum_{l=1}^{m} \prod_{j=1}^{d} \langle h_{ij}, g_{lj} \rangle_k \tag{36}$$

for $h(\mathbf{x}) = \sum_{i=1}^{n} \prod_{j=1}^{d} h_{ij}(\mathbf{x}^j)$, $g(\mathbf{x}) = \sum_{l=1}^{m} \prod_{j=1}^{d} g_{lj}(\mathbf{x}^j)$. The component functions h_{ij}, g_{lj} all belong to the univariate RKHS $\mathcal{H}_k$.

.

Alternatively, we can also consider the basis expansion form of the functions in $\mathcal{H}_{k^d}$ (similar to Theorem B.3). The tensor product kernel k^d has the following Mercer expansion (which can be formally verified using its Mercer expansion):

$$k^d(\mathbf{s}, \mathbf{t}) = \sum_{\mathbf{j} \in (\mathbb{N}^+)^d} \prod_{k=1}^{d} \lambda_{\mathbf{j}^k} \psi_{\mathbf{j}}(\mathbf{s}) \psi_{\mathbf{j}}(\mathbf{t}), \text{ with } \lambda_{\mathbf{j}^k} \in \mathbb{R}. \tag{37}$$

We have the following equivalent characterization:

Proposition B.6. *The inner product presented in Proposition B.5 is equivalent to the following one expressed in basis expansion form:*

$$\langle h, g \rangle_{k^d} = \sum_{\mathbf{j} \in (\mathbb{N}^+)^d} \left(\prod_{k=1}^{d} \lambda_{\mathbf{j}^k} \right)^{-1} h_{\mathbf{j}} g_{\mathbf{j}} \tag{38}$$

for h, g in the multivariate RKHS $\mathcal{H}_{k^d}$ with the basis expansion

$$h = \sum_{\mathbf{j} \in (\mathbb{N}^+)^d} h_{\mathbf{j}} \psi_{\mathbf{j}}, g = \sum_{\mathbf{j} \in (\mathbb{N}^+)^d} g_{\mathbf{j}} \psi_{\mathbf{j}}.$$

The multivariate basis $\psi_{\mathbf{j}}(\mathbf{x}) = \prod_{k=1}^{d} \phi_{\mathbf{j}^k}(\mathbf{x}^{\mathbf{j}^k})$ is the product of the eigenfunctions (as defined in (28)) of the univariate kernel k.

Lemma B.7. *A ball in $S_1([0,1]^d)$ is equivalent to a multivariate Sobolev-type ellipsoid. Formally,*

$$\begin{aligned}&\left\{h \in L_2\big([0,1]^d\big) \mid \sum_{\|\mathbf{a}\|_\infty \le 1} \big\|D^{\mathbf{a}} h\big\|^2_{L_2([0,1]^d)} \le Q^2\right\} \\ &= \left\{h = \sum_{\mathbf{j}\in(\mathbb{N}^+)^d} \beta_{\mathbf{j}}\psi_{\mathbf{j}} \mid \sum_{\mathbf{j}\in(\mathbb{N}^+)^d} \left(\prod_{k=1}^d \mathbf{j}^k\right)^2 \beta_{\mathbf{j}}^2 \le Q^2\right\}\end{aligned} \tag{39}$$

where the multivariate basis $\psi_{\mathbf{j}} = \prod_{k=1}^d \phi_{\mathbf{j}^k}$ is the product of the cosine functions (ϕ_j defined in (32)).

Proof. The natural d-dimensional tensor product extension of $W_1([0,1])$ space is the RKHS of the kernel:

$$\begin{aligned}k^d(\mathbf{s},\mathbf{t}) &= \prod_{m=1}^d k\big(\mathbf{s}^m,\mathbf{t}^m\big) \\ &= \big\{\sinh(1)\big\}^{-d} \prod_{m=1}^d \cosh\big(\min\big(\mathbf{s}^m,\mathbf{t}^m\big)\big)\cosh\big(1-\max\big(\mathbf{s}^m,\mathbf{t}^m\big)\big)\end{aligned} \tag{40}$$

The inner product, according to Proposition B.5, can be explicitly written as:

$$\begin{aligned}\langle h,g\rangle_{k^d} &= \sum_{k=1}^n\sum_{l=1}^m\prod_{j=1}^d \langle h_{kj},g_{lj}\rangle_{W_1([0,1])} \\ &= \sum_{k=1}^n\sum_{l=1}^m\prod_{j=1}^d\left(\int_0^1 h_{kj}(\tau)g_{lj}(\tau)d\tau + \int_0^1 h'_{kj}(\tau)g'_{lj}(\tau)d\tau\right)\end{aligned} \tag{41}$$

for $h(\mathbf{x}) = \sum_{k=1}^n \prod_{j=1}^d h_{kj}(\mathbf{x}^j)$, $g(\mathbf{x}) = \sum_{l=1}^m \prod_{j=1}^d g_{lj}(\mathbf{x}^j)$. The component functions h_{kj}, g_{lj} all belong to $W_1([0,1])$. Then the RKHS-norm (induced by the inner product) for a function $h \in \mathcal{H}_{k^d}$ is:

$$\begin{aligned}\|h\|_{k^d} &= \sum_{k=1}^n\sum_{l=1}^n\prod_{j=1}^d\left(\int_0^1 h_{kj}(\tau)h_{lj}(\tau)d\tau + \int_0^1 h'_{kj}(\tau)h'_{lj}(\tau)d\tau\right) \\ &\overset{\text{(i)}}{=} \sum_{\|\mathbf{a}\|_\infty\le 1}\|D^{\mathbf{a}}h\|^2_{L_2([0,1]^d)}.\end{aligned} \tag{42}$$

The above step (i) can be checked directly using Fubini's theorem. We present the explicit calculation for the case when $h(\mathbf{x}) = \prod_{j=1}^d h_j(\mathbf{x}^j)$ and $d=2$:

$$\begin{aligned}&\|h\|_{k^2} \\ &= \int_0^1 h_1^2(\tau_1)d\tau_1 \cdot \int_0^1 h_2^2(\tau_2)d\tau_2 + \int_0^1 \big(h_1(\tau_1)\big)^2 d\tau_1 \cdot \int_0^1 \big(h'_2(\tau_2)\big)^2 d\tau_2\end{aligned}$$

$$
\begin{aligned}
&+ \int_0^1 \big(h_1'(\tau_1)\big)^2 d\tau_1 \cdot \int_0^1 h_2^2(\tau_2) d\tau_2 + \int_0^1 \big(h_1'(\tau_1)\big)^2 d\tau_1 \cdot \int_0^1 \big(h_2'(\tau_2)\big)^2 d\tau_2 \\
&= \int_{[0,1]^2} h^2(\tau_1,\tau_2) d\tau_1 d\tau_2 + \int_{[0,1]^2} \left(\frac{\partial}{\partial \tau_2} h(\tau_1,\tau_2)\right)^2 d\tau_1 d\tau_2 \\
&\quad + \int_{[0,1]^2} \left(\frac{\partial}{\partial \tau_1} h(\tau_1,\tau_2)\right)^2 d\tau_1 d\tau_2 + \int_{[0,1]^2} \left(\frac{\partial^2}{\partial \tau_1 \partial \tau_2} h(\tau_1,\tau_2)\right)^2 d\tau_1 d\tau_2 \qquad (43)
\end{aligned}
$$

Briefly, the $W_1([0,1])$ space is an example of a univariate RKHS; the $S_1([0,1])$ space, when equipped with a proper inner product, is the tensor product extension of $W_1([0,1])$. Moreover, Proposition B.6 implies an equivalent way to express the RKHS inner product and its induced norm. Specifically, we know that

$$
\sum_{\|\mathbf{a}\|_\infty \le 1} \|D^{\mathbf{a}} h\|^2_{L_2([0,1]^d)} = \|h\|_{k^d} = \sum_{\mathbf{j} \in (\mathbb{N}^+)^d} \left(\prod_{k=1}^d \mathbf{j}^k\right)^2 \beta_{\mathbf{j}}^2 \qquad (44)
$$

for $h = \sum_{\mathbf{j} \in (\mathbb{N}^+)^d} \beta_{\mathbf{j}} \psi_{\mathbf{j}}$ (Proposition B.6 gives the second equality, $\lambda_{\mathbf{j}^k} = (\mathbf{j}^k)^{-1}$). Recall that the multivariate basis $\psi_{\mathbf{j}} = \prod_{k=1}^d \phi_{\mathbf{j}^k}$ is the product of the cosine functions (defined in (32)). □

B.3. Proof of Corollary 6.3

In the main text, we discussed the asymptotic order of the eigenvalues of tensor product kernels. Now we give a direct and concise proof of those results.

Proof. Since the Mercer expansion

$$
\sum_{j=1}^\infty j^{-2s} \phi_j(x) \phi_j(z) \qquad (45)
$$

converges to K absolutely and uniformly, we can switch the order of product and taking limit when engaging with the product kernel. Formally

$$
\begin{aligned}
K^{prod}(\mathbf{x},\mathbf{z}) &= \prod_{k=1}^d K\big(\mathbf{x}^k, \mathbf{z}^k\big) \\
&= \prod_{k=1}^d \left(\sum_{j=1}^\infty j^{-2s} \phi_j\big(\mathbf{x}^k\big) \phi_j\big(\mathbf{z}^k\big)\right) \\
&= \sum_{j_1=1}^\infty \cdots \sum_{j_d=1}^\infty \left(\prod_{k=1}^d j_k\right)^{-2s} \left(\prod_{k=1}^d \phi_{j_k}\big(\mathbf{x}^k\big)\right)\left(\prod_{k=1}^d \phi_{j_k}\big(\mathbf{z}^k\big)\right) \\
&= \sum_{\mathbf{j} \in \mathbb{N}^+} \left(\prod_{k=1}^d \mathbf{j}^k\right)^{-2s} \psi_{\mathbf{j}}(\mathbf{x}) \psi_{\mathbf{j}}(\mathbf{z}) =: \sum_{\mathbf{j} \in \mathbb{N}^+} \lambda_{\mathbf{j}} \psi_{\mathbf{j}}(\mathbf{x}) \psi_{\mathbf{j}}(\mathbf{z}) \qquad (46)
\end{aligned}
$$

To obtain the Mercer expansion of K^{prod} with non-increasing λ_j, we need to reorder the sum over d-tuples into a sum over $\mathbb{N}^+$. Observing the definition of $\lambda_{\mathbf{j}}$, we know the unravelling rule should be the product of grid index $\mathbf{j}$: $c_{\mathbf{j}} = \prod_{k=1}^{d} \mathbf{j}^k$. We use this rule c to unravel both $\lambda_{\mathbf{j}}$ and $\psi_{\mathbf{j}}$ into sequences λ_j, ψ_j.

$$K^{prod}(\mathbf{x}, \mathbf{z}) = \sum_{j=1}^{\infty} \lambda_j \psi_j(\mathbf{x}) \psi_j(\mathbf{z}). \tag{47}$$

The magnitude of λ_j is given in Corollary C.4, which is a result of applying the average order of divisor functions. □

Appendix C: Unravelling and approximation results

In the rest of the paper, we will switch from concrete example spaces to more abstract Sobolev ellipsoid-type spaces. The (univariate) Sobolev ellipsoid has been a benchmark model in the literature of sieve estimators: We just showed how it relates to multivariate spaces. In the multivariate case, we will be engaging with a true function f^0 that belongs to the multivariate Sobolev "ellipsoid":

$$f^0 \in \left\{ f = \sum_{\mathbf{j} \in (\mathbb{N}^+)^d} \beta_{\mathbf{j}} \psi_{\mathbf{j}} \;\middle|\; \sum_{\mathbf{j} \in (\mathbb{N}^+)^d} \left(\prod_{k=1}^{d} \mathbf{j}^k \right)^{2s} \beta_{\mathbf{j}}^2 \le Q^2 \right\}. \tag{48}$$

for some product basis $\psi_{\mathbf{j}}$. In particular, we assume the regression function can be expanded as an infinite linear combination of a set of basis functions $\psi_{\mathbf{j}}$ indexed by d-tuples. At the same time, we require $\beta_{\mathbf{j}}$ to converge to zero at a fast enough rate as the product of index $\mathbf{j}$ goes to infinity. The function space in (48) is the same as a ball in some multivariate RKHS (as illustrated in Lemma B.7). We also introduced another parameter s that determines the decay rate of $\beta_{\mathbf{j}}$, which is often interpreted as a smoothness parameter ([56], Chapter 2).

C.1. Magnitude of unravelled series

In this section we will first quantify the asymptotic behavior of unravelled series c_j, which is depicted in the right panel of Fig. 1. We will use these results to reduce Sobolev ellipsoids indexed by D-tuples (48) to those indexed by natural numbers. This will directly lead to some useful approximation results in multivariate tensor product spaces.

In general, we cannot give a closed form for the unravelled sequence C_j as function of j (in Algorithm 1 we gave an algorithm to generate finitely many elements). However, it is still possible to derive some results on the magnitude of c_j as a function of j. To this end, we first introduce the concept of a divisor function.

Definition C.1. We use $\tau_D(\cdot) : \mathbb{N}^+ \to \mathbb{N}^+$ to denote the D-th divisor function, which counts the number of unique ways to factor n as a product of D positive

integers (where order matters). Formally,

$$\tau_D(n) = \sum_{\substack{(\mathbf{j}^1,\dots,\mathbf{j}^D)\in(\mathbb{N}^+)^D \\ \prod_{k=1}^D \mathbf{j}^k = n}} 1 \tag{49}$$

The divisor function τ_D distinguishes the order of factorization: For example $\tau_2(4) = 3$ because there are 3 ways to write 4 as a product of 2 numbers: $4 = 1 \times 4 = 4 \times 1 = 2 \times 2$. In the exposition of this section, we also need to engage with the following partial sum of divisor functions.

Definition C.2. We define the sequence $T_D(x)$ to be the sum of the D-divisor function evaluated at the first $\lfloor x \rfloor$ positive natural numbers, that is

$$T_D(x) = \sum_{n\le x} \tau_D(n). \tag{50}$$

Clearly, $T_D(x)$ is the number of D-tuples $\mathbf{j} = (\mathbf{j}^1,\dots,\mathbf{j}^D) \in (\mathbb{N}^+)^D$ with $\prod_{k=1}^D \mathbf{j}^k \le x$. The number x is not necessarily an integer: the summation index $n \le x$ should be interpreted as $\{1,2,\dots,\lfloor x \rfloor\}$.

The first several elements in c_j (depicted in Fig. 1) are $1,2,2,3,3,4,4,4,\dots$. As our readers may notice, each natural number n shows up exactly $\tau_2(n)$ times: if we know (on average) how many ways there are to factor a positive integer, we can sketch the general magnitude of the unravelled sequence as well. The following lemma formalizes such an idea.

Lemma C.3. *Define* $c_{\mathbf{j}} = \prod_{k=1}^D \mathbf{j}^k$ *as a function on the* D*-tuple* $\mathbf{j} = (\mathbf{j}^1,\dots,\mathbf{j}^D) \in (\mathbb{N}^+)^D$*. Let* c_j *be the c-unravelling sequence of* $c_{\mathbf{j}}$ *(see Definition 6.2). Then, for* D *fixed, we know its asymptotic magnitude is:*

$$c_j = \Theta\big(j \log^{-(D-1)} j\big) \tag{51}$$

Proof. All the elements of c_j are positive integers since they are products of positive integers. And every positive integer shows up in c_j at least once. We also observe that there are repeated elements in c_j: For any positive integer m, it shows up exactly $\tau_D(m)$ times in the sequence c_j.

To determine the increase rate of c_j, it is enough to determine the largest b_j such that

$$T_D(b_j) = \sum_{m=1}^{b_j} \tau_D(m) \le j. \tag{52}$$

The unravelling sequence c_j increases at the same rate as b_j. To quantify the summation on the LHS, we need to use the following result from number theory:

$$\sum_{m=1}^{x} \tau_D(m) = \frac{\log^{D-1} x}{(D-1)!} x + O\big(x \log^{D-2} x\big), \tag{53}$$

where the big O notation indicates $x \to \infty$ (but D is fixed). If we divide both sides by x, then we know: on average, there are $(\log x)^{D-1}$ ways to factorize

a natural number into a product of D natural numbers. This result has been established in the literature of number theory, we give more discussion and references in Appendix F. For the special case when $D = 2$, there are sharper results available, e.g. Theorem 3.2 in Tenenbaum [47].

Let $b_j = \lfloor (D-1)! j \log^{-(D-1)} j \rfloor$. Plug b_j into (53):

$$\begin{aligned}\sum_{m=1}^{b_j} \tau_D(m) &= b_j \log^{D-1} b_j + O\big(b_j \log^{D-2} b_j\big) \\ &= \Theta\big(j(\log j)^{-(D-1)} \log^{D-1}\{j(\log j)^{-(D-1)}\}\big) = \Theta(j)\end{aligned} \tag{54}$$

It is direct to check that if $b_j = q_j j \log^{-(D-1)} j$ for any positive $q_j \to \infty$, $b_j \log^{D-1} b_j$ would diverge at a rate faster than j. So we know the largest b_j we can take is of order $j \log^{-(D-1)} j$, which concludes our proof. □

Corollary C.4. *Let $c_{\mathbf{j}} = \prod_{k=1}^{D} (\mathbf{j}^k)^s$ be a function defined on the D-tuple $\mathbf{j} = (\mathbf{j}^1, \ldots, \mathbf{j}^D) \in (\mathbb{N}^+)^D$ for some $s > 0$. Let c_j be the c-unravelling sequence of $c_{\mathbf{j}}$. Then we know*

$$c_j = \Theta\big(\big(j \log^{-(D-1)} j\big)^s\big) \tag{55}$$

(the notation $(\mathbf{j}^k)^s$ means the s-th power of the j-th entry of vector $\mathbf{j}$).

The next theorem is the main result in this section, which uses Lemma C.3 or Corollary C.4.

Theorem C.5. *Let $W(s, Q, \{\psi_{\mathbf{j}}\})$ be the multivariate product Sobolev space:*

$$W\big(s, Q, \{\psi_{\mathbf{j}}\}\big) = \Big\{ f = \sum_{\mathbf{j} \in (\mathbb{N}^+)^D} \beta_{\mathbf{j}} \psi_{\mathbf{j}}, \text{ for some } \beta_{\mathbf{j}} \in \mathbb{R} \;\Big|\; \sum_{\mathbf{j} \in (\mathbb{N}^+)^D} c_{\mathbf{j}}^{2s} \beta_{\mathbf{j}}^2 \le Q^2 \Big\}, \tag{56}$$

where $c_{\mathbf{j}} = \prod_{k=1}^{D} \mathbf{j}^k$ for $\mathbf{j} = (\mathbf{j}^1, \ldots, \mathbf{j}^D) \in (\mathbb{N}^+)^D$. Denote (ψ_j) as the c-unravelling sequence of $\{\psi_{\mathbf{j}}\}$.

Then there exists two constants $C_i(s, D)$, $i \in \{1, 2\}$ such that

$$\begin{aligned}&\Big\{ f = \sum_{j=1}^{\infty} \beta_j \psi_j, \text{ for some } \beta_j \in \mathbb{R} \;\Big|\; \sum_{j=1}^{\infty} \Big(\frac{j}{\log^{D-1} j \vee 1} \Big)^{2s} \beta_j^2 \le C_1(s, D) Q^2 \Big\} \\ \subset\;& W\big(s, Q, \{\psi_{\mathbf{j}}\}\big) \\ \subset&\Big\{ f = \sum_{j=1}^{\infty} \beta_j \psi_j, \text{ for some } \beta_j \in \mathbb{R} \;\Big|\; \sum_{j=1}^{\infty} \Big(\frac{j}{\log^{D-1} j \vee 1} \Big)^{2s} \beta_j^2 \le C_2(s, D) Q^2 \Big\}\end{aligned} \tag{57}$$

In plain(er) language, Theorem C.5 states that: The multivariate function space $W(s, Q, \{\psi_{\mathbf{j}}\})$ can be sandwiched between two formally simpler function spaces. These "bread" function spaces in (57) are still multivariate function spaces, but the basis functions (ψ_j) are listed in a sequence. In contrast, $W(s, Q, \{\psi_{\mathbf{j}}\})$ has basis functions indexed by D-tuples.

Proof. The multivariate ellipsoid $W(s, Q, \{\psi_{\mathbf{j}}\})$ is exactly the same space as:

$$\left\{ f = \sum_{j=1}^{\infty} \beta_j \psi_j \;\middle|\; \sum_{j=1}^{\infty} c_j^{2s} \beta_j^2 \leq Q^2 \right\}, \tag{58}$$

where c_j, β_j, ψ_j are the c-unravelling sequences of $c_{\mathbf{j}}, \beta_{\mathbf{j}}, \psi_{\mathbf{j}}$, respectively. In this step we performed nothing but a change of notation.

According to Corollary C.4, c_j is asymptotically of the same order as $(j \log^{-(D-1)} j)^{2s}$ as $j \to \infty$. Define $b_j = (\frac{j}{\log^{D-1} j \vee 1})^{2s}$, then we know that there exist constants C_1, C_2 (that only depends on s, and D) such that $C_1 b_j \leq c_j \leq C_2 b_j$ for all $j \in \mathbb{N}^+$. Plugging this in to (58) concludes our proof. □

As a direct result, we have the following corollary for the space S_1 in the main text.

Corollary C.6. *Let $S_1(Q)$ be the a ball in $S_1([0,1]^d)$:*

$$\left\{ f \in L_2\big([0,1]^d\big) \mid \sum_{\|\mathbf{a}\|_\infty \leq 1} \left\| D^{\mathbf{a}} f \right\|^2_{L_2([0,1]^d)} \leq Q^2 \right\} \tag{59}$$

Then we know it is sandwiched between two Sobolev-ellipsoids of single indices:

$$\begin{aligned} &\left\{ f = \sum_{j=1}^{\infty} \beta_j \psi_j, \text{ for some } \beta_j \in \mathbb{R} \mid \sum_{j=1}^{\infty} \left(\frac{j}{\log^{D-1} j \vee 1} \right)^{2s} \beta_j^2 \leq C_1(s, D) Q^2 \right\} \\ \subset \quad & S_1(Q) \\ \subset &\left\{ f = \sum_{j=1}^{\infty} \beta_j \psi_j, \text{ for some } \beta_j \in \mathbb{R} \mid \sum_{j=1}^{\infty} \left(\frac{j}{\log^{D-1} j \vee 1} \right)^{2s} \beta_j^2 \leq C_2(s, D) Q^2 \right\}, \end{aligned} \tag{60}$$

where (ψ_j) is the c^{CPR}-unravelling sequence of $\{\psi_{\mathbf{j}}\}$. The multivariate basis $\psi_{\mathbf{j}} = \prod_{k=1}^d \phi_{\mathbf{j}^k}$ is the product of the cosine functions (ϕ_j defined in (32)).

Proof. Combine Lemma B.7 and Theorem C.5. □

C.2. Approximation in dense tensor product models

In this section, we will use the results in Theorem C.5 to derive some approximation results that are crucial to understand the performance of our sieve estimators. Before we go into more detail, we provide some intuitive discussion of why Theorem C.5 can simplify our analysis. Let's denote the three function spaces in (57) as W_1, W_2 and W_3 ($W_1 \subset W_2 \subset W_3$). To study the problem of approximation/estimation of functions in W_2, it is equivalent – up to a constant – to study the corresponding problems in W_1 or W_3. The regression problem under the assumption $f^0 \in W_2$ is easier than assuming $f^0 \in W_3$ but harder than $f^0 \in W_1$. Therefore the generalization error of any estimators for truth

$f^0 \in W_2$ should be of the same order as $f^0 \in W_1$ or W_3. Similar statements also hold for minimax rates analysis.

Ellipsoids related to a (univariate) series c_j can be treated much more directly than those related to the D-tuple $c_{\mathbf{j}}$. For readers who are familiar with classical projection estimators (e.g. Tsybakov [49]), the following approximation results may appear familiar.

In the remainder of our discussion, we will use $\mathcal{X} \subset \mathbb{R}$ to denote a subset of real line and use ν to denote a (finite) Borel measure on $\mathcal{X}$. We do not need to specify either $\mathcal{X}$ or ν accurately: Often we just need $\mathcal{X}^d$ to be large enough to cover the support of feature distribution ρ_X, and in many important cases $\nu =$ uniform measure is enough for our purposes.

Lemma C.7. *Suppose that function f^* has an expansion $f^* = \sum_{j=1}^{\infty} \beta_j^* \psi_j$ with respect to a set of ν-orthonormal system, i.e. $\langle \psi_j, \psi_i \rangle_{L_2(\nu)} = \delta_{ij}$. Assume $\|\psi_j\|_\infty \le M$ for all j. If the expansion coefficients satisfy the following ellipsoid-type condition:*

$$f^* = \sum_{j=1}^{\infty} \beta_j^* \psi_j \in \left\{ f = \sum_{j=1}^{\infty} \beta_j \psi_j \in L_2(\nu) \;\Big|\; \sum_{j=1}^{\infty} \left(\frac{j}{\log^{D-1} j \vee 1} \right)^{2s} \beta_j^2 \le Q^2 \right\}, \tag{61}$$

with some $s > 1/2$. Then there exist a sequence of functions

$$f_n^* = \sum_{j=1}^{J_n} \beta_{nj}^* \psi_j \;\; \text{with } J_n = \left\lfloor \left(\log^{D-1} n\right)^{2s/(2s+1)} n^{1/(2s+1)} \right\rfloor, \quad n = 2, 3, \ldots \tag{62}$$

satisfy the following:

- *There is a constant $C(M, s, D, Q)$, such that for any n:*

$$\|f_n^*\|_\infty \le C(M, s, D, Q) \tag{63}$$

- *For any measure ρ_X that is absolute continuous to ν with a bounded density:*

$$\begin{aligned} \|f_n^* - f^*\|_{2,\rho_X}^2 &= \int \left\{ f_n^*(\mathbf{z}) - f^*(\mathbf{z}) \right\}^2 d\rho_X(\mathbf{z}) \\ &\le C(s, D, \rho_X, Q) \left(\frac{\log^{D-1} n}{n} \right)^{\frac{2s}{2s+1}} \end{aligned} \tag{64}$$

Proof. • We first prove the uniform bound in the $\|\cdot\|_\infty$-norm. According to our discussion Appendix B.2, a Sobolev-ellipsoid like (61) can be seen as a ball in an RKHS. That is, the functions f^*, f_n^* all belong to an RKHS with reproducing kernel

$$k(\mathbf{s}, \mathbf{t}) = \sum_{j=1}^{\infty} \lambda_j \psi_j(\mathbf{s}) \psi_j(\mathbf{t}), \tag{65}$$

where $\lambda_j = (\frac{\log^{D-1} j \vee 1}{j})^{2s}$. Denote the RKHS inner product as $\langle \cdot, \cdot \rangle_k$:

$$\begin{aligned} \|f_n^*\|_\infty &= \sup_{\mathbf{x}} f_n^*(\mathbf{x}) = \sup_x \big\langle f_n^*, k(\mathbf{x}, \cdot) \big\rangle_k \\ &\le \|f_n^*\|_k \sup_x \|k(\mathbf{x}, \cdot)\|_k \\ &\overset{(1)}{\le} QC(M, s, D). \end{aligned} \tag{66}$$

In step (1), we need the explicit representation of the RKHS norm (Theorem B.3). The RKHS norm of kernel k (centered at x) is

$$\|k(\mathbf{x}, \cdot)\|_k = \sum_{j=1}^{\infty} \big(\lambda_j \psi_j(\mathbf{x})\big)^2 / \lambda_j \le M^2 \sum_{j=1}^{\infty} \lambda_j = C(M, s, D)$$

- Next we prove the bound in ρ_X-2-norm. Let U denote a bound on the density of ρ_X (with respect to ν). We define f_n^* to be the projection of f^* (under $L_2(\rho_X)$ inner product) onto the linear space spanned by $\{\psi_j, j = 1, 2, \ldots, J_n\}$:

$$f_n^* = \underset{g \in span(\psi_1, \ldots, \psi_{J_n})}{\arg\min} \|g - f^*\|_{2,\rho_X}^2 \tag{67}$$

Then we have

$$\begin{aligned} \|f_n^* - f^*\|_{2,\rho_X}^2 &\le \left\| \sum_{j=1}^{J_n} \beta_j^* \psi_j - f^* \right\|_{2,\rho_X}^2 \\ &\le U \left\| \sum_{j=1}^{J_n} \beta_j^* \psi_j - f^* \right\|_{2,\nu}^2 = U \sum_{J_n+1}^{\infty} \big(\beta_j^*\big)^2 \\ &\le U \lambda_{J_n} \sum_{J_n+1}^{\infty} \big(\beta_j^*\big)^2 / \lambda_j \le U \lambda_{J_n} Q^2 \end{aligned} \tag{68}$$

We just need to determine the magnitude of λ_{J_n}:

$$\begin{aligned} &\lambda_{J_n} \le c J_n^{-2s} \big(\log^{D-1} J_n\big)^{2s} \\ &= c \Big\{ \big(\log^{D-1} n\big)^{\frac{2s}{2s+1}} n^{1/(2s+1)} \Big\}^{-2s} \Big[\log^{D-1} \Big\{ \big(\log^{D-1} n\big)^{\frac{2s}{2s+1}} n^{1/(2s+1)} \Big\} \Big]^{2s} \\ &\le C(s, D) n^{-\frac{2s}{2s+1}} \big(\log^{D-1} n\big)^{-\frac{4s^2}{2s+1} + 2s} = C(s, D) \left(\frac{\log^{D-1} n}{n} \right)^{\frac{2s}{2s+1}}, \end{aligned} \tag{69}$$

which concludes our proof. □

C.3. Covering number of S_1 spaces

After establishing the approximation results, we can derive the covering number of a ball in $S_1([0,1]^d)$ space. Although the covering number results are not directly used to prove our estimators' performance, we find them can be helpful to more intuitively understand the size of S_1 in contrast to the isotropic Sobolev spaces (see main text (11)).

Proposition C.8. *A unit ball in $S_1 = S_1([0,1]^d)$ space*

$$S_1(\textit{unit ball}) = \left\{ f \in L_2\big([0,1]^d\big) \mid \int \big(D^{\mathbf{a}} f(\mathbf{x})\big)^2 d\mathbf{x} \le 1 \textit{ for all } \|\mathbf{a}\|_\infty \le 1 \right\} \tag{70}$$

has a covering number of order

$$\log N\big(\delta, S_1(\textit{unit ball})\big) \asymp \delta^{-1} \log^{d-1}(1/\delta). \tag{71}$$

Proof. We first apply Corollary C.6 to reduce the problem to solving the covering number of

$$\mathcal{F}_{single} = \left\{ f = \sum_{j=1}^{\infty} \theta_j \psi_j \mid \sum_{j=1}^{\infty} \left(\frac{j}{\log^{d-1} j \vee 1} \right)^2 \theta_j^2 \le C(d) \right\}, \tag{72}$$

where (ψ_j) is the unravelled sequence of product cosine basis.

Note that $\mathcal{F}_{single}$ is a subspace of $L_2([0,1]^d)$ (equipped with Lebesgue measure). The basis functions ψ_js are orthonormal, therefore each ϵ-covering of the function space has a one-to-one correspondence to a covering of the following subspace of $\ell^2(\mathbb{N}^+)$:

$$\mathcal{E} = \left\{ (\theta_j)_{j=1}^{\infty} \mid \sum_{j=1}^{\infty} \frac{\theta_j^2}{\mu_j} \le 1 \right\}, \tag{73}$$

with $\mu_j = (\log^{d-1} j \vee 1/j)^2$.

We are going to show that

$$\log N(\delta, \mathcal{E}) \asymp \delta^{-1} \log^{d-1}(1/\delta) \quad \text{for all suitably small } \delta > 0 \tag{74}$$

The rest of our argument is standard (e.g., Example 5.12 of [58]).

Let J be the smallest integer such that $\mu_J \le \delta^2$, and consider the truncated ellipsoid

$$\widetilde{\mathcal{E}} = \{\theta \in \mathcal{E} \mid \theta_j = 0 \text{ for all } j \ge J+1\} \tag{75}$$

We claim that any δ-cover of this truncated ellipsoid, say $\{\theta^1, \ldots, \theta^N\}$, forms a $\sqrt{2}\delta$-cover of the full ellipsoid. Indeed, for any $\theta \in \mathcal{E}$, we have

$$\sum_{j=J+1}^{\infty} \theta_j^2 \le \mu_J \sum_{j=J+1}^{\infty} \frac{\theta_j^2}{\mu_j} \le \delta^2 \tag{76}$$

and hence

$$\min_{k \in [N]} \big\|\theta - \theta^k\big\|_2^2 = \min_{k \in [N]} \sum_{j=1}^{J} \big(\theta_j - \theta_j^k\big)^2 + \sum_{j=J+1}^{\infty} \theta_j^2 \le 2\delta^2 \tag{77}$$

Consequently, it suffices to upper bound the cardinality N of this covering of $\widetilde{\mathcal{E}}$. Since $\delta^2 \le \mu_J$ for all $j \in \{1, \ldots, J\}$, if we view $\widetilde{\mathcal{E}}$ as a subset of $\mathbb{R}^J$, then it

contains the (2-norm) δ-ball $\mathbb{B}_2^J(\delta)$, and hence $\text{vol}(\widetilde{\mathcal{E}} + \mathbb{B}_2^J(\delta/2)) \leq \text{vol}(2\widetilde{\mathcal{E}})$ (vol() stands for volume). Consequently, by Lemma 5.7 of [58], we have

$$N \leq \left(\frac{2}{\delta}\right)^J \frac{\text{vol}(\widetilde{\mathcal{E}} + \mathbb{B}_2^J(\delta/2))}{\text{vol}(\mathbb{B}_2^J(1))} \leq \left(\frac{4}{\delta}\right)^J \frac{\text{vol}(\widetilde{\mathcal{E}})}{\text{vol}(\mathbb{B}_2^J(1))} \tag{78}$$

By standard formulae for the volume of ellipsoids, we have $\frac{\text{vol}(\widetilde{\mathcal{E}})}{\text{vol}(\mathbb{B}_2^J(1))} = \prod_{j=1}^J \sqrt{\mu_j}$. Putting together the pieces, we find that

$$\begin{aligned} \log N &\leq J \log(4/\delta) + \frac{1}{2}\sum_{j=1}^J \log \mu_j \\ &= J \log(4/\delta) - \sum_{j=1}^J \log j + (d-1)\log\log j \\ &\overset{\text{(i)}}{\leq} J\log(4/\delta) + J - J\log J + (d-1)J\log\log J \\ &= (\log 4 + 1)J + J\big\{\log(1/\delta) - \log J + (d-1)\log\log J\big\} \end{aligned} \tag{79}$$

where step (i) used the inequality $\sum_{j=1}^J \log j \geq J\log J - J$. By definition:

$$\begin{aligned} \mu_J &\leq \delta^2 \\ \Rightarrow J^{-1}\log^{d-1} J &\leq \delta \\ \Rightarrow \log J - (d-1)\log\log J &\geq \log(1/\delta) \end{aligned} \tag{80}$$

So we can bound (79) by

$$\log N \leq (\log 4 + 1)J \tag{81}$$

It is direct to verify that J cannot be larger than $C\delta^{-1}\log^{d-1}(1/\delta)$ – constant C cannot be replace by any decreasing function $C(\delta)$ – so we conclude that

$$\log N \lesssim \delta^{-1}\log^{d-1}(1/\delta) \tag{82}$$

For the lower bound, we note that the ellipsoid $\mathcal{E}$ contains the truncated ellipsoid $\widetilde{\mathcal{E}}$, which (when viewed as a subset of $\mathbb{R}^J$) contains the ball $\mathbb{B}_2^J(\delta)$. Thus, we have

$$\log N\left(\frac{\delta}{2}, \mathcal{E}\right) \geq \log N\left(\frac{\delta}{2}, \mathbb{B}_2^d(\delta)\right) \geq J\log 2 \tag{83}$$

where the final inequality uses the lower bound (5.9) from Example 5.8 in [58]. Given the inequality $J \geq C\delta^{-1}\log^{d-1}(1/\delta)$, we have established the lower bound in our original claim (74). □

C.4. Approximation in sparse tensor product models

In Section C.2 we investigated the approximation error under dense tensor product models ($d = D$). In this section we will switch to the sparse/ higher dimensional setting where $d > D$.

Now we present some more general conditions on the product basis and sparse nonparametric models. They can be seen as generalization of Condition 7.1 and Condition 7.2 in the main text.

Notation: recall that $\mathcal{X} \subset \mathbb{R}$ is a subset of real line and ν is a Borel measure on $\mathcal{X}$.

Condition C.9. Let ϕ_j be an orthonormal system of univariate functions, that is, $\langle \phi_i, \phi_j \rangle_{L_2(\nu)} = \delta_{ij}$. Assume $\phi_1 = 1$, $\|\phi_j\|_\infty \le M$ for all $j = 1, 2, \ldots$. Consider their natural d-dimensional product extension $\psi_{\mathbf{j}}(\mathbf{x}) = \prod_{k=1}^d \phi_{\mathbf{j}^k}(\mathbf{x}^k)$, denote (ψ_j) to be the c-unravelling sequence of $\{\psi_{\mathbf{j}}\}$. The unravelling rule $c_{\mathbf{j}}$ is defined as

$$c_{\mathbf{j}} = \begin{cases} \prod_{k=1}^d \mathbf{j}^k, & \text{if at most } D' \text{ entries of } \mathbf{j} \text{ are greater than 1} \\ \infty, & \text{otherwise} \end{cases} \tag{84}$$

Condition C.10. There exists a D-variate function $f^* : \mathcal{X}^D \to \mathbb{R}$ such that:

1. (feature sparsity) There is set of indices $\{k_1, \ldots, k_D\} \subset \{1, 2, \ldots, d\}$ such that for any $\mathbf{u} \in \mathcal{X}^d$,
$$f^0(\mathbf{u}) = f^*\big(\mathbf{u}^{k_1}, \mathbf{u}^{k_2}, \ldots, \mathbf{u}^{k_D}\big). \tag{85}$$
2. (smoothness assumptions) The function f^* satisfies the following ellipsoid condition:
$$f^* \in \left\{ f = \sum_{j=1}^\infty \beta_j \diamondsuit_j \;\Big|\; \sum_{j=1}^\infty \left(\frac{j}{\log^{D-1} j \vee 1} \right)^{2s} \beta_j^2 \le Q^2 \right\}. \tag{86}$$
The function sequence $(\diamondsuit_j)$ is the $\triangle$-unravelling of $\diamondsuit_{\mathbf{j}} = \prod_{l=1}^D \phi_{\mathbf{j}^l}(\mathbf{u}^{k_l})$, $\mathbf{j} \in (\mathbb{N}^+)^D$. And the unravelling rule is defined by $\triangle_{\mathbf{j}} = \prod_{l=1}^D \mathbf{j}^l$.

The first part in Condition C.10 is a feature sparsity assumption. Although f^0 formally is a function of d-dimensional vector $\mathbf{x}$ (d can be large), this assumption states that it can be completely described using a small subset of the dimensions of $\mathbf{x}$ (specifically, we assume it depends on D out of the d dimensions).

The second part in Condition C.10 is in nature a smoothness assumption, but expressed in a basis expansion/Sobolev ellipsoid fashion. The basis functions $\diamondsuit_{\mathbf{j}}$ and unravelling rules $\triangle_{\mathbf{j}}$ only engage with the informative features $(\mathbf{u}^{k_1}, \mathbf{u}^{k_2}, \ldots, \mathbf{u}^{k_D})$. According to Lemma C.7, if we use the first $J_n^{\text{oracle}} = \lfloor (\log^{D-1} n)^{2s/(2s+1)} n^{1/(2s+1)} \rfloor$ functions of $\diamondsuit_j$, we can construct a sequence of approximation functions $f_n^{\text{oracle}} = \sum_{j=1}^{J_n} \beta_{nj}^{\text{oracle}} \diamondsuit_j$ of f^* that satisfy

$$\|f_n^{\text{oracle}} - f^*\|_{2,\rho_X}^2 = O\left(\frac{\log^{D-1} n}{n} \right)^{\frac{2s}{2s+1}}. \tag{87}$$

However, in practice, we unfortunately do not have *a priori* accessible information of which D dimensions of $\mathbf{x}$ are important. We thus cannot just use the oracle basis $\diamondsuit_j$ that only depend on the D relevant dimensions. The basis functions we use in (14) take the form of $\psi_{\mathbf{j}} = \prod_{k=1}^{d} \phi_{\mathbf{j}^k}(\mathbf{x}^k)$, involving d univariate functions as described in Condition C.9. We are interested in how many functions we need to include in the sequence of ψ_j, such that we can achieve the same approximation error as f_n^{oracle}. The following Lemma tells us this number is exponential in the active dimension D (which we treat as a fixed number) but only polynomial in the ambient dimension d (which may formally increase with the sample size n). The polynomial dependence in d is important both theoretically and in practice.

Lemma C.11. *Assume f^0 satisfies Condition C.10. Denote (ψ_j) as the sequence of product basis functions in Condition C.9. If the working dimension D' in Condition C.9 is greater than or equal to the active dimension D in Condition C.10, then:*

- *The true regression function f^0 can be expanded with respect to ψ_j as well, that is,*
$$f^0 = \sum_{j=1}^{\infty} \beta_j^0 \psi_j, \text{ for } \beta_j^0 \in \mathbb{R}. \tag{88}$$
- *There exists a sequence of functions $f_{\beta_n^0} = \sum_{j=1}^{J_n} \beta_{nj}^0 \psi_j$ with*
$$J_n \leq C(s, D) d^{D'} n^{1/(2s+1)} \log^{D'-1} n$$
such that
$$\|f_{\beta_n^0}\|_\infty \leq C(M, s, D, Q) \tag{89}$$
and
$$\|f_{\beta_n^0} - f^0\|_{2,\rho_X}^2 \leq C(s, D, \rho_X, Q) \left(\frac{\log^{D-1} n}{n} \right)^{\frac{2s}{2s+1}}. \tag{90}$$

Proof. We introduce the mapping $1_{d\to D} : \mathbb{R}^d \to \mathbb{R}^D$ that only keeps the relevant dimensions of a feature $\mathbf{x}$:
$$1_{d\to D}(\mathbf{x}) = \left(\mathbf{x}^{k_1}, \ldots, \mathbf{x}^{k_D}\right), \tag{91}$$
where $k_1, \ldots, k_D$ are the informative dimension indices defined in Condition C.10. By assumption, the true regression function can be written as:
$$f^0(\mathbf{x}) = f^*\left(1_{d\to D}(\mathbf{x})\right) = \sum_{j=1}^{\infty} \beta_j^* \diamondsuit_j \left(1_{d\to D}(\mathbf{x})\right) = \sum_{j=1}^{\infty} \beta_j^* \diamondsuit_j \circ 1_{d\to D}(\mathbf{x}) \tag{92}$$
Each of the basis functions above, $\diamondsuit_j \circ 1_{d\to D}$, varies at most in D dimensions. The function set $\{\psi_j\}$ in Condition C.9 includes all the function product functions varying in at most D' dimensions. Since $\diamondsuit_j \circ 1_{d\to D}$ are also product functions, we

conclude $\{\diamondsuit_j \circ 1_{d\to D}, j \in \mathbb{N}\} \subset \{\psi_j, j \in \mathbb{N}\}$. Therefore f^0 also has an expansion with respect to ψ_j as in (88).

Approximating f^* satisfying Condition C.10 (equivalently, f^0), using the oracle basis $\diamondsuit_j$ in the ellipsoid assumption (86), is already studied in Lemma C.7. We know that we need the first $J_n^{\text{oracle}} = (\log^{D-1} n)^{2s/(2s+1)} n^{1/(2s+1)}$ basis elements from $\{\diamondsuit_j\}$ in order to achieve the desired approximation error. We claim that

$$\{\diamondsuit_1, \ldots, \diamondsuit_{J_n^{\text{oracle}}}\} \subset \left\{ \diamondsuit_{\mathbf{j}}, \mathbf{j} \in \left(\mathbb{N}^+\right)^D \;\middle|\; c_{\mathbf{j}} = \prod_{k=1}^{D} \mathbf{j}^k \le C(s,D) R_n \right\}, \tag{93}$$

where $R_n = n^{1/(2s+1)} (\log n)^{-(D-1)/(2s+1)}$. To see this, we need to apply the number theory results we used to establish the equivalence between ellipsoids. According to Lemma C.3 we know that

$$T_D\big(C(s,D)R_n\big) = C(s,D) n^{1/(2s+1)} \log^{\frac{(D-1)2s}{2s+1}} n + \text{ lower order terms} \ge J_n^{\text{oracle}}. \tag{94}$$

(recall that T_D is defined in Definition C.2). However, in practice we do not know the oracle features, so we can only work with ψ_j or $\psi_{\mathbf{j}}$ (not $\diamondsuit_j$ or $\diamondsuit_{\mathbf{j}}$). To approximate f^0 well, we need to choose J_n large enough so that all the functions below are included:

$$\left\{ \psi_{\mathbf{j}}, \mathbf{j} \in \left(\mathbb{N}^+\right)^d \;\middle|\; c_{\mathbf{j}} = \prod_{k=1}^{d} \mathbf{j}^k \le C(s,D) R_n \text{ and at most } D \text{ of } \mathbf{j}^k > 1 \right\}. \tag{95}$$

This ensures all the functions in the RHS of (93) are included (strictly speaking, their d-dimensional extensions are included). By our assumption that $D' > D$, we only need to select J_n large enough so that the following basis functions are all included:

$$\begin{aligned} &\left\{ \psi_{\mathbf{j}}, \mathbf{j} \in \left(\mathbb{N}^+\right)^d \;\middle|\; c_{\mathbf{j}} = \prod_{k=1}^{d} \mathbf{j}^k \le C(s,D) R_n \text{ and at most } D' \text{ of } \mathbf{j}^k > 1 \right\} \\ &= \bigcup_{m=1}^{\lfloor C(s,D)R_n \rfloor} \left\{ \psi_{\mathbf{j}} \;\middle|\; c_{\mathbf{j}} = \prod_{k=1}^{d} \mathbf{j}^k = m \text{ and at most } D' \text{ of } \mathbf{j}^k > 1 \right\} \end{aligned} \tag{96}$$

How many elements are there in (96)? We give the following bound:

$$\begin{aligned} &\#\text{ of elements in (96)} \le \\ &\underbrace{\sum_{m=1}^{C(s,D)R_n}}_{(I)} \underbrace{C_{D'}^d}_{(II)} \cdot \underbrace{\tau_{D'}(m)}_{(III)} = C_{D'}^d \sum_{m=1}^{C(s,D)R_n} \tau_{D'}(m) \\ &\le C_{D'}^d T_{D'}\big(C(s,D)R_n\big) \overset{(1)}{\le} C\big(s,D,D'\big) d^{D'} n^{1/(2s+1)} \log^{D'-1} n. \end{aligned} \tag{97}$$

In (1) we used Lemma F.3 to bound $T_{D'}(C(s,D)R_n)$ and the well-known bound on the binomial coefficients $C_{D'}^d \leq C(D')d^{D'}$. To help our readers understand the above calculation, we have the following comments on each term:

- (I): consider all the $\mathbf{j}$ whose product is m;
- (II): choose D' dimensions;
- (III): factorize m into a product of D'numbers;

Unravelling the functions set in (96) will give us at most the first

$$C\big(s,D,D'\big)d^{D'}n^{1/(2s+1)}\log^{D'-1}n$$

elements in (ψ_j). To achieve the desired approximation error bound, we do not need to use any additional basis elements. □

Appendix D: Theoretical guarantees of least-square sieve estimators

The proof of Theorem 5.1 is standard after establishing the approximation results like Lemma C.7. Recall that we used $\epsilon_i = Y_i - f^0(\mathbf{X}_i)$ to denote the noise variable, f^0 the true regression function, and f_n^{OLS} the ordinary least-square estimator over a sieve space. In this section, we will also use W_i to denote independent Rademacher random variables: $\text{pr}(W_i = 1) = \text{pr}(W_i = -1) = 0.5$.

D.1. Proof of Theorem 5.1

Proof. We first apply Corollary C.6 to reduce the problem to an estimation problem when f^0 belonging to the function space:

$$\mathcal{F}_{single} = \left\{ f = \sum_{j=1}^{\infty} \theta_j \psi_j \mid \sum_{j=1}^{\infty} \left(\frac{j}{\log^{d-1} j \vee 1} \right)^2 \theta_j^2 \leq C(d,Q) \right\}, \tag{98}$$

where (ψ_j) is the unravelled sequence of product cosine basis.

Assuming $f^0 \in S_1$ implies $f^0 \in \mathcal{F}_{single}$ with a large enough (but not depending on n) $C(d,Q)$.

The rest of proof consists of four steps:

1. We first derive bounds on a (local) Rademacher process using Dudley's integral (Theorem D.1), noting that a sieve linear space is a VC-subgraph.
2. From the bounds on Rademacher process, we can derive bounds on some relevant sub-Gaussian multiplier process (Theorem D.2).
3. After obtaining bounds on the multiplier process, we can use them to derive bounds on the distance between f_n^{OLS} and some deterministic oracle functions f_n^* (the peeling argument, Theorem D.4).
4. A final triangular inequality relates the estimation error and the approximation error.

Applying Theorem D.1, D.2 and D.4 sequentially with

$$J_n = \lfloor \left(\log^{d-1} n\right)^{2/3} n^{1/3} \rfloor$$

and

$$\delta_n = n^{-1/3} \log^{(d-1)/3+1/2} n,$$

we know that

$$\|f_n^{OLS} - f_n^*\|_{2,\rho_X} = O_p(\delta_n). \tag{99}$$

Here we used the deterministic oracle functions f_n^* defined in Lemma C.7 (setting $f^* = f^0, s = 1$).

The distance between f_n^{OLS} and f^0 can be decomposed as:

$$\|f_n^{OLS} - f^0\|_{2,\rho_X} \le \|f_n^{OLS} - f_n^*\|_{2,\rho_X} + \|f_n^* - f^0\|_{2,\rho_X} \tag{100}$$

According to Lemma C.7, we know that the approximation error $\|f_n^* - f^0\|_{2,\rho_X}$ is bounded by $n^{-1/3}(\log n)^{(d-1)/3}$. So we conclude that

$$\left\|f_n^{OLS} - f^0\right\|_{2,\rho_X} = O_P(\delta_n) = O_P\left(n^{-1/3} \log^{(d-1)/3+1/2} n\right) \tag{101}$$

□

D.2. Technical results for Theorem 5.1

Theorem D.1. *Let $\mathcal{F}_n(\delta)$ denote the local, linear space centered at oracle f_n^*:*

$$\mathcal{F}_n(\delta) = \left\{ f \in L_2(\rho_X) \mid f(\cdot) = \sum_{j=1}^{J_n} \beta_j \psi_j(\cdot), \|f\|_\infty \le M, \left\|f - f_n^*\right\|_{2,\rho_X} \le \delta \right\}, \tag{102}$$

with $J_n = Cn^{1/3} \log^{2(d-1)/3}(n)$. Then we have the following bound on the Rademacher process indexed by functions in $\mathcal{F}_n(\delta) - f_n^* = \{g - f_n^* \mid g \in \mathcal{F}_n(\delta)\}$:

$$E\left[\sup_{f \in \mathcal{F}_n(\delta) - f_n^*} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n W_i f(\mathbf{X}_i) \right| \right] \le \phi_n(\delta). \tag{103}$$

The function ϕ_n is defined as

$$\phi_n(\delta) = C J_n^{1/2} \delta \sqrt{\log\left(\frac{1}{\delta}\right)} \left(1 + \frac{J_n^{1/2} \delta \sqrt{\log(1/\delta)}}{\sqrt{n}\delta^2 M}\right) \tag{104}$$

The same bound also holds for the process related to $f(\mathbf{X}_i)(f^0 - f_n^*)(\mathbf{X}_i)$:

$$E\left[\sup_{f \in \mathcal{F}_n(\delta) - f_n^*} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n W_i f(\mathbf{X}_i)\left(f^0 - f_n^*\right)(\mathbf{X}_i) \right| \right] \le \phi_n(\delta) \tag{105}$$

Proof. We first note that $\mathcal{F}_n(\delta)$ is a subset of a J_n-dimensional linear space $\mathcal{F}_n$

$$\mathcal{F}_n = \left\{ f \in L_2(\rho_X) \mid f(\cdot) = \sum_{j=1}^{J_n} \beta_j \psi_j(\cdot), \|f\|_\infty \leq M \right\}, \tag{106}$$

which is a VC-subgraph of dimension no more than $J_n + 2$. (Regarding a finite-dimensional linear space being a VC-subgraph, see [53], Lemma 2.6.15 or [58], Proposition 4.20).

It is also known that for a VC-subgraph function class, we can bound its covering number with a function of its VC-dimension ([53], Theorem 2.6.7). In our case, we have

$$\sup_Q N\big(\epsilon \|F\|_{2,Q}, \mathcal{F}_n, L_2(Q)\big) \leq C J_n (16e)^{J_n} \left(\frac{1}{\epsilon}\right)^{2(J_n-1)}, \tag{107}$$

where N is the covering number of a function space and the supremum is over all discrete measures. Eq. (107) implies:

$$\begin{aligned} &\sup_Q \log\big(N\big(\epsilon \|F\|_{2,Q}, \mathcal{F}_n, L_2(Q)\big)\big) \\ &\leq \log C + 2J_n \log(4\sqrt{e}/\epsilon) + \log J_n \\ &\overset{(i)}{\leq} 2J_n \log(C/\epsilon). \end{aligned} \tag{108}$$

In step (i), one may need the elementary fact that $J_n^{(2J_n)^{-1}} \leq \sqrt{e}$.

This means that the (local) Dudley integral has the following bound:

$$\begin{aligned} &J\big(\delta, \mathcal{F}_n(\delta) - f_n^*, L_2\big) \\ &:= \sup_Q \int_0^\delta \sqrt{1 + \log N\big(\varepsilon \|F\|_{2,Q}, \mathcal{F}_n(\delta) - f_n^*, L_2(Q)\big)} d\varepsilon \\ &\leq \sup_Q \int_0^\delta \sqrt{1 + \log N\big(\varepsilon \|F\|_{2,Q}, \mathcal{F}_n, L_2(Q)\big)} d\varepsilon \\ &\lesssim \int_0^\delta \sqrt{J_n \log\left(\frac{C}{\epsilon}\right)} d\epsilon \\ &\lesssim \sqrt{J_n}\left(\int_{\sqrt{-\log \delta}}^\infty \exp(-\tau^2) d\tau + \delta\sqrt{\log(1/\delta)}\right) \\ &\lesssim J_n^{1/2} \delta \sqrt{\log\left(\frac{1}{\delta}\right)} \end{aligned} \tag{109}$$

Next, we relate the Rademacher process with the function space's covering number integral (Theorem 2.1, [52]):

$$E\left[\sup_{f \in \mathcal{F}_n(\delta) - f_n^*} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n W_i f(\mathbf{X}_i) \right| \right]$$

$$\lesssim J\big(\delta, \mathcal{F}_n(\delta) - f_n^*, L_2\big)\left(1 + \frac{J(\delta, \mathcal{F}_n(\delta) - f_n^*, L_2)}{\sqrt{n}\delta^2 \|F\|_{2,\rho_X}}\right)\|F\|_{2,\rho_X}$$
$$\lesssim J_n^{1/2}\delta\sqrt{\log\left(\frac{1}{\delta}\right)}\left(1 + \frac{J_n^{1/2}\delta\sqrt{\log(\frac{1}{\delta})}}{\sqrt{n}\delta^2 M}\right)$$
$$=: \phi_n(\delta) \tag{110}$$

The above argument can be repeated for the other multiplier process (104). We can treat $f(\mathbf{X}_i)(f^0 - f_n^*)(\mathbf{X}_i)$ as a single function $g(\mathbf{X}_i)$, and the supremum is taken over $g \in (\mathcal{F}_n(\delta) - f_n^*)(f^0 - f_n^*)$. This function space is just multiplying each function in $\mathcal{F}_n(\delta) - f_n^*$ by a non-random, uniformly bounded function., which is still a VC-subgraph class. □

Theorem D.2. *Under the same notation as in Theorem D.1.*

$$E\left[\sup_{f \in \mathcal{F}_n(\delta) - f_n^*}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \epsilon_i f(\mathbf{X}_i)\right|\right] \lesssim (J_n \log n)^{1/2}\delta_n \tag{111}$$

Here δ_n is any positive non-increasing sequence of form $\delta_n = n^{-a}\log^b n$, with some $a \in (0, 1/2), b > 0$, such that $(J_n/n)^{1/2} \lesssim \delta_n(\log(1/\delta_n))^{-1/2}$. Recall that $\epsilon_i = Y_i - f^0(X_i)$.

Proof. We note that under our choice of δ_k, the term

$$\left(1 + \frac{J_n^{1/2}\delta_k\sqrt{\log(\frac{1}{\delta_k})}}{\sqrt{n}\delta_k^2 M}\right) \le (1 + C/M) \tag{112}$$

can be bounded by a constant not depending on n for all $1 \le k \le n$. Therefore, applying Theorem D.1 we have:

$$\begin{aligned} &E\left[\sup_{f \in \mathcal{F}_n(\delta_k) - f_n^*}\left|\sum_{i=1}^k W_i f(\mathbf{X}_i)\right|\right] \\ &\lesssim J_n^{1/2}\sqrt{k}\delta_k\sqrt{\log\left(\frac{1}{\delta_k}\right)} \\ &\lesssim J_n^{1/2}k^{1/2-a}\log^{b+1/2} k \end{aligned} \tag{113}$$

for any $1 \le k \le n$. Then we can apply Theorem 1 of [19] to bound the sub-Gaussian process of interest using a function of the bounds of its corresponding Rademacher process. For our readers' ease of reference, we include the cited theorem here.

Theorem D.3. *Suppose $\mathbf{X}_i$, ϵ_i are all IID random variables and $\mathbf{X}_i$ are independent of ϵ_i. Let $\{\mathcal{G}_k\}_{k=1}^n$ be a sequence of function classes such that $\mathcal{G}_k \supset \mathcal{G}_n$ for any $1 \le k \le n$. Assume further that there exists a nondecreasing concave function $\phi_n : \mathbb{R}_{\ge 0} \to \mathbb{R}_{\ge 0}$ with $\phi_n(0) = 0$ such that*

$$E\sup_{f \in \mathcal{G}_k}\left|\sum_{i=1}^k W_i f(\mathbf{X}_i)\right| \le \phi_n(k)$$

holds for all $1 \le k \le n$. *Then*

$$E \sup_{f \in \mathcal{G}_n} \left| \sum_{i=1}^{n} \epsilon_i f(\mathbf{X}_i) \right| \le 4 \int_0^\infty \phi_n \left(\sum_{i=1}^{n} P\big(|\epsilon_i| > t\big) \right) \mathrm{d}t.$$

Apply Theorem D.3, we have

$$\begin{aligned} & E\left[\sup_{f \in \mathcal{F}_n(\delta_n) - f_n^*} \left| \sum_{i=1}^{n} \epsilon_i f(\mathbf{X}_i) \right| \right] \\ & \lesssim \int_0^\infty J_n^{1/2} n^{1/2-a} \mathrm{pr}^{1/2-a}(|\epsilon_1| > t) \log^{b+1/2} n dt \\ & \lesssim \int_0^\infty \mathrm{pr}^{1/2-a}(|\epsilon_1| > t) dt \cdot (n J_n \log n)^{1/2} \delta_n \end{aligned} \tag{114}$$

The quantity $\|\epsilon_1\|_{2/(1-2a),1} = \int_0^\infty pr^{(1-2a)/2}(|\epsilon_1| > t)dt$ term is known as the $L_{2/(1-2a),1}$-moment of ϵ. In general, any random variable having finite $\|\epsilon\|_{p+\Delta}$ (for any $\Delta > 0$) also has a finite $\|\epsilon\|_{p,1}$-moment. For sub-Gaussian noise ϵ_1, all moments exist. More background regarding $L_{p,1}$-moment, see Chapter 10 of [27]. □

Theorem D.4. *Let* $\mathcal{F}$ *be a class of function and* f_n^* *a non-random function in* $L_2(\rho_X)$. *Suppose that for any* $f \in \mathcal{F}$, $\|f - f_n^*\|_\infty \le B^*$. *If* $\mathcal{F}$ *is convex and*

$$\begin{aligned} & E \sup_{f \in \mathcal{F}: \|f - f_n^*\|_{2,\rho_X} \le \delta_n} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \epsilon_i \big(f - f_n^*\big)(\mathbf{X}_i) \right| \le \phi_n(\delta_n) \\ & E \sup_{f \in \mathcal{F}: \|f - f_n^*\|_{2,\rho_X} \le \delta_n} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} W_i \big(f - f_n^*\big)(\mathbf{X}_i) \right| \le \phi_n(\delta_n) \\ & E \sup_{f \in \mathcal{F}: \|f - f_n^*\|_{2,\rho_X} \le \delta_n} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} W_i \big(f - f_n^*\big)(\mathbf{X}_i) \big(f^0 - f_n^*\big)(\mathbf{X}_i) \right| \le \phi_n(\delta_n) \end{aligned} \tag{115}$$

for some ϕ_n *such that* $\delta \mapsto \phi_n(\delta)/\delta$ *is non-increasing. Then*

$$\big\| f_n^{OLS} - f_n^* \big\|_{2,\rho_X} = O_p(\delta_n) \tag{116}$$

for any δ_n *such that* $\phi_n(\delta_n) \le \sqrt{n}\delta_n^2$.

Proof. Recall that f_n^{OLS} is defined as the empirical loss minimizer:

$$f_n^{OLS} = \arg\min_{f \in \mathcal{F}} P_n \big(Y - f(\mathbf{X})\big)^2 \tag{117}$$

In what follows, we denote by $B(f_n^*, \delta_n)$ the collection of functions in $\mathcal{F}$ contained by $L_2(P)$ ball of radius less than or equal to δ, centered at f_n^*. The event we eventually want to control is given by:

$$\big\{ \|f_n^{OLS} - f_n^*\|_{2,\rho_X} \ge 2^M \delta_n \big\} \tag{118}$$

and we aim to show that the probability of the above event $\to 0$ as $M \to \infty$.

To relate this event with the empirical process assumptions in the lemma, we apply the peeling mechanism:

$$\begin{aligned}
&\{\|f_n^{OLS} - f_n^*\|_{2,\rho_X} \geq 2^M \delta_n\} \\
&\quad \subseteq \Big\{ \inf_{f \in B(f_n^*, 2^M\delta_n)^c} P_n\big(Y - f(X)\big)^2 \leq P_n\big(Y - f_n^*(X)\big)^2 \Big\} \\
&\quad = \Big\{ \inf_{f \in B(f_n^*, 2^M\delta_n)^c} P_n\big(Y - f(X)\big)^2 - P_n\big(Y - f_n^*(X)\big)^2 \leq 0 \Big\} \\
&\quad = \bigcup_{j=M}^{\infty} \Big\{ \inf_{2^j\delta_n \leq \|f - f_n^*\|_{2,\rho_X} \leq 2^{j+1}\delta_n} P_n\big(Y - f(X)\big)^2 - P_n\big(Y - f_n^*(X)\big)^2 \leq 0 \Big\}.
\end{aligned} \tag{119}$$

Therefore

$$\begin{aligned}
&\mathrm{pr}\big(\|f_n^{OLS} - f_n^*\|_{2,\rho_X} \geq 2^M\delta_n\big) \\
&\leq \sum_{j=M}^{\infty} \mathrm{pr}\Big(\inf_{2^j\delta_n \leq \|f - f_n^*\|_{2,\rho_X} \leq 2^{j+1}\delta_n} P_n\big(Y - f(X)\big)^2 - P_n\big(Y - f_n^*(X)\big)^2 \leq 0 \Big) \\
&= \sum_{j=M}^{\infty} \mathrm{pr}\Bigg(\inf_{2^j\delta_n \leq \|f - f_n^*\|_{2,\rho_X} \leq 2^{j+1}\delta_n} \mathbb{K}_n\big(f, f_n^*, f^0\big) \\
&\qquad\qquad \leq -P\big(f_n^* - f\big)^2 - 2P\big(f^0 - f_n^*\big)\big(f_n^* - f\big) \Bigg)
\end{aligned} \tag{120}$$

Here we used the notation

$$\begin{aligned}
\mathbb{K}_n\big(f, f_n^*, f^0\big) =& P_n\big(Y - f(X)\big)^2 - P_n\big(Y - f_n^*(X)\big)^2 \\
&- P\big(f_n^* - f\big)^2 - 2P\big(f^0 - f_n^*\big)\big(f_n^* - f\big)
\end{aligned} \tag{121}$$

We can further bound (120) as following:

$$\begin{aligned}
&\mathrm{pr}\big(\|f_n^{OLS} - f_n^*\| \geq 2^M\delta_n\big) \\
&\quad \overset{(1)}{\leq} \sum_{j=M}^{\infty} \mathrm{pr}\Big(\inf_{2^j\delta_n \leq \|f - f_n^*\|_{2,\rho_X} \leq 2^{j+1}\delta_n} \mathbb{K}_n\big(f, f_n^*, f^0\big) \leq -P\big(f_n^* - f\big)^2 \Big) \\
&\quad \leq \sum_{j=M}^{\infty} \mathrm{pr}\Big(\inf_{2^j\delta_n \leq \|f - f_n^*\|_{2,\rho_X} \leq 2^{j+1}\delta_n} \mathbb{K}_n\big(f, f_n^*, f^0\big) \leq -2^{2j}\delta_n^2 \Big) \\
&\quad \leq \sum_{j=M}^{\infty} \mathrm{pr}\Big(\sup_{2^j\delta_n \leq \|f - f_n^*\|_{2,\rho_X} \leq 2^{j+1}\delta_n} |\sqrt{n}\mathbb{K}_n\big(f, f_n^*, f^0\big)| \geq \sqrt{n}2^{2j}\delta_n^2 \Big) \\
&\quad \leq \sum_{j=M}^{\infty} E\Big[\sup_{2^j\delta_n \leq \|f - f_n^*\|_{2,\rho_X} \leq 2^{j+1}\delta_n} |\sqrt{n}\mathbb{K}_n\big(f, f_n^*, f^0\big)| \Big] / \big(\sqrt{n}2^{2j}\delta_n^2\big) \\
&\quad \leq \sum_{j=M}^{\infty} E\Big[\sup_{\|f - f_n^*\|_{2,\rho_X} \leq 2^{j+1}\delta_n} |\sqrt{n}\mathbb{K}_n\big(f, f_n^*, f^0\big)| \Big] / \big(\sqrt{n}2^{2j}\delta_n^2\big)
\end{aligned} \tag{122}$$

In step (1) we used the property that $\mathcal{F}$ is convex. In such a case, $P(f^0 - f_n^*)(f_n^* - f) > 0$. In fact, for any $0 < \delta < 1$:

$$\begin{aligned}
&P\big(f^0 - f_n^*\big)^2 \overset{(2)}{\leq} P\big(f^0 - (1-\delta)f_n^* - \delta f_n^{OLS}\big)^2 = P\big(f^0 - f_n^* + \delta\big(f_n^* - f_n^{OLS}\big)\big)^2 \\
&\quad = P\big(f^0 - f_n^*\big)^2 + 2\delta P\big(f^0 - f_n^*\big)\big(f_n^* - f_n^{OLS}\big) + \delta^2 P\big(f_n^* - f_n^{OLS}\big)^2 \\
&\Rightarrow 2P\big(f^0 - f_n^*\big)\big(f_n^* - f_n^{OLS}\big) \geq -\delta P\big(f_n^* - f_n^{OLS}\big)^2
\end{aligned}$$

and thus we conclude $P(f^0 - f_n^*)(f_n^* - f_n^{OLS} - L) \geq 0$ by taking $\delta \to 0$. In step (2) we used the definition of f_n^* as a $L^2(P)$-projection of f^0 onto $\mathcal{F}$. Also note that since $\mathcal{F}$ is convex, $(1-\delta)f_n^* + \delta f_n^{OLS}$ is also an element of $\mathcal{F}$.

Now we rearrange the expression in (122) to relate it with the empirical processes in our assumption.

$$\begin{aligned}
&P_n\big(Y - f(X)\big)^2 - P_n\big(Y - f_n^*(X)\big)^2 \\
&= P_n\big(Y - f^0(X)\big)^2 + P_n\big(f^0(X) - f(X)\big)^2 \\
&\qquad + 2P_n\big(Y - f^0(X)\big)\big(f^0(X) - f(X)\big) - P_n\big(Y - f^0(X)\big)^2 \\
&\qquad - P_n\big(f^0(X) - f_n^*(X)\big)^2 - 2P_n\big(Y - f^0(X)\big)\big(f^0(X) - f_n^*(X)\big) \\
&= P_n\big(f^0(X) - f(X)\big)^2 - P_n\big(f^0(X) - f_n^*(X)\big)^2 \\
&\qquad + 2P_n\big(Y - f^0(X)\big)\big(f_0^*(X) - f(X)\big) \qquad (123) \\
&= P_n\big(f^0(X) - f_n^*(X)\big)^2 + P_n\big(f_n^*(X) - f(X)\big)^2 \\
&\qquad + 2P_n\big(f^0(X) - f_n^*(X)\big)\big(f_0^*(X) - f(X)\big) - P_n\big(f^0(X) - f_n^*(X)\big)^2 \\
&\qquad + 2P_n\epsilon\big(f_n^*(X) - f(X)\big) \\
&= P_n\big(f_n^*(X) - f(X)\big)^2 + 2P_n\big(f^0(X) - f_n^*(X)\big)\big(f_0^*(X) - f(X)\big) \\
&\qquad + 2P_n\epsilon\big(f_n^*(X) - f(X)\big)
\end{aligned}$$

Subtract $P(f_n^* - f)^2 + 2P(f^0 - f_n^*)(f_n^* - f)$ on both sides, we have:

$$\begin{aligned}
&\mathbb{K}_n\big(f, f_n^*, f^0\big) = P_n\big(f_n^*(X) - f(X)\big)^2 + 2P_n\big(f^0(X) - f_n^*(X)\big)\big(f_0^*(X) - f(X)\big) \\
&\qquad + 2P_n\epsilon\big(f_n^*(X) - f(X)\big) - P\big(f_n^* - f\big)^2 - 2P\big(f^0 - f_n^*\big)\big(f_n^* - f\big) \\
&\Rightarrow \sqrt{n}\mathbb{K}_n\big(f, f_n^*, f^0\big) \\
&\qquad = \mathbb{G}_n\big(f_n^*(X) - f(X)\big)^2 + 2\mathbb{G}_n\big(f^0(X) - f_n^*(X)\big)\big(f_n^*(X) - f(X)\big) \\
&\qquad\quad + 2\mathbb{G}_n\epsilon\big(f_n^*(X) - f(X)\big), \qquad (124)
\end{aligned}$$

where $\mathbb{G}_n$ stands for the empirical process: $\mathbb{G}_n g(X, \epsilon) = \sqrt{n}(P_n - P)g(X, \epsilon)$.

Now we can continue (122):

$$\begin{aligned}
&\mathrm{pr}\big(\|f_n^{OLS} - f_n^*\| \geq 2^M \delta_n\big) \\
&\leq 2\sum_{j=M}^{\infty} \big(\sqrt{n}2^{2j}\delta_n^2\big)^{-1} \left(E\Big[\sup_{\|f - f_n^*\|_{2,\rho_X} \leq 2^{j+1}\delta_n} \big|\mathbb{G}_n\big(f_n^* - f\big)^2\big| \Big] \right.
\end{aligned}$$

$$+ E\Big[\sup_{\|f-f_n^*\|_{2,\rho_X}\le 2^{j+1}\delta_n} |\mathbb{G}_n\big(f^0-f_n^*\big)\big(f_n^*-f\big)|\Big]$$
$$+ E\Big[\sup_{\|f-f_n^*\|_{2,\rho_X}\le 2^{j+1}\delta_n} |\mathbb{G}_n\epsilon\big(f_n^*-f\big)|\Big]\Bigg) \tag{125}$$

Using a typical symmetrization argument (Section 3.3 of [37] for a detailed presentation), we have

$$E\Big[\sup_{\|f-f_n^*\|_{2,\rho_X}\le 2^{j+1}\delta_n} |\mathbb{G}_n\big(f_n^*-f\big)^2|\Big] \le 2E\left[\sup_{\|f-f_n^*\|_{2,\rho_X}\le 2^{j+1}\delta_n}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n W_i\big(f_n^*(X_i)-f(X_i)\big)^2\right|\right] \tag{126}$$

and

$$E\Big[\sup_{\|f-f_n^*\|_{2,\rho_X}\le 2^{j+1}\delta_n} |\mathbb{G}_n\big(f^0-f_n^*\big)\big(f_n^*-f\big)|\Big] \le 2E\left[\sup_{\|f-f_n^*\|_{2,\rho_X}\le 2^{j+1}\delta_n}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n W_i\big(f^0(X_i)-f_n^*(X_i)\big)\big(f_n^*(X_i)-f(X_i)\big)\right|\right] \tag{127}$$

By our assumption that $\mathcal{F}-f_n^*\in L_\infty(B^*)$, we know $\|f_n^*(X_i)-f(X_i)\|_\infty\le B^*$. Therefore, $(f_n^*(X_i)-f(X_i))^2$ is a $2B^*$-Lipschitz function of $f_n^*(X_i)-f(X_i)$. Apply Talagrand's contraction principal (Proof of Lemma 8.17 in [37] and the citation therein), we have

$$E\left[\sup_{\|f-f_n^*\|_{2,\rho_X}\le 2^{j+1}\delta_n}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n W_i\big(f_n^*(X_i)-f(X_i)\big)^2\right|\right] \le 2B^*E\left[\sup_{\|f-f_n^*\|_{2,\rho_X}\le 2^{j+1}\delta_n}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n W_i\big(f_n^*(X_i)-f(X_i)\big)\right|\right] \tag{128}$$

So the quantities in (125) can be further bounded by:

$$\begin{aligned}
&\mathrm{pr}\big(\|f_n^{OLS}-f_n^*\|\ge 2^M\delta_n\big)\\
&\le 8B^*\sum_{j=M}^\infty\big(\sqrt{n}2^{2j}\delta_n^2\big)^{-1}\left(E\left[\sup_{\|f-f_n^*\|_{2,\rho_X}\le 2^{j+1}\delta_n}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n W_i\big(f(X_i)-f_n^*(X_i)\big)\right|\right]\right.\\
&\quad + E\left[\sup_{\|f-f_n^*\|_{2,\rho_X}\le 2^{j+1}\delta_n}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n W_i\big(f^0(X_i)-f_n^*(X_i)\big)\big(f_n^*(X_i)-f(X_i)\big)\right|\right]\\
&\quad \left. + E\left[\sup_{\|f-f_n^*\|_{2,\rho_X}\le 2^{j+1}\delta_n}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n \epsilon_i\big(f_n^*(X_i)-f(X_i)\big)\right|\right]\right)\\
&\le 24B^*\sum_{j=M}^\infty\frac{\phi_n(2^{j+1}\delta_n)}{\sqrt{n}2^{2j}\delta_n^2}\overset{(3)}{\le} 24B^*\sum_{j=M}^\infty\frac{2^{j+1}\phi_n(\delta_n)}{\sqrt{n}2^{2j}\delta_n^2}\le 48B^*\sum_{j=M}^\infty 2^{-j}.
\end{aligned} \tag{129}$$

In step (3) we used the condition that $\phi_n(\delta)/\delta$ is a non-increasing function of δ. As $M \to \infty$,

$$\mathrm{pr}\big(\|f_n^{OLS} - f_n^*\|_{2,\rho_X} \geq 2^M \delta_n\big) \leq 48B^* \sum_{j=M}^{\infty} 2^{-j} \to 0, \tag{130}$$

for any deterministic sequences δ_n such that $\phi_n(\delta_n) \leq \sqrt{n}\delta_n^2$. □

Appendix E: Theoretical guarantees of penalized sieve estimators

To present the statistical guarantees of l_1-penalized sieve estimators, we are going to employ the following steps:

1. We will give nonparametric oracle inequalities to control the "training-design error" of the estimators and the deviation of the estimated regression coefficients (Corollary E.5).
2. We will use the information of the regression coefficients to derive a metric entropy bound on the function space the estimator lies in (Lemma E.8).
3. We will control the difference between the training and testing errors of the estimate using results from empirical process theory (Theorem E.10).

E.1. Nonparametric oracle inequalities

We first define the concept of the compatibility constant, which is an important component in the oracle inequalities and widely used in the analysis of penalized methods. In the rest of the section, for a $\beta = (\beta_1, \ldots, \beta_J)^\top \in \mathbb{R}^J$, we define its related function f_β as

$$f_\beta = \sum_{j=1}^{J} \beta_j \psi_j, \tag{131}$$

where (ψ_j) is the sequence of functions in Condition C.9.

Definition E.1. For a given matrix Σ of size $J \times J$, constant L, and an index set $S \subset \{1, 2, \ldots, J\}$, we define the (Σ, L, S)-compatibility constant $\phi_\Sigma(L, S)$ to be

$$\phi_\Sigma^2(L, S) = \min_{\beta} \left\{ \frac{|S|\beta^\top \Sigma \beta}{\|\beta_S\|_1^2} : \|\beta_{-S}\|_1 \leq L\|\beta_S\|_1 \right\}, \tag{132}$$

where $-S$ is the complementary set of S in $\{1, 2, \ldots, J\}$. The notation $\beta_S \in \mathbb{R}^J$ is a shorthand for the "restriction" of a vector $\beta \in \mathbb{R}^J$ on the index set S: $(\beta_S)_j = \beta_j$ if $j \in S$, otherwise $(\beta_S)_j = 0$.

The following oracle inequality is a generalization of Theorem 2.2 in Van de Geer [51]. In our case, the true regression function does not have to be linear.

Theorem E.2. *Let $(\mathbf{X}_i, Y_i)$, $i = 1, 2, \ldots, n$ denote the n IID samples. We use f^0 to denote the true conditional mean function and define $\epsilon_i = Y_i - f^0(\mathbf{X}_i)$.*

Let $J_n \geq 1$ be the number of basis function used in estimation. Let (ψ_j) be the unravelled sequence described in Condition C.9. Let λ_ϵ be a number satisfying:

$$\sup_{1\leq j\leq J_n}\left|\frac{1}{n}\sum_{i=1}^{n}\psi_j(\mathbf{X}_i)\epsilon_i\right| \leq \lambda_\epsilon \tag{133}$$

Let $0 \leq \delta < 1$ and define for $\lambda > \lambda_\epsilon > 0$:

$$\underline{\lambda} = \lambda - \lambda_\epsilon, \bar{\lambda} = \lambda + \lambda_\epsilon + \delta\underline{\lambda}, \text{ and } L = \frac{\bar{\lambda}}{(1-\delta)\lambda}. \tag{134}$$

We use $\hat{\beta}_n = (\beta_1^{PLS}, \ldots, \beta_{J_n}^{PLS})^\top$ to denote the minimizer of the penalized problem (14).

Then for any $\beta \in \mathbb{R}^{J_n}$ and any set $S \subset \{1, 2, \ldots, J_n\}$:

$$2\delta\underline{\lambda}\|\hat{\beta}_n - \beta\|_1 + \left\|f_{\hat{\beta}_n} - f^0\right\|_n^2 \leq \left\|f_\beta - f^0\right\|_n^2 + \frac{\bar{\lambda}^2|S|}{\phi^2_{\hat{\Sigma}}(L,S)} + 4\lambda\|\beta_{-S}\|_1 \tag{135}$$

where $\phi^2_{\hat{\Sigma}}(L,S)$ is the $(\hat{\Sigma}, L, S)$-compatibility constant and the $\hat{\Sigma}$ is the empirical covariance matrix: $\hat{\Sigma}_{ij} = \frac{1}{n}\sum_{k=1}^{n}\psi_i(\mathbf{X}_k)\psi_j(\mathbf{X}_k)$.

Proof. We define $2\diamondsuit = \|f_{\hat{\beta}_n} - f^0\|_n^2 - \|f_\beta - f^0\|_n^2 + \|f_{\hat{\beta}_n} - f_\beta\|_n^2$. The empirical norm $\|\cdot\|_n$ can also be written in matrix form:

$$\begin{aligned}\|f_{\hat{\beta}_n} - f^0\|_n^2 =& \frac{1}{n}\sum_{i=1}^{n}\left(f_{\hat{\beta}_n}(\mathbf{X}_i) - f^0(\mathbf{X}_i)\right)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{J_n}\beta_j^{PLS}\psi_j(\mathbf{X}_i) - f^0(\mathbf{X}_i)\right)^2 \\ &= \left\langle\hat{\Psi}\hat{\beta}_n - f^0(X), \hat{\Psi}\hat{\beta}_n - f^0(X)\right\rangle/n.\end{aligned} \tag{136}$$

The design matrix, $\hat{\Psi}$, has entries $\hat{\Psi}_{i,j} = \psi_j(\mathbf{X}_i)$. And $f^0(X) = (f^0(\mathbf{X}_1), \ldots, f^0(\mathbf{X}_n))^\top \in \mathbb{R}^n$ is the evaluation vector of f^0 at n features vectors $\{\mathbf{X}_i\}_{i=1}^n$. We will use the above equivalence later.

Similar to the proof in the literature [51], we consider two cases for $\diamondsuit$:

- If $\diamondsuit \leq -\delta\underline{\lambda}\|\hat{\beta}_n - \beta\|_1 + 2\lambda\|\beta_{-S}\|_1$. Then we have

$$\begin{aligned}&2\delta\underline{\lambda}\|\hat{\beta}_n - \beta\|_1 + \|f_{\hat{\beta}_n} - f^0\|_n^2 \\ &\quad= 2\delta\underline{\lambda}\|\hat{\beta}_n - \beta\|_1 + 2\diamondsuit + \|f_\beta - f^0\|_n^2 - \|f_{\hat{\beta}_n} - f_\beta\|_n^2 \\ &\quad\leq \|f_\beta - f^0\|_n^2 - \|f_{\hat{\beta}_n} - f_\beta\|_n^2 + 4\lambda\|\beta_{-S}\|_1 \\ &\quad\leq \|f_\beta - f^0\|_n^2 + 4\lambda\|\beta_{-S}\|_1\end{aligned} \tag{137}$$

- In the case when $\diamondsuit > -\delta\underline{\lambda}\|\hat{\beta}_n - \beta\|_1 + 2\lambda\|\beta_{-S}\|_1$, we start with the following *two point inequality* (Lemma 6.1 in Van de Geer [51]):

$$\left\langle\hat{\Psi}(\beta - \hat{\beta}_n), Y - \hat{\Psi}\hat{\beta}_n\right\rangle/n \leq \lambda\|\beta\|_1 - \lambda\|\hat{\beta}_n\|_1 \tag{138}$$

Using the results in the beginning of this proof, we know $\diamondsuit$ can be expanded as:

$$\diamondsuit = \langle \hat{\Psi}\hat{\beta}_n, \hat{\Psi}\hat{\beta}_n \rangle / n - \langle \hat{\Psi}\hat{\beta}_n, f^0(X) \rangle / n + \langle \hat{\Psi}\beta, f^0(X) \rangle / n - \langle \hat{\Psi}\hat{\beta}_n, \hat{\Psi}\beta \rangle / n \tag{139}$$

Then Eq. (138) implies that:

$$\diamondsuit \leq \langle \hat{\Psi}\hat{\beta}_n, \boldsymbol{\epsilon} \rangle / n - \langle \hat{\Psi}\beta, \boldsymbol{\epsilon} \rangle / n + \lambda \|\beta\|_1 - \lambda \|\hat{\beta}_n\|_1, \tag{140}$$

The ϵ vector stores the noise variables: $\epsilon_i = Y_i - f^0(\mathbf{X}_i)$. The rest of the proof follows identically to that of Theorem 2.2 in Van de Geer [51, page 21], replacing the $(\hat{\beta} - \beta)^\top \hat{\Sigma}(\hat{\beta} - \beta^0)$ term there by $\diamondsuit$. □

The following lemmas tell us that the random compatibility constant $\phi_{\hat{\Sigma}}(L, S)$ is bounded away from zero with high probability.

Lemma E.3. *Let Σ be the population $J_n \times J_n$ covariance matrix $\Sigma_{ij} = E[\psi_i(\mathbf{X})\psi_j(\mathbf{X})]$, where (ψ_j) is the unravelled function sequence defined in Condition C.9. Assume the feature density function $p_X(\mathbf{x}) = d\rho_X / d\nu^d \geq u > 0$ is bounded away from 0. Here ν^d is the d-dimension product measure of ν in Condition C.9.*

Then we know Σ has a compatibility constant that does not depend on L, S: $\phi_\Sigma^2 \geq u$.

Proof. For any $\beta \in \mathbb{R}^{J_n}$:

$$\begin{aligned} \beta^\top \Sigma \beta &= \sum_{1 \leq i,j \leq J_n} \beta_i \beta_j E\big[\psi_i(\mathbf{X})\psi_j(\mathbf{X})\big] = E\left[\left(\sum_{j=1}^{J_n} \psi_j(\mathbf{X})\beta_j\right)^2\right] \\ &\geq u \int \left(\sum_{j=1}^{J_n} \psi_j(\mathbf{x})\beta_j\right)^2 d\mathbf{x} \overset{(1)}{=} u\|\beta\|_2^2. \end{aligned} \tag{141}$$

In step (1) we used the orthonomality of ψ_j stated in Condition C.9. At the same time, we have $\|\beta_S\|_1^2 \leq \|\beta\|_2^2 |S|$. Checking the definition of compatibility (Definition E.1), we conclude for any L, S, the matrix Σ has a uniform compatibility constant ϕ_Σ greater than $\sqrt{u}$ (meaning that this lower bound does not depend on either L or S). □

Lemma E.4. *Under the same conditions as in Lemma E.3, we know the empirical matrix $\hat{\Sigma}$ has a compatibility constant $\phi_{\hat{\Sigma}}^2(L, S) \geq u/2$, with probability at least $1 - J_n^2 \exp(-na^2/2M^{4D'})$, $a = u(L+1)^{-2}|S|^{-1}/2$.*

Proof. We first consider the difference between two quadratic forms related to the two covariance matrices:

$$|\beta^\top \hat{\Sigma}\beta - \beta^\top \Sigma \beta| = \left| \sum_{1 \leq i,j \leq J_n} \beta_i \beta_j (\hat{\Sigma}_{ij} - \Sigma_{ij}) \right| \leq \|\beta\|_1^2 \|\hat{\Sigma} - \Sigma\|_\infty \tag{142}$$

By the definition of compatibility constant ϕ_Σ, for any β such that $\|\beta_{-S}\|_1 \leq L\|\beta_S\|_1$, we have

$$\|\beta\|_1 \leq (L+1)\|\beta_S\|_1 \leq (L+1)\sqrt{|S|\beta^\top\Sigma\beta}/\phi_\Sigma \tag{143}$$

Plugging this into (142), we have:

$$\begin{aligned} &| \beta^\top\hat{\Sigma}\beta - \beta^\top\Sigma\beta | \leq (L+1)^2\|\hat{\Sigma}-\Sigma\|_\infty |S|\beta^\top\Sigma\beta/\phi_\Sigma^2 \\ \iff &\left|\frac{\beta^\top\hat{\Sigma}\beta}{\beta^\top\Sigma\beta} - 1\right| \leq (L+1)^2\|\hat{\Sigma}-\Sigma\|_\infty |S|/\phi_\Sigma^2 \end{aligned} \tag{144}$$

By a typical application of Hoeffding's inequality (every entry in $\hat{\Sigma}$ is a bounded random variable), we know with probability at least $1 - J_n^2\exp(-na^2/2M^{4D'})$, where $a = u(L+1)^{-2}|S|^{-1}/2$, that

$$\|\hat{\Sigma}-\Sigma\|_\infty \leq \left(2(L+1)^2|S|/u\right)^{-1} \tag{145}$$

This means, with the same probability we have

$$\left|\frac{\beta^\top\hat{\Sigma}\beta}{\beta^\top\Sigma\beta} - 1\right| \leq \frac{1}{2} \tag{146}$$

Therefore, for all any β such that $\|\beta_{-S}\|_1 \leq L\|\beta_S\|_1$, we have that:

$$\frac{|S|\beta^\top\hat{\Sigma}\beta}{\|\beta_S\|_1^2} \geq \frac{|S|\beta^\top\Sigma\beta}{2\|\beta_S\|_1^2}, \tag{147}$$

with high probability. By the definition of the compatibility constant, we can read out

$$\phi_{\hat{\Sigma}}^2(L,S) \geq \phi_\Sigma^2/2 \tag{148}$$

which concludes our proof. □

Corollary E.5. *Let $\lambda_\epsilon = [2\log(2J_n)/\{C(C_{sub},M,D')n\}]^{1/2}$ and assume ϵ_i to be uniform sub-Gaussian noise. Then, under the same conditions as in Theorem E.2, for any $\beta \in \mathbb{R}^{J_n}$ whose support is $S \subset \{1,2,\ldots,J_n\}$, we have*

$$\lambda_\epsilon\|\hat{\beta}_n - \beta\|_1 + \left\|f_{\hat{\beta}_n} - f^0\right\|_n^2 \leq \frac{3}{2}\left\|f_\beta - f^0\right\|_2^2 + \frac{49\lambda_\epsilon^2|S|}{2u} \tag{149}$$

with probability larger than $1 - 1/(2J_n) - J_n^2\exp(-na^2/2M^{4D'}) - \exp(-cn\|f_\beta - f^0\|_2^2/M_0^2)$, where $a = u(L+1)^{-2}|S|^{-1}/2$. The definition of f_β is stated in (131).

Proof. First we show that for the chosen λ_ϵ, the following holds with high probability:

$$\sup_{1\leq j\leq J_n}\left|\frac{1}{n}\sum_{i=1}^n \psi_j(\mathbf{X}_i)\epsilon_i\right| \leq \lambda_\epsilon. \tag{150}$$

Since (ϵ_i) are sub-Gaussian random variables (with a parameter not depending on $\mathbf{X}_i$), we know there exists a constant C_{sub} such that

$$\|\epsilon_i\|_{L^p} = \left\{E\left(|\epsilon_i|^p\right)\right\}^{1/p} \leq C_{sub}\sqrt{p} \quad \text{for all } p \geq 1 \tag{151}$$

For reference, see e.g. Proposition 2.5.2 in Vershynin [54]. Since the basis functions ψ_j are also uniformly bounded (by $M^{D'}$), we have

$$\|\psi_j(\mathbf{X}_i)\epsilon_i\|_{L^p} \leq C_{sub}M^{D'}\sqrt{p} \quad \text{for all } p \geq 1. \tag{152}$$

This means $\psi_j(\mathbf{X}_i)\epsilon_i$ is also sub-Gaussian. Applying a union bound and Hoeffding's inequality for sub-Gaussian variables (e.g. Theorem 2.6.3 in Vershynin [54]), we get:

$$\begin{aligned}\text{pr}\left\{\sup_{1\leq j\leq J_n}\left|\frac{1}{n}\sum_{i=1}^{n}\psi_j(\mathbf{X}_i)\epsilon_i\right| \geq t\right\} &\leq \sum_{j=1}^{J_n}\text{pr}\left\{\left|\frac{1}{n}\sum_{i=1}^{n}\psi_j(\mathbf{X}_i)\epsilon_i\right| \geq t\right\} \\ &\leq 2J_n\exp\left\{-C\left(C_{sub}, M, D'\right)nt^2\right\}\end{aligned} \tag{153}$$

Taking $t = \lambda_\epsilon = [2\log(2J_n)/\{C(C_{sub}, M, D')n\}]^{1/2}$, we see that

$$\text{pr}\left\{\sup_{1\leq j\leq J_n}\left|\frac{1}{n}\sum_{i=1}^{n}\psi_j(\mathbf{X}_i)\epsilon_i\right| \leq \lambda_\epsilon\right\} \geq 1 - 1/(2J_n) \tag{154}$$

This is what we claimed in the beginning of the proof.

Next, we bound the difference between $\|f_\beta - f^0\|_n^2$ and $\|f_\beta - f^0\|_2^2$ for any fixed f_β satisfying $\|f_\beta\|_\infty < 2\|f^0\|_\infty$. First, the difference $(f_\beta(\mathbf{X}_i) - f^0(\mathbf{X}_i))^2$ is a bounded variable, therefore it is sub-Gaussian with parameter $9M_0^2$, where M_0 bounds $\|f^0\|_\infty$. The centered version, $\{f_\beta(\mathbf{X}_i) - f^0(\mathbf{X}_i)\}^2 - \|f_\beta - f^0\|_2^2$ is also sub-Gaussian with parameter CM_0^2 (see e.g. Lemma 2.6.8 in Vershynin [54]). Again applying Hoeffding's inequality we see that

$$\begin{aligned}&\text{pr}\left[\left|\frac{1}{n}\sum_{i=1}^{n}\left\{f_\beta(\mathbf{X}_i) - f^0(\mathbf{X}_i)\right\}^2 - \|f_\beta - f^0\|_2^2\right| \geq t\right] \leq \exp\left(-cnt^2/M_0^2\right) \\ &\qquad \Rightarrow \text{pr}\left(\left|\frac{\|f_\beta - f^0\|_n^2}{\|f_\beta - f^0\|_2^2} - 1\right| \geq \frac{1}{2}\right) \leq \exp\left(-cn\|f_\beta - f^0\|_2^2/M_0^2\right)\end{aligned} \tag{155}$$

We know, with probability larger than $1 - \exp(-cn\|f_\beta - f^0\|_2^2/M_0^2)$

$$\frac{\|f_\beta - f^0\|_n^2}{\|f_\beta - f^0\|_2^2} \leq \frac{3}{2} \tag{156}$$

Combining (154), (156), Lemma E.4 and Theorem E.2, we can conclude our proof. □

E.2. Theoretical guarantees under sparse tensor product models

In this section, we will combine the oracle inequalities developed in the last section with approximation results to derive performance guarantees of the l_1-penalized sieve estimator.

Recall the following notation: d is the overall ambient dimension of our features $\mathbf{X}_i$, D is the number of explanatory features related to the outcome Y (the active dimension), s is the smoothness parameter of f^0 (Condition C.10), and J_n is the number of basis functions in the lasso problem (14). The constant C_{sub} is the sub-Gaussian parameter for the noise variables, u is the lower bound of the feature density function and M_0 is a bound on the $\|\cdot\|_\infty$-norm of f^0.

Corollary E.6. *Let $f_{\hat{\beta}_n}$ be the penalized sieve estimate of f^0, and $f_{\beta_n^0}$ be the approximation of f^0 as in Lemma C.11. Choose the penalization hyperparameter as $\lambda_\epsilon = [2\log(2J_n)/\{C(C_{sub}, M, D')n\}]^{1/2}$. Under the same conditions as in Theorem 7.3, we have the following two bounds*

$$\begin{aligned} \big\|f_{\hat{\beta}_n} - f^0\big\|_n^2 &\le C\big(C_{sub}, M, D', \rho_X, f^0\big)\log(J_n)\left(\frac{\log^{D-1}(n)}{n}\right)^{\frac{2s}{2s+1}} \\ \|\hat{\beta}_n - \beta_n^0\|_1 &\le C\big(C_{sub}, M, D', \rho_X, f^0\big)(\log J_n/n)^{1/2} n^{1/(2s+1)} (\log n)^{2s(D-1)/(2s+1)} \end{aligned} \tag{157}$$

with probability at least

$$\begin{aligned} &1 - 1/(2J_n) - J_n^2 \exp\big(-na^2/2M^{4D'}\big) \\ &\quad - \exp\big(-C\big(s, D, \rho_X, f^0\big)(\log n)^{(D-1)2s/(2s+1)} n^{1/(2s+1)}\big), \end{aligned}$$

where $a = u(L+1)^{-2}|S|^{-1}/2$.

Proof. To get the bounds above, we only need to combine the oracle inequality in Corollary E.5 with the approximation results in Lemma C.11.

In Lemma C.11, we discussed that so long as J_n is large enough, we can find a function $f_{\beta_n^0}$ that approximates f^0 well enough. Plugging the results of Lemma C.11 into the oracle inequality (149), we have:

$$\begin{aligned} &\lambda_\epsilon \|\hat{\beta}_n - \beta\|_1 + \big\|f_{\hat{\beta}_n} - f^0\big\|_n^2 \\ &\qquad \le C\big(s, D, \rho_X, f^0\big)\left(\frac{\log^{D-1} n}{n}\right)^{\frac{2s}{2s+1}} + \frac{49\lambda_\epsilon^2 |S_n|}{2u}, \end{aligned} \tag{158}$$

here $|S_n|$ is the cardinality of non-zero elements in β_n^0. Although formally $f_{\beta_n^0}$ is a linear combination of $J_n = C(s, D) d^{D'} n^{1/(2s+1)} \log^{D'-1} n$ basis functions, the size of its support is much smaller (thanks to the feature sparsity conditions). In fact, $f_{\beta_n^0}$ only needs to engage with the informative dimensions of the features. In Lemma C.7, we showed that $|S_n|$ can be bounded by $(\log^{D-1} n)^{2s/(2s+1)} n^{1/(2s+1)}$.

Plugging this in the above inequality gives:

$$\lambda_\epsilon \|\hat{\beta}_n - \beta\|_1 + \left\|f_{\hat{\beta}_n} - f^0\right\|_n^2 \leq C\big(s, D, \rho_X, f^0\big)\left(\frac{\log^{D-1} n}{n}\right)^{\frac{2s}{2s+1}} + C\big(C_{sub}, M, D', \rho_X\big)\log(J_n)\left(\frac{\log^{D-1} n}{n}\right)^{\frac{2s}{2s+1}}. \tag{159}$$

This gives us the results regarding the $\|\cdot\|_n$-norm distance and l_1-distance stated in Corollary E.6 (the second term will dominate for large n). □

At this point, we already established bounds on the training-design error (expressed as the $\|\cdot\|_n$-norm). However, for most prediction problems we are interested in the testing error (quantified in the $\|\cdot\|_{2,\rho_X}$-norm). For arbitrarily flexible estimators, a low training-design error does not imply a strong generalization ability. However, according to Corollary E.6, the coefficient $\hat{\beta}_n$ lives in a small $\|\cdot\|_1$-ball centered around the oracle β_n^0 with high probability. From this we can also develop some bounds on metric entropy of the space in which $f_{\hat{\beta}_n}$ takes value. These will in turn link the expected distance to the empirical distance.

In the following discussion we will use the concept of metric entropy of a function space. For more comprehensive discussion, see Chapter 2 of van de Geer [50].

Definition E.7. Let Q be a measure on $\mathcal{X}$ and let $\mathcal{G}$ be a function space $\mathcal{G} \subset L_2(\mathcal{X}; Q)$. Consider for each $\delta > 0$, a collection of functions $g_1, \ldots, g_N$, such that for each $g \in \mathcal{G}$, there is a $j = j(g) \in \{1, \ldots, N\}$, such that

$$\left(\int_{\mathcal{X}} \big(g(x) - g_j(x)\big)^2 dQ(x)\right)^{1/2} \leq \delta. \tag{160}$$

Let $N(\delta, \mathcal{G}, Q)$ be the smallest value of N for which such a covering by balls with radius δ and centers $g_1, \ldots, g_N$ exists. Then $N(\delta, \mathcal{G}, Q)$ is called the covering number (under measure Q) and $H(\delta, \mathcal{G}, Q) = \log N(\delta, \mathcal{G}, Q)$ is called the metric entropy of $\mathcal{G}$ (under measure Q).

One of the function spaces $\mathcal{G}_n$ we are going to consider is

$$\mathcal{G}_n = \mathcal{G}_n\big(\psi_j, \beta_n^0, r_n\big) = \left\{ f = \sum_{j=1}^{J_n} \beta_j \psi_j \mid \beta = (\beta_1, \ldots, \beta_{J_n})^\top \in B_1\big(\beta_n^0, r_n\big) \right\}, \tag{161}$$

with $r_n = r_n(s, D) = (\log J_n / n)^{1/2} n^{1/(2s+1)} \log^{2s(D-1)/(2s+1)}(n)$. This radius is of the same order as the RHS in (157). The set $B_1(\beta, r) \subset \mathbb{R}^{J_n}$ is the $\|\cdot\|_1$-ball of radius r centered at β, formally

$$B_1(\beta, r) = \big\{\gamma \in \mathbb{R}^{J_n} \mid \|\gamma - \beta\|_1 \leq r\big\} \tag{162}$$

For a specified sequence of r_n and J_n, $\mathcal{G}_n$ is a deterministic sequence of function spaces.

In the rest of this section, we will derive some bounds on the metric entropy of $\mathcal{G}_n$ and apply some maximal inequalities to relate the testing-design errors to the training-design errors. We will show that the metric entropy of the function space $\mathcal{G}_n$ (equipped with $\|\cdot\|_n$-norm) is of the same order as the metric entropy of $B_1(\beta_n, r_n)$ (equipped with Euclidean $\|\cdot\|_2$-norm). Since the latter is known in the literature (e.g, Lemma 3 in Raskutti et al. [33]), we have the following results:

Lemma E.8. *Let* $r_n = r_n(s, D) = (\log J_n/n)^{1/2} n^{1/(2s+1)} (\log n)^{2s(D-1)/(2s+1)}$. *Then for the* $\mathcal{G}_n$ *defined in* (161), *we have*

$$H(\delta, \mathcal{G}_n, P_n) \leq C(M) r_n^2 \delta^{-2} \log J_n \tag{163}$$

Proof. We first rewrite the empirical norm in matrix notation, for any $\beta = (\beta_1, \ldots, \beta_{J_n})^\top \in \mathbb{R}^{J_n}$:

$$\|f_\beta\|_n = \left\{ \frac{1}{n} \sum_{i=1}^n f_\beta^2(\mathbf{X}_i) \right\}^{1/2} = \frac{1}{\sqrt{n}} \left\{ \sum_{i=1}^n \left(\sum_{j=1}^{J_n} \beta_j \psi_j(\mathbf{X}_i) \right)^2 \right\}^{1/2} = \left\| \frac{1}{\sqrt{n}} \hat{\Psi} \beta \right\|_2, \tag{164}$$

where $\hat{\Psi}$ is the design matrix: $\hat{\Psi}_{ij} = \psi_j(\mathbf{X}_i)$.

Therefore, if there is a δ-cover of the set $\{n^{-1/2}\hat{\Psi}\beta, \beta \in B_1(\beta_n^0, r_n)\} \subset \mathbb{R}^n$ under the Euclidean $\|\cdot\|_2$-norm, we can directly construct one for $\mathcal{G}_n$ under $\|\cdot\|_n$-norm. There are available bounds on the covering number of the $n^{-1/2}\hat{\Psi}\beta$ when β belongs to a l_1-ball. Specifically, we can apply Lemma 4 of Raskutti et al. [33]:

$$H\big(\delta, \big\{n^{-1/2}\hat{\Psi}\beta, \beta \in B_1\big(\beta_n^0, r_n\big)\big\}, \|\cdot\|\big) \leq C(M) r_n^2 \delta^{-2} \log J_n. \tag{165}$$

This concludes our proof. □

To relate the training and testing errors, we need to consider a function space closely related to $\mathcal{G}_n$:

$$\tilde{\mathcal{G}}_n^2 = \big(\mathcal{G}_n - f^0\big)^2 \tag{166}$$

We summarize several properties of it in the following lemma.

Lemma E.9. *Let* $r_n = (\log J_n/n)^{1/2} n^{1/(2s+1)} (\log n)^{2s(D-1)/(2s+1)}$ *and* $\delta_n < r_n$. *Then for the function space* $\tilde{\mathcal{G}}_n^2$ *we know:*

$$\begin{aligned} &\sup_{g \in \tilde{\mathcal{G}}_n^2} \|g\|_\infty \leq C\big(M, D', s, D, Q\big), \\ &pr\Big\{ \sup_{g \in \tilde{\mathcal{G}}_n^2} \|g\|_n \leq C\big(M, D', s, D, Q\big) r_n \Big\} \overset{n\to\infty}{\longrightarrow} 1 \text{ and} \\ &\int_{\delta_n}^{r_n} H^{1/2}\big(u, \tilde{\mathcal{G}}_n^2, P_n\big) du \leq C\big(M, D', s, D, Q\big) r_n (\log J_n)^{1/2} \log(1/\delta_n). \end{aligned} \tag{167}$$

Proof. • We first derive the bound on the $\|\cdot\|_\infty$-norm. By definition, every element g in $\tilde{\mathcal{G}}_n^2$ can be expressed as $g = (f - f^0)^2$ for some $f \in \mathcal{G}_n$. To bound $\|g\|_\infty$, it is enough to bound $\|f - f^0\|_\infty$.

$$\begin{aligned}\|f - f^0\|_\infty &\le \|f - f_{\beta_n^0}\|_\infty + \|f_{\beta_n^0} - f^0\|_\infty \\ &\le C\big(M, D'\big)r_n + \|f_{\beta_n^0}\|_\infty + \|f^0\|_\infty \le C\big(M, D', s, D, Q\big).\end{aligned} \tag{168}$$

- We now bound the empirical norm $\|\cdot\|_n$. For any $f \in \mathcal{G}_n$, we can define a function $g = f - f^0$. And we know:

$$\begin{aligned}\|g\|_n &\le \|f - f_{\beta_n^0}\|_n + \|f_{\beta_n^0} - f^0\|_n \\ &\le C\big(M, D'\big)r_n + C(s, D, \rho_X, Q)\big(\log^{D-1} n/n\big)^{s/2s+1} \text{ w.h.p.}\end{aligned} \tag{169}$$

The first term is using the explicit form of f and $f_{\beta_n^0}$. The second bound is based on the approximation results in Lemma C.11 and the probability bound in (156). Since $r_n = (\log J_n/n)^{1/2} n^{1/(2s+1)} (\log n)^{2s(D-1)/(2s+1)}$, the order of the first term in (169) is larger than the second one's. For $g^2 \in \tilde{\mathcal{G}}_n^2$,

$$\begin{aligned}\|g^2\|_n^2 &= \frac{1}{n}\sum_{i=1}^n \big\{f(\mathbf{X}_i) - f^0(\mathbf{X}_i)\big\}^4 \\ &\le C\big(M, D', s, D, Q\big)\|g\|_n^2 \le C\big(M, D', s, D, Q\big)r_n^2.\end{aligned} \tag{170}$$

So we conclude that for any $g^2 \in \tilde{\mathcal{G}}_n^2$, $\|g^2\|_n \le C(M, D', s, D, Q)r_n$ with probability going to 1.

- Now we derive the bound on the integrated metric entropy. For any $h_1, h_2 \in \tilde{\mathcal{G}}_n^2$, there exist $f_1, f_2 \in \mathcal{G}_n$ such that $h_i = (f_i - f^0)^2$, $i \in \{1, 2\}$. So we know that

$$\begin{aligned}\|h_1 - h_2\|_n^2 &= \|\big(f_1 - f^0\big)^2 - \big(f_2 - f^0\big)^2\|_n^2 \\ &= \frac{1}{n}\sum_{i=1}^n \big[\big\{f_1(\mathbf{X}_i) + f_2(\mathbf{X}_i) - 2f^0(\mathbf{X}_i)\big\}\big\{f_1(\mathbf{X}_i) - f_2(\mathbf{X}_i)\big\}\big]^2 \\ &\le C\big(M, D', s, D, Q\big)\frac{1}{n}\sum_{i=1}^n \big\{f_1(\mathbf{X}_i) - f_2(\mathbf{X}_i)\big\}^2 \\ &= C\big(M, D', s, D, Q\big)\|f_1 - f_2\|_n^2\end{aligned} \tag{171}$$

Now we know that if we have a δ-covering of $\mathcal{G}_n$ with center points $\{f_k\}$, then the functions $\{(f_k - f^0)^2\}$ form a $C(M, D', s, D, Q)\delta$-covering of $\tilde{\mathcal{G}}_n^2$. Since we already have an entropy bound on $\mathcal{G}_n$ stated in Lemma E.8, we have one for $\tilde{\mathcal{G}}_n^2$ of the same order as well. The integrated entropy can be bounded as follows:

$$\begin{aligned}\int_{\delta_n}^{r_n} H^{1/2}\big(\tau, \tilde{\mathcal{G}}_n^2, P_n\big)d\tau &\le C\big(M, D', s, D, Q\big)\int_{\delta_n}^{r_n} (\log J_n)^{1/2} r_n \tau^{-1} d\tau \\ &\le C\big(M, D', s, D, Q\big)(\log J_n)^{1/2} r_n \log(1/\delta_n),\end{aligned} \tag{172}$$

when $r_n \le 1$. □

Theorem E.10. *Let*

$$\begin{aligned}\delta_n &= C(s,D)\log(J_n)n^{-\frac{2s}{2s+1}}\log^{\frac{2s(D-1)}{2s+1}+1}(n),\\ r_n &= C(s,D)(\log J_n/n)^{1/2}n^{1/(2s+1)}(\log n)^{2s(D-1)/(2s+1)}.\end{aligned} \tag{173}$$

And let $\hat{\beta}_n$ denote the minimizer of the penalized problem (14). Then, under the same conditions as in Theorem 7.3, we have

$$\limsup_{n\to\infty} pr\big(\big|\|f_{\hat{\beta}_n}-f^0\|_n^2-\|f_{\hat{\beta}_n}-f^0\|_2^2\big|\geq\delta_n\big)=0 \tag{174}$$

Proof. We first need to apply the symmetrization trick (e.g. Corollary 3.4 in van de Geer [50])

$$\begin{aligned}&\mathrm{pr}\big(\big|\|f_{\hat{\beta}_n}-f^0\|_n^2-\|f_{\hat{\beta}_n}-f^0\|_2^2\big|\geq\delta_n\big)\\ &=\mathrm{pr}\big\{\big|(P_n-P)\big(f_{\hat{\beta}_n}-f^0\big)^2\big|\geq\delta_n\big\}\\ &\leq\mathrm{pr}\Big\{\sup_{g\in\tilde{\mathcal{G}}_n^2}\big|(P_n-P)g\big|\geq\delta_n\Big\}+\mathrm{pr}(f_{\hat{\beta}_n}\notin\mathcal{G}_n)\\ &\leq 4\mathrm{pr}\left\{\sup_{g\in\tilde{\mathcal{G}}_n^2}\left|\frac{1}{n}\sum_{i=1}^n W_i g(\mathbf{X}_i)\right|\geq\delta_n/4\right\}+\mathrm{pr}(f_{\hat{\beta}_n}\notin\mathcal{G}_n)\end{aligned} \tag{175}$$

The W_i variables above are independent and identically distributed Rademacher variables ($\mathrm{pr}(W_i=1)=\mathrm{pr}(W_i=-1)=0.5$). They are bounded (therefore sub-Gaussian) random variables. The probability $\mathrm{pr}(f_{\hat{\beta}_n}\notin\mathcal{G}_n)$ has been investigated in Corollary E.6. To bound the first term in (175), we need to apply some maximal inequalities (e.g., Corollary 8.3 or Lemma 3.2 in van de Geer [50]). These results require that $r_n>\delta_n$ and

$$\sqrt{n}\delta_n\geq C\left(\int_{\delta_n/8}^{r_n}H^{1/2}\big(\tau,\tilde{\mathcal{G}}_n^2,P_n\big)d\tau\vee r_n\right). \tag{176}$$

We already checked these properties in Lemma E.9. So we conclude that with probability going to 1, the difference between the training and testing error is no larger than δ_n. □

Proof of Theorem 7.3. To show the testing error stated in Theorem 7.3, we just need to combine the results in Theorem E.10 and Corollary E.6. □

Appendix F: The average order of divisor functions

In this section we will present a derivation of the average order of D-divisor functions that was used in the proof of Lemma C.3. This result is known to mathematicians working on number theory, and is usually considered as a direct generalization of the $D=2$ case. However, most standard references only include the special (but important) case when $D=2$. For the purpose of completeness, we replicate a proof based on an unpublished online note by Graham Jameson, from Lancaster University. For other references of similar results, see Huybrechs et al. [23] (Proposition 6) and Dobrovol'skii and Roshchenya [6].

Lemma F.1. *We have the following recurrence relation for T_D (over D):*

$$T_D(x) = \sum_{n \le x} T_{D-1}\left(\frac{x}{n}\right) \tag{177}$$

Proof. Fix $n \le x$. The number of D-tuples $(\mathbf{j}^1, \mathbf{j}^2, \ldots, \mathbf{j}^{D-1}, n)$ with $n \prod_{k=1}^{D-1} \mathbf{j}^k \le x$ is the number of $(D-1)$-tuples with $\prod_{k=1}^{D-1} \mathbf{j}^k \le x/n$, that is, $T_{D-1}(x/n)$. Hence $T_D(x) = \sum_{n \le x} T_{D-1}(x/n)$. □

Lemma F.2. *Define*

$$A_D(x) = \sum_{n \le x} \frac{x}{n} \log^D(x/n) \tag{178}$$

then we have

$$\frac{1}{D+1} x \log^{D+1} x \le A_D(x) \le \frac{1}{D+1} x \log^{D+1} x + x \log^D x \tag{179}$$

Proof. Let $f(t) = (x/t) \log^D(x/t)$ for $1 \le t \le x$ (also $f(t) = 0$ for $t > x$). Then $f(t)$ is decreasing and non-negative, and

$$\int_1^x f(t)dt = \left[\frac{x}{D+1} \log^{D+1}(u)\right]_1^x = \frac{x \log^{D+1}(x)}{D+1} \tag{180}$$

The statement follows, by using the following basic integral estimate (Proposition 1.4.2 of Jameson [24]): Let $f(t)$ be a decreasing, non-negative function for $t \ge 1$. Write $S(x) = \sum_{n \le x} f(n)$ and $I(x) = \int_1^x f(t)dt$. Then for all $x \ge 1$,

$$I(x) \le S(x) \le I(x) + f(1) \tag{181}$$

□

Lemma F.3. *For any fixed $D \ge 2$,*

$$T_D(x) = \frac{1}{(D-1)!} x \log^{D-1} x + O\big(x \log^{D-2} x\big). \tag{182}$$

The $O(\cdot)$ in (182) is for $x \to \infty$.

Proof. Induction on D. The case $D = 2$ is known to be true (Theorem 3.2 Tenenbaum [47]). Assume (182) for D, with the error term denoted by $q_D(x)$. Then by (177),

$$T_{D+1}(x) = \sum_{n \le x} T_D\left(\frac{x}{n}\right) = I(x) + Q(x) \tag{183}$$

where

$$\begin{aligned} I(x) &= \frac{1}{(D-1)!} \sum_{n \le x} \frac{x}{n} \log^{D-1} \frac{x}{n} = \frac{1}{(D-1)!} A_{D-1}(x) \\ Q(x) &= \sum_{n \le x} q_D\left(\frac{x}{n}\right) \sim \sum_{n \le x} \frac{x}{n} \log^{D-2} \frac{x}{n} = A_{D-2}(x) \end{aligned} \tag{184}$$

By (179),

$$I(x) = \frac{1}{D!} x \log^D x + O\big(x \log^{D-1} x\big) \tag{185}$$

and $Q(x) = O(x \log^{D-1} x)$. Hence (182) holds for $D+1$. □

Acknowledgments

The authors would like to thank the anonymous referees, an Associate Editor and the Editor for their constructive comments that improved the quality of this paper.

Funding

The authors gratefully acknowledge NIH grant R01HL137808.

References

[1] Akgül, A., E. K. Akgül, and S. Korhan (2020). New reproducing kernel functions in the reproducing kernel Sobolev spaces. *AIMS Mathematics 5*(1), 482–496. MR4140481

[2] Benkeser, D. and M. Van Der Laan (2016). The highly adaptive lasso estimator. In *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 689–696. IEEE.

[3] Bungartz, H.-J. and M. Griebel (2004). Sparse grids. *Acta Numerica 13*(1), 147–269. MR2249147

[4] Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. *Handbook of Econometrics 6*, 5549–5632.

[5] Cucker, F. and S. Smale (2002). On the mathematical foundations of learning. *Bulletin of the American Mathematical Society 39*(1), 1–49. MR1864085

[6] Dobrovol'skii, N. M. and A. L. Roshchenya (1998). Number of lattice points in the hyperbolic cross. *Matematicheskie Zametki 63*(3), 363–369. MR1631932

[7] Dũng, D., V. Temlyakov, and T. Ullrich (2018). *Hyperbolic Cross Approximation*. Springer. MR3887571

[8] Efromovich, S. (2008). *Nonparametric Curve Estimation: Methods, Theory, and Applications*. Springer Science & Business Media. MR1705298

[9] Efromovich, S. (2010). Orthogonal series density estimation. *Wiley Interdisciplinary Reviews: Computational Statistics 2*(4), 467–476.

[10] Eubank, R. and P. Speckman (1990). Curve fitting by polynomial-trigonometric regression. *Biometrika 77*, 1–9. MR1049403

[11] Fasshauer, G. E. and M. J. McCourt (2015). *Kernel-based Approximation Methods Using Matlab*, Volume 19. World Scientific Publishing Company.

[12] Friedman, J. H. (1991). Multivariate adaptive regression splines. *The Annals of Statistics 19*(1), 1–67. MR1091842

[13] Friedman, J. H. and W. Stuetzle (1981). Projection pursuit regression. *Journal of the American Statistical Association 76*(376), 817–823. MR0650892

[14] Gao, F., G. Wahba, R. Klein, and B. Klein (2001). Smoothing spline anova for multivariate Bernoulli observations with application to ophthalmology data. *Journal of the American Statistical Association 96*(453), 127–160. MR1952725

[15] Glynn, A. N. and K. M. Quinn (2010). An introduction to the augmented inverse propensity weighted estimator. *Political Analysis 18*(1), 36–56.

[16] Grisoni, F., V. Consonni, M. Vighi, S. Villa, and R. Todeschini (2016). Investigating the mechanisms of bioconcentration through qsar classification trees. *Environment international 88*, 198–205.

[17] Gu, C. (2013). *Smoothing Spline ANOVA Models*, Volume 297. Springer. MR3025869

[18] Hamidieh, K. (2018). A data-driven statistical model for predicting the critical temperature of a superconductor. *Computational Materials Science 154*, 346–354.

[19] Han, Q. and J. A. Wellner (2019). Convergence rates of least squares regression estimators with heavy-tailed errors. *The Annals of Statistics 47*(4), 2286–2319. MR3953452

[20] Haris, A., D. Witten, and N. Simon (2016). Convex modeling of interactions with strong heredity. *Journal of Computational and Graphical Statistics 25*(4), 981–1004. MR3572025

[21] Hastie, T., R. Tibshirani, and M. Wainwright (2015). Statistical learning with sparsity. *Monographs on Statistics and Applied Probability 143*, 143. MR3616141

[22] Horowitz, J., J. Klemelä, and E. Mammen (2006). Optimal estimation in additive regression models. *Bernoulli 12*(2), 271–298. MR2218556

[23] Huybrechs, D., A. Iserles, et al. (2011). From high oscillation to rapid approximation iv: Accelerating convergence. *IMA Journal of Numerical Analysis 31*(2), 442–468. MR2813179

[24] Jameson, G. J. O. (2003). *The Prime Number Theorem*, Volume 53. Cambridge University Press. MR1976226

[25] Kennedy, E. H. (2022). Semiparametric doubly robust targeted double machine learning: a review. *arXiv preprint arXiv:2203.06469*.

[26] Kühn, T., W. Sickel, and T. Ullrich (2015). Approximation of mixed order Sobolev functions on the d-torus: asymptotics, preasymptotics, and d-dependence. *Constructive Approximation 42*(3), 353–398. MR3416161

[27] Ledoux, M. and M. Talagrand (2011). *Probability in Banach Spaces*. Classics in Mathematics. MR2814399

[28] Lin, X., G. Wahba, D. Xiang, F. Gao, R. Klein, and B. Klein (2000). Smoothing spline anova models for large data sets with Bernoulli observations and the randomized gacv. *The Annals of Statistics 28*(6), 1570–1600. MR1835032

[29] Lin, Y. (2000). Tensor product space ANOVA models. *The Annals of Statistics 28*(3), 734–755. MR1792785

[30] Liu, Z. and T. Stengos (1999). Non-linearities in cross-country growth regressions: a semiparametric approach. *Journal of Applied Econometrics 14*(5), 527–538.
[31] Nguyen, V. K. and W. Sickel (2016). Isotropic and dominating mixed besov spaces-a comparison. *arXiv preprint arXiv:1601.04000*. MR3682620
[32] Raskutti, G., M. J Wainwright, and B. Yu (2012). Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *Journal of Machine Learning Research 13*(2). MR2913704
[33] Raskutti, G., M. J. Wainwright, and B. Yu (2011). Minimax rates of estimation for high-dimensional linear regression over l_q-balls. *IEEE Transactions on Information Theory 57*(10), 6976–6994. MR2882274
[34] Richard, B. (1961). Adaptive control processes: A guided tour. *Princeton, New Jersey, USA*.
[35] Rosner, B. (2015). *Fundamentals of Biostatistics*. Cengage Learning.
[36] Schmeisser, H.-J. (2007). Recent developments in the theory of function spaces with dominating mixed smoothness. *Nonlinear Analysis, Function Spaces and Applications*, 145–204. MR2657119
[37] Sen, B. (2018). A gentle introduction to empirical process theory and applications. *Lecture Notes, Columbia University 11*, 28–29.
[38] Sickel, W. and T. Ullrich (2009). Tensor products of Sobolev–Besov spaces and applications to approximation from the hyperbolic cross. *Journal of Approximation Theory 161*(2), 748–786. MR2563079
[39] Sickel, W. and T. Ullrich (2011). Spline interpolation on sparse grids. *Applicable Analysis 90*(3-4), 337–383. MR2780900
[40] Simon, N., J. Friedman, T. Hastie, and R. Tibshirani (2011). Regularization paths for cox's proportional hazards model via coordinate descent. *Journal of Statistical Software 39*(5), 1–13.
[41] Steinwart, I. and A. Christmann (2008). *Support Vector Machines*. Springer Science & Business Media. MR2796580
[42] Steinwart, I. and C. Scovel (2012). Mercer's theorem on general domains: On the interaction between measures, kernels, and RKHSs. *Constructive Approximation 35*(3), 363–417. MR2914365
[43] Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, 1040–1053. MR0673642
[44] Tan, K. M. (2019). Layer-wise learning strategy for nonparametric tensor product smoothing spline regression and graphical models. *Journal of Machine Learning Research 20*(119). MR4002873
[45] Temlyakov, V. (2017). On the entropy numbers of the mixed smoothness function classes. *Journal of Approximation Theory 217*, 26–56. MR3628948
[46] Temlyakov, V. (2018). *Multivariate Approximation*. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press. MR3837133
[47] Tenenbaum, G. (2015). *Introduction to Analytic and Probabilistic Number Theory*, Volume 163. American Mathematical Soc. MR3363366
[48] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological) 58*(1),

267–288. MR1379242
[49] Tsybakov, A. (2008). *Introduction to Nonparametric Estimation*. Springer Science & Business Media. MR2724359
[50] van de Geer, S. (2000). *Empirical Processes in M-estimation*, Volume 6. Cambridge University Press.
[51] Van de Geer, S. A. (2016). *Estimation and Testing Under Sparsity*. Springer. MR3526202
[52] Van Der Vaart, A. and J. A. Wellner (2011). A local maximal inequality under uniform entropy. *Electronic Journal of Statistics 5*(2011), 192. MR2792551
[53] van der Vaart, A. W. and J. A. Wellner (1996). *Weak Convergence*, pp. 16–28. New York, NY: Springer New York.
[54] Vershynin, R. (2018). *High-dimensional Probability: An Introduction with Applications in Data Science*, Volume 47. Cambridge University Press. MR3837109
[55] Vybiral, J. (2006). Function spaces with dominating mixed smoothness. *Dissertationes Math. 436*, 73 pp. MR2231066
[56] Wahba, G. (1990). *Spline Models for Observational Data*. SIAM. MR1045442
[57] Wahba, G., Y. Wang, C. Gu, R. Klein, and B. Klein (1995). Smoothing spline anova for exponential families, with application to the wisconsin epidemiological study of diabetic retinopathy: the 1994 Neyman memorial lecture. *The Annals of Statistics 23*(6), 1865–1895. MR1389856
[58] Wainwright, M. J. (2019). *High-dimensional Statistics: A Non-asymptotic Viewpoint*, Volume 48. Cambridge University Press. MR3967104
[59] Wasserman, L. (2006). *All of Nonparametric Statistics*. Springer Science & Business Media. MR2172729
[60] Waugh, S. G. (1995). *Extending and benchmarking Cascade-Correlation: extensions to the Cascade-Correlation architecture and benchmarking of feed-forward supervised artificial neural networks*. Ph. D. Thesis, University of Tasmania.
[61] Xiang, Y. and N. Simon (2020). A flexible framework for nonparametric graphical modeling that accommodates machine learning. In *International Conference on Machine Learning*, pp. 10442–10451. PMLR.
[62] Yang, Y. (2007). Consistency of cross validation for comparing regression procedures. *The Annals of Statistics 35*(6), 2450–2473. MR2382654
[63] Yang, Y. and S. T. Tokdar (2015). Minimax-optimal nonparametric regression in high dimensions. *The Annals of Statistics 43*(2), 652–674. MR3319139
[64] Zhang, T. and N. Simon (2022). A sieve stochastic gradient descent estimator for online nonparametric regression in Sobolev ellipsoids. *The Annals of Statistics 50*(5), 2848–2871. MR4500627
[65] Zhang, T. and N. Simon (2023). An online projection estimator for nonparametric regression in reproducing kernel Hilbert spaces. *Statistica Sinica 33*(1), 127. MR4527765