OPTIMAL ESTIMATION OF SPARSE HIGH-DIMENSIONAL ADDITIVE MODELS$^{\#}$

KARL GREGORY[1], ENNO MAMMEN[2] AND MARTIN WAHL[3]

Abstract. In this paper we discuss the estimation of a nonparametric component f_1 of a nonparametric additive model $Y = f_1(X_1) + ... + f_q(X_q) + \varepsilon$. We allow the number q of additive components to grow to infinity and we make sparsity assumptions about the number of nonzero additive components. We compare this estimation problem with that of estimating f_1 in the oracle model $Z = f_1(X_1) + \varepsilon$, for which the additive components $f_2, \dots, f_q$ are known. We construct a two-step presmoothing-and-resmoothing estimator of f_1 in the additive model and state finite-sample bounds for the difference between our estimator and some smoothing estimators $\tilde{f}_1^{\text{oracle}}$ in the oracle model which satisfy mild conditions. In an asymptotic setting these bounds can be used to show asymptotic equivalence of our estimator and the oracle estimators; the paper thus shows that, asymptotically, under strong enough sparsity conditions, knowledge of $f_2, \dots, f_q$ has no effect on estimation efficiency. Our first step is to estimate all of the components in the additive model with undersmoothing using a group-Lasso estimator. We then construct pseudo responses $\hat{Y}$ by evaluating a desparsified modification of our undersmoothed estimator of f_1 at the design points. In the second step the smoothing method of the oracle estimator $\tilde{f}_1^{\text{oracle}}$ is applied to a nonparametric regression problem with "responses" $\hat{Y}$ and covariates X_1. Our mathematical exposition centers primarily on establishing properties of the presmoothing estimator. We also present simulation results demonstrating close-to-oracle performance of our estimator in practical applications. The main results of the paper are also important for understanding the behavior of the presmoothing estimator when the resmoothing step is omitted.

Date: March 24, 2016.

2010 *Mathematics Subject Classification.* Primary 62G08; secondary 62G20.

Key words and phrases. nonparametric curve estimation, additive models, penalization, Lasso, variable selection, dimension reduction.

[1] Research Training Group RTG 1953, University of Mannheim, L9,7, 68131 Mannheim, Germany.

[2] Institute of Applied Mathematics, Heidelberg University, Im Neuenheimer Feld 205, 69120 Heidelberg, Germany and National Research University Higher School of Economics, Russian Federation.

[3] Institute of Mathematics, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany.

1. Introduction

In this paper we study the estimation of an additive component in a high-dimensional sparse additive model. We compare this estimation problem with estimation in a nonparametric sub-model that contains only a single nonparametric component and we show that the two estimation problems are asymptotically equivalent. Our central argument is based on the construction of a class of two-step estimators that achieve the operating characteristics achieved by arbitrarily chosen smoothing estimators in the model with a single nonparametric component. We will prove finite-sample bounds for the difference between these two estimators. In an asymptotic framework, these bounds imply asymptotic equivalence of the two estimators under weak conditions. In addition to their theoretical value, these estimators are also of direct practical value, which we illustrate in simulations.

Our approach is analogous to that by which it is shown, in semiparametric modeling, that optimal estimation of a finite-dimensional parameter θ is asymptotically equivalent to optimal estimation in the hardest parametric sub-model containing only the parameter θ. This corresponds to our studying the estimation of an additive component f_1 in an additive model with additive components $f_1, \ldots, f_q$ as compared to the estimation of f_1 in the classical nonparametric regression model in which f_1 is the sole component. We refer to the latter model as the oracle model because estimation in this model is equivalent to estimation in the additive model when the functions $f_2, \ldots, f_q$ are known.

When we study estimation in semiparametric models, we typically have at our disposal an estimator for the parametric sub-model which is asymptotically normal and unbiased and of which the asymptotic covariance matrix achieves a lower bound. Thus, in order to establish the asymptotic efficiency of an estimator for the parametric component of a semiparametric model, it suffices to show that it is asymptotically normal and unbiased and that its asymptotic covariance matrix achieves the same lower bound as that achieved by the estimator in the parametric sub-model.

In contrast, in nonparametric estimation, we typically do not have any single asymptotically optimal estimator for the sub-model containing only f_1. This is because there are many different types of smoothing estimators, such as regression splines, kernel estimators, smoothing splines, and orthogonal series, which are not naturally comparable to one another and which have distinct asymptotic variances and biases, where the biases, moreover, are typically non-vanishing. Thus, there is no benchmark optimal estimator of f_1 in the single-component submodel to which we can compare estimators of f_1 in the additive model.

We circumvent this problem by showing that for every smoothing estimator $\tilde{f}_1^{\text{oracle}}$ in the oracle model, there exists a corresponding estimator $\tilde{f}_1$ in the additive model such that $\|\tilde{f}_1 - \tilde{f}_1^{\text{oracle}}\|_\infty = o_P(\delta_n)$, where $\|\cdot\|_\infty$ is the supremum norm and where δ_n is the pointwise rate of convergence of $\tilde{f}_1^{\text{oracle}}$ to f_1. For this result we make some weak assumptions on $\tilde{f}_1^{\text{oracle}}$ that hold for all classical smoothers and which we shall outline shortly. In particular, we get that, given a kernel estimator or smoothing spline in the oracle model with bias $\delta_n b(x_1)$ and asymptotic variance $\delta_n^2\sigma^2(x_1)$ at a point x_1, we get an estimator in the additive model with bias $\delta_n b(x_1) + o(\delta_n)$ and asymptotic variance $\delta_n^2\sigma^2(x_1) + o(\delta_n^2)$. Furthermore, asymptotic minimax results in the oracle model directly carry over to the additive model. This holds because the lower bound of the oracle model trivially also applies in the additive model. Upper bounds of the oracle model also remain valid in the additive model because the minimax estimator in the oracle model has an asymptotically equivalent counterpart in the additive model, according to our theory.

We prescribe a two-step construction of the estimator $\tilde{f}_1$. In the first step all the components of the additive model are estimated with undersmoothing—that is with low bias and high variance—resulting in a pilot estimator $\hat{f}_1$ of f_1 that is intentionally too wiggly. In the second step we apply the smoothing operation used in the calculation of $\tilde{f}_1^{\text{oracle}}$ to the nonparametric regression problem where $\hat{f}_1(X_1^i)$ is regressed on the values of the first covariate X_1^i, $i = 1, ..., n$. The resulting resmoothed estimator $\tilde{f}_1$ is our proposed estimator for f_1.

Our main result will state finite-sample properties for the presmoothing or pilot estimator $\hat{f}_1$. It is important that this is needed only for one specification of $\hat{f}_1$. We will argue that these finite-sample properties imply that, asymptotically,

$$\|\tilde{f}_1 - \tilde{f}_1^{\text{oracle}}\|_\infty = o_P(\delta_n) \tag{1.1}$$

for a whole class of smoothing estimators $\tilde{f}_1^{\text{oracle}}$ in the oracle model. Thus, we have developed an asymptotic optimality theory for sparse high-dimensional additive models.

We now formally express our estimation problem. Let

$$Y = f(X) + \epsilon = \sum_{j=1}^{q} f_j(X_j) + \epsilon$$

with responses Y and covariates $X = (X_1, ..., X_q)$ taking values in $[0,1]^q$. For identifiability, we assume that $\mathbf{E}[f_j(X_j)] = 0$ for $j = 2, \ldots, q$. We assume that ϵ is a Gaussian random variable independent of X with expectation 0 and variance σ^2. Moreover, we assume that we observe n independent copies $(Y^1, X^1), \ldots, (Y^n, X^n)$ of (Y, X), i.e.,

$$Y^i = \sum_{j=1}^{q} f_j(X_j^i) + \epsilon^i, \quad i = 1, \ldots, n. \tag{1.2}$$

We aim at estimating f_1 globally as well as locally at some point x_0. We compare the additive model (1.2) with the oracle model

$$Z^i = f_1(X_1^i) + \epsilon^i, \quad i = 1, \ldots, n, \tag{1.3}$$

where X_1^i and ε_i are the same variables as in the additive model (1.2). We will choose $\hat{f}_1$ such that for an undersmoothed estimator $\hat{f}_1^{\text{oracle}}$ in the oracle model it holds that

$$\|\hat{f}_1 - \hat{f}_1^{\text{oracle}}\|_\infty = o_P(\delta_n). \tag{1.4}$$

Define now $\tilde{\hat{f}}_1^{\text{oracle}}$ as the estimator obtained from applying the smoothing operation of $\tilde{f}_1^{\text{oracle}}$ to the regression problem with covariate values X_1^i and "response" values $\hat{f}_1^{\text{oracle}}(X_1^i)$. Similarly, $\tilde{\hat{f}}_1$ is defined as the estimator resulting from the smoothing operation of $\tilde{f}_1^{\text{oracle}}$ applied to the regression problem with covariate values X_1^i and response values $\hat{f}_1(X_1^i)$. Our main assumption on $\tilde{f}_1^{\text{oracle}}$ is that

$$\|\tilde{\hat{f}}_1^{\text{oracle}} - \tilde{f}_1^{\text{oracle}}\|_\infty = o_P(\delta_n). \tag{1.5}$$

This is a natural assumption that is valid for many smoothing estimators. It says that a smoothing operation applied after an undersmoothing of the data is asymptotically equivalent to a single application of the smoothing. For our next argument we need that the smoothing operation of $\tilde{\hat{f}}_1^{\text{oracle}}$ has the following continuity property for all $\delta > 0$ and a constant $C > 0$:

> A change in the responses by a maximal amount (1.6)
> less than δ does not lead to a change larger than $C\delta$ in the resulting smoother.

This gives, with (1.4), that

$$\begin{aligned} \|\tilde{\hat{f}}_1 - \tilde{\hat{f}}_1^{\text{oracle}}\|_\infty &\leq C \max_{1 \leq i \leq n} |\hat{f}_1(X_1^i) - \hat{f}_1^{\text{oracle}}(X_1^i)| \\ &\leq C \|\hat{f}_1 - \hat{f}_1^{\text{oracle}}\|_\infty = o_P(\delta_n). \end{aligned}$$

Thus, from (1.5) we get that

$$\|\tilde{\hat{f}}_1 - \tilde{f}_1^{\mathrm{oracle}}\|_\infty = o_P(\delta_n). \tag{1.7}$$

We now choose $\tilde{f}_1 = \tilde{\hat{f}}_1$. Because of (1.7) this estimator is asymptotically equivalent to $\tilde{f}_1^{\mathrm{oracle}}$. The main mathematical difficulty in our line of arguments lies in the choice of $\hat{f}_1$ and $\hat{f}_1^{\mathrm{oracle}}$ and in the proof of (1.4) for this choice. Once (1.4) is established for our choice of $\hat{f}_1$, we get relatively easily that for all smoothers $\tilde{f}_1^{\mathrm{oracle}}$ in the oracle model satisfying (1.5) and (1.6), the estimator $\tilde{f}_1 = \tilde{\hat{f}}_1$ is asymptotically equivalent to the oracle estimator $\tilde{f}_1^{\mathrm{oracle}}$. This is our asymptotic optimality theory for additive models.

This theoretical program has been carried out in [15] for additive models with a fixed number of functions q. In this paper we will go far beyond this restriction and allow the total number of functions q as well as the number s_0 of nonzero functions to grow with n, allowing also the case in which $q > n$.

The discussion of additive models goes back to the influential work of C.J. Stone, who pointed out that additive nonparametric models efficiently circumvent the poor accuracy of high-dimensional regression functions and yet still provide high flexibility for statistical modeling; see [30]. In recent years, estimation of nonparametric high-dimensional sparse additive models has been considered in a series of papers. Earlier references are [22], [1], [35], [16], and [29], where L_1-penalty based methods have been used for variable selection in additive models. For a related paper on model choice in nonparametric regression, see [4]. For sparse models in functional linear regression see [21], and for sparse models in varying coefficient models, see [26]. Rates of convergence for a fixed number of non-zero components have been discussed in [22] and [16]. Rates of convergence for settings that allow for an increasing number of non-zero components were studied in [25], [27], [31], and [19]. The latter paper also includes more general additive models where the summands are not necessarily functions of differing one-dimensional arguments. The paper [18] proposes a two-step procedure in which variables are selected in a first step and a rate-optimal estimator is implemented in the second step. In [13] sure independence screening is proposed for ultra-high dimensional additive models.

All of these papers discuss only variable selection and/or optimal rates of convergence. None of them presents any asymptotic distribution results for the proposed estimators, which severely restricts their range of statistical application. In particular, there are no procedures in the current literature for the construction of valid confidence regions or tests of hypotheses in the high-dimensional sparse additive model. An asymptotic distribution theory for the Lasso estimator is complex because model choice is implicitly embedded in the construction of the estimator. For high-dimensional parametric models modifications for the Lasso estimator have been proposed that allow a complete asymptotic distribution theory. The method is to replace in the least-squares estimator each orthogonal projection of a covariate onto the other covariates with a projection of relaxed orthogonality (using the Lasso) and then to subtract an estimate of the resulting bias, which is constructed with Lasso estimates of the parameters. The result is a non-sparse estimator, and it has been called the desparsified Lasso for this reason. Influential discussions of this method include [3] and [2], [36], [32], and [17]. We will use this method in the nonparametric context of additive models. Desparsification has also been used in nonparametrics in [23] for the discussion of undersmoothing estimators in additive models. Their estimator of a component f_1 is based on fits of the model $f_1(x_1) + f_2(x_2, x_1) + ... + f_q(x_q, x_1)$ with $\mathrm{E}[f_k(X_k, x_1)]$ for all x_1. In our model we assume that error variables are homoscedastic. Things change in the case of heteroscedastic errors, as has been pointed out in [11] for fixed q.

In this paper we will use the desparsification technique in the definition of our presmoothing estimator $\hat{f}_1$, whereby its resmoothed version $\tilde{f}_1$ evaluated at a point will be asymptotically normally distributed. This will allow for the construction of pointwise confidence intervals and of global confidence bands for f_1 based on the resmoothed estimator $\tilde{f}_1$. Moreover, these confidence intervals will have oracle width due to the asymptotic oracle properties of the resmoothed estimator $\tilde{f}_1$.

The paper is organized as follows. In the next section we will describe the construction of our two-step procedure. Section 3 contains our main results and a simulation study. Section 4 gives an outline of the structure of the proofs. The main part of the proofs can be found in Section 5. More details of the proofs are collected in the supplementary material of this paper.

1.1. **Notation.**

The space $L^2(\mathbb{P}^X)$ is a Hilbert space with the inner product $\langle g, h\rangle = \mathbb{E}[g(X)h(X)]$ and the corresponding norm $\|g\| = \sqrt{\langle g, g\rangle}$. Let $\|\cdot\|_\infty$ denote the supremum norm on $L^\infty(\mathbb{P}^X)$. Let $\langle\cdot,\cdot\rangle_n$ denote the empirical inner product defined by

$$\langle g, h\rangle_n = \frac{1}{n}\sum_{i=1}^{n} g(X^i)f(X^i)$$

and let $\|\cdot\|_n$ denote the corresponding empirical norm. Let $\|\cdot\|_2$ denote the Euclidean norm. Furthermore, let $\mathbf{Y}=(Y^1,\dots,Y^n)^T$, $\boldsymbol{\epsilon}=(\epsilon^1,\dots,\epsilon^n)^T$, and for $f\in L^2(\mathbb{P}^X)$, $\mathbf{f}=(f(X^1),\dots,f(X^n))^T$. By C we denote a constant depending only on the (minor) quantities t, t_1, c_1, C_0, and C_1. We make use of the convention that the constant C need not represent the same value at each occurrence.

2. The estimator

2.1. **Piecewise polynomials.** For $j=1,\dots,q$, let $t_j\geq 0$ and $m_j\geq 1$ be integers and let U_j be the space of piecewise polynomials in the variable $x_j\in[0,1]$ of maximal degree t_j defined on the intervals

$$I_{jk}=\left(\frac{k}{m_j},\frac{k+1}{m_j}\right],$$

$k=0,\dots,m_j-1$. Thus each function $g_j\in U_j$ has the property that, restricted to each interval I_{jk}, it is a polynomial of degree at most t_j. Let Q_l, $l\geq 0$ be the sequence of the Legendre polynomials (see, e.g., the book by Whittaker and Watson [34] for the definition and fundamental properties of the Legendre polynomials). Then the shifted and rescaled polynomials $R_l(x)=\sqrt{2l+1}Q_l(2x-1)$, $x\in[0,1]$, are orthonormal with respect to the inner product induced by the Lebesgue measure on $[0,1]$. For $k=0,\dots,m_j-1$ and $l=1,\dots,t_j+1$, we now define

$$b_{j,k(t_j+1)+l}(x_j)=\sqrt{m_j}R_{l-1}\left(m_j\left(x_j-\frac{k}{m_j}\right)\right)$$

for $x\in I_{jk}$ (and equal to zero otherwise). Hence

$$b_{j,k(t_j+1)+1},\dots,b_{j,k(t_j+1)+t_j+1}$$

is an orthonormal basis of the functions in U_j which are zero outside the interval I_{jk}, and we conclude that

$$b_{j,1},\dots,b_{j,m_j(t_j+1)}$$

is an orthonormal basis of U_j with respect to the Lebesgue measure. The space of piecewise polynomials has a number of important properties, among which we mention that

$$\|g_j\|_\infty^2\leq(t_j+1)^2m_j\int_0^1 g_j^2(x_j)dx_j \tag{2.1}$$

for each $g_j\in U_j$ (see, e.g., [6, Equation (7)]).

In the following we suppose that

$$m_2=\dots=m_q$$

and that

$$t_2 = \cdots = t_q,$$

that is, we suppose that $U_2, \ldots, U_q$ are defined with piecewise polynomials of the same order and on the same intervals. We let $m = \max_j m_j$ and $t = \max_j t_j$. Moreover, let

$$V_1 = U_1$$

and for $j = 2, \ldots, q$, define from $U_2, \ldots, U_q$ the centered function spaces

$$V_j = \{g_j \in U_j : \mathbb{E}\left[g_j(X_j)\right] = 0\}.$$

In the sequel we will consider the spaces $U_1, \ldots, U_q$ as subspaces of $L^2(\mathbb{P}^X) \cap L^\infty(\mathbb{P}^X)$. We let $d_j = \dim V_j$ and $d = \max_j d_j$. Hence (under Assumption 1), we have $d_1 = m_1(t_1 + 1)$ and $d_2 = \cdots = d_p = m_2(t_2 + 1) - 1$. We let

$$V = \sum_{j=1}^{q} V_j$$

and abbreviate the space of additive functions with components coming from $V_2, \ldots, V_q$ as

$$V_{-1} = \sum_{j=2}^{q} V_j.$$

Finally, let $\Pi_{-1} : L^2(\mathbb{P}^X) \to V_{-1}$ be the orthogonal projection from $L^2(\mathbb{P}^X)$ to V_{-1} given by

$$\Pi_{-1} f = \operatorname*{argmin}_{g \in V_{-1}} \|f - g\|^2.$$

2.2. **The Lasso estimators.** To reconstruct the desparsified Lasso estimator in the additive model context, we will need Lasso estimators of $f_1, \ldots, f_q$ as well as a Lasso version of the projection of the V_1 basis functions onto V_{-1}.

We first define the nonparametric Lasso estimator

$$\hat{f}^L = \sum_{j=1}^{q} \hat{f}_j^L$$

of f by

$$\left(\hat{f}_1^L, \ldots, \hat{f}_q^L\right) = \operatorname*{argmin}_{g_j \in V_j} \left\{ \left\| Y - \sum_{j=1}^{q} g_j \right\|_n^2 + 2\lambda \sum_{j=1}^{q} \|g_j\|_n \right\},$$

where $\lambda > 0$ is some tuning parameter. This estimator will be used to correct the bias resulting from the replacement in the least-squares estimator of the orthogonal projection of the V_1 basis functions onto V_{-1} with a projection of relaxed orthogonality.

For $k = 1, \dots, d_1$, we define the nonparametric Lasso estimator

$$\hat{\Pi}^L_{-1} b_{1k} = \sum_{j=2}^{q} (\hat{\Pi}^L_{-1} b_{1k})_j \in V_{-1}$$

of $\Pi_{-1} b_{1k}$ by

$$\left((\hat{\Pi}^L_{-1} b_{1k})_2, \dots, (\hat{\Pi}^L_{-1} b_{1k})_q\right) = \operatorname*{argmin}_{g_j \in V_j} \left\{ \left\| b_{1k} - \sum_{j=2}^{q} g_j \right\|_n^2 + 2\eta \sum_{j=2}^{q} \|g_j\|_n \right\},$$

where $\eta > 0$ is some tuning parameter. Moreover, we extend $\hat{\Pi}^L_{-1}$ linearly to all of V_1 as follows:

$$\hat{\Pi}^L_{-1} : V_1 \to V_{-1}, \sum_{k=1}^{d_1} \alpha_k b_{1k} \mapsto \sum_{k=1}^{d_1} \alpha_k \hat{\Pi}^L_{-1} b_{1k},$$

which can be seen as an empirical version of Π_{-1} restricted to V_1. We use $\hat{\Pi}^L_{-1}$ as the projection from V_1 to V_{-1} of relaxed orthogonality in the desparsified Lasso construction.

Remark 1. In practice, the Lasso estimators should be based on the spaces

$$V_j^n = \left\{ g_j \in U_j : \frac{1}{n} \sum_{i=1}^{n} g_j(X_j^i) = 0 \right\}$$

instead of V_j ($j = 2, \dots, q$), which is achieved by centering each basis function b_{jk} by its empirical mean $\langle b_{jk}, 1 \rangle_n$. However, since the difference between the centering $\langle b_{jk}, 1 \rangle$ and $\langle b_{jk}, 1 \rangle_n$ is of order $n^{-1/2}$, we choose, in our analysis, to proceed using the spaces V_j instead of V_j^n in order to avoid cumbersome technicalities.

2.3. The presmoothing estimator.

Let $\phi_{11}, \dots, \phi_{1d_1}$ be any basis of V_1 and let us denote by $\hat{\boldsymbol{\beta}}_1^L \in \mathbb{R}^{d_1}$ the vector of coefficients of $\hat{f}_1^L$ with respect to the basis $\phi_{11}, \dots, \phi_{1d_1}$. We now define

$$\hat{f}_1 = \sum_{k=1}^{d_1} \hat{\beta}_{1k} \phi_{1k},$$

where $\hat{\boldsymbol{\beta}}_1 = (\hat{\beta}_{11}, \dots, \hat{\beta}_{1d_1})$ is defined by

$$\hat{\boldsymbol{\beta}}_1 = \hat{\boldsymbol{\beta}}_1^L + \left(\mathbf{Z}_1^T \mathbf{X}_1\right)^{-1} \mathbf{Z}_1^T (\mathbf{Y} - \hat{\mathbf{f}}^L) = \left(\frac{1}{n} \mathbf{Z}_1^T \mathbf{X}_1\right)^{-1} \frac{1}{n} \mathbf{Z}_1^T (\mathbf{Y} - \hat{\mathbf{f}}^L_{-1}),$$

where

$$\mathbf{X}_1 = \big(\phi_{1k}(X_1^i)\big)_{1\le i\le n, 1\le k\le d_1}$$

and

$$\mathbf{Z}_1 = \Big((\phi_{1k} - \hat{\Pi}_{-1}^L \phi_{1k})(X^i)\Big)_{1\le i\le n, 1\le k\le d_1}.$$

So far, we have constructed an estimator only for the case in which $(1/n)\mathbf{Z}_1^T\mathbf{X}_1$ is invertible. However, since we will show that $(1/n)\mathbf{Z}_1^T\mathbf{X}_1$ is invertible with high probability, it is not necessary for our theoretical considerations to define the estimator in the case in which this matrix is not invertible.

Our presmoothing estimator is chosen to be piecewise polynomial. For finite-sample applications the discontinuity of this estimator at knot points may affect the shape of the second-step estimator. This could be avoided by using least-squares splines or other alternative smoothing methods in the presmoothing step instead of piecewise polynomials. We conjecture that the whole theory of this paper would go through with the choice of least-squares splines in the presmoothing step, but at the cost of considerable inflation in our notation. For this reason we pursue our theory with piecewise polynomials, which makes our exposition more transparent. In our simulations we explore the performance of our estimator when least-squares splines are chosen in the presmoothing step.

3. Main results

3.1. Assumptions.

Our main results make use of the following assumptions.

Assumption 1. *Suppose that for $j = 1, \dots, q$, X_j takes values in $[0,1]$ and has a density p_j with respect to the Lebesgue measure on $[0,1]$ which satisfies $c_1 \le p_j \le 1/c_1$ for some constant $c_1 > 0$. Moreover, suppose that for $j = 2 \dots, q$, (X_1, X_j) has a density p_{1j} with respect to the Lebesgue measure on $[0,1]^2$ which is bounded from above by $1/c_1$.*

Assumption 2 introduces a geometric quantity ρ_0 which governs the degree of collinearity between the spaces V_1 and V_{-1}. The closer ρ_0 is to 1, the harder it is to distinguish the effects of X_1 from those of $X_2, \dots, X_q$.

Assumption 2. *Suppose that there is a constant $0 \le \rho_0 < 1$ such that for all $g_1 \in V_1$,*

$$\|\Pi_{-1} g_1\| \le \rho_0 \|g_1\|.$$

Note that ρ_0 can also be defined as the minimal angle between V_1 and V_{-1} (see, e.g., [33]).

Assumption 3. *Suppose that there exist some $r_1, r_2 > 0$ and a subset $J_0 \subseteq \{1, \ldots, q\}$ with $1 \in J_0$ and $|J_0| \leq s_0$ such that for each $j \in J_0$ there is a $g_j^* \in V_j$ satisfying*

$$\|f_1 - g_1^*\|_\infty \leq C_0 d_1^{-r_1}$$

if $j = 1$ and

$$\|f_j - g_j^*\|_\infty \leq C_0 d_2^{-r_2}$$

otherwise for some constant $C_0 > 0$. Moreover, setting

$$g^* = \sum_{j \in J_0} g_j^*,$$

suppose that

$$\|f - g^*\|_\infty \leq C_0 \left(d_1^{-r_1} + s_0 d_2^{-r_2}\right).$$

Assumption 4 states that the projection of each basis function of V_1 onto the space V_{-1} may be approximated sufficiently well by its projection onto a subspace of V_{-1} of s_1 or fewer additive components.

Assumption 4. *For each $k = 1, \ldots, d_1$, suppose that there is a subset $J_k \subseteq \{2, \ldots, q\}$ with $|J_k| \leq s_1$, such that there is a decomposition*

$$\Pi_{J_k} b_{1k} - \Pi_{-1} b_{1k} = \sum_{j=2}^{q} v_j$$

with $v_j \in V_j$ satisfying

$$\sum_{j=2}^{q} \|v_j\| \leq C_1 \sqrt{s_1} \sqrt{\frac{d}{n}}$$

for some constant $C_1 > 0$. Finally, suppose that $d \leq n$ and

$$\eta \geq \sqrt{\frac{d}{n}}.$$

Assumption 5 (Theoretical compatibility conditions)**.** *Suppose that there is a real number $0 < \phi \leq 1$ such that*

$$\sum_{j \in J_0} \|g_j\|^2 \leq \left\| \sum_{j=1}^{q} g_j \right\|^2 / \phi^2$$

for all $(g_1, \ldots, g_q) \in (V_1, \ldots, V_q)$ satisfying

$$\sum_{j=1}^{q} \|g_j\| \leq 8\sqrt{3} \sum_{j \in J_0} \|g_j\|. \tag{3.1}$$

Moreover, for $k = 1, \ldots, d_1$, suppose that

$$\sum_{j\in J_k} \|g_j\|^2 \le \Big\| \sum_{j=2}^{q} g_j \Big\|^2 / \phi^2$$

for all $(g_2, \ldots, g_q) \in (V_2, \ldots, V_q)$ satisfying

$$\sum_{j=2}^{q} \|g_j\| \le 8\sqrt{3} \sum_{j\in J_k} \|g_j\|.$$

Let $0 \le \psi \le 1$ be the largest number such that

$$\sum_{j\in J_k} \|g_j\|^2 \le \left\| \sum_{j\in J_k} g_j \right\|^2 / \psi^2$$

for all $g_j \in V_j$, $j \in J_k$ and all $k = 1, \ldots, d_1$. By Assumption 5, we know that $\psi \ge \phi > 0$. Note that while the constant ϕ plays an important role in the analysis of the Lasso estimators, the (weaker) constant ψ will be used in the analysis of the norms of the $\Pi_{-1} b_{1k}$.

3.2. **Main result: bound for the presmoothing estimator.** In this section, we state our main result, which is a precise, finite-sample statement of equation (1.4). Recall the convention that C denotes a constant depending only on the quantities t, t_1, c_1, C_0, and C_1 and that C is not necessarily the same at each occurrence. For our main result we suppose that the Lasso estimators are defined with

$$\lambda = 2\sigma\sqrt{\frac{d}{n}} + 2\sigma\sqrt{\frac{2x + 2\log q}{n}}$$

and

$$\eta = C\left(\sqrt{\frac{d(x + \log d_1 + \log q)}{n}} + \frac{\sqrt{s_1} d(x + \log d_1 + \log q)}{\psi n}\right)$$

where $x > 1$. Moreover, we also introduce

$$\delta = \frac{C}{\psi}\sqrt{\frac{s_1 d(x + \log d + \log q)}{n}}.$$

Note that an explicit expression of the constants C can be found in Appendix E in the supplementary material. Our theory requires two conditions on the dimensions d_1 and d. First, we need that

$$\left(\frac{s_1\delta}{\psi^2} + \frac{s_1\sqrt{d_1}\eta}{\psi\phi} + \frac{s_1 d_1 \eta^2}{\phi^2}\right) \le (1 - \rho_0)^2 / C, \tag{3.2}$$

where C is the constant in Proposition 6. This condition will ensure that an empirical version $\hat{\rho}_0$ of ρ_0 is strictly less than 1 with high probability. Second, we need that

$$\max(s_0, s_1)\left(\frac{d}{\sqrt{n}} + \sqrt{\frac{d(x+\log q)}{n}} + \frac{d(x+\log q)}{n}\right) \leq \phi^2/C \tag{3.3}$$

where C comes from Proposition 12. This second condition is needed in the analysis of the Lasso estimators. It implies that certain empirical compatibility conditions are satisfied with high probability. Note that setting $x = y = \log q$ and considering the geometric quantities ρ_0, ϕ, ψ as constants, these two conditions are satisfied if

$$s_1\sqrt{s_1}\sqrt{\frac{d\,(\log q + \log d)}{n}} \quad \text{and} \quad \max(s_0, s_1)\frac{d\,(\log q + \log d)}{\sqrt{n}}$$

are bounded by a constant. In Section 4.2 we decompose $\hat{f}_1 - \hat{f}_1^{\text{oracle}}$ into three main terms: the approximation error term, the improved Lasso bias term, and the variance term. To these terms correspond the following three error terms:

$$\Delta_1 = \frac{1}{\psi(1-\rho_0)}\left(s_1 d_1^{-r_1} + s_1 s_0 d_2^{-r_2}\right),$$

and

$$\Delta_2 = \frac{1}{\psi(1-\rho_0)}\Big((\eta/\lambda)\sqrt{s_1 d_1}\left(d_1^{-r_1} + s_0 d_2^{-r_2}\right)^2 + s_0\sqrt{s_1}\sqrt{d_1}\lambda\eta\Big),$$

and, for $y > 0$,

$$\Delta_3 = \frac{1}{\psi(1-\rho_0)}\sqrt{\frac{s_1(\log d_1 + y)}{n}}.$$

We now have:

Theorem 1. *Suppose that Assumptions 1-5 hold. Moreover, suppose that* (3.2) *and* (3.3) *are satisfied. Then we have that*

$$\mathbb{P}\Big(\|\hat{f}_1 - \hat{f}_1^{\text{oracle}}\|_\infty \geq C\left(\Delta_1 + \Delta_2 + \Delta_3\right)\Big) \leq 4\exp(-x) + \exp(-y).$$

3.3. Application of the main results to the resmoothing step.

We now consider the resmoothing step discussed in the introduction which makes use of the presmoothed data. Our main result, presented in Theorem 1, established that

$$\|\hat{f}_1 - \hat{f}_1^{\text{oracle}}\|_\infty \leq C(\Delta_1 + \Delta_2 + \Delta_3)$$

with probability greater than or equal to $1 - 4\exp(-x) - \exp(-y)$. We now consider several classes of estimators $\tilde{f}_1^{\text{oracle}}$ for the oracle model.

As explained in the introduction, we want to construct a two-step estimator $\tilde{\hat{f}}_1$ for f_1 in the additive model for which $\|\tilde{\hat{f}}_1 - \tilde{f}_1^{\mathrm{oracle}}\|_\infty$ is small. This requires verifying two things: that $\|\tilde{\hat{f}}_1^{\mathrm{oracle}} - \tilde{f}_1^{\mathrm{oracle}}\|_\infty$ is small and that the second smoothing step is continuous with respect to changes in the inputs as per (1.6).

We start by discussing two estimators for which $\tilde{\hat{f}}_1^{\mathrm{oracle}} = \tilde{f}_1^{\mathrm{oracle}}$ and thus $\|\tilde{\hat{f}}_1^{\mathrm{oracle}} - \tilde{f}_1^{\mathrm{oracle}}\|_\infty = 0$ trivially holds. For such estimators only (1.6) has to be verified. We will see that the equality $\tilde{\hat{f}}_1^{\mathrm{oracle}} = \tilde{f}_1^{\mathrm{oracle}}$ holds when $\tilde{f}_1^{\mathrm{oracle}}$ is a least-squares piecewise polynomial smoother or a least-squares spline estimator and when this type of smoothing is used in the second step of the two-step estimator $\tilde{\hat{f}}_1^{\mathrm{oracle}}$. For these two classes of estimators there are two conditions which are necessary for the equality $\tilde{\hat{f}}_1^{\mathrm{oracle}} = \tilde{f}_1^{\mathrm{oracle}}$ to hold. The first is that the B-splines or polynomials, respectively, in the second step have the same order as the polynomials in the first step. The second is that the grid of the first step is a sub-grid of the second step. Under these two conditions the equality follows from the projection interpretation of the estimators. Note that the projection of a vector x onto a linear space E_1 is equivalent to the projection of x^* onto E_1 if x^* is the projection of x onto a linear space E_2 and E_1 is a linear subspace of E_2. For the additive model, define $\tilde{f}_1^{\mathrm{pol}}$ or $\tilde{f}_1^{\mathrm{spl}}$, respectively, as the two-step estimators where least-squares polynomial or B-spline fitting has been used in the second step. We get the following result (For a proof see Appendix D in the supplementary material).

Theorem 2. *Suppose that the assumptions of Theorem 1 hold and let $\tilde{f}_1^{pol}$ and $\tilde{f}_1^{spl}$ be two-step estimators for which least-squares polynomial and B-spline fitting, respectively, with an equidistant partition of m^* intervals has been used in the second step. Furthermore, suppose that the order of the polynomials or B-splines used in the second step is the same as that of the polynomials used in the first step and suppose that the number of intervals m_1 used in the first step is a multiple of m^*. Then*

$$\begin{aligned} \|\tilde{f}_1^{pol} - \tilde{f}_1^{oracle,pol}\|_\infty &\leq C(\Delta_1 + \Delta_2 + \Delta_3), \\ \|\tilde{f}_1^{spl} - \tilde{f}_1^{oracle,spl}\|_\infty &\leq C(\Delta_1 + \Delta_2 + \Delta_3) \end{aligned}$$

with probability greater than or equal to $1 - 4\exp(-y) - \exp(-y)$. Here $\tilde{f}_1^{oracle,pol}$ and $\tilde{f}_1^{oracle,spl}$ are the one-step estimators in the oracle model based on least-squares polynomial or B-spline fitting, repectively.

The assumption that the grid used in the first step is based on subdividing the grid used in the second step greatly simplifies the proof. But by using more refined arguments, it can be shown that this assumption is not necessary.

For an asymptotic interpretation let us assume that

$$\log\log q = o(\log n), \;\; q \to \infty, \tag{3.4}$$

$$s_0 = O(n^{\gamma_0}), \quad s_1 = O(n^{\gamma_1}) \tag{3.5}$$

for some constants $0 \le \gamma_0 < 1/2$ and $0 \le \gamma_1 \le 1/4$. Then we have (see proof in Appendix D in the supplementary material) that for $\beta > 0$, the following is true: For the preliminary estimator m_1 and m can be chosen such that with $x = y = \log q$

$$\Delta_1 + \Delta_2 + \Delta_3 = o(n^{-\beta}), \tag{3.6}$$

if

$$\left(1+\frac{1}{r_2}\right)\gamma_0 + \left(\frac{1}{2}+\frac{1}{2r_1}+\frac{1}{r_2}\right)\gamma_1 \;<\; 1-\left(1+\frac{1}{2r_1}+\frac{1}{r_2}\right)\beta \tag{3.7}$$

$$2(\gamma_0 \vee \gamma_1) + \frac{2}{r_1}\gamma_1 \;<\; 1 - \frac{2}{r_1}\beta, \tag{3.8}$$

$$\frac{2}{r_2}(\gamma_0 \wedge \gamma_1) + \left(2+\frac{2}{r_2}\right)(\gamma_0 \vee \gamma_1) \;<\; 1 - \frac{2}{r_2}\beta. \tag{3.9}$$

Equation (3.6) implies that $\|\tilde{f}_1^{\mathrm{pol}} - \tilde{f}_1^{\mathrm{oracle,pol}}\|_\infty = o_P(n^{-\beta})$ and $\|\tilde{f}_1^{\mathrm{spl}} - \tilde{f}_1^{\mathrm{oracle,spl}}\|_\infty = o_P(n^{-\beta})$. This result can be applied to check whether an estimator in the additive model exists that is asymptotically equivalent to a rate-optimal spline or polynomial estimator in the oracle model. For rate-optimal estimation in the oracle model, the number of intervals m_1 should be a constant times $n^{1/(2r_1+1)}$, which results in a pointwise rate of $n^{-r_1/(2r_1+1)}$. To establish the existence of an asymptotically oracle-equivalent estimator, we thus have to show that (3.7)–(3.9) hold with $\beta = r_1/(2r_1+1)$. Inequalities (3.7)–(3.9) hold for $\gamma_0, \gamma_1 \ge 0$ small enough as long as the right hand sides of the inequalities are positive. The right hand sides are positive as long as $r_2 \ge 2r_1/(2r_1+1) = 2\beta$ and $r_1 > 1/2$. In particular, we see that for r_1 choices of r_2 with $r_2 < r_1$ are allowed. Thus we do not require the nuisance additive components $f_2, ..., f_q$ to be as smooth as f_1.

We now discuss local polynomial estimators. The degree of the local polynomial estimator is denoted by k. Define $(\tilde{a}_0, ..., \tilde{a}_k)$ as the minimum of

$$\sum_{i=1}^{n} \left[Z^i - a_0 - ... - a_k(X_1^i - x)^k\right]^2 K_h(X_1^i - x)$$

over $(a_0, ..., a_k) \in \mathbb{R}^{k+1}$ and set $\tilde{f}_1^{j,\mathrm{oracle,lpol}}(x) = \tilde{a}_j$. This is an estimator of the j-th derivative of f_1 in the oracle model. Here, $K_h(u) = h^{-1}K(h^{-1}u)$ is a kernel with kernel function K and bandwidth h. Similarly, we define $\tilde{f}_1^{j,\mathrm{lpol}}(x) = \tilde{\hat{a}}_j$, where now $(\tilde{\hat{a}}_0, ..., \tilde{\hat{a}}_k)$ minimizes

$$\sum_{i=1}^{n} \left[\hat{Y}^i - a_0 - ... - a_k(X_1^i - x)^k\right]^2 K_h(X_1^i - x)$$

with $\hat{Y}^i = \hat{f}_1(X_1^i)$. For this class of estimators we have the following result.

Theorem 3. *Suppose that the assumptions of Theorem 1 hold. Suppose further that the kernel K is a probability density function with bounded support, $[-1, 1]$ say, that it has an absolutely bounded derivative and that the bandwidth h fulfills $c_1 n^{-\eta_1} \leq h \leq c_2 n^{-\eta_2}$ for some $c_1, c_2 > 0$ and $0 < \eta_2 \leq \eta_1 < 1/3$. Furthermore, assume that for a value $\rho_1 \leq k + 1$ the function f has an absolutely bounded derivative of order ρ_1. Then it holds for $j = 0, \ldots, k$ that*

$$\begin{aligned} h^j \|\tilde{f}_1^{j,oracle,lpol} - \tilde{f}_1^{j,lpol}\|_\infty \ \leq \ & C\big[\Delta_1 + \Delta_2 + \Delta_3 + d_1^{-\rho_1} \\ & +(d_1 h)^{-1}(nh)^{-1/2}(\sqrt{\log(n)} + \sqrt{z})\big] \end{aligned}$$

uniformly for all h with $c_1 n^{-\eta_1} \leq h \leq c_2 n^{-\eta_2}$ with probability greater than or equal to $1 - 4\exp(-x) - \exp(-y) - \exp(-z)$.

Applying this theorem with $j = 0$ and $k = \rho_1 - 1$ and with a choice of h of optimal order $n^{-1/(2\rho_1+1)}$, we can show that the two-step estimator is asymptotically equivalent to a local polynomial estimator in the oracle model if (3.7)–(3.9) holds with $\beta = \rho_1/(2\rho_1 + 1)$. Furthermore, one can argue in the same way as in the discussion after Theorem 2 to get asymptotic oracle equivalence of the two-step estimator and the oracle estimator.

We conclude this section by discussing a minimax theorem. To simplify notation we formulate this theorem asymptotically. For the first additive component we assume that

$$f_1 \in \mathcal{S} = \left\{ g : [0, 1] \to \mathbb{R} : \int_0^1 g^{(\rho_1)}(x)^2 \, dx \leq C_{\mathcal{S}} \right\}, \tag{3.10}$$

where $\rho_1 \geq 1$ and $C_{\mathcal{S}} > 0$. By the Sobolev embedding theorem this implies that for all $f_1 \in \mathcal{S}$ there is a $g_1^* \in V_1$ satisfying

$$\|f_1 - g_1^*\|_\infty \leq C_0 d_1^{-r_1}$$

with $r_1 = \rho_1 - 1/2$, so that the first part of Assumption 3 is satisfied.

We now define a class $\mathcal{F}_n = \mathcal{F}_n(\rho_1, C_{\mathcal{S}}, c_1, C_0, C_1, r_2, \phi, \rho_0, \gamma_0, C^0, \gamma_1, C^1)$ of tuples $(f_1, \dots, f_q, p)$ of additive components $f_1, \dots, f_q$ and densities p of $(X_1, \dots, X_q)$, where it is assumed that these functions fulfill Assumptions 1-5 with constants c_1, C_0, C_1, r, $r_1 = \rho_1 - 1/2$, ϕ, ρ_0 and s_0, s_1 with $s_0 \leq C^0 n^{\gamma_0}$ and $s_1 \leq C^1 n^{\gamma_1}$, $q \leq \omega_n$ and where $f_1 \in \mathcal{S}$. Here ω_n is a fixed sequence with $\log\log \omega_n = o(\log n)$. We now state our minimax theorem.

Theorem 4. *Suppose that for some constants r_2, ρ_1, γ_0, and γ_1,* (3.7)–(3.9) *hold with $\beta = \rho_1/(2\rho_1 + 1)$ and $r_1 = \rho_1 - 1/2$. Then there exists an estimator $\tilde{f}_1$ in the additive model with*

$$n^{2\rho_1/(2\rho_1+1)} \kappa(p_1)^{-1} \mathbf{E}_\epsilon \left[\int_0^1 (\tilde{f}_1(x) - f_1(x))^2 \, dx \right] = 1 + o_P(1)$$

uniformly over $(f_1, \dots, f_q, p) \in \mathcal{F}_n(\rho_1, C_{\mathcal{S}}, c_1, C_0, C_1, r_2, \phi, \rho_0, \gamma_0, C^0, \gamma_1, C^1)$ for positive constants $C_{\mathcal{S}}, c_1, C_0, C_1, \phi, C^0, C^1 > 0$ and $0 \leq \rho_0 < 1$. Here

$$\kappa(p_1) = \left\{ (2\rho_1 + 1) C_{\mathcal{S}} \left(\frac{\sigma^2 \rho_1}{\pi(\rho_1 + 1)} \int_0^1 p_1^{-1}(x) \, dx \right)^{2\rho_1} \right\}^{1/(2\rho_1+1)}$$

and $\mathbf{E}_\epsilon$ denotes the conditional expectation, given $X_1, \dots, X_q$.

The proof of this theorem is similar to those of the previous two theorems; see also the proof of Theorem 6 in [15]. The minimax estimator can be chosen as two-step estimator according to the construction presented in this paper. The value $\kappa(p_1)$ is the asymptotic minimax risk in the oracle model, which has been established in [10]; see also the discussion in [15], where a minimax theorem for additive models was proved for the case in which q is fixed. Theorem 4 states that, under our assumptions, the asymptotic minimax risk for estimators in the oracle model can be achieved in the additive model. This holds for $\gamma_0, \gamma_1 \geq 0$ small enough as long as the right hand sides of (3.7)–(3.9) are positive. Such choices of γ_0, γ_1 exist for all values r_2, ρ_1 with $\rho_1 \geq 2$ and $r_2 > (2\rho_1^2 - \rho_1)/(2\rho_1^2 - 1)$. Thus we have the same asymptotic minimax bound in the additive model as in the oracle model as long as γ_0, γ_1 are small enough. We conjecture that the minimax result continues to hold under weaker sparsity conditions and also under conditions that include the case $\rho_1 = 1$. Note that our theory gives bounds for L_∞ norms between the pilot estimators but for the stated minimax theorem only L_2 norms are needed.

3.4. **Simulation results.** We generated data sets of n independent observations from the model

$$Y = \sum_{j=1}^{q} (1/j) f(X_j) \mathbf{1}(j \le s_0) + \varepsilon, \tag{3.11}$$

where ε is a Gaussian error term independent of X with zero mean and unit variance. We then used the proposed presmoothing and resmoothing two-step procedure to estimate and make pointwise inference on $f_1 := f$. We chose cubic B-splines in both the presmoothing and resmoothing steps, as these are a common choice in practice. Simulations were run at all combinations of $n = 100, 500, 1000$, $q = 50, 200$ and $f(x)$ equal to

$$\begin{aligned} \text{sine}(x) &= 2\text{sin}(2x) \\ \text{line}(x) &= x \\ \text{expo}(x) &= \exp(-x) - (2/5)\text{sinh}(5/2) \\ \text{quad}(x) &= x^2 - 25/12, \end{aligned}$$

which come from [25].

The number of non-null components s_0 is set to $s_0 = \lceil q/20 \rceil$. The non-null functions are scaled such that they have decreasing magnitude. This is to underscore the fact that we do not require any so-called "beta-min" conditions for our procedure to work; that is, there is no lower bound which the norms of our non-null functions must exceed in order for our procedure to produce valid pointwise confidence intervals.

The covariates $X_1, \dots, X_q$ are generated such that X_j is marginally uniformly distributed on $(-2.5, 2.5)$ for $j = 1, \dots, q$ and such that the correlation matrix of $(X_1, \dots, X_q)$ is a block diagonal matrix with blocks of size $s_1 = s_0 = \lceil q/20 \rceil$, where the off-diagonals in each block are equal to 0.9. The high correlation among the covariates will make the functions harder to estimate, as the effects of the different variables will be harder to distinguish. This is an important setting to explore, as in some areas of application it is common that the active covariates are highly intercorrelated.

We construct 95% confidence intervals for $f_1(x)$ over a range of x values based on the oracle estimator $\tilde{f}_1^{\text{oracle}}(x)$, the presmoothing estimator $\hat{f}_1(x)$, and the resmoothed final estimator $\tilde{f}_1(x)$. To make fair comparisons with the oracle, the true variance of the error term is used in constructing both the oracle and the pre- and resmoothed confidence intervals.

The Lasso tuning parameters λ and η are each chosen via 10-fold cross-validation—however, not in every simulation run, as the computation time is quite high; instead, at each n, q combination, a small simulation is run in which the λ and η values are chosen via cross-validation, and then the averages of the λ and η choices over the small simulation are used in the full-size simulation.

Instead of piecewise polynomials, we chose to use cubic B-splines in both the presmoothing and resmoothing steps, as this is a common choice in practice.

The simulations were carried out at different smoothnesses of the presmoothing estimator as well as of the oracle and resmoothing estimators. We observed that more extreme undersmoothing in the presmoothing step lead to closer-to-oracle coverage of the resmoothed final estimator.

Figure 1 displays the estimation and coverage results for the $n = 100$, $q = 50$, $f = \text{sine}$ simulation when the dimension of the presmoother was $d_{\text{pre}} = 75$ and that of the oracle and resmoothed final estimator was $d_{\text{re/orcl}} = 40$. The top panel displays, for a single simulated data set, the pointwise confidence intervals for $f_1(x)$ across a range of x values based on the presmoothing estimator, the oracle estimator, and the resmoothed estimator. The middle panel displays the averages of the upper and lower bounds of each of these three intervals over 500 simulated data sets. We see that the presmoothing intervals are much wider than the oracle and resmoothed intervals, and that the oracle and resmoothed intervals are very similar to each other in width and behavior. The bottom panel plots the coverage of the pointwise confidence intervals across the range of x values. The oracle confidence interval has coverage close to the nominal coverage of 0.95 across the range of x values, and this coverage is nearly matched by the presmoothing confidence interval and the resmoothed confidence interval. Thus, the confidence interval based on our estimator has width and coverage very close to that based on the oracle estimator, for which all the other components are known.

Tables 1–3 give the coverage results over all the $n = 100, 500, 1000$, $q = 50, 200$ and $f = \text{sine}, \text{line}, \text{expo}, \text{quad}$ simulations at the values $x = -1.5, 0, 1$ for different degrees of undersmoothing in the presmoothing step. For the larger sample sizes $n = 500, 1000$, the coverages of the confidence intervals based on the resmoothed estimator are very close to those of the oracle confidence intervals. For $n = 100$, the coverage of the resmoothing confidence interval is in some cases somewhat less than oracle coverage. This is not surprising, as in the $n = 100$, $q = 200$ case, there are twice as many unknown functions as there are observations.

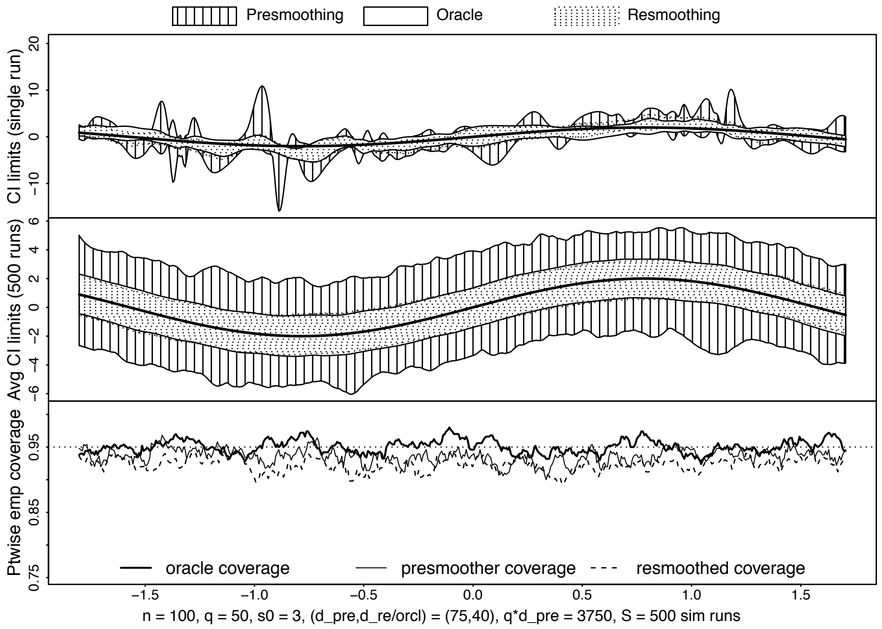

FIGURE 1. Results from the $n = 100$, $q = 50$, $f =$ sine simulation with extreme undersmoothing in the presmoothing estimator. (Upper) Pointwise confidence intervals based on the presmoothing, resmoothing, and oracle estimator for $f_1(x)$ over a range of x values for a single simulated data set. (Middle) Average over 500 simulation runs of upper and lower bounds of pointwise confidence intervals based on the three estimators with true function overlaid. (Lower) Empirical coverage over the 500 simulation runs of the pointwise confidence intervals based on the three estimators.

Moreover, the correlations between the covariates are very high, so that the influences of the different covariates are difficult to distinguish. Even in this setting, the proposed estimator performs quite reliably.

$\hat{f}_1$ *moderately undersmoothed*				sine			line			expo			quad	
		$x =$	-1.5	0	1	-1.5	0	1	-1.5	0	1	-1.5	0	1
$n = 100$	$q = 50$, $s_0 = 3$	orcl	0.95	0.94	0.94	0.94	0.96	0.93	0.91	0.91	0.92	0.93	0.93	0.94
			1.86	*1.86*	*1.87*	*1.88*	*1.88*	*1.87*	*1.88*	*1.87*	*1.86*	*1.88*	*1.88*	*1.87*
	$(d_{\mathrm{pre}} = 35)$	pre-s	0.91	0.90	0.89	0.93	0.96	0.93	0.84	0.94	0.89	0.91	0.93	0.94
			2.43	*2.45*	*2.49*	*2.44*	*2.48*	*2.45*	*2.49*	*2.49*	*2.46*	*2.45*	*2.47*	*2.46*
	$(d_{\mathrm{re/orcl}} = 23)$	re-s	0.92	0.89	0.90	0.90	0.95	0.91	0.83	0.89	0.86	0.90	0.90	0.93
			1.81	*1.82*	*1.82*	*1.84*	*1.84*	*1.83*	*1.84*	*1.82*	*1.82*	*1.84*	*1.84*	*1.83*
	$q = 200$, $s_0 = 10$	orcl	0.94	0.96	0.95	0.94	0.93	0.94	0.92	0.92	0.92	0.95	0.93	0.95
			2.12	*2.14*	*2.18*	*2.11*	*2.18*	*2.12*	*2.11*	*2.16*	*2.11*	*2.11*	*2.14*	*2.12*
	$(d_{\mathrm{pre}} = 35)$	pre-s	0.88	0.84	0.89	0.94	0.98	0.96	0.73	0.92	0.88	0.92	0.93	0.91
			2.49	*2.48*	*2.50*	*2.46*	*2.49*	*2.45*	*2.45*	*2.51*	*2.43*	*2.47*	*2.47*	*2.48*
	$(d_{\mathrm{re/orcl}} = 28)$	re-s	0.86	0.86	0.89	0.91	0.98	0.95	0.71	0.90	0.83	0.91	0.90	0.88
			2.07	*2.09*	*2.13*	*2.07*	*2.13*	*2.08*	*2.06*	*2.12*	*2.06*	*2.07*	*2.10*	*2.07*
$n = 500$	$q = 50$, $s_0 = 3$	orcl	0.95	0.95	0.93	0.93	0.95	0.93	0.94	0.93	0.96	0.92	0.95	0.95
			1.14	*1.14*	*1.14*	*1.15*	*1.14*	*1.14*	*1.14*	*1.15*	*1.14*	*1.14*	*1.14*	*1.14*
	$(d_{\mathrm{pre}} = 75)$	pre-s	0.95	0.94	0.93	0.92	0.94	0.93	0.89	0.92	0.91	0.91	0.92	0.95
			1.56	*1.55*	*1.56*	*1.55*	*1.55*	*1.56*	*1.56*	*1.55*	*1.57*	*1.55*	*1.55*	*1.55*
	$(d_{\mathrm{re/orcl}} = 44)$	re-s	0.91	0.93	0.90	0.91	0.94	0.90	0.88	0.90	0.88	0.90	0.86	0.92
			1.13	*1.13*	*1.13*	*1.14*	*1.13*	*1.13*	*1.13*	*1.13*	*1.13*	*1.13*	*1.13*	*1.13*
	$q = 200$, $s_0 = 10$	orcl	0.95	0.95	0.95	0.95	0.96	0.93	0.94	0.91	0.92	0.93	0.94	0.94
			1.21	*1.22*	*1.22*	*1.22*	*1.23*	*1.23*	*1.22*	*1.23*	*1.22*	*1.23*	*1.23*	*1.23*
	$(d_{\mathrm{pre}} = 75)$	pre-s	0.88	0.86	0.89	0.94	0.97	0.95	0.86	0.91	0.90	0.93	0.94	0.92
			1.54	*1.55*	*1.55*	*1.56*	*1.57*	*1.55*	*1.55*	*1.54*	*1.57*	*1.56*	*1.56*	*1.57*
	$(d_{\mathrm{re/orcl}} = 50)$	re-s	0.85	0.85	0.88	0.94	0.97	0.94	0.86	0.91	0.88	0.90	0.91	0.92
			1.20	*1.21*	*1.21*	*1.21*	*1.22*	*1.22*	*1.21*	*1.22*	*1.21*	*1.22*	*1.23*	*1.22*
$n = 1000$	$q = 50$, $s_0 = 3$	orcl	0.94	0.95	0.96	0.95	0.95	0.97	0.94	0.93	0.92	0.93	0.95	0.94
			1.03	*1.03*	*1.03*	*1.03*	*1.04*	*1.02*	*1.03*	*1.04*	*1.03*	*1.03*	*1.03*	*1.03*
	$(d_{\mathrm{pre}} = 125)$	pre-s	0.93	0.94	0.94	0.91	0.93	0.94	0.89	0.92	0.91	0.93	0.93	0.93
			1.41	*1.42*	*1.41*	*1.41*	*1.41*	*1.40*	*1.41*	*1.42*	*1.42*	*1.41*	*1.42*	*1.42*
	$(d_{\mathrm{re/orcl}} = 70)$	re-s	0.92	0.92	0.92	0.90	0.96	0.94	0.88	0.89	0.87	0.91	0.89	0.90
			1.03	*1.02*	*1.02*	*1.02*	*1.03*	*1.01*	*1.02*	*1.03*	*1.03*	*1.03*	*1.02*	*1.03*
	$q = 200$, $s_0 = 10$	orcl	0.93	0.95	0.95	0.96	0.94	0.95	0.92	0.93	0.92	0.95	0.94	0.93
			1.11	*1.12*	*1.13*	*1.12*	*1.13*	*1.12*	*1.12*	*1.11*	*1.12*	*1.13*	*1.11*	*1.12*
	$(d_{\mathrm{pre}} = 125)$	pre-s	0.91	0.88	0.90	0.94	0.95	0.96	0.86	0.90	0.89	0.93	0.93	0.91
			1.42	*1.42*	*1.42*	*1.42*	*1.42*	*1.43*	*1.41*	*1.40*	*1.43*	*1.43*	*1.41*	*1.42*
	$(d_{\mathrm{re/orcl}} = 82)$	re-s	0.90	0.89	0.88	0.93	0.95	0.95	0.85	0.88	0.87	0.92	0.90	0.90
			1.11	*1.12*	*1.12*	*1.11*	*1.12*	*1.11*	*1.11*	*1.10*	*1.11*	*1.13*	*1.11*	*1.11*

TABLE 1. Coverage and average width in italics of confidence intervals based on oracle, presmoothing, and resmoothed estimators at points $x = -1.5, 0, 1$ for the sine, line, expo and quad functions for $n = 100, 500, 1000$ and $q = 50, 200$ over 500 simulation runs. Dimension d_{pre} used in presmoothing and $d_{\mathrm{re/orcl}}$ for the oracle and the resmoothed estimator shown.

$\hat{f}_1$ *quite undersmoothed*			sine			line			expo			quad		
		$x =$	−1.5	0	1	−1.5	0	1	−1.5	0	1	−1.5	0	1
$n = 100$	$q = 50$, $s_0 = 3$	orcl	0.94	0.95	0.95	0.93	0.95	0.95	0.92	0.93	0.90	0.93	0.94	0.95
			2.34	*2.29*	*2.26*	*2.27*	*2.29*	*2.27*	*2.30*	*2.26*	*2.27*	*2.28*	*2.28*	*2.29*
	$(d_{\text{pre}} = 50)$	pre-s	0.92	0.89	0.92	0.96	0.99	0.98	0.82	0.92	0.93	0.95	0.92	0.94
			3.83	*3.46*	*3.36*	*3.30*	*3.61*	*3.32*	*3.49*	*3.39*	*3.33*	*3.69*	*3.46*	*3.45*
	$(d_{\text{re/orcl}} = 31)$	re-s	0.90	0.89	0.91	0.93	0.97	0.95	0.79	0.91	0.85	0.93	0.90	0.94
			2.31	*2.26*	*2.23*	*2.24*	*2.26*	*2.23*	*2.26*	*2.22*	*2.23*	*2.24*	*2.24*	*2.25*
	$q = 200$, $s_0 = 10$	orcl	0.94	0.96	0.94	0.95	0.92	0.95	0.94	0.93	0.94	0.95	0.94	0.94
			2.70	*2.76*	*2.69*	*2.70*	*2.70*	*2.70*	*2.72*	*2.75*	*2.72*	*2.68*	*2.67*	*2.71*
	$(d_{\text{pre}} = 50)$	pre-s	0.91	0.88	0.90	0.95	0.99	0.98	0.58	0.86	0.84	0.87	0.93	0.90
			3.43	*3.47*	*3.33*	*3.45*	*3.54*	*3.37*	*3.44*	*3.48*	*3.47*	*3.48*	*3.39*	*3.31*
	$(d_{\text{re/orcl}} = 39)$	re-s	0.90	0.88	0.86	0.93	0.98	0.98	0.60	0.84	0.78	0.88	0.87	0.88
			2.67	*2.72*	*2.65*	*2.67*	*2.66*	*2.66*	*2.68*	*2.72*	*2.68*	*2.65*	*2.63*	*2.67*
$n = 500$	$q = 50$, $s_0 = 3$	orcl	0.94	0.94	0.94	0.94	0.95	0.95	0.92	0.94	0.94	0.95	0.94	0.95
			1.31	*1.33*	*1.32*	*1.32*	*1.33*	*1.32*	*1.33*	*1.33*	*1.33*	*1.31*	*1.34*	*1.33*
	$(d_{\text{pre}} = 100)$	pre-s	0.94	0.90	0.91	0.93	0.96	0.94	0.87	0.91	0.89	0.92	0.92	0.91
			1.82	*1.82*	*1.83*	*1.82*	*1.80*	*1.83*	*1.81*	*1.82*	*1.82*	*1.81*	*1.85*	*1.84*
	$(d_{\text{re/orcl}} = 57)$	re-s	0.92	0.91	0.92	0.93	0.95	0.93	0.86	0.90	0.87	0.92	0.89	0.92
			1.30	*1.31*	*1.31*	*1.31*	*1.32*	*1.31*	*1.31*	*1.31*	*1.32*	*1.30*	*1.33*	*1.31*
	$q = 200$, $s_0 = 10$	orcl	0.96	0.96	0.94	0.94	0.94	0.93	0.94	0.92	0.94	0.94	0.92	0.94
			1.43	*1.43*	*1.43*	*1.42*	*1.44*	*1.44*	*1.44*	*1.45*	*1.44*	*1.42*	*1.43*	*1.44*
	$(d_{\text{pre}} = 100)$	pre-s	0.91	0.91	0.91	0.96	0.98	0.97	0.82	0.94	0.95	0.94	0.96	0.95
			1.81	*1.82*	*1.81*	*1.82*	*1.84*	*1.82*	*1.83*	*1.83*	*1.82*	*1.81*	*1.82*	*1.84*
	$(d_{\text{re/orcl}} = 66)$	re-s	0.87	0.89	0.92	0.95	0.98	0.97	0.77	0.93	0.92	0.93	0.96	0.94
			1.42	*1.42*	*1.42*	*1.41*	*1.43*	*1.43*	*1.43*	*1.44*	*1.43*	*1.41*	*1.42*	*1.43*
$n = 1000$	$q = 50$, $s_0 = 3$	orcl	0.94	0.94	0.94	0.95	0.93	0.93	0.94	0.94	0.96	0.94	0.95	0.94
			1.30	*1.32*	*1.32*	*1.31*	*1.31*	*1.31*	*1.31*	*1.30*	*1.31*	*1.31*	*1.30*	*1.32*
	$(d_{\text{pre}} = 200)$	pre-s	0.94	0.92	0.92	0.94	0.95	0.92	0.92	0.93	0.94	0.92	0.92	0.95
			1.84	*1.84*	*1.83*	*1.85*	*1.82*	*1.83*	*1.85*	*1.83*	*1.84*	*1.86*	*1.85*	*1.85*
	$(d_{\text{re/orcl}} = 109)$	re-s	0.93	0.91	0.90	0.92	0.94	0.92	0.89	0.89	0.92	0.93	0.91	0.93
			1.30	*1.31*	*1.31*	*1.31*	*1.30*	*1.30*	*1.30*	*1.30*	*1.31*	*1.30*	*1.30*	*1.31*
	$q = 200$, $s_0 = 10$	orcl	0.95	0.94	0.94	0.93	0.93	0.96	0.94	0.94	0.94	0.95	0.95	0.93
			1.43	*1.43*	*1.43*	*1.44*	*1.43*	*1.42*	*1.43*	*1.43*	*1.44*	*1.44*	*1.44*	*1.45*
	$(d_{\text{pre}} = 200)$	pre-s	0.88	0.88	0.90	0.98	0.98	0.99	0.84	0.95	0.98	0.93	0.92	0.92
			1.84	*1.84*	*1.84*	*1.86*	*1.84*	*1.83*	*1.84*	*1.83*	*1.85*	*1.84*	*1.85*	*1.86*
	$(d_{\text{re/orcl}} = 129)$	re-s	0.87	0.88	0.90	0.98	0.98	0.98	0.84	0.94	0.96	0.93	0.90	0.90
			1.42	*1.42*	*1.43*	*1.43*	*1.43*	*1.42*	*1.42*	*1.43*	*1.43*	*1.44*	*1.43*	*1.45*

TABLE 2. Coverage and average width in italics of confidence intervals based on oracle, presmoothing, and resmoothed estimators at points $x = -1.5, 0, 1$ for the sine, line, expo and quad functions for $n = 100, 500, 1000$ and $q = 50, 200$ over 500 simulation runs. Dimension d_{pre} used in presmoothing and $d_{\text{re/orcl}}$ for the oracle and the resmoothed estimator shown.

$\hat{f}_1$ *extremely undersmoothed*			sine			line			expo			quad		
		$x =$	−1.5	0	1	−1.5	0	1	−1.5	0	1	−1.5	0	1
$n = 100$	$q = 50,\ s_0 = 3$	orcl	0.95	0.95	0.94	0.94	0.95	0.95	0.93	0.91	0.93	0.94	0.94	0.94
			2.79	*2.74*	*2.72*	*2.76*	*2.78*	*2.75*	*2.78*	*2.76*	*2.71*	*2.81*	*2.80*	*2.73*
	$(d_{\text{pre}} = 75)$	pre-s	0.92	0.94	0.95	0.97	0.97	0.98	0.83	0.94	0.95	0.94	0.97	0.96
			7.44	*7.27*	*7.65*	*7.22*	*6.42*	*7.27*	*10.74*	*6.62*	*6.45*	*8.88*	*6.70*	*7.15*
	$(d_{\text{re/orcl}} = 40)$	re-s	0.92	0.92	0.92	0.93	0.98	0.97	0.78	0.93	0.88	0.93	0.93	0.92
			2.76	*2.71*	*2.69*	*2.73*	*2.75*	*2.72*	*2.75*	*2.73*	*2.68*	*2.79*	*2.77*	*2.70*
	$q = 200,\ s_0 = 10$	orcl	0.94	0.95	0.95	0.95	0.96	0.96	0.93	0.93	0.93	0.94	0.94	0.93
			3.90	*3.81*	*3.88*	*3.92*	*4.05*	*3.92*	*3.83*	*3.97*	*3.86*	*3.83*	*3.91*	*3.83*
	$(d_{\text{pre}} = 75)$	pre-s	0.92	0.88	0.92	0.97	0.98	0.98	0.70	0.88	0.85	0.88	0.92	0.90
			6.63	*6.41*	*6.73*	*9.89*	*7.09*	*6.84*	*8.01*	*7.37*	*7.04*	*6.43*	*6.80*	*6.72*
	$(d_{\text{re/orcl}} = 57)$	re-s	0.90	0.90	0.89	0.93	0.98	0.97	0.67	0.88	0.79	0.87	0.87	0.87
			3.87	*3.78*	*3.84*	*3.90*	*4.02*	*3.89*	*3.80*	*3.94*	*3.83*	*3.80*	*3.88*	*3.80*
$n = 500$	$q = 50,\ s_0 = 3$	orcl	0.97	0.94	0.95	0.93	0.94	0.93	0.94	0.94	0.95	0.93	0.95	0.95
			1.84	*1.81*	*1.81*	*1.84*	*1.82*	*1.84*	*1.83*	*1.86*	*1.80*	*1.83*	*1.82*	*1.82*
	$(d_{\text{pre}} = 200)$	pre-s	0.94	0.94	0.91	0.97	0.97	0.97	0.92	0.96	0.97	0.94	0.95	0.93
			3.00	*2.88*	*2.85*	*2.89*	*2.88*	*2.93*	*2.88*	*2.89*	*2.83*	*2.92*	*2.91*	*2.94*
	$(d_{\text{re/orcl}} = 100)$	re-s	0.94	0.92	0.92	0.96	0.97	0.96	0.90	0.94	0.94	0.92	0.93	0.94
			1.83	*1.80*	*1.80*	*1.83*	*1.81*	*1.83*	*1.82*	*1.85*	*1.79*	*1.82*	*1.81*	*1.81*
	$q = 200,\ s_0 = 10$	orcl	0.95	0.96	0.94	0.95	0.95	0.93	0.94	0.94	0.95	0.94	0.94	0.93
			2.12	*2.12*	*2.15*	*2.13*	*2.12*	*2.15*	*2.13*	*2.14*	*2.11*	*2.12*	*2.15*	*2.16*
	$(d_{\text{pre}} = 200)$	pre-s	0.90	0.89	0.94	0.98	1.00	0.99	0.80	0.97	0.99	0.93	0.99	0.96
			2.91	*2.86*	*2.86*	*2.85*	*2.86*	*2.93*	*2.86*	*2.89*	*2.85*	*2.83*	*2.93*	*2.96*
	$(d_{\text{re/orcl}} = 129)$	re-s	0.88	0.90	0.94	0.98	0.99	0.99	0.75	0.97	0.94	0.93	0.99	0.97
			2.11	*2.11*	*2.14*	*2.12*	*2.11*	*2.14*	*2.12*	*2.13*	*2.10*	*2.12*	*2.14*	*2.15*
$n = 1000$	$q = 50,\ s_0 = 3$	orcl	0.96	0.94	0.96	0.96	0.94	0.95	0.94	0.94	0.94	0.95	0.95	0.94
			1.55	*1.54*	*1.53*	*1.55*	*1.54*	*1.56*	*1.53*	*1.54*	*1.55*	*1.53*	*1.54*	*1.57*
	$(d_{\text{pre}} = 300)$	pre-s	0.93	0.92	0.93	0.97	0.96	0.96	0.90	0.94	0.93	0.91	0.94	0.94
			2.37	*2.35*	*2.36*	*2.34*	*2.33*	*2.36*	*2.33*	*2.37*	*2.38*	*2.37*	*2.34*	*2.38*
	$(d_{\text{re/orcl}} = 147)$	re-s	0.94	0.90	0.93	0.95	0.94	0.96	0.90	0.93	0.93	0.94	0.92	0.94
			1.55	*1.54*	*1.52*	*1.54*	*1.54*	*1.56*	*1.53*	*1.54*	*1.55*	*1.52*	*1.53*	*1.56*
	$q = 200,\ s_0 = 10$	orcl	0.95	0.95	0.93	0.93	0.96	0.93	0.94	0.95	0.93	0.94	0.94	0.93
			1.62	*1.62*	*1.62*	*1.48*	*1.47*	*1.48*	*1.65*	*1.64*	*1.63*	*1.47*	*1.47*	*1.47*
	$(d_{\text{pre}} = 300)$	pre-s	0.88	0.89	0.93	0.99	0.98	0.99	0.78	0.97	0.99	0.92	0.97	0.94
			2.40	*2.37*	*2.36*	*2.11*	*2.11*	*2.09*	*2.41*	*2.35*	*2.34*	*2.10*	*2.10*	*2.09*
	$(d_{\text{re/orcl}} = 162)$	re-s	0.88	0.90	0.93	0.97	0.99	0.98	0.76	0.98	0.96	0.92	0.93	0.92
			1.61	*1.62*	*1.62*	*1.48*	*1.46*	*1.47*	*1.64*	*1.63*	*1.62*	*1.46*	*1.47*	*1.46*

TABLE 3. Coverage and average width in italics of confidence intervals based on oracle, presmoothing, and resmoothed estimators at points $x = -1.5, 0, 1$ for the sine, line, expo and quad functions for $n = 100, 500, 1000$ and $q = 50, 200$ over 500 simulation runs. Dimension d_{pre} used in presmoothing and $d_{\text{re/orcl}}$ for the oracle and the resmoothed estimator shown.

4. The mathematical approach

4.1. **The geometric representation.** In this section, we suppose that the event holds on which V_1 and $\{\mathbf{g}_j : g_j \in V_1\} \subseteq \mathbb{R}^n$, where $\mathbf{g}_j = (g_j(X_j^1), \dots, g_j(X_j^n))^T$, have the same dimension. Then, we choose $\phi_{11}, \dots, \phi_{1d_1}$ to be the orthonormal basis of V_1 with respect to the empirical inner product obtained by applying the Gram-Schmidt orthogonalization (with respect to the empirical inner product) to the basis $b_{11}, \dots, b_{1d_1}$. Clearly, this basis is still local in the sense that

$$\phi_{1,k(t_1+1)+1}, \dots, \phi_{1,k(t_1+1)+t_1+1}$$

is a basis of the functions in V_1 which are zero outside the interval I_{jk}. Note that in the case of local constant functions, i.e. $t_1 = 0$, the above procedure simply normalizes the basis functions according to the empirical norm $\phi_{1k} = b_{1k}/\|b_{1k}\|_n$. We again restrict our analysis to the event that $(1/n)\mathbf{Z}_1^T\mathbf{X}_1$ is invertible. We have

$$\frac{1}{n}\mathbf{Z}_1^T\mathbf{X}_1 = \left(\langle \phi_{1k} - \hat{\Pi}_{-1}^L\phi_{1k}, \phi_{1l}\rangle_n\right)_{k,l=1}^{d_1}.$$

The matrix $\frac{1}{n}\mathbf{Z}_1^T\mathbf{X}_1$ can be considered as a linear map on the coefficients in $\mathbb{R}^{d_1}$. Equivalently, it can be considered as a linear map from V_1 into itself. Therefore, let $\hat{\Pi}_1$ be the linear map defined by

$$\hat{\Pi}_1 f = \sum_{k=1}^{d_1} \langle \phi_{1k}, f\rangle_n \phi_{1k},$$

where $f \in L^2(\mathbb{P}^X)$ (resp. $\in \mathbb{R}^n$). Since $\phi_{11}, \dots, \phi_{1d_1}$ is an orthonormal basis of V_1 with respect to the empirical inner product, $\hat{\Pi}_1$ is the orthogonal projection from $L^2(\mathbb{P}^X)$ (resp. $\mathbb{R}^n$) to V_1 (resp. $\{\mathbf{g}_1 : g_1 \in V_1\}$) with respect to the empirical inner product (resp. Euclidean inner product). Now, let

$$g_\alpha = \sum_{k=1}^{d_1} \alpha_k \phi_{1k} \in V_1.$$

Then $\frac{1}{n}\mathbf{Z}_1^T\mathbf{X}_1$ sends $\alpha = (\alpha_1, \dots, \alpha_{d_1})^T$ to the coefficient vector

$$\begin{aligned}\left(\sum_{l=1}^{d_1} \langle \phi_{1k} - \hat{\Pi}_{-1}^L\phi_{1k}, \phi_{1l}\rangle_n \alpha_l\right)_{k=1}^{d_1} &= \left(\langle (I - \hat{\Pi}_{-1}^L)\phi_{1k}, g_\alpha\rangle_n\right)_{k=1}^{d_1} \\ &= \left(\langle (I - \hat{\Pi}_1\hat{\Pi}_{-1}^L)\phi_{1k}, g_\alpha\rangle_n\right)_{k=1}^{d_1}.\end{aligned}$$

Now, the linear operator $\hat{\Pi}_1\hat{\Pi}_{-1}^L : V_1 \to V_1$ has an adjoint operator (its transpose) $(\hat{\Pi}_1\hat{\Pi}_{-1}^L)^* : V_1 \to V_1$, and thus this coefficient vector can be written as

$$\Big(\langle \phi_{1k}, (I - (\hat{\Pi}_1 \hat{\Pi}_{-1}^L)^*) g_\alpha \rangle_n \Big)_{k=1}^{d_1}$$

which are the coefficients of

$$(I - (\hat{\Pi}_1 \hat{\Pi}_{-1}^L)^*) g_\alpha.$$

Thus, considered as a map from V_1 into itself, we have that $(\frac{1}{n}\mathbf{Z}_1^T \mathbf{X}_1)^{-1}$ is equal to

$$(I - (\hat{\Pi}_1 \hat{\Pi}_{-1}^L)^*)^{-1}.$$

In particular,

$$\hat{f}_1 = (I - (\hat{\Pi}_1 \hat{\Pi}_{-1}^L)^*)^{-1} g_\alpha,$$

where

$$\alpha = \Big(\langle \phi_{1k} - \hat{\Pi}_{-1}^L \phi_{1k}, \mathbf{Y} - \hat{f}_{-1}^L \rangle_n \Big)_{k=1}^{d_1}.$$

We can go slightly further. $\hat{\Pi}_{-1}^L$ is a map from V_1 to V_{-1}, but it can also be considered as a map from V_1 to $\mathbb{R}^n$. In both cases, it has an adjoint operator $(\hat{\Pi}_{-1}^L)^*$ such that

$$\Big(\langle \phi_{1k} - \hat{\Pi}_{-1}^L \phi_{1k}, \mathbf{Y} - \hat{f}^L \rangle_n \Big)_{k=1}^{d_1} = \Big(\langle \phi_{1k}, (I - (\hat{\Pi}_{-1}^L)^*)(\mathbf{Y} - \hat{f}^L) \rangle_n \Big)_{k=1}^{d_1}.$$

This is the coefficient vector of the function

$$\hat{\Pi}_1 (I - (\hat{\Pi}_{-1}^L)^*)(\mathbf{Y} - \hat{f}^L).$$

We conclude:

Proposition 1. *If V_1 and $\{\mathbf{g}_j : g_j \in V_1\} \subseteq \mathbb{R}^n$ have the same dimension and if $I - (\hat{\Pi}_1 \hat{\Pi}_{-1}^L)^* : V_1 \to V_1$ is invertible, then we have*

$$\hat{f}_1 = (I - (\hat{\Pi}_1 \hat{\Pi}_{-1}^L)^*)^{-1} \hat{\Pi}_1 (I - (\hat{\Pi}_{-1}^L)^*)(\mathbf{Y} - \hat{f}_{-1}^L).$$

4.2. **The main decomposition.** We continue the discussion of the previous section and present a decomposition of $\hat{f}_1 - \hat{f}_1^{\text{oracle}}$ which gives rise the terms Δ_1, Δ_2, Δ_3 appearing in the main result in Theorem 1. Let

$$\hat{f}_1^{\text{oracle}} = \hat{\Pi}_1 (\mathbf{f}_1 + \boldsymbol{\epsilon}),$$

which has coefficient vector

$$\hat{\boldsymbol{\beta}}_1^{\text{oracle}} = \left(\mathbf{X}_1^T \mathbf{X}_1\right)^{-1} \mathbf{X}_1^T (\mathbf{f}_1 + \boldsymbol{\epsilon}).$$

For any $g \in V$, we may write the difference between the coefficient vectors of the presmoothing and undersmoothed oracle estimators as

$$\begin{aligned}\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_1^{\text{oracle}} &= \left(\left(\frac{1}{n}\mathbf{Z}_1^T\mathbf{X}_1\right)^{-1}\frac{1}{n}\mathbf{Z}_1^T - \left(\frac{1}{n}\mathbf{X}_1^T\mathbf{X}_1\right)^{-1}\frac{1}{n}\mathbf{X}_1^T\right)\boldsymbol{\epsilon} \\ &+ \left(\frac{1}{n}\mathbf{Z}_1^T\mathbf{X}_1\right)^{-1}\frac{1}{n}\mathbf{Z}_1^T(\mathbf{g}_{-1} - \hat{\mathbf{f}}_{-1}) \\ &+ \left(\frac{1}{n}\mathbf{Z}_1^T\mathbf{X}_1\right)^{-1}\frac{1}{n}\mathbf{Z}_1^T(\mathbf{f} - \mathbf{g}) \\ &+ \boldsymbol{\beta}_1 - \left(\frac{1}{n}\mathbf{X}_1^T\mathbf{X}_1\right)^{-1}\frac{1}{n}\mathbf{X}_1^T\mathbf{f}_1,\end{aligned}$$

where $g = g_1 + g_{-1}$, $g_1 \in V_1$, $g_{-1} \in V_{-1}$ and $\boldsymbol{\beta}_1$ is the coefficient vector of g_1. Using Proposition 1, we can also formulate this decomposition in terms of functions

$$\begin{aligned}\hat{f}_1 - \hat{f}_1^{\text{oracle}} &= (I - (\hat{\Pi}_1\hat{\Pi}_{-1}^L)^*)^{-1}\hat{\Pi}_1(I - (\hat{\Pi}_{-1}^L)^*)\boldsymbol{\epsilon} - \hat{\Pi}_1\boldsymbol{\epsilon} \\ &+ (I - (\hat{\Pi}_1\hat{\Pi}_{-1}^L)^*)^{-1}\hat{\Pi}_1(I - (\hat{\Pi}_{-1}^L)^*)(g_{-1} - \hat{f}_{-1}) \\ &+ (I - (\hat{\Pi}_1\hat{\Pi}_{-1}^L)^*)^{-1}\hat{\Pi}_1(I - (\hat{\Pi}_{-1}^L)^*)(f - g) \\ &+ g_1 - \hat{\Pi}_1 f_1.\end{aligned}$$

Theorems 5–7 presented in Section 4.4 establish bounds for the terms in this decomposition.

4.3. **Events.** We now define several events upon which the inequalities presented in the following sections will hold. We show in Appendix E in the supplementary material that these events occur with high probability. First, we define

$$\mathcal{E}_0 = \mathcal{A}_0 \cap \mathcal{E}_{\phi,J_0},$$

where

$$\mathcal{A}_0 = \left\{2\max_{j=1,\dots,q}\sup_{0\neq g_j\in V_j}\frac{|\langle\epsilon, g_j\rangle_n|}{\|g_j\|_n} \le \lambda\right\}$$

and $\mathcal{E}_{\phi,J_0}$ is the compatibility condition event defined as the event on which

$$\sum_{j\in J_0}\|g_j\|_n^2 \le 3\Big\|\sum_{j=1}^q g_j\Big\|_n^2/\phi^2$$

for all $(g_1, \ldots, g_q) \in (V_1, \ldots, V_q)$ satisfying

$$\sum_{j=1}^{q} \|g_j\|_n \leq 8 \sum_{j \in J_0} \|g_j\|_n. \tag{4.1}$$

Note that $\mathcal{E}_0$ is needed in the analysis of the Lasso estimator of f. We also define

$$\mathcal{E}_1 = \bigcap_{k=1}^{d_1} \left(\mathcal{A}_k \cap \mathcal{E}_{\phi, J_k}\right),$$

where for $k = 1, \ldots, d_1$,

$$\mathcal{A}_k = \left\{ 2 \max_{j=2,\ldots,q} \sup_{0 \neq g_j \in V_j} \frac{|\langle b_{1k} - \Pi_{-1} b_{1k}, g_j \rangle_n|}{\|g_j\|_n} \leq \eta \right\}$$

and $\mathcal{E}_{\phi, J_k}$ is the compatibility condition event defined as the event on which

$$\sum_{j \in J_k} \|g_j\|_n^2 \leq 3 \Big\| \sum_{j=2}^{q} g_j \Big\|_n^2 / \phi^2$$

for all $(g_2, \ldots, g_q) \in (V_2, \ldots, V_q)$ satisfying

$$\sum_{j=2}^{q} \|g_j\|_n \leq 8 \sum_{j \in J_k} \|g_j\|_n.$$

Note that $\mathcal{E}_1$ is needed in the analysis of the Lasso estimators of the $\Pi_{-1} b_{1k}$. Finally, for $0 < \delta \leq 1/2$, we define the empirical norm approximation event

$$\mathcal{E}_2 = \mathcal{E}_{\delta,1} \cap \bigcap_{k,l=1}^{d_1} \left\{ |\langle \Pi_{J_k} b_{1k}, \Pi_{J_l} b_{1l} \rangle - \langle \Pi_{J_k} b_{1k}, \Pi_{J_l} b_{1l} \rangle_n| \leq \delta \|\Pi_{J_k} b_{1k}\| \|\Pi_{J_l} b_{1l}\| \right\},$$

where

$$\mathcal{E}_{\delta,1} = \bigcap_{j=1}^{q} \left\{ (1-\delta)\|g_j\|^2 \leq \|g_j\|_n^2 \leq (1+\delta)\|g_j\|^2 \text{ for all } g_j \in V_j \right\}.$$

The event $\mathcal{E}_{\delta,1}$ specifies the closeness of the empirical norm $\|g_j\|_n$ to the $L^2(\mathbb{P}^{X_j})$ norm $\|g_j\|$ of $g_j \in V_j$ for $j = 1, \ldots, q$, thus specifying the cost at which we may switch between the empirical and the true norm of a function in our analysis. Moreover, the event $\mathcal{E}_{\delta,1}$ implies an equivalence between the spaces and V_1 and $\{\mathbf{g}_j : g_j \in V_1\} \subseteq \mathbb{R}^n$ to which we alluded in the beginning of Section 4.1.

4.4. **The main result revisited.** In this section, we state upper bounds for different terms appearing in the decomposition presented in Section 4.2. Moreover, we show how these bounds lead to a proof of Theorem 1.

Theorem 5. *Suppose that Assumptions 1-5 hold. Moreover, suppose that* (3.2) *is satisfied. If* $\mathcal{E}_0 \cap \mathcal{E}_1 \cap \mathcal{E}_2$ *holds, then we have*

$$\begin{aligned}&\left\|(I-(\hat{\Pi}_1\hat{\Pi}_{-1}^L)^*)^{-1}\hat{\Pi}_1(I-(\hat{\Pi}_{-1}^L)^*)(g_{-1}^*-\hat{f}_{-1})\right\|_\infty\\&\leq \frac{C}{(1-\rho_0)\psi}\left((\eta/\lambda)\sqrt{s_1 d_1}\left(d_1^{-r_1}+s_0 d_2^{-r_2}\right)^2+s_0\sqrt{s_1}\sqrt{d_1}\lambda\eta/\phi^2\right).\end{aligned}$$

The proof of this theorem invokes a bound on $\|(I-(\hat{\Pi}_1\hat{\Pi}_{-1}^L)^*)^{-1}g_1\|_\infty$ for $g_1 \in V_1$ (Corollary 3) and makes use of a nonparametric version of the KKT equations.

Theorem 6. *Suppose that Assumptions 1-5 hold. Moreover, suppose that* (3.2) *is satisfied. If* $\mathcal{E}_1 \cap \mathcal{E}_2$ *holds, then we have*

$$\begin{aligned}&\left\|(I-(\hat{\Pi}_1\hat{\Pi}_{-1}^L)^*)^{-1}\hat{\Pi}_1(I-(\hat{\Pi}_{-1}^L)^*)(f-g^*)\right\|_\infty\\&\leq \frac{C}{(1-\rho_0)\psi}\left(s_1 d_1^{-r_1}+s_1 s_0 d_2^{-r_2}\right).\end{aligned}$$

Moreover, if $\mathcal{E}_2$ *holds, then*

$$\|g_1^*-\hat{\Pi}_1 f_1\|_\infty \leq C d_1^{-r_1}.$$

The proof of this theorem invokes the same bound on $\|(I-(\hat{\Pi}_1\hat{\Pi}_{-1}^L)^*)^{-1}g_1\|_\infty$ for $g_1 \in V_1$ and uses the approximation properties of g_* formulated in Assumption 3.

Theorem 7. *Suppose that Assumptions 1-5 hold. Moreover, suppose that* (3.2) *is satisfied. If* $\mathcal{E}_1 \cap \mathcal{E}_2$ *holds, then we have for all* $y \geq 0$,

$$\begin{aligned}\mathbf{P}_\epsilon\Bigg(&\|(I-(\hat{\Pi}_1\hat{\Pi}_{-1}^L)^*)^{-1}\hat{\Pi}_1(I-(\hat{\Pi}_{-1}^L)^*)\boldsymbol{\epsilon}-\hat{\Pi}_1\boldsymbol{\epsilon}\|_\infty\\&\geq \frac{C}{(1-\rho_0)\psi}\sqrt{\frac{s_1(\log d_1+y)}{n}}\Bigg)\leq \exp(-y),\end{aligned}$$

where $\mathbf{P}_\epsilon$ *denotes the probability with respect to* $\epsilon^1,\dots,\epsilon^n$ *for given, fixed values of* $X^1,\dots,X^n$.

A proof of Theorems 5-7 is given in the next section. Let us see how Theorem 1 can be deduced from these theorems combined with a lower bound for the probabilities of events. Using the main decomposition of

Section 4.2 and recalling the definitions of the error terms in Section 3.2, we obtain under the assumption made in Theorems 5-7 that

$$\mathbb{P}\left(\|\hat{f}_1 - \hat{f}_1^{\text{oracle}}\|_\infty \geq C\left(\Delta_1 + \Delta_2 + \Delta_3\right)\right) \leq \mathbb{P}\left((\mathcal{E}_0 \cap \mathcal{E}_1 \cap \mathcal{E}_2)^c\right) + \exp(-y).$$

In order to obtain Theorem 1 from this inequality, the last step is the following concentration result proven in Appendix E in the supplementary material.

Theorem 8. *For $x > 1$, let λ, η, and δ be as defined in Section 3.2. Moreover, suppose that* (3.3) *is satisfied. Then we have*

$$\mathbb{P}\left((\mathcal{E}_0 \cap \mathcal{E}_1 \cap \mathcal{E}_2)^c\right) \leq 4\exp(-x).$$

5. Proofs

5.1. **Preliminary results on the Lasso estimators.**

5.1.1. *The nonparametric Lasso estimator.* In this section, we state a risk bound for the Lasso estimator $\hat{f}^L$ which is suitable to our purposes. In order to bound the approximation error terms we need a risk bound for the Lasso estimator in the undersmoothed case. From now on we will denote the nonparametric Lasso penalty by

$$\text{pen}_\lambda(g) = 2\lambda \sum_{j=1}^{q} \|g_j\|_n, \text{ for any } g \in V.$$

Applying the work by Bickel, Ritov, and Tsybakov [5], we obtain:

Proposition 2. *Suppose that Assumption 3 holds. If $\mathcal{E}_0$ holds, then we have for each $g \in V_{J_0}$,*

$$\|\hat{f}^L - f\|_n^2 + \text{pen}_\lambda(\hat{f}^L - g) \leq 4\|f - g\|_n^2 + 240 s_0 \lambda^2/\phi_*^2.$$

In particular, if we choose g^* from Assumption 3, then we have on $\mathcal{E}_0 \cap \mathcal{E}_2$,

$$\|\hat{f}^L - f\|_n^2 + \text{pen}_\lambda(\hat{f}^L - g^*) \leq 4C_0^2(d_1^{-r_1} + s_0 d_2^{-r_2})^2 + 240 s_0 \lambda^2/\phi_*^2.$$

5.1.2. *The Lasso projection of relaxed orthogonality.* In this section, we state risk bounds for the Lasso estimators $\hat{\Pi}_{-1}^L b_{1l}$ of $\Pi_{-1} b_{1l}$. The analysis is analogous to that of the Lasso estimator $\hat{f}^L$ of f. We only have to replace Y by b_{1l}, f by $\Pi_{-1} b_{1l}$, and ϵ by $b_{1l} - \Pi_{-1} b_{1l}$. Note that for all $g \in V_{-1}$, we have

$$\langle b_{1l} - \Pi_{-1} b_{1l}, g\rangle = 0.$$

Let

$$\operatorname{pen}_\eta(g) = 2\eta \sum_{j=2}^{q} \|g_j\|_n, \text{ for any } g \in V_{-1}.$$

The following result is similar to the result above. Note that a proof of Propositions 2 and 3 is given in Appendix B in the supplementary material.

Proposition 3. *Suppose that Assumption 4 holds. Let $l \in \{1, \dots, d_1\}$. If $\mathcal{A}_l \cap \mathcal{E}_{\phi, J_l}$ holds, then for each $g \in V_{J_l}$,*

$$\|\hat{\Pi}^L_{-1} b_{1l} - \Pi_{-1} b_{1l}\|_n^2 + \operatorname{pen}_\eta\left(\hat{\Pi}^L_{-1} b_{1l} - g\right) \le 4\|\Pi_{-1} b_{1l} - g\|_n^2 + 240 s_1 \eta^2/\phi_*^2.$$

In particular, choosing $g = \Pi_{J_l} b_{1l}$ gives on $\mathcal{A}_l \cap \mathcal{E}_{\phi, J_l} \cap \mathcal{E}_{\delta,1}$

$$\|\hat{\Pi}^L_{-1} b_{1l} - \Pi_{-1} b_{1l}\|_n^2 + \operatorname{pen}_\eta\left(\hat{\Pi}^L_{-1} b_{1l} - \Pi_{J_l} b_{1l}\right) \le \left(4C_1^2 + (240/\phi_*^2)\right) s_1 \eta^2.$$

5.1.3. *Approximate orthogonality.* In the case of the empirical Lasso projection, the KKT conditions have the following form:

Lemma 1 (Nonparametric KKT conditions)**.** *Let $l \in \{1, \dots, d_1\}$. For all $j = 2, \dots, q$ and all $g_j \in V_j$, we have*

$$|\langle g_j, b_{1l} - \hat{\Pi}^L_{-1} b_{1l}\rangle_n| \le \eta \|g_j\|_n.$$

Equivalently, for all $g \in V_{-1}$, we have

$$2|\langle g, b_{1l} - \hat{\Pi}^L_{-1} b_{1l}\rangle_n| \le \operatorname{pen}_\eta(g).$$

A proof of Lemma 1 is given in Appendix C in the supplementary material. Let us derive a first consequence of Lemma 1. Since $\hat{\Pi}^L_{-1} b_{1k} \in V_{-1}$, Lemma 1 implies

$$\begin{aligned} &|\langle \hat{\Pi}^L_{-1} b_{1k}, b_{1l}\rangle_n - \langle \hat{\Pi}^L_{-1} b_{1k}, \hat{\Pi}^L_{-1} b_{1l}\rangle_n| \\ &= |\langle \hat{\Pi}^L_{-1} b_{1k}, b_{1l} - \hat{\Pi}^L_{-1} b_{1l}\rangle_n| \\ &\le (1/2) \operatorname{pen}_\eta(\hat{\Pi}^L_{-1} b_{1k}). \end{aligned} \tag{5.1}$$

5.2. **Evaluating Π_{-1} and $\hat{\Pi}^L_{-1}$ at the basis functions.** Recall the convention that C denotes a constant depending only on the quantities t, t_1, c_1, C_0, and C_1 and that C is not necessarily the same at each occurrence. Explicit constants can be derived from the proofs in Appendix A in the supplementary material.

Proposition 4. *Suppose that Assumptions 1 and 4 hold. Then, for $k = 1, \dots, d_1$, we have*

$$\|\Pi_{J_k} b_{1k}\| \le \frac{C}{\psi}\sqrt{\frac{s_1}{d_1}}$$

and

$$\|\Pi_{-1} b_{1k}\| \le \frac{C}{\psi}\sqrt{\frac{s_1}{d_1}} + C\sqrt{\frac{s_1 d}{n}}.$$

A proof of Proposition 4 is given in Appendix A in the supplementary material. It is based on the definition of the quantity ψ, Assumption 4, and the following lemma:

Lemma 2. *Suppose that Assumption 1 holds. Then, for each $j = 2, \dots, q$ and $k = 1, \dots, d_1$, we have*

$$\|\Pi_{V_j} b_{1k}\| \le C\frac{1}{\sqrt{d_1}}.$$

Note that Proposition 4 deals only with the basis functions b_{1k}. However, it can also be stated for functions in V_1 having their support in one of the intervals $I_{1k'}$: suppose that $g_1 \in V_1$ satisfies $\operatorname{supp}(g_1) \subseteq I_{1k'}$ for some $k' \in \{1, \dots, m_1\}$. Then g_1 is a linear combination of the $t_1 + 1$ basis functions $b_{1,k'(t_1+1)+1}, \dots, b_{1,k'(t_1+1)+t_1+1}$. Applying the triangle inequality, Proposition 4, the Cauchy-Schwarz inequality, and Assumption 1, we obtain

$$\|\Pi_{-1} g_1\| \le \left(\frac{C}{\psi}\sqrt{\frac{s_1}{d_1}} + C\sqrt{\frac{s_1 d}{n}}\right)\sqrt{\frac{t_1+1}{c_1}}\|g_1\|,$$

where C is the constant in Proposition 4. Moreover, we have:

Proposition 5. *Suppose that Assumptions 1 and 4 hold. If $\mathcal{E}_{\delta,1}$ holds, then we have*

$$\operatorname{pen}_\eta(\Pi_{J_k} b_{1k}) \le \frac{C}{\psi^2}\frac{s_1\eta}{\sqrt{d_1}}.$$

A proof of Proposition 5 is given in Appendix A in the supplementary material. Combining Propositions 4 and 5 with Proposition 3, we obtain:

Corollary 1. *Suppose that Assumptions 1, 4, and 5 hold. Suppose that $\mathcal{E}_1 \cap \mathcal{E}_2$ holds. Then we have*

$$\|\hat{\Pi}^L_{-1} b_{1k}\|_n \le \frac{C}{\psi}\sqrt{\frac{s_1}{d_1}} + \frac{C}{\phi}\sqrt{s_1}\eta$$

and

$$\operatorname{pen}_\eta(\hat{\Pi}^L_{-1} b_{1k}) \le \frac{C}{\psi^2}\frac{s_1\eta}{\sqrt{d_1}} + \frac{C}{\phi^2}s_1\eta^2.$$

Moreover, if $g_1 \in V_1$ satisfied $\operatorname{supp}(g_1) \subseteq I_{1k'}$ for some $k' \in \{1, \dots, m_1\}$, then we have

$$\|\hat{\Pi}^L_{-1} g_1\|_n \le \left(\frac{C}{\psi}\sqrt{\frac{s_1}{d_1}} + \frac{C}{\phi}\sqrt{s_1}\eta\right)\|g_1\|_n.$$

5.3. **Geometric properties of Π_{-1} and $\hat{\Pi}^L_{-1}$.** A main quantity in the analysis of the estimator is the following empirical counterpart of ρ_0:

$$\hat{\rho}_0 = \sup_{g_1 \in V_1 : \|g_1\|_n \leq 1} \|\hat{\Pi}^L_{-1} g_1\|_n.$$

We have:

Proposition 6. *Suppose that Assumptions 1, 2, 4, and 5 hold. If $\mathcal{E}_1 \cap \mathcal{E}_2$ holds, then we have*

$$\hat{\rho}_0^2 \leq \rho_0^2 + C\left(\frac{s_1\delta}{\psi^2} + \frac{s_1\sqrt{d_1}\eta}{\psi\phi} + \frac{s_1 d_1 \eta^2}{\phi^2}\right).$$

In order that $\hat{\rho}_0 < 1$, we suppose that the second summand on the right-hand side of Proposition 6 satisfies

$$C\left(\frac{s_1\delta}{\psi^2} + \frac{s_1\sqrt{d_1}\eta}{\psi\phi} + \frac{s_1 d_1 \eta^2}{\phi^2}\right) < (1-\rho_0)^2/4. \tag{5.2}$$

If (5.2) is satisfied, then we have

$$\hat{\rho}_0 \leq \rho_0 + (1-\rho_0)/2 = (1+\rho_0)/2 < 1$$

and thus

$$\frac{1}{1-\hat{\rho}_0} \leq \frac{2}{1-\rho_0}. \tag{5.3}$$

Since the operator norm of an bounded linear operator and its adjoint operator are equal, Proposition 6 implies:

Corollary 2. *Suppose that Assumptions 1, 2, 4, and 5 hold. Moreover, suppose that* (5.2) *is satisfied. If $\mathcal{E}_1 \cap \mathcal{E}_2$ holds, then for each $g_1 \in V_1$ we have*

$$\|(\hat{\Pi}_1 \hat{\Pi}^L_{-1})^* g_1\|_n \leq \hat{\rho}_0 \|g_1\|_n \leq ((1+\rho_0)/2)\|g_1\|_n.$$

Proof of Proposition 6. Let

$$g_1 = \sum_{k=1}^{d_1} \alpha_k b_{1k}.$$

By Assumption 2, we have

$$\|\Pi_{-1} g_1\|^2 \leq \rho_0^2 \|g_1\|^2.$$

Thus on $\mathcal{E}_{\delta,1}$,

$$\|\Pi_{-1} g_1\|^2 \leq \rho_0^2 \|g_1\|_n^2/(1-\delta) \leq (1+2\delta)\rho_0^2 \|g_1\|_n^2,$$

where we used that $\delta \le 1/2$. Hence,

$$\begin{aligned}\|\hat{\Pi}^L_{-1} g_1\|_n^2 &= \|\Pi_{-1} g_1\|^2 + \|\hat{\Pi}^L_{-1} g_1\|_n^2 - \|\Pi_{-1} g_1\|^2 \\ &\le \rho_0^2 \|g_1\|_n^2 + 2\delta \|g_1\|_n^2 + \|\hat{\Pi}^L_{-1} g_1\|_n^2 - \|\Pi_{-1} g_1\|^2, \qquad (5.4)\end{aligned}$$

and it remains to consider the last two terms. Now,

$$\|\hat{\Pi}^L_{-1} g_1\|_n^2 = \sum_{k=1}^{d_1} \sum_{l=1}^{d_1} \alpha_k \alpha_l \langle \hat{\Pi}^L_{-1} b_{1k}, \hat{\Pi}^L_{-1} b_{1l} \rangle_n,$$

$$\|\Pi_{-1} g_1\|^2 = \sum_{k=1}^{d_1} \sum_{l=1}^{d_1} \alpha_k \alpha_l \langle \Pi_{-1} b_{1k}, \Pi_{-1} b_{1l} \rangle,$$

and thus

$$\begin{aligned}&\|\hat{\Pi}^L_{-1} g_1\|_n^2 - \|\Pi_{-1} g_1\|^2 \\ &= \sum_{k=1}^{d_1} \sum_{l=1}^{d_1} \alpha_k \alpha_l \left(\langle \hat{\Pi}^L_{-1} b_{1k}, \hat{\Pi}^L_{-1} b_{1l} \rangle_n - \langle \Pi_{J_k} b_{1k}, \Pi_{J_l} b_{1l} \rangle_n \right) \qquad (5.5) \\ &+ \sum_{k=1}^{d_1} \sum_{l=1}^{d_1} \alpha_k \alpha_l \left(\langle \Pi_{J_k} b_{1k}, \Pi_{J_l} b_{1l} \rangle_n - \langle \Pi_{J_k} b_{1k}, \Pi_{J_l} b_{1l} \rangle \right) \qquad (5.6) \\ &+ \sum_{k=1}^{d_1} \sum_{l=1}^{d_1} \alpha_k \alpha_l \left(\langle \Pi_{J_k} b_{1k}, \Pi_{J_l} b_{1l} \rangle - \langle \Pi_{-1} b_{1k}, \Pi_{-1} b_{1l} \rangle \right). \qquad (5.7)\end{aligned}$$

First, consider the term (5.5). Using the identity

$$\langle a', b' \rangle_n - \langle a, b \rangle_n = \langle a' - a, b' - b \rangle_n + \langle a, b' - b \rangle_n + \langle a' - a, b \rangle_n,$$

we get

$$\begin{aligned}&\langle \hat{\Pi}^L_{-1} b_{1k}, \hat{\Pi}^L_{-1} b_{1l} \rangle_n - \langle \Pi_{J_k} b_{1k}, \Pi_{J_l} b_{1l} \rangle_n \\ &= \langle \hat{\Pi}^L_{-1} b_{1k} - \Pi_{J_k} b_{1k}, \hat{\Pi}^L_{-1} b_{1l} - \Pi_{J_l} b_{1l} \rangle_n \\ &+ \langle \Pi_{J_k} b_{1k}, \hat{\Pi}^L_{-1} b_{1l} - \Pi_{J_l} b_{1l} \rangle_n + \langle \hat{\Pi}^L_{-1} b_{1k} - \Pi_{J_k} b_{1k}, \Pi_{J_l} b_{1l} \rangle_n.\end{aligned}$$

Plugging in the formulas

$$\|\Pi_{J_k} b_{1k}\|_n \le \frac{C}{\psi} \sqrt{\frac{s_1}{d_1}}$$

and

$$\|\hat{\Pi}^L_{-1} b_{1k} - \Pi_{J_k} b_{1k}\|_n \le \frac{C}{\phi} \sqrt{s_1} \eta,$$

which hold on $\mathcal{E}_1 \cap \mathcal{E}_2$ by (A.5), (A.6) in the supplementary material, and Proposition 3, we get

$$|\langle \hat{\Pi}_{-1}^L b_{1k}, \hat{\Pi}_{-1}^L b_{1l}\rangle_n - \langle \Pi_{J_k} b_{1k}, \Pi_{J_l} b_{1l}\rangle_n| \le C\left(\frac{1}{\phi\psi}\frac{s_1\eta}{\sqrt{d_1}} + \frac{s_1\eta^2}{\phi^2}\right).$$

Hence, if $\mathcal{E}_1 \cap \mathcal{E}_2$ holds, then the term (5.5) can be bounded by

$$\begin{aligned} &C\sum_{k=1}^{d_1}\sum_{l=1}^{d_1}|\alpha_k||\alpha_l|\left(\frac{1}{\phi\psi}\frac{s_1\eta}{\sqrt{d_1}} + \frac{s_1\eta^2}{\phi^2}\right) \\ &\le C\left(\frac{1}{\phi\psi}\frac{s_1\sqrt{d_1}\eta}{\sqrt{d_1}} + \frac{s_1 d_1\eta^2}{\phi^2}\right)\|\alpha\|_2^2, \end{aligned}$$

where we applied the Cauchy-Schwarz inequality in the last step. Next, the last term in (5.7) can be bounded similarly. As above, we have

$$\begin{aligned} &\langle \Pi_{-1}b_{1k}, \Pi_{-1}b_{1l}\rangle - \langle \Pi_{J_k}b_{1k}, \Pi_{J_l}b_{1l}\rangle \\ &= \langle \Pi_{-1}b_{1k} - \Pi_{J_k}b_{1k}, \Pi_{-1}b_{1l} - \Pi_{J_l}b_{1l}\rangle \\ &+ \langle \Pi_{J_k}b_{1k}, \Pi_{-1}b_{1l} - \Pi_{J_l}b_{1l}\rangle + \langle \Pi_{-1}b_{1k} - \Pi_{J_k}b_{1k}, \Pi_{J_l}b_{1l}\rangle \end{aligned}$$

and thus, using Assumption 4 and Proposition 4,

$$|\langle \Pi_{-1}b_{1k}, \Pi_{-1}b_{1l}\rangle - \langle \Pi_{J_k}b_{1k}, \Pi_{J_l}b_{1l}\rangle| \le \frac{C}{\psi}\frac{s_1\eta}{\sqrt{d_1}} + C_1^2 s_1\eta^2.$$

Hence, the term (5.7) can be bounded by

$$\left(\frac{C}{\psi}s_1\sqrt{d_1}\eta + C_1^2 s_1 d_1\eta^2\right)\|\alpha\|_2^2.$$

Finally, consider the middle term (5.6). If $\mathcal{E}_2$ holds, then

$$4|\langle \Pi_{J_k}b_{1k}, \Pi_{J_l}b_{1l}\rangle_n - \langle \Pi_{J_k}b_{1k}, \Pi_{J_l}b_{1l}\rangle| \le 4\delta\|\Pi_{J_k}b_{1k}\|\|\Pi_{J_l}b_{1l}\|$$

which, by using Proposition 4, is bounded by

$$\frac{C}{\psi^2}\frac{\delta s_1}{d_1}.$$

Hence, (5.6) can be bounded by

$$\frac{C}{\psi^2}\delta s_1\|\alpha\|_2^2$$

Inserting the bounds for (5.5)-(5.7) into (5.4), we obtain

$$\|\hat{\Pi}_{-1}^L g_1\|_n^2 \le \rho_0^2\|g_1\|_n^2 + 2\delta\|g_1\|_n^2 + C\left(\frac{s_1\delta}{\psi^2} + \frac{s_1\sqrt{d_1}\eta}{\psi\phi} + \frac{s_1 d_1\eta^2}{\phi^2}\right)\|\alpha\|_2^2.$$

Finally, if $\mathcal{E}_{\delta,1}$ holds, then

$$\|g_1\|_n^2 \ge (1-\delta)\|g_1\|^2 \ge (c/2)\|\alpha\|_2^2, \tag{5.8}$$

and the claim follows. □

Proposition 7. *Suppose that Assumptions 1, 2, 4, and 5 hold. If $\mathcal{E}_1 \cap \mathcal{E}_2$ holds, then for each $g_1 \in V_1$ we have*

$$\|\hat{\Pi}_1(\hat{\Pi}_1\hat{\Pi}_{-1}^L)^* g_1\|_\infty \leq C\left(\frac{\sqrt{s_1}}{\psi} + \frac{\sqrt{s_1 d_1}\eta}{\phi}\right)\|g_1\|_n.$$

In particular, if additionally (5.2) *is satisfied, then*

$$\|\hat{\Pi}_1(\hat{\Pi}_1\hat{\Pi}_{-1}^L)^* g_1\|_\infty \leq \frac{C}{\psi}\sqrt{s_1}\|g_1\|_n.$$

Remark 2. The proof adapts the following argument valid in the population setting: if $\psi_{11}, \dots, \psi_{1d_1}$ is an orthonormal basis of V_1 with respect to $\|\cdot\|$, then

$$\begin{aligned}
\|\Pi_1\Pi_{-1}g_1\|_\infty &\leq C\sqrt{d_1}\max_{k=1,\dots,d_1}|\langle\psi_{1k}, \Pi_{-1}g_1\rangle| \\
&= C\sqrt{d_1}\max_{k=1,\dots,d_1}|\langle\Pi_{-1}\psi_{1k}, \Pi_{-1}g_1\rangle| \\
&\leq C\sqrt{d_1}\max_{k=1,\dots,d_1}\|\Pi_{-1}\psi_{1k}\|\|\Pi_{-1}g_1\| \\
&\leq C\sqrt{s_1}\|\Pi_{-1}g_1\|.
\end{aligned}$$

Proof. Suppose that $\mathcal{E}_{\delta,1}$ holds. Then let $\phi_{11}, \dots, \phi_{1d_1}$ be the empirical orthonormal basis of V_1 constructed in Section 4.1. Between the supremum norm of the coefficient vector and the supremum norm of the corresponding function, we have the following relation (see Appendix C.2 in the supplementary material for the proof): let $g_1 = \sum_{k=1}^{d_1}\alpha_k\phi_{1k}$. If the event $\mathcal{E}_{\delta,1}$ occurs, then

$$\|g_1\|_\infty \leq C\sqrt{d_1}\|\alpha\|_\infty, \tag{5.9}$$

where C is a constant depending on c_1 and t_1. A proof of (5.9) is given in Appendix C in the supplementary material. If $\mathcal{E}_{\delta,1}$ holds, then (5.9) implies

$$\begin{aligned}
\|\hat{\Pi}_1(\hat{\Pi}_1\hat{\Pi}_{-1}^L)^* g_1\|_\infty &\leq C\sqrt{d_1}\max_{k=1,\dots,d_1}|\langle\phi_{1k}, (\hat{\Pi}_1\hat{\Pi}_{-1}^L)^* g_1\rangle_n| \\
&\leq C\sqrt{d_1}\max_{k=1,\dots,d_1}|\langle\hat{\Pi}_1\hat{\Pi}_{-1}^L\phi_{1k}, g_1\rangle_n| \\
&= C\sqrt{d_1}\max_{k=1,\dots,d_1}|\langle\hat{\Pi}_{-1}^L\phi_{1k}, g_1\rangle_n| \\
&\leq C\sqrt{d_1}\max_{k=1,\dots,d_1}\|\hat{\Pi}_{-1}^L\phi_{1k}\|_n\|g_1\|_n.
\end{aligned}$$

By Corollary 1, we have on $\mathcal{E}_1 \cap \mathcal{E}_2$,

$$\|\hat{\Pi}_{-1}^L \phi_{1k}\|_n \leq \frac{C}{\psi}\sqrt{\frac{s_1}{d_1}} + \frac{C}{\phi}\sqrt{s_1}\eta,$$

and the claim follows. □

Corollary 3. *Suppose that Assumptions 1, 2, 4, and 5 hold. If* (5.2) *is satisfied and if $\mathcal{E}_1 \cap \mathcal{E}_2$ holds, then for each $g_1 \in V_1$, we have*

$$\left\|(I - (\hat{\Pi}_1\hat{\Pi}_{-1}^L)^*)^{-1} g_1\right\|_\infty \leq \|g_1\|_\infty + \frac{C}{(1-\rho_0)\psi}\sqrt{s_1}\|g_1\|_n.$$

Proof. First note that (5.2) is satisfied and if $\mathcal{E}_1 \cap \mathcal{E}_2$ holds, then Proposition 6 implies that $\hat{\rho}_0 \leq (1+\rho_0)/2 < 1$. Hence, by Corollary 2,

$$(I - (\hat{\Pi}_1\hat{\Pi}_{-1}^L)^*)^{-1} = \sum_{m\geq 0}((\hat{\Pi}_1\hat{\Pi}_{-1}^L)^*)^m = I + \sum_{m\geq 1}\hat{\Pi}_1((\hat{\Pi}_1\hat{\Pi}_{-1}^L)^*)^m.$$

and

$$\left\|(I - (\hat{\Pi}_1\hat{\Pi}_{-1}^L)^*)^{-1} g_1\right\|_\infty \leq \|g_1\|_\infty + \sum_{m\geq 1}\|\hat{\Pi}_1((\hat{\Pi}_1\hat{\Pi}_{-1}^L)^*)^m g_1\|_\infty.$$

Applying Proposition 7 and then Corollary 2, this can be bounded by

$$\begin{aligned}
&\|g_1\|_\infty + \frac{C}{\psi}\sqrt{s_1}\sum_{m\geq 1}\|((\hat{\Pi}_1\hat{\Pi}_{-1}^L)^*)^{m-1} g_1\|_n \\
&\leq \|g_1\|_\infty + \frac{C}{\psi}\sqrt{s_1}\sum_{m\geq 1}\hat{\rho}_0^m\|g_1\|_n \\
&= \|g_1\|_\infty + \frac{C}{(1-\hat{\rho}_0)\psi}\sqrt{s_1}\|g_1\|_n \\
&\leq \|g_1\|_\infty + \frac{2C}{(1-\rho_0)\psi}\sqrt{s_1}\|g_1\|_n,
\end{aligned}$$

and the claim follows. □

5.4. **Proof of Theorem 5.** Recall that

$$(I - (\hat{\Pi}_1\hat{\Pi}_{-1}^L)^*)^{-1}\hat{\Pi}_1(I - (\hat{\Pi}_{-1}^L)^*)(g_{-1}^* - \hat{f}_{-1}^L) = (I - (\hat{\Pi}_1\hat{\Pi}_{-1}^L)^*)^{-1} g_{\alpha_1},$$

where

$$\alpha_1 = \left(\langle \phi_{1k} - \hat{\Pi}_{-1}^L\phi_{1k}, g_{-1}^* - \hat{f}_{-1}^L\rangle_n\right)_{k=1}^{d_1}$$

and

$$g_{\alpha_1} = \sum_{k=1}^{d_1}\alpha_{1k}\phi_{1k}.$$

Suppose that $\mathcal{E}_0 \cap \mathcal{E}_1 \cap \mathcal{E}_2$ holds. Applying Corollary 3 and Equation (5.9), we obtain that

$$\begin{aligned}\left\|(I - (\hat{\Pi}_1\hat{\Pi}_{-1}^L)^*)^{-1} g_{\alpha_1}\right\|_\infty &\le \frac{C}{(1-\rho_0)\psi}\sqrt{s_1}\|g_{\alpha_1}\|_\infty \\ &\le \frac{C}{(1-\rho_0)\psi}\sqrt{s_1 d_1}\|\alpha_1\|_\infty.\end{aligned}$$

Applying the fact that ϕ_{1k} is a linear combination of at most t_1+1 basis functions (b_{1l}), the Cauchy-Schwarz inequality, and Equation (5.8), we get

$$\|\alpha_1\|_\infty \le C \max_{l=1,\dots,d_1} \left|\langle b_{1k} - \hat{\Pi}_{-1}^L b_{1k}, g_{-1}^* - \hat{f}_{-1}^L\rangle_n\right|.$$

Hence, by Lemma 1, we conclude that

$$\left\|(I - (\hat{\Pi}_1\hat{\Pi}_{-1}^L)^*)^{-1} g_{\alpha_1}\right\|_\infty \le \frac{C}{(1-\rho_0)\psi}\sqrt{s_1 d_1}\,\mathrm{pen}_\eta(\hat{f}_{-1}^L - g_{-1}^*).$$

Moreover, by Proposition 2,

$$\begin{aligned}\mathrm{pen}_\eta(\hat{f}_{-1}^L - g_{-1}^*) &\le (\eta/\lambda)\,\mathrm{pen}_\lambda(\hat{f}^L - g^*) \\ &\le C\left((\eta/\lambda)\left(d_1^{-r_1} + s_0 d_2^{-r_2}\right)^2 + s_0\lambda\eta/\phi^2\right),\end{aligned}$$

and the claim follows. □

5.5. **Proof of Theorem 6.**

Recall that

$$(I - (\hat{\Pi}_1\hat{\Pi}_{-1}^L)^*)^{-1}\hat{\Pi}_1(I - (\hat{\Pi}_{-1}^L)^*)(f - g^*) = (I - (\hat{\Pi}_1\hat{\Pi}_{-1}^L)^*)^{-1} g_{\alpha_2},$$

where

$$\alpha_2 = \left(\langle \phi_{1k} - \hat{\Pi}_{-1}^L\phi_{1k}, f - g^*\rangle_n\right)_{k=1}^{d_1}$$

and

$$g_{\alpha_2} = \sum_{k=1}^{d_1}\alpha_{2k}\phi_{1k}.$$

Suppose that $\mathcal{E}_1 \cap \mathcal{E}_2$ holds. Applying Corollary 3 and (5.9), we obtain as above

$$\left\|(I - (\hat{\Pi}_1\hat{\Pi}_{-1}^L)^*)^{-1} g_{\alpha_2}\right\|_\infty \le \frac{C}{(1-\rho_0)\psi}\sqrt{s_1 d_1}\|\alpha_2\|_\infty.$$

Now,

$$\begin{aligned}\|\alpha_2\|_\infty &\le \max_{k=1,\dots,d_1}|\langle\phi_{1k}, f - g^*\rangle_n| + \max_{k=1,\dots,d_1}|\langle\hat{\Pi}_{-1}^L\phi_{1k}, f - g^*\rangle_n| \\ &\le C\|f - g^*\|_\infty \max_{k=1,\dots,d_1}\|b_{1k}\|_n/\sqrt{d_1} + \max_{k=1,\dots,d_1}\|\hat{\Pi}_{-1}^L\phi_{1k}\|_n\|f - g^*\|_n,\end{aligned}$$

and the first claim follows from Assumption 3, Corollary 1, and the bound

$$\|b_{1k}\|_n^2 \leq (1+\delta)\|b_{1k}\|^2 \leq 2/c_1.$$

The second bound can be proven by using the first part of Assumption 3 and the fact that piecewise polynomial smoothing preserves the sup norm. The details of the latter argument can be found in the proof of Theorem 2. □

5.6. **Proof of Theorem 7.** For the variance term, we will return to the representation of the estimator through the coefficient vector. The function

$$(I - (\hat{\Pi}_1 \hat{\Pi}_{-1}^L)^*)^{-1} \hat{\Pi}_1 (I - (\hat{\Pi}_{-1}^L)^*) \boldsymbol{\epsilon} - \hat{\Pi}_1 \boldsymbol{\epsilon} \tag{5.10}$$

has coefficient vector

$$\left(\mathbf{I} - \frac{1}{n}\boldsymbol{\Delta}_1^T \mathbf{X}_1\right)^{-1} \frac{1}{n}\left(\mathbf{X}_1 - \boldsymbol{\Delta}_1\right)^T \boldsymbol{\epsilon} - \frac{1}{n}\mathbf{X}_1^T \boldsymbol{\epsilon},$$

where

$$\boldsymbol{\Delta}_1 = \left((\hat{\Pi}_{-1}^L \phi_{1k})(X^i)\right)_{1\leq i\leq n, 1\leq k\leq d_1}.$$

Recall that $\phi_{11}, \ldots, \phi_{1d_1}$ is the empirical orthonormal basis of V_1 constructed in Section 4.1. Setting

$$U_k = \|\phi_{1k}\|_\infty \cdot e_k^T \left(\left(\mathbf{I} - \frac{1}{n}\boldsymbol{\Delta}_1^T \mathbf{X}_1\right)^{-1} \frac{1}{n}\left(\mathbf{X}_1 - \boldsymbol{\Delta}_1\right)^T \boldsymbol{\epsilon} - \frac{1}{n}\mathbf{X}_1^T \boldsymbol{\epsilon}\right),$$

where e_k is the kth standard basis vector, we see that the supremum of the function in (5.10) can bounded as follows:

$$\|(I - (\hat{\Pi}_1 \hat{\Pi}_{-1}^L)^*)^{-1} \hat{\Pi}_1 (I - (\hat{\Pi}_{-1}^L)^*) \boldsymbol{\epsilon} - \hat{\Pi}_1 \boldsymbol{\epsilon}\|_\infty \leq (t_1 + 1) \max_{k=1,\ldots,d_1} U_k,$$

The following result implies Theorem 7:

Proposition 8. *Suppose that Assumptions 1, 2, 4, and 5 hold. Suppose that (5.2) is satisfied. Moreover, suppose that $\mathcal{E}_1 \cap \mathcal{E}_2$ holds. Then*

$$\mathbf{E}_\epsilon U_k^2 \leq \frac{C}{\psi^2 (1-\rho_0)^2} \frac{s_1}{n},$$

where $\mathbf{E}_\epsilon$ denotes the expectation with respect to $\epsilon^1, \ldots, \epsilon^n$ for given, fixed values of $X^1, \ldots, X^n$. In particular, since each U_k is Gaussian (conditional on $X^1, \ldots, X^n$), we obtain that for all $y \geq 0$,

$$\mathbf{P}_\epsilon \left(\max_{k=1,\ldots,d_1} U_k \geq \frac{1}{(1-\rho_0)\psi} \sqrt{\frac{Cs_1(2\log d_1 + 2y)}{n}}\right) \leq \exp(-y),$$

where $\mathbf{P}_\epsilon$ denotes the probability with respect to $\epsilon^1, \ldots, \epsilon^n$ for given, fixed values of $X^1, \ldots, X^n$.

We have

$$\begin{aligned}\mathbf{E}_\epsilon U_k^2 = \frac{\|\phi_{1k}\|_\infty^2}{n} \\ \cdot \left(e_k^T \left(\mathbf{I} - \frac{1}{n}\mathbf{\Delta}_1^T \mathbf{X}_1 \right)^{-1} \frac{1}{n}(\mathbf{X}_1 - \mathbf{\Delta}_1)^T(\mathbf{X}_1 - \mathbf{\Delta}_1) \left(\mathbf{I} - \frac{1}{n}\mathbf{X}_1^T \mathbf{\Delta}_1 \right)^{-1} e_k - 1 \right).\end{aligned} \tag{5.11}$$

If $\mathcal{E}_{\delta,1}$ holds, then (5.9) gives

$$\|\phi_{1k}\|_\infty^2 \le C d_1.$$

Hence, it remains to show that the term in the brackets is bounded by Cs_1/d_1. The proof of this result is a bit technical (since the term in the brackets is quite long). However, the main idea in the proof can be seen by analyzing the following similar but more simple term:

$$e_k^T \left(\mathbf{I} - \frac{1}{n}\mathbf{\Delta}_1^T \mathbf{\Delta}_1 \right)^{-1} e_k - e_k^T e_k. \tag{5.12}$$

Let us restrict ourselves to the event that the operator norm of $(1/n)\mathbf{\Delta}_1^T\mathbf{\Delta}_1$ is bounded by ρ (see Lemma 4). Then, using that $(1/n)\mathbf{\Delta}_1^T\mathbf{\Delta}_1$ is symmetric and positive semi-definite, one can show that (5.12) is bounded by

$$\frac{e_k^T \frac{1}{n}\mathbf{\Delta}_1^T \mathbf{\Delta}_1 e_k}{1-\rho} = \frac{\|\hat{\Pi}_{-1}^L \phi_{1k}\|_n^2}{1-\rho},$$

and Corollary 1 implies that (5.12) is bounded by Cs_1/d_1, as claimed. In order to generalize this analysis to the term in the brackets of (5.11), we first derive some lemmas:

Lemma 3. *For each $\alpha \in \mathbb{R}^{d_1}$, we have*

$$\left\| \frac{1}{n}\mathbf{X}_1^T \mathbf{\Delta}_1 \alpha \right\|_2^2 \le \hat{\rho}_0^2 \|\alpha\|_2^2.$$

Proof. We have

$$\begin{aligned}&\left\| \frac{1}{n}\mathbf{X}_1^T \mathbf{\Delta}_1 \alpha \right\|_2^2 \\ &= \sum_{l=1}^{d_1} \left(\sum_{k=1}^{d_1} \langle \phi_{1l}, \hat{\Pi}_{-1}^L \phi_{1k} \rangle_n \alpha_k \right)^2 = \sum_{l=1}^{d_1} \langle \phi_{1l}, \hat{\Pi}_{-1}^L g_\alpha \rangle_n^2 = \|\hat{\Pi}_1 \hat{\Pi}_{-1}^L g_\alpha\|_n^2\end{aligned}$$

Hence, by the definition of $\hat{\rho}_0$,

$$\left\| \frac{1}{n}\mathbf{X}_1^T \mathbf{\Delta}_1 \alpha \right\|_2^2 = \|\hat{\Pi}_1 \hat{\Pi}_{-1}^L g_\alpha\|_n^2 \le \|\hat{\Pi}_{-1}^L g_\alpha\|_n^2 \le \hat{\rho}_0^2 \|g_\alpha\|_n^2 = \hat{\rho}_0^2 \|\alpha\|_2^2,$$

and the claim follows. □

Lemma 4. *Suppose that Assumptions 1, 4, and 5 hold. If $\mathcal{E}_1 \cap \mathcal{E}_2$ holds, then we have for each $\alpha \in \mathbb{R}^{d_1}$,*

$$\left\| \frac{1}{n} \mathbf{\Delta}_1^T \mathbf{\Delta}_1 \alpha \right\|_2^2 \leq \left((1+\epsilon)\hat{\rho}_0^2 + C(1+1/\epsilon) \left(\frac{s_1\sqrt{d_1}\eta}{\psi^2} + \frac{s_1 d_1 \eta^2}{\phi^2} \right)^2 \right) \|\alpha\|_2^2,$$

where $\epsilon > 0$ is arbitrary. In particular, if (5.2) *is satisfied and if we choose $\epsilon = 1$, then we have*

$$\left\| \frac{1}{n} \mathbf{\Delta}_1^T \mathbf{\Delta}_1 \alpha \right\|_2^2 \leq C \|\alpha\|_2^2.$$

Proof. We have

$$\begin{aligned} &\left\| \frac{1}{n} \mathbf{\Delta}_1^T \mathbf{\Delta}_1 \alpha \right\|_2^2 \\ &= \sum_{l=1}^{d_1} \left(\sum_{k=1}^{d_1} \langle \hat{\Pi}_{-1}^L \phi_{1l}, \hat{\Pi}_{-1}^L \phi_{1k} \rangle_n \alpha_k \right)^2 \\ &\leq (1+\epsilon) \sum_{l=1}^{d_1} \left(\sum_{k=1}^{d_1} \langle \phi_{1l}, \hat{\Pi}_{-1}^L \phi_{1k} \rangle_n \alpha_k \right)^2 \\ &+ (1+1/\epsilon) \sum_{l=1}^{d_1} \left(\sum_{k=1}^{d_1} \langle \phi_{1l} - \hat{\Pi}_{-1}^L \phi_{1l}, \hat{\Pi}_{-1}^L \phi_{1k} \rangle_n \alpha_k \right)^2 . \end{aligned}$$

Now, the first term is equal to

$$(1+\epsilon) \left\| \frac{1}{n} \mathbf{X}_1^T \mathbf{\Delta}_1 \alpha \right\|_2^2 \leq (1+\epsilon)\hat{\rho}_0^2 \|\alpha\|_2^2,$$

by Lemma 3. Applying (5.1) and Corollary 1, we have on $\mathcal{E}_1 \cap \mathcal{E}_2$ (see Appendix C.3 in the supplementary material for the details),

$$|\langle \phi_{1l} - \hat{\Pi}_{-1}^L \phi_{1l}, \hat{\Pi}_{-1}^L \phi_{1k} \rangle_n| \leq C \left(\frac{1}{\psi^2} \frac{s_1 \eta}{\sqrt{d_1}} + \frac{s_1 \eta^2}{\phi^2} \right). \tag{5.13}$$

Hence, applying the Cauchy-Schwarz inequality, the second term can be bounded by

$$C^2(1+1/\epsilon) d_1 \left(\frac{s_1 \eta}{\psi^2} + \frac{s_1 \sqrt{d_1} \eta^2}{\phi^2} \right)^2 \|\alpha\|_2^2,$$

and the claim follows. □

Lemma 5. *Suppose that Assumptions 1, 4, and 5 hold. If $\mathcal{E}_1 \cap \mathcal{E}_2$ holds, then we have*

$$\left\| \frac{1}{n} \mathbf{X}_1^T \mathbf{\Delta}_1 e_k \right\|_2^2 \leq \frac{C}{\psi^2} \frac{s_1}{d_1} + \frac{C}{\phi^2} s_1 \eta^2,$$

and

$$\left\| \frac{1}{n} \mathbf{\Delta}_1^T \mathbf{\Delta}_1 e_k \right\|_2^2 \leq C \left(\frac{s_1}{\psi^2 d_1} + \frac{s_1 \eta^2}{\phi^2} + \frac{s_1^2 \eta^2}{\psi^4} + \frac{s_1^2 d_1 \eta^4}{\phi^4} \right).$$

In particular, if (5.2) *is satisfied, then the two upper bounds become* $(C/\psi^2) s_1/d_1$.

Proof. We have

$$\left\| \frac{1}{n} \mathbf{X}_1^T \mathbf{\Delta}_1 e_k \right\|_2^2 = \|\hat{\Pi}_1 \hat{\Pi}_{-1}^L \phi_{1k}\|_n^2 \leq \|\hat{\Pi}_{-1}^L \phi_{1k}\|_n^2$$

and thus Corollary 1 gives the first claim. Next, we have

$$\begin{aligned}
\left\| \frac{1}{n} \mathbf{\Delta}_1^T \mathbf{\Delta}_1 e_k \right\|_2^2 &= \sum_{l=1}^{d_1} \langle \hat{\Pi}_{-1}^L \phi_{1l}, \hat{\Pi}_{-1}^L \phi_{1k} \rangle_n^2 \\
&\leq 2 \sum_{l=1}^{d_1} \langle \phi_{1l}, \hat{\Pi}_{-1}^L \phi_{1k} \rangle_n^2 + 2 \sum_{l=1}^{d_1} \langle \phi_{1l} - \hat{\Pi}_{-1}^L \phi_{1l}, \hat{\Pi}_{-1}^L \phi_{1k} \rangle_n^2 \\
&= 2 \|\hat{\Pi}_1 \hat{\Pi}_{-1}^L \phi_{1k}\|_n^2 + 2 \sum_{l=1}^{d_1} \langle \phi_{1l} - \hat{\Pi}_{-1}^L \phi_{1l}, \hat{\Pi}_{-1}^L \phi_{1k} \rangle_n^2,
\end{aligned}$$

and thus (5.13) and Corollary 1 imply the second claim. □

Proof of Proposition 8. As argued above, it remains to show that the term in the brackets of (5.11) is bounded by $C s_1/d_1$. First, this term is equal to

$$e_k^T \left(\mathbf{I} - \frac{1}{n} \mathbf{\Delta}_1^T \mathbf{X}_1 \right)^{-1} \left(\mathbf{I} - \frac{1}{n} \mathbf{X}_1^T \mathbf{\Delta}_1 - \frac{1}{n} \mathbf{\Delta}_1^T \mathbf{X}_1 + \frac{1}{n} \mathbf{\Delta}_1^T \mathbf{\Delta}_1 \right) \left(\mathbf{I} - \frac{1}{n} \mathbf{X}_1^T \mathbf{\Delta}_1 \right)^{-1} e_k - 1,$$

which, by using the identity

$$1 = e_k^T \left(\mathbf{I} - \frac{1}{n} \mathbf{\Delta}_1^T \mathbf{X}_1 \right)^{-1} \left(\mathbf{I} - \frac{1}{n} \mathbf{\Delta}_1^T \mathbf{X}_1 \right) \left(\mathbf{I} - \frac{1}{n} \mathbf{X}_1^T \mathbf{\Delta}_1 \right) \left(\mathbf{I} - \frac{1}{n} \mathbf{X}_1^T \mathbf{\Delta}_1 \right)^{-1} e_k,$$

can be rewritten as

$$\begin{aligned}
&e_k^T \left(\mathbf{I} - \frac{1}{n} \mathbf{\Delta}_1^T \mathbf{X}_1 \right)^{-1} \frac{1}{n} \mathbf{\Delta}_1^T \mathbf{\Delta}_1 \left(\mathbf{I} - \frac{1}{n} \mathbf{X}_1^T \mathbf{\Delta}_1 \right)^{-1} e_k \\
&- e_k^T \left(\mathbf{I} - \frac{1}{n} \mathbf{\Delta}_1^T \mathbf{X}_1 \right)^{-1} \frac{1}{n} \mathbf{\Delta}_1^T \mathbf{X}_1 \frac{1}{n} \mathbf{X}_1^T \mathbf{\Delta}_1 \left(\mathbf{I} - \frac{1}{n} \mathbf{X}_1^T \mathbf{\Delta}_1 \right)^{-1} e_k. \quad (5.14)
\end{aligned}$$

By Lemma 3, the operator norm of $(1/n)\mathbf{X}_1^T\mathbf{\Delta}_1$ is bounded by $\hat{\rho}_0$. From now on suppose that $\mathcal{E}_1 \cap \mathcal{E}_2$ holds and that (5.2) is satisfied. Then Proposition 6 implies that $\hat{\rho}_0 < 1$. Combining this with Lemma 3, we get

$$\left(\mathbf{I} - \frac{1}{n}\mathbf{X}_1^T\mathbf{\Delta}_1\right)^{-1} = \sum_{r\geq 0}\left(\frac{1}{n}\mathbf{X}_1^T\mathbf{\Delta}_1\right)^r.$$

First, consider the second term of (5.14). It is equal to

$$\sum_{r,s\geq 1} e_k^T\left(\frac{1}{n}\mathbf{\Delta}_1\mathbf{X}_1\right)^r\left(\frac{1}{n}\mathbf{X}_1^T\mathbf{\Delta}_1\right)^s e_k.$$

Plugging-in Lemma 3 and Lemma 5, this is bounded by

$$\frac{C}{\psi^2}\frac{s_1}{d_1}\left(\sum_{r,s\geq 1}\hat{\rho}_0^{r+s-2}\right) \leq \frac{C}{\psi^2(1-\rho_0)^2}\frac{s_1}{d_1},$$

where we also applied (5.3). Similarly, the first term is equal to

$$\frac{1}{n}e_k^T\mathbf{\Delta}_1^T\mathbf{\Delta}_1 e_k + \sum_{r+s\geq 1} e_k^T\left(\frac{1}{n}\mathbf{X}_1^T\mathbf{\Delta}_1\right)^r\frac{1}{n}\mathbf{\Delta}_1^T\mathbf{\Delta}_1\left(\frac{1}{n}\mathbf{\Delta}_1^T\mathbf{X}_1\right)^s e_k,$$

which, by Corollary 1 and Lemmas 3-5, is bounded by

$$\frac{C}{\psi^2}\frac{s_1}{d_1}\left(1 + \sum_{r+s\geq 2}\hat{\rho}_0^{r+s-2}\right) \leq \frac{C}{\psi^2(1-\rho_0)^2}\frac{s_1}{d_1},$$

and the claim follows. □

Acknowledgements. Support by Deutsche Forschungsgemeinschaft through the Research Training Group RTG 1953, funded by the German Research Foundation (DFG) is gratefully acknowledged. Research of the second author was prepared within the framework of a subsidy granted to the HSE by the Government of the Russian Federation for the implementation of the Global Competitiveness Program.

References

[1] M. Avalos, Y. Grandvalet, and C. Ambroise. Parsimonious additive models. *Comput. Stat. Data Anal.*, 51:2851–2870, 2007.

[2] A. Belloni and V. Chernozhukov. Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19:521–547, 2013.

[3] A. Belloni, V. Chernozhukov, and C. B. Hansen. Inference on treatment effects after selection amongst high-dimensional controls. *Rev. Econ. Stud.*, 81:608–650, 2013.

[4] K. Bertin and G. Lecue. Selection of variables and dimension reduction in high-dimensional non-parametric regression. *Electron. J. Stat.*, 2:1224–1241, 2008.

[5] P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *Ann. Statist.*, 37:1705–1732, 2009.

[6] L. Birgé and P. Massart. From model selection to adaptive estimation. In *Festschrift for Lucien Le Cam*, pages 55–87. Springer, New York, 1997.

[7] L. Birgé and P. Massart. Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli*, 4:329–375, 1998.

[8] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities. A nonasymptotic theory of independence.* Oxford University Press, Oxford, 2013.

[9] C.R. de Boor. A bound on the L_∞-norm of L_2-approximation by splines in terms of global mesh ratio. *Math. Comput.*, 30:765–771, 1976.

[10] S. Efroimovich. On nonparametric regression for iid observations in a general setting. *Ann. Statist.*, 24:1126–1144, 1996.

[11] S. Efroimovich. Nonparametric regression with the scale depending on auxiliary variable. *Ann. Statist.*, 41:1542–1568, 2013.

[12] S. Ehrenfeld. Complete class theorems in experimental design. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 1954–1955, vol. I*, pages 57–67. University of California Press, Berkeley and Los Angeles, 1956.

[13] J. Fan, Y. Feng, and R. Song. Nonparametric independence screening in sparse ultra- high-dimensional additive models. *J. Amer. Statist. Assoc.*, 106:544–557, 2011.

[14] C. Giraud. *Introduction to high-dimensional statistics.* CRC Press, Boca Raton, 2015.

[15] J. Horowitz, J. Klemelä, and E. Mammen. Optimal estimation in additive regression models. *Bernoulli*, 12:271–298, 2006.

[16] J. Huang, J. L. Horowitz, and F. Wei. Variable selection in nonparametric additive models. *Ann. Statist.*, 38:2282–2313, 2010.

[17] A. Javanmard and A. Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *Mach. Learn. Res.*, 15:2869–2909, 2014.

[18] K. Kato. Two-step estimation of high dimensional additive models. Technical Report 1207.5313, ArXiv, 2012.

[19] V. Koltchinskii and M. Yuan. Sparse recovery in large ensembles of kernel machines. In R. A. Servedio and T. Zhang, editors, *Colt*, pages 229–238. Omnipress, Madison, WI., 2008.

[20] M. Ledoux. *The concentration of measure phenomenon.* American Mathematical Society, Providence, 2001.

[21] E.R. Lee and B. U. Park. Sparse estimation in functional linear regression. *J. Multivariate Anal.*, 105:1–17, 2012.

[22] Y. Lin and H. H. Zhang. Component selection and smoothing in multivariate nonparametric regression. *Ann. Statist.*, 34:2272–2297, 2006.

[23] J. Lu, M. Kolar, and H. Liu. Post-regularization confidence bands for high dimensional nonparametric models with local sparsity. Technical Report arXiv: 1503.02978, 2015.

[24] P. Massart. *Concentration Inequalities and Model Selection.* Springer, Berlin, 2007.

[25] L. Meier, S. van de Geer, and P. Bühlmann. High-dimensional additive modeling. *Ann. Statist.*, 37:3779–3821, 2009.

[26] H. S. Noh and B. U. Park. Sparse varying coefficient models for longitudinal data. *Statistica Sinica*, 20:1183–1202, 2010.

[27] G. Raskutti, M. J. Wainwright, and B. Yu. Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *J. Mach. Learn. Res.*, 13:389–427, 2012.

[28] H. Rauhut. Compressive sensing and structured random matrices. In *Theoretical Foundations and Numerical Methods for Sparse Recovery*, Radon Ser. Comput. Appl. Math., 9, pages 1–92. Walter de Gruyter, Berlin, 2010.

[29] P. Ravikumar, J. Lafferty, H. Liu, and L. Wasserman. Sparse additive models. *J. Royal Statist. Assoc. B*, 71:1009–1030, 2009.

[30] C. J. Stone. Additive regression and other nonparametric models. *Ann. Statist.*, 13:689–705, 1985.

[31] T. Suzuki and M. Sugiyama. Fast learning rate of multiple kernel learning: trade-off between sparsity and smoothness. *Ann. Statist.*, 41:1381–1405, 2013.

[32] S.A. van de Geer, P. Bühlmann, Y. Ritov, and R. Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Stat.*, 42:1166–1202, 2014.

[33] M. Wahl. A theory of nonparametric regression in the presence of complex nuisance components. *Preprint. http://arxiv.org/abs/1403.1088*, 2014.

[34] E. T. Whittaker and G. N. Watson. *A course of modern analysis.* Cambridge Mathematical Library. Cambridge University Press, Cambridge, reprint of the fourth (1927) edition, 1996.

[35] M. Yuan. Nonnegative garrote component selection in functional ANOVA models. *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics (AISTATS-07)*, pages 656–662, 2007.

[36] C.-H. Zhang and S. S. Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Stat. Soc. B*, 76:217–242, 2013.

[37] S. Zhou, X. Shen, and D.A. Wolfe. Local asymptotics for regression splines and confidence regions. *Ann. Statist.*, 26:1760–1782, 1998.