The Annals of Statistics
2016, Vol. 44, No. 6, 2564–2593
DOI: 10.1214/15-AOS1422

MINIMAX OPTIMAL RATES OF ESTIMATION IN HIGH DIMENSIONAL ADDITIVE MODELS

BY MING YUAN[1] AND DING-XUAN ZHOU[2]

University of Wisconsin-Madison and City University of Hong Kong

We establish minimax optimal rates of convergence for estimation in a high dimensional additive model assuming that it is approximately sparse. Our results reveal a behavior universal to this class of high dimensional problems. In the *sparse regime* when the components are sufficiently smooth or the dimensionality is sufficiently large, the optimal rates are identical to those for high dimensional linear regression and, therefore, there is no additional cost to entertain a nonparametric model. Otherwise, in the so-called *smooth regime*, the rates coincide with the optimal rates for estimating a univariate function and, therefore, they are immune to the "curse of dimensionality."

1. Introduction. With the recent advances in science and technology, high dimensional regression problems have become ubiquitous in a multitude of areas—genomics, medical imaging and finance are a few well-known examples. A considerable amount of research effort has been devoted to the understanding of challenges brought about by the high dimensionality, and development of statistical methodology to counter them. Most of the existing work focuses on high dimensional linear regression where a number of approaches such as the nonnegative garrote [Breiman (1995)], the Lasso [Tibshirani (1996)], the SCAD [Fan and Li (2001)] and the Dantzig selector [Candes and Tao (2007)] have been developed to exploit sparsity, or perform variable selection; and much progress has also been made to understand to what extent a high dimensional regression coefficient vector can be reliably estimated; see, for example, Koltchinskii (2011), Bühlmann and van de Geer (2011) and references therein.

Linear models, however, could be too restrictive in many applications. As a more flexible alternative, high dimensional additive models have attracted much attention in the past several years. See, for example, Lin and Zhang (2006), Yuan (2007), Koltchinskii and Yuan (2008), Ravikumar et al. (2009), Meier, van de Geer and Bühlmann (2009), Huang, Horowitz and Wei (2010), Koltchinskii and Yuan

Received August 2015; revised November 2015.

[1]Supported in part by NSF Career Award DMS-13-21692, FRG Grant DMS-12-65202 and NIH Grant 1-U54AI117924-01.

[2]Supported in part by the Research Grants Council of Hong Kong [Project No. CityU 104012] and by National Natural Science Foundation of China [Project No. 11461161006].

MSC2010 subject classifications. Primary 62G08, 62F12; secondary 62J07.

Key words and phrases. Convergence rate, method of regularization, minimax optimality, reproducing kernel Hilbert space, Sobolev space.

(2010), Fan, Feng and Song (2011), Raskutti, Wainwright and Yu (2012) and Cui et al. (2013) among others. Let $\{(X_i, Y_i) : i = 1, \ldots, n\}$ be independent copies of a random couple (X, Y) following a regression model

$$Y = f(X) + \varepsilon, \tag{1.1}$$

where the error ε follows a $\mathcal{N}(0, \sigma^2)$ distribution. The additive model amounts to the assumption that

$$f(x_1, \ldots, x_d) = f_1(x_1) + \cdots + f_d(x_d), \tag{1.2}$$

where the component functions f_js are modeled nonparametrically; see, for example, Stone (1985) or Hastie and Tibshirani (1990). Here, we assume that they reside in certain reproducing kernel Hilbert spaces (RKHS); see, for example, Aronszajn (1950) and Wahba (1990).

To fix ideas, assume that X follows a distribution Π supported on a product space $\mathcal{X}^d$ for some compact subset $\mathcal{X}$ of $\mathbb{R}$, and that all component functions come from a common RKHS of functions on $\mathcal{X}$, denoted by $(\mathcal{H}_1, \|\cdot\|_{\mathcal{H}_1})$. It is clear that the additive model (1.2) can be identified with space

$$\begin{aligned}\mathcal{H}_d &:= \mathcal{H}_1 \oplus \cdots \oplus \mathcal{H}_1 \\ &= \{g : \mathcal{X}^d \to \mathbb{R} | g(x_1, \ldots, x_d) = g_1(x_1) + \cdots + g_d(x_d), \\ &\quad\ \text{and } g_1, \ldots, g_d \in \mathcal{H}_1\}.\end{aligned}$$

Obviously, linear models can be viewed as a trivial special case of (1.2) by taking $\mathcal{H}_1$ to be the collection of all univariate linear functions defined over $\mathcal{X}$. Another canonical example of $\mathcal{H}_1$ is the mth $(2m > k)$ order Sobolev space $\mathcal{W}_2^m([0, 1]^k)$ defined on a unit interval $(\mathcal{X} = [0, 1]^k)$. See, for example, Wahba (1990) for further examples.

We note that for a general $g \in \mathcal{H}_d$, the additive representation given by (1.2) may not be unique. Define the (quasi-)norm $\|f\|_{\ell_q(\mathcal{H}_d)}$ $(q > 0)$ by

$$\begin{aligned}&\|g\|_{\ell_q(\mathcal{H}_d)} \\ &\quad = \inf\{\|(\|g_1\|_{\mathcal{H}_1}, \ldots, \|g_d\|_{\mathcal{H}_1})^\top\|_{\ell_q} : g_1(x_1) + \cdots + g_d(x_d) = g(x_1, \ldots, x_d) \\ &\qquad \text{and } g_1, \ldots, g_d \in \mathcal{H}_1\}.\end{aligned}$$

In other words, $\|f\|_{\ell_q(\mathcal{H}_d)}$ is the ℓ_q norm of the vector of RKHS norms of its component functions minimized over all of its additive representations. In particular, when $q = 2$, $\|\cdot\|_{\ell_2(\mathcal{H}_d)}$ can be viewed as a RKHS norm. More specifically, let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a Mercer kernel generating the RKHS $(\mathcal{H}_1, \|\cdot\|_{\mathcal{H}_1})$ and write

$$K_d((x_1, \ldots, x_d)^\top, (x_1', \ldots, x_d')^\top) = K(x_1, x_1') + \cdots + K(x_d, x_d').$$

It is not hard to see that K_d is the generating kernel of the RKHS $(\mathcal{H}_d, \|\cdot\|_{\ell_2(\mathcal{H}_d)})$. Another special case of the $\ell_q(\mathcal{H}_d)$ norm defined above is the case when $q \downarrow 0$.

$\|\cdot\|_{\ell_0(\mathcal{H}_d)}$ can be interpreted as the smallest number of additive components needed to express a function from $\mathcal{H}_d$.

When the dimension d is large, it is of particular interest to consider the case when f resides in an $\ell_q(\mathcal{H}_d)$ ball for $0 < q < 1$:

$$\mathcal{B}_R(\ell_q(\mathcal{H}_d)) = \{g \in \mathcal{H}_d : \|g\|^q_{\ell_q(\mathcal{H}_d)} \le R\}.$$

Write

$$\|g\|_{L_2(\Pi)} = \left(\int_{\mathcal{X}^d} g^2(x)\, d\Pi(x)\right)^{1/2}.$$

We are interested in the minimax optimal rate of convergence for estimating f in terms of the squared $\|\cdot\|_{L_2(\Pi)}$ norm. In particular, when the eigenvalues of the K decays polynomially, that is, its kth largest eigenvalue is of the order $k^{-2\alpha}$, our results imply that the minimax optimal rate for estimating $f \in \mathcal{B}_R(\ell_q(\mathcal{H}_d))$ is given by

$$\mathcal{R}(n, d) = \left(\frac{\log d}{n}\right)^{1-q/2} + n^{-2\alpha/r2\alpha+1}, \tag{1.3}$$

up to a constant scaling factor. The optimal rate of convergence given by (1.3) exhibits an interesting two-regime dichotomy as illustrated in Figure 1.

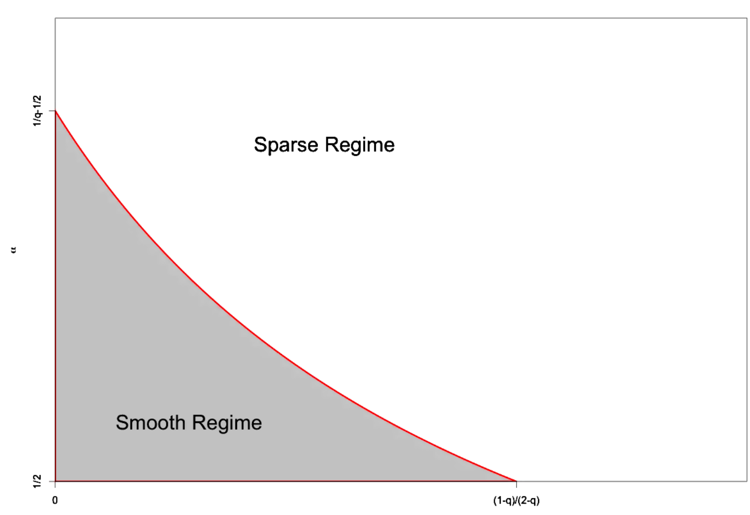

FIG. 1. *When the smoothness index* α *and dimensionality measured by* $\log\log d/\log n$ *falls in the* smooth region *in the figure above, the optimal rate is given by* $n^{-2\alpha/(2\alpha+1)}$ *which is determined solely by the smoothness index. On the other hand, if they fall into the* sparse regime, *then the optimal rate is given by* $(n^{-1}\log d)^{1-q/2}$ *which is determined entirely by the dimensionality.*

More specifically, when the component functions are not sufficiently smooth in the sense that

$$\alpha < \frac{1}{q} - \frac{1}{2},$$

the second term on the right-hand side of (1.3) is dominated by the first one if d is ultra-large:

$$d > \exp[n^{(2/(2-q))(1/(2\alpha+1)-q/2)}],$$

and hence the minimax optimal rate becomes

$$\mathcal{R}(n,d) \asymp \left(\frac{\log d}{n}\right)^{1-q/2}, \tag{1.4}$$

where we write for two positive sequences $a_{n,d}$ and $b_{n,d}$, $a_{n,d} \asymp b_{n,d}$ if $a_{n,d}/b_{n,d}$ is bounded away from both zero and infinity. The rate given by (1.4) happens to be the minimax optimal rate for estimating a d dimensional linear regression when assuming the vector of regression coefficient comes from a ℓ_q ball in $\mathbb{R}^d$; see, for example, Ye and Zhang (2010) or Raskutti, Wainwright and Yu (2011). On the other hand, when

$$d \le \exp[n^{(2/(2-q))(1/(2\alpha+1)-q/2)}],$$

the optimal rate is given by

$$\mathcal{R}(n,d) \asymp n^{-2\alpha/(2\alpha+1)}.$$

This rate coincides with the optimal rate for estimating f if we know in advance that it actually comes from a single component space $\mathcal{H}_1$, for example, $f_2 = \cdots = f_d = 0$, rather than the d-variate function space $\mathcal{H}_d$; see, for example, Stone (1980, 1982) and Tsybakov (2009). Similar phenomenon depending on the dimensionality d has also been observed earlier for high dimensional additive models under exact sparsity ($q = 0$); see, for example, Koltchinskii and Yuan (2010), Raskutti, Wainwright and Yu (2012) and Suzuki and Sugiyama (2013). Our results suggest that such phenomenon is more universal and applies in general to the approximate sparse case.

It is also worth pointing out that such a regime-switch in d vanishes when the component functions are sufficiently smooth in that

$$\alpha \ge \frac{1}{q} - \frac{1}{2},$$

a phenomenon absent in the case of exact sparsity ($q = 0$). In this situation, the second term on the right-hand side of (1.3) is always dominated by the first one and, therefore, the optimal rate is always

$$\mathcal{R}(n,d) \asymp \left(\frac{\log d}{n}\right)^{1-q/2}.$$

In other words, we pay no extra price, in terms of rates of convergence, for entertaining a generally nonparametric additive model (1.2) when compared with the much more restrictive linear models, regardless of the value of d.

Although we focus on additive models, our general framework is also closely related to multiple kernel learning or "aggregation" of kernel machines, a popular technique in machine learning to combine multiple kernels instead of using a single one in order to achieve improved prediction performance. These type of problems have been studied previously by Bousquet and Herrmann (2003), Crammer, Keshet and Singer (2003), Lanckriet et al. (2004), Micchelli and Pontil (2005), Srebro and Ben-David (2006), Bach (2008) and Suzuki and Sugiyama (2013) among others. It is expected that our results here could lead to further understanding of these problems as well.

The rest of the paper is organized as follows. We first review some basic concepts and properties of reproducing kernel Hilbert spaces in Section 2. Section 3 presents the main results. All proofs are relegated to Section 4.

2. Reproducing Kernel Hilbert Spaces. We begin with a brief review of some of the basic facts about RKHS, which we shall make repeated use later on. Interested readers are referred to Aronszajn (1950) and Wahba (1990) for further details. In particular, we shall focus on the jth component space, for example, the RKHS defined on the jth coordinate of $X \in \mathcal{X}^d$.

2.1. *Kernel and RKHS.* Recall that K is a symmetric positive semi-definite, square integrable function on $\mathcal{X} \times \mathcal{X}$. It can be uniquely identified with the Hilbert space $\mathcal{H}_1$ that is the completion of

$$\{K(x, \cdot) : x \in \mathcal{X}\}$$

under the inner product

$$\left\langle \sum_i c_i K(x_i, \cdot), \sum_j c'_j K(x'_j, \cdot) \right\rangle_K = \sum_{i,j} c_i c'_j K(x_i, x'_j).$$

In the rest of the section, we shall write $\mathcal{H}_1$ and $\mathcal{H}(K)$ interchangeably with the latter notion emphasizing the one-to-one correspondence between a kernel and a RKHS. Most, if not all, the commonly used kernels are bounded, which we shall assume in what follows. In fact, without loss of generality, we shall assume in the rest of the paper that $\sup_x K(x, x) = 1$. Note that, for any $h \in \mathcal{H}(K)$,

$$\|h\|_\infty := \sup_{x\in\mathcal{X}} |h(x)| = \sup_{x\in\mathcal{X}} |\langle h, K(x, \cdot)\rangle_K| \le \sup_x \|K(x, \cdot)\|_K \|h\|_K, \tag{2.1}$$

by the Cauchy–Schwarz inequality. Recall that

$$\|K(x, \cdot)\|_K^2 = \langle K(x, \cdot), K(x, \cdot)\rangle_K = K(x, x) \le 1.$$

Thus,

$$\|h\|_\infty \le \|h\|_K,$$

a convenient fact that we shall use repeatedly in the later analysis.

By the spectral theorem, K admits the following eigenvalue decomposition:

$$K(x, x') = \sum_{k\ge 1} \lambda_{jk}\varphi_{jk}(x)\varphi_{jk}(x'), \tag{2.2}$$

where $\lambda_{j1} \ge \lambda_{j2} \ge \cdots \ge 0$ are its eigenvalues and $\{\varphi_{jk} : k \ge 1\}$ are the corresponding eigenfunctions such that

$$\langle \varphi_{jk}, \varphi_{jk'}\rangle_{L_2(\Pi_j)} = \delta_{kk'}.$$

Here, Π_j is the jth marginal distribution of Π, and $\delta_{kk'}$ is the Kronecker delta. Note that the decomposition (2.2) depends on the jth marginal distribution Π_j through eigenfunctions φ_{jk}s. It is well known that the RKHS-norm of any $h \in \mathcal{H}(K)$ can be written as

$$\|h\|_K^2 = \sum_{k\ge 1} \frac{1}{\lambda_{jk}} \langle h, \varphi_{jk}\rangle^2_{L_2(\Pi_j)},$$

which means that the "smoothness" of functions in $\mathcal{H}(K)$ is determined by the rate of decay of the eigenvalues λ_{jk}, and the unit balls in the RKHS $\mathcal{H}(K)$ are ellipsoids in the space $L_2(\Pi_j)$ with "axes" $\sqrt{\lambda_{jk}}$. For example, it is well known that if Π_j is the Lebesgue measure on [0, 1], then $\lambda_{jk} \asymp k^{-2\alpha}$ for $\mathcal{W}_2^\alpha$.

2.2. *Complexity of RKHS.* How well we can recover a function from a particular RKHS is fundamentally related to the capacity of the unit ball in $\mathcal{H}(K)$:

$$\mathcal{B}_1(\mathcal{H}(K)) := \{h \in \mathcal{H}(K) : \|h\|_K \le 1\}.$$

See, for example, Yang and Barron (1999). In particular, the capacity of $\mathcal{B}_1(\mathcal{H}(K))$ can be measured by its covering number $\mathcal{N}(\mathcal{B}_1(\mathcal{H}(K)), \delta, \|\cdot\|_\infty)$ where $\|\cdot\|_\infty$ is defined in (2.1). Recall that for $\delta > 0$ and a set $\mathcal{F}$ of continuous functions on a metric space $\mathcal{X}$, the covering number $\mathcal{N}(\mathcal{F}, \delta, \|\cdot\|_\infty)$ with respect to the $\|\cdot\|_\infty$ metric is defined as the smallest integer m such that

$$\mathcal{F} = \bigcup_{i=1}^m \{f \in \mathcal{F} : \|f - f^{(i)}\|_\infty \le \delta\}$$

for some $\{f^{(i)}\}_{i=1}^m \subset \mathcal{F}$. In particular, if $\lambda_{jk} = O(k^{-2\alpha})$ and $\sup_{j,k} \|\varphi_{jk}\|_\infty < \infty$, then

$$\log \mathcal{N}(\mathcal{B}_1(\mathcal{H}(K)), \delta, \|\cdot\|_\infty) \le c\delta^{-1/\alpha} \qquad \forall \delta > 0, \tag{2.3}$$

for some constant $c > 0$. This holds, for example, for Sobolev spaces of order α.

For our purposes, we are also interested in certain data-dependent estimates of the complexity of a function class, namely, Rademacher and Gaussian complexities; see, for example, Bartlett and Mendelson (2002). Write

$$R_{jn}(u) := \sup_{h \in \mathcal{B}_1(\mathcal{H}(K)): \|h\|_{L_2(\Pi_j)} \le u} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i h(x_{ij}) \right|, \tag{2.4}$$

where σ_is are i.i.d. Rademacher variables, that is, $\mathbb{P}(\sigma_i = 1) = \mathbb{P}(\sigma_i = -1) = 1/2$. The following bound of R_{jn} will become useful for our later analysis.

LEMMA 2.1. *Assume that $\lambda_{jk} \le c_1 k^{-2\alpha}$ and $\sup_{j,k} \|\varphi_{jk}\|_{L_\infty} < c_2$ for some constants $c_1, c_2 > 0$. Then there exists a constant $c > 0$ depending on α, c_1 and c_2 only such that for any $\beta > 0$, with probability at least $1 - d^{-\beta}$,*

$$R_{jn}(u) \le cn^{-1/2}\left(u^{1-1/(2\alpha)} + u\sqrt{\beta \log d} + \frac{\beta \log d}{\sqrt{n}} + e^{-d}\right)$$

uniformly for all $u \in [0, 1]$.

Another quantity of interests to us is the "empirical" Gaussian complexity of the unit ball in $\mathcal{H}(K)$:

$$\widehat{Z}_{jn}(u) := \sup_{h \in \mathcal{B}_1(\mathcal{H}(K)): \|h\|_{L_2(\Pi_{jn})} \le u} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i h(x_{ij}) \right| \tag{2.5}$$

where Π_{jn} is the jth marginal of the empirical distribution Π_n. Similar to Lemma 2.1, we have the following bound for $\widehat{Z}_{jn}$.

LEMMA 2.2. *Assume that $\lambda_{jk} \le c_1 k^{-2\alpha}$ and $\sup_{j,k} \|\varphi_{jk}\|_{L_\infty} < c_2$ for some constants $c_1, c_2 > 0$. Then there exists a constant $c > 0$ depending on α, c_1 and c_2 only such that for any $\beta > 0$, with probability at least $1 - d^{-\beta}$,*

$$\widehat{Z}_{jn}(u) \le cn^{-1/2}\left(u^{1-1/(2\alpha)} + u\sqrt{\beta \log d} + e^{-d}\right)$$

uniformly for all $u \in [0, 1]$.

Both Lemmas 2.1 and 2.2 appears to be standard and follow from a standard peeling argument [see, e.g., van de Geer (2000)]. Although they are useful for our analysis, we are unable to find these specific results in the literature. For completeness, we present their proofs in Section 4.2.

3. Main results. In what follows, we shall assume that there exists a constant $\eta_q > 1$ such that

$$\eta_q^{-1} \|g\|_{L_2(\Pi)}^2 \le \sum_{j=1}^{d} \|g_j\|_{L_2(\Pi_j)}^2 \le \eta_q \|g\|_{L_2(\Pi)}^2 \tag{3.1}$$

for any $g \in \mathcal{B}_R(\ell_q(\mathcal{H}_d))$, where

$$g(x_1, \ldots, x_d) = g_1(x_1) + \cdots + g_d(x_d)$$

and

$$\|g\|_{\ell_q(\mathcal{H}_d)}^q = \sum_{j=1}^d \|g_j\|_{\mathcal{H}_1}^q.$$

Condition (3.1) is a nonparametric version of the restricted eigenvalue conditions commonly used in analyzing sparse estimation in high dimensional linear regression; see, for example, Bickel, Ritov and Tsybakov (2009). It is worth noting that different from the usual restricted eigenvalue conditions in linear regression, Condition (3.1) is on the *distribution* of X rather than the design matrix, or observations $X_1, \ldots, X_n$. The condition is satisfied in particular when Π is a product measure.

To fix ideas, in the rest of the paper, we shall also assume that there exist a constant $c_\lambda > 1$ and a nonincreasing sequence of nonnegative numbers $\lambda_1 \geq \lambda_2 \geq \cdots$ such that

$$c_\lambda^{-1}\lambda_k \leq \lambda_{jk} \leq c_\lambda \lambda_k, \tag{3.2}$$

for all $j = 1, 2, \ldots, d$ and $k \geq 1$. In addition, similar to the treatment of high dimensional linear models [see, e.g., Raskutti, Wainwright and Yu (2011)], we shall assume in the rest of the paper that $c_0 n^{q/2} \leq d \leq e^n$ for some universal constant $c_0 > 0$ to ensure nontrivial probabilistic bounds. This, in particular, is true in high dimensional settings where $n < d < e^n$.

We are now in position to present the main results. We first state a minimax lower bound.

THEOREM 3.1. *Assume that $\lambda_k = k^{-2\alpha}$ for some $\alpha > 1/2$. Under the regression model (1.1) where $f \in \mathcal{B}_R(\ell_q(\mathcal{H}_d))$ and the covariate X follows a distribution Π such that (3.1) and (3.2) hold, and the eigenfunctions $\{\varphi_{jk} : j = 1, \ldots, d, k \geq 1\}$ are uniformly bounded, there exists a constant $c > 0$ depending on σ^2, α, R, c_λ and η_q only such that*

$$\lim_{n\to\infty} \inf_{\tilde f} \sup_{f \in \mathcal{B}_R(\ell_q(\mathcal{H}_d))} \mathbb{P}\left\{\|\tilde f - f\|_{L_2(\Pi)}^2 \geq c\left[\left(\frac{\log d}{n}\right)^{1-q/2} + n^{-2\alpha/(2\alpha+1)}\right]\right\} > 0.$$

The lower bound is established via Fano's lemma; see, for example, Cover and Thomas (1991). We relegate its proof to Section 4. Next, we show that the rates given in the lower bound in the previous theorem is attainable. In particular, we consider the least squares estimator:

$$\widehat{f} = \operatorname*{argmin}_{g \in \mathcal{B}_R(\ell_q(\mathcal{H}_d))} \left\{\frac{1}{n}\sum_{i=1}^n [Y_i - g(X_i)]^2\right\}. \tag{3.3}$$

The next result shows that $\widehat{f}$ is indeed minimax rate optimal.

THEOREM 3.2. *Assume that* $\lambda_k = k^{-2\alpha}$ *for some* $\alpha > 1/2$. *Under the regression model* (1.1) *where* $f \in \mathcal{B}_R(\ell_q(\mathcal{H}_d))$ *and the covariate* X *follows a distribution* Π *such that* (3.1) *and* (3.2) *hold, and the eigenfunctions* $\{\varphi_{jk} : j = 1, \dots, d, k \geq 1\}$ *are uniformly bounded, there exists a constant* $c > 0$ *depending on* σ^2, α, R, c_λ *and* η_q *only such that for any* $\beta > 0$ *with probability at least* $1 - d^{-\beta}$,

$$\|\widehat{f} - f\|^2_{L_2(\Pi)} \leq c(\beta + 1)\left[\left(\frac{\log d}{n}\right)^{1-q/2} + n^{-2\alpha/(2\alpha+1)}\right] \tag{3.4}$$

and

$$\|\widehat{f} - f\|^2_{L_2(\Pi_n)} \leq c(\beta + 1)\left[\left(\frac{\log d}{n}\right)^{1-q/2} + n^{-2\alpha/(2\alpha+1)}\right], \tag{3.5}$$

where $\widehat{f}$ *is the least squares estimator defined by* (3.3).

The proof of Theorem 3.2 is also presented in Section 4. It relies on several basic facts of the empirical processes theory such as symmetrization inequalities and contraction inequalities for Rademacher processes that can be found in the books of Ledoux and Talagrand (1991) and van der Vaart and Wellner (1996). We also use Talagrand's concentration inequality for empirical processes; see, for example, Talagrand (1996) and Bousquet (2002).

Theorems 3.1 and 3.2 together immediately imply that the minimax optimal rate for estimating $f \in \mathcal{B}_R(\ell_q(\mathcal{H}_d))$ is

$$\|\widehat{f} - f\|^2_{L_2(\Pi)} \asymp \left(\frac{\log d}{n}\right)^{1-q/2} + n^{-2\alpha/(2\alpha+1)}.$$

This result connects with two strands of literature—estimating high dimensional linear regression assuming that the coefficient vector belongs to an ℓ_q ball, and estimating a high dimensional additive model assuming that the underlying function comes from a $\ell_0(\mathcal{H}_d)$ ball. In the case of linear regression, it is known that the ℓ_1 penalty or the Lasso [Tibshirani (1996)] leads to rate optimal estimators under suitable regularity conditions; see, for example, Ye and Zhang (2010). A similar phenomenon has also been observed for the high dimensional additive models where it is shown that a mixed ℓ_1 norm penalty of the form

$$a_n^2 \sum_{j=1}^{d} \|g_j\|_{\mathcal{H}_1} + a_n \sum_{j=1}^{d} \|g_j\|_{L_2(\Pi_{jn})} \tag{3.6}$$

can lead to rate optimal estimators with appropriate choices of the tuning parameter $a_n > 0$; see, for example, Koltchinskii and Yuan (2010) and Raskutti, Wainwright and Yu (2012). The use of a mixed ℓ_1 penalty of the form (3.6) highlights the difference between linear models and additive models. When dealing with nonparametric component functions, we need to penalize both the RKHS norm and L_2

norm; the former ensures smoothness of the estimate whereas the latter is needed for thresholding redundant components, and hence inducing sparsity.

A natural question is whether or not a similar strategy will lead to minimax rate optimal estimators under an $\ell_q(\mathcal{H}_d)$ ball for general $0 < q \le 1$. Somewhat surprisingly, the answer appears to be negative in general, and we give here a heuristic argument why. The challenge occurs in the smooth regime where

$$\alpha < \frac{1}{q} - \frac{1}{2} \quad \text{and} \quad d \le \exp\bigl[n^{(2/2-q)(1/(2\alpha+1)-q/2)}\bigr].$$

Recall that the corresponding minimax optimal rate of convergence in the smooth regime is given by

$$n^{-2\alpha/(2\alpha+1)}.$$

As pointed out before, this is the best possible rate of convergence even if there is only one nonzero component. And to achieve this rate, we need to choose

$$a_n \gtrsim n^{-\alpha/(2\alpha+1)}, \tag{3.7}$$

because, if a_n is smaller, then in the particular case of one nonzero component, the minimax optimal rate cannot be attained; see, for example, Tsybakov (2009) or Koltchinskii and Yuan (2010). Now for a general f from the unit $\ell_q(\mathcal{H}_d)$ ball, we will need a diverging number of nonzero components to approximate it. More precisely, as we shall show in the proofs, we may need estimate up to

$$\left\lceil \left(\frac{n}{\log d}\right)^{q/2} \right\rceil$$

nonzero components to balance the approximation error and estimation error due to estimating the nonzero component functions. If we choose a_n to be of the order given by (3.7), then each component can only be estimated with squared L_2 error of the order of

$$a_n^2 \gtrsim n^{-2\alpha/(2\alpha+1)},$$

leading to an overall rate of convergence no smaller than, up to a multiplicative constant,

$$\left(\frac{n}{\log d}\right)^{q/2} n^{-2\alpha/(2\alpha+1)},$$

at least under the assumption that Π is a product measure. This rate is obviously suboptimal. As a result, in the smooth regime, no matter what value a_n is, we cannot attain the minimax optimal rate of convergence through a mixed ℓ_1 penalty of the form (3.6).

As a working model, we assume that the eigenvalues decay at the same polynomial rate across components, and the eigenfunctions φ_{jk}s are bounded, which hold true for Sobolev kernels among other commonly used kernels. It is of interest to consider more general settings, for example, when the eigenfunctions are unbounded, or if the eigenvalues decay at different rates, or if the eigenvalues for some components decay even exponentially. It is conceivable that our analysis could be extended to deal with more general situations. But as in the single kernel case, treating these more general cases is typically more tedious and technical, and we shall leave them for future studies.

4. Proofs.

4.1. *Proof of main results.* We now prove the main results Theorems 3.1 and 3.2. For brevity, we shall also assume that $\sigma^2 = 1$ and $R = 1$ in the proofs. The more general case follows an identical arguments with different constants.

4.1.1. *Lower bounds.* We establish the lower bound via Fano's lemma. To this end, we need to construct a set of functions

$$\mathcal{G} := \{g^1, \dots, g^M\} \subset \mathcal{B}_1(\ell_q(\mathcal{H}_d))$$

that are sufficiently apart from each other. Let N be a natural number whose value will be specified later. For a matrix $A \in \{-1, 0, 1\}^{d\times N}$, denote by s_A the number of its nonzero rows, that is,

$$s_A = \operatorname{card}\{i : A_{i\cdot} \neq \mathbf{0}\},$$

where $A_{i\cdot}$ is the ith row vector of A. Write

$$g_A(x_1, \dots, x_d) = N^{-1/2} s_A^{-1/q} \sum_{j=1}^{d} \sum_{k=1}^{N} a_{jk} \lambda_{j,N+k}^{1/2} \varphi_{j,N+k}(x_j).$$

It is clear that

$$\begin{aligned}\|g_A\|_{\ell_q(\mathcal{H}_d)}^q &\le N^{-q/2} s_A^{-1} \sum_{j=1}^{d} \left\| \sum_{k=1}^{N} a_{jk} \lambda_{j,N+k}^{1/2} \varphi_{j,N+k}(x_j) \right\|_{\mathcal{H}_1}^q \\ &= s_A^{-1} \sum_{j=1}^{d} \left(N^{-1} \sum_{k=1}^{N} a_{jk}^2 \right)^{q/2}.\end{aligned}$$

Because $a_{jk}^2 \in \{0, 1\}$, this can be further bounded by

$$\|g_A\|_{\ell_q(\mathcal{H}_d)}^q \le s_A^{-1} \sum_{j=1}^{d} \mathbb{I}(A_{i\cdot} \neq 0) = 1,$$

which implies that $g_A \in \mathcal{B}_1(\ell_q(\mathcal{H}_d))$.

We now describe how to generate the set $\mathcal{G}$. In particular, we consider functions of the form g_A with $A \in \{\pm 1, 0\}^{d\times N}$ as described before. We first choose s rows of A to be nonzero, and set the rest of the rows of A to be zero. The value of s will become clear later. To this end, we appeal to the Vershamov–Gilbert lemma which states that we can find a set $\{\theta_1, \dots, \theta_{M_1}\} \subset \{0,1\}^d$ such that:

(a) $\|\theta_k\|_{\ell_1} = s$ for $1 \le k \le M_1$;
(b) for any $k \ne k'$, $\|\theta_k - \theta_{k'}\|_{\ell_1} \ge s/2$;
(c) $\log M_1 \ge \frac{1}{4} s \log(d/s)$.

See, for example, Massart (2007). For a given θ, we set zero the rows of A if the corresponding coordinate of θ is zero. In the next step, we fill in the remaining rows of A with ± 1. Again, by the Vershamov–Gilbert lemma, there exists a set $\{\Gamma_1, \dots, \Gamma_{M_2}\} \in \{\pm 1\}^{s\times N}$ such that:

(a′) for any $k \ne k'$, $\|\Gamma_k - \Gamma_{k'}\|_{\mathrm{F}}^2 \ge Ns/2$;
(b′) $\log M_2 \ge Ns/8$.

For a given Γ, we shall fill in the nonzero rows of A by Γ, leading to a collection

$$\mathcal{G} = \{g_{A(\theta_j, \Gamma_k)} : 1 \le j \le M_1, 1 \le k \le M_2\},$$

where $A(\theta, \Gamma)$ is a $d \times N$ matrix whose ith row is zero if the ith entry of θ is zero, and the collection of the nonzero rows of A is given by Γ. In what follows, for brevity, we shall write

$$\mathcal{G} = \{g_{A_k} : 1 \le k \le M\},$$

where $M = M_1 M_2$ and

$$\mathcal{A} = \{A_k : 1 \le k \le M\}$$

is the collection of $d \times N$ matrices of the form $A(\theta_j, \Gamma_k)$. By (c) and (b′),

$$\log M \ge \frac{1}{4} s \log(d/s) + \frac{1}{8} Ns.$$

Note that, for any two matrices $A, B \in \{-1, 0, 1\}^{d\times N}$ such that $s_A = s_B =: s$, we have

$$\begin{aligned}
\|g_A - g_B\|_{L_2(\Pi)}^2 &= N^{-1} s^{-2/q} \int_{\mathcal{X}^d} \left(\sum_{j=1}^d \sum_{k=1}^N (a_{jk} - b_{jk}) \lambda_{j,N+k}^{1/2} \varphi_{j,N+k}(x_j) \right)^2 \\
&\quad \times d\Pi\big((x_1, \dots, x_d)^\top\big) \\
&\ge \eta_q^{-1} N^{-1} s^{-2/q} \sum_{j=1}^d \left\| \sum_{k=1}^N (a_{jk} - b_{jk}) \lambda_{j,N+k}^{1/2} \varphi_{j,N+k} \right\|_{L_2(\Pi_j)}^2 \\
&= \eta_q^{-1} N^{-1} s^{-2/q} \sum_{j=1}^d \sum_{k=1}^N \lambda_{j,N+k} (a_{jk} - b_{jk})^2,
\end{aligned}$$

where the inequality follows from (3.1). By (3.2), this can be further lower-bounded by

$$\begin{aligned}\|g_A - g_B\|^2_{L_2(\Pi)} &\geq c_\lambda^{-1}\eta_q^{-1}N^{-1}s^{-2/q}\sum_{j=1}^{d}\sum_{k=1}^{N}\lambda_{N+k}(a_{jk}-b_{jk})^2\\ &\geq c_\lambda^{-1}\eta_q^{-1}N^{-1}s^{-2/q}\lambda_{2N}\sum_{j=1}^{d}\sum_{k=1}^{N}(a_{jk}-b_{jk})^2\\ &= c_\lambda^{-1}\eta_q^{-1}2^{-2\alpha}N^{-1-2\alpha}s^{-2/q}\|A-B\|^2_{\mathrm{F}}.\end{aligned}$$

By construction, for any $A \neq A' \in \mathcal{A}$,

$$\|A - A'\|^2_{\mathrm{F}} \geq Ns/2,$$

and hence,

$$\|g_A - g_{A'}\|^2_{L_2(\Pi)} \geq c_\lambda^{-1}\eta_q^{-1}2^{-1-2\alpha}N^{-2\alpha}s^{1-2/q}.$$

On the other hand, for any $A \in \mathcal{A}$,

$$\begin{aligned}\|g_A\|^2_{L_2(\Pi)} &= N^{-1}s^{-2/q}\int_{\mathcal{X}^d}\left(\sum_{j=1}^{d}\sum_{k=1}^{N}a_{jk}\lambda_{j,N+k}^{1/2}\varphi_{j,N+k}(x_j)\right)^2 d\Pi\big((x_1,\dots,x_d)^\top\big)\\ &\leq \eta_q N^{-1}s^{-2/q}\sum_{j=1}^{d}\left\|\sum_{k=1}^{N}a_{jk}\lambda_{j,N+k}^{1/2}\varphi_{j,N+k}\right\|^2_{L_2(\Pi_j)}\\ &= \eta_q N^{-1}s^{-2/q}\sum_{j=1}^{d}\sum_{k=1}^{N}\lambda_{j,N+k}a_{jk}^2\\ &\leq c_\lambda\eta_q N^{-1}s^{-2/q}\sum_{j=1}^{d}\sum_{k=1}^{N}\lambda_{N+k}a_{jk}^2\\ &\leq c_\lambda\eta_q N^{-1}s^{-2/q}\lambda_N\sum_{j=1}^{d}\sum_{k=1}^{N}a_{jk}^2\\ &= c_\lambda\eta_q N^{-2\alpha}s^{1-2/q}.\end{aligned}$$

Following a standard argument, the lower bound can be reduced to the error probability in a multi-way hypothesis test; see, for example, Tsybakov (2009). More specifically, let Θ be a random variable uniformly distributed on $\{1,\dots,M\}$. Then it can be deduced that

$$\inf_{\tilde{f}}\sup_{f\in\mathcal{B}_1(\ell_q(\mathcal{H}_d))}\mathbb{P}\left\{\|\tilde{f}-f\|^2_{L_2(\Pi)} \geq \frac{1}{4}\min_{A\neq A'\in\mathcal{A}}\|g_A - g_{A'}\|^2_{L_2(\Pi)}\right\} \geq \inf_{\widehat{\Theta}}\mathbb{P}\{\widehat{\Theta}\neq\Theta\},$$

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

$$\mathbb{P}\{\widehat{\Theta} \neq \Theta | X_1, \dots, X_n\} \geq 1 - \frac{1}{\log M}\big[\mathbb{I}_{X_1,\dots,X_n}(Y_1, \dots, Y_n; \Theta) + \log 2\big], \tag{4.1}$$

where $\mathbb{I}_{X_1,\dots,X_n}(Y_1, \dots, Y_n; \Theta)$ is the mutual information between Θ and $Y_1, \dots, Y_n$ with $X_1, \dots, X_n$ being held fixed. It is not hard to derive

$$\begin{aligned} &\mathbb{E}_{X_1,\dots,X_n}\big[\mathbb{I}_{X_1,\dots,X_n}(Y_1, \dots, Y_n; \Theta)\big] \\ &\quad \leq \binom{M}{2}^{-1} \sum_{A \neq A' \in \mathcal{A}} \mathbb{E}_{X_1,\dots,X_n} \mathcal{K}(\mathbf{P}_{g_A} \| \mathbf{P}_{g_{A'}}) \\ &\quad \leq \frac{n}{2}\binom{M}{2}^{-1} \sum_{A \neq A' \in \mathcal{A}} \mathbb{E}_{X_1,\dots,X_n} \|g_A - g_{A'}\|^2_{L_2(\Pi_n)}, \end{aligned}$$

where $\mathcal{K}(\cdot\|\cdot)$ denote the Kullback–Leibler distance, $\mathbf{P}_g$ stands for conditional distribution of $\{Y_i : 1 \leq i \leq n\}$ given $\{X_i : 1 \leq i \leq n\}$ and the true regression function in (1.1) is given by $f = g$, and for any $g : \mathcal{X}^d \to \mathbb{R}$,

$$\|g\|^2_{L_2(\Pi_n)} = \frac{1}{n}\sum_{i=1}^{n}\big[g(X_i)\big]^2.$$

Thus,

$$\begin{aligned} \mathbb{E}_{X_1,\dots,X_n}\big[\mathbb{I}_{X_1,\dots,X_n}(Y_1, \dots, Y_n; \Theta)\big] &\leq \frac{n}{2}\binom{M}{2}^{-1} \sum_{A \neq A' \in \mathcal{A}} \|g_A - g_{A'}\|^2_{L_2(\Pi)} \\ &\leq \frac{n}{2} \max_{A \neq A' \in \mathcal{A}} \|g_A - g_{A'}\|^2_{L_2(\Pi)} \\ &\leq 2n \max_{A \in \mathcal{A}} \|g_A\|^2_{L_2(\Pi)} \\ &\leq 2c_\lambda \eta_q n N^{-2\alpha} s^{1-2/q}. \end{aligned}$$

Now, from (4.1), we get

$$\begin{aligned} &\inf_{\tilde{f}} \sup_{f \in \mathcal{B}_1(\ell_q(\mathcal{H}_d))} \mathbb{P}\big\{\|\tilde{f} - f\|_2^2 \geq c_\lambda^{-1}\eta_q^{-1}2^{-2-2\alpha}N^{-2\alpha}s^{1-2/q}\big\} \\ &\quad \geq \inf_{\widehat{\Theta}} \mathbb{P}\{\widehat{\Theta} \neq \Theta\} \\ &\quad \geq 1 - \frac{\mathbb{E}_{X_1,\dots,X_n}[\mathbb{I}_{X_1,\dots,X_n}(Y_1, \dots, Y_n; \Theta)] + \log 2}{\log M} \\ &\quad \geq 1 - \frac{2c_\lambda \eta_q n N^{-2\alpha} s^{1-2/q} + \log 2}{(1/4)s\log(d/s) + (1/8)Ns}. \end{aligned}$$

Taking $N = 1$ and

$$s = C_1\left(\frac{n}{\log d}\right)^{q/2}$$

for a sufficiently small constant $C_1 > 0$ yields

$$\inf_{\tilde{f}} \sup_{f\in\mathcal{B}_1(\ell_q(\mathcal{H}_d))} \mathbb{P}\left\{\|\tilde{f} - f\|_2^2 \geq C_2\left(\frac{\log d}{n}\right)^{1-q/2}\right\} \geq 3/4, \tag{4.2}$$

for some constant $C_2 > 0$ depending on α, η_q and c_λ only. On the other hand, if $\alpha \leq 1/q - 1/2$, taking

$$s = 1 \quad \text{and} \quad N = C_1 n^{1/(2\alpha+1)}$$

for a sufficiently small constant $C_1 > 0$ yields

$$\inf_{\tilde{f}} \sup_{f\in\mathcal{B}_1(\ell_q(\mathcal{H}_d))} \mathbb{P}\{\|\tilde{f} - f\|_2^2 \geq C_2 n^{-2\alpha/(2\alpha+1)}\} \geq 3/4. \tag{4.3}$$

Combining (4.2) and (4.3), we have

$$\inf_{\tilde{f}} \sup_{f\in\mathcal{B}_1(\ell_q(\mathcal{H}_d))} \mathbb{P}\left\{\|\tilde{f} - f\|_2^2 \geq C_2\left[\left(\frac{\log d}{n}\right)^{1-q/2} + n^{-2\alpha/(2\alpha+1)}\right]\right\} \geq 3/4,$$

which completes the proof.

4.1.2. *Upper bounds*. We now prove the upper bounds given in Theorem 3.2. By definition,

$$\frac{1}{n}\sum_{i=1}^n [Y_i - \widehat{f}(X_i)]^2 \leq \frac{1}{n}\sum_{i=1}^n [Y_i - f(X_i)]^2,$$

which immediately implies that

$$\frac{1}{n}\sum_{i=1}^n [\widehat{f}(X_i) - f(X_i)]^2 \leq \frac{2}{n}\sum_{i=1}^n \varepsilon_i[\widehat{f}(X_i) - f(X_i)]. \tag{4.4}$$

Write $\Delta_j = \widehat{f}_j - f_j$ and $\Delta = \widehat{f} - f$. It is clear that $\Delta = \sum_{j=1}^d \Delta_j$.

Our main strategy is to derive upper and lower bounds for the right- and left-hand side of (4.4), respectively, and then put them together to derive (3.4).

Step 1. *Bounding the right-hand side of* (4.4). Observe that

$$\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i \Delta_j(x_{ij})\right| \leq \|\Delta_j\|_{\mathcal{H}_1} \widehat{Z}_{jn}\left(\frac{\|\Delta_j\|_{L_2(\Pi_{jn})}}{\|\Delta_j\|_{\mathcal{H}_1}}\right),$$

where $\widehat{Z}_{jn}$ is defined by (2.5). By Lemma 2.2, this can be further bounded by

$$C_1 n^{-1/2}\left(\|\Delta_j\|_{L_2(\Pi_{jn})}^{1-1/(2\alpha)} \|\Delta_j\|_{\mathcal{H}_1}^{1/(2\alpha)} + \|\Delta_j\|_{L_2(\Pi_{jn})}\sqrt{(\beta+1)\log d} + e^{-d}\|\Delta_j\|_{\mathcal{H}_1}\right)$$

for some constant $C_1 > 0$, with probability at least $1 - d^{-(\beta+1)}$. By union bound, with probability $1 - d^{-\beta}$,

$$\begin{aligned}
&\frac{2}{n}\sum_{i=1}^{n}\varepsilon_i\big[\widehat{f}(X_i) - f(X_i)\big] \\
&\quad \le 2\sum_{j=1}^{d}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\Delta_j(x_{ij})\right| \\
&\quad \le 2C_1 n^{-1/2}\sum_{j=1}^{d}\|\Delta_j\|_{L_2(\Pi_{jn})}^{1-1/(2\alpha)}\|\Delta_j\|_{\mathcal{H}_1}^{1/(2\alpha)} \qquad (4.5)\\
&\qquad + 2C_1 n^{-1/2}\sqrt{(\beta+1)\log d}\sum_{j=1}^{d}\|\Delta_j\|_{L_2(\Pi_{jn})} \\
&\qquad + 2C_1 n^{-1/2}e^{-d}\sum_{j=1}^{d}\|\Delta_j\|_{\mathcal{H}_1}.
\end{aligned}$$

We denote by $\mathcal{E}_1$ the event that the above inequality holds. We now bound the three terms on the rightmost side separately.

We first derive a bound for

$$n^{-1/2}\sum_{j=1}^{d}\|\Delta_j\|_{\mathcal{H}_1}^{1/(2\alpha)}\|\Delta_j\|_{L_2(\Pi_{jn})}^{1-1/(2\alpha)}.$$

We treat the cases of $2/(2\alpha+1) \ge q$ and $2/(2\alpha+1) < q$ separately.

Case 1: $2/(2\alpha+1) \ge q$. By Young's inequality, for a constant $\zeta > 1$ whose value will be specified later,

$$\begin{aligned}
&n^{-1/2}\|\Delta_j\|_{\mathcal{H}_1}^{1/(2\alpha)}\|\Delta_j\|_{L_2(\Pi_{jn})}^{1-1/(2\alpha)} \\
&\quad \le \zeta^{-4\alpha/(2\alpha-1)}\|\Delta_j\|_{L_2(\Pi_{jn})}^2 + \zeta^{4\alpha/(2\alpha+1)}n^{-2\alpha/(2\alpha+1)}\|\Delta_j\|_{\mathcal{H}_1}^{2/(2\alpha+1)}.
\end{aligned}$$

Note that for any $q \le q' \le 2$,

$$\begin{aligned}
\sum_{j=1}^{d}\|\Delta_j\|_{\mathcal{H}_1}^{q'} &\le 2\left(\sum_{j=1}^{d}\|\widehat{f}_j\|_{\mathcal{H}_1}^{q'} + \sum_{j=1}^{d}\|f_j\|_{\mathcal{H}_1}^{q'}\right) \\
&\le 2\left(\sum_{j=1}^{d}\|\widehat{f}_j\|_{\mathcal{H}_1}^{q} + \sum_{j=1}^{d}\|f_j\|_{\mathcal{H}_1}^{q}\right) \\
&\le 4.
\end{aligned}$$

In particular, we get

$$\sum_{j=1}^{d} \|\Delta_j\|_{\mathcal{H}_1}^{2/(2\alpha+1)} \le 4.$$

Hence,

$$\begin{aligned}&\sum_{j=1}^{d} n^{-1/2}\|\Delta_j\|_{\mathcal{H}_1}^{1/(2\alpha)}\|\Delta_j\|_{L_2(\Pi_{jn})}^{1-1/(2\alpha)}\\ &\qquad\le \zeta^{-4\alpha/(2\alpha-1)}\|\Delta_j\|_{L_2(\Pi_{jn})}^2 + 4\zeta^{4\alpha/(2\alpha+1)}n^{-2\alpha/(2\alpha+1)}.\end{aligned} \tag{4.6}$$

Case 2: $2/(2\alpha+1) < q$. Write

$$\begin{aligned}&n^{-1/2}\sum_{j=1}^{d} \|\Delta_j\|_{\mathcal{H}_1}^{1/(2\alpha)}\|\Delta_j\|_{L_2(\Pi_{jn})}^{1-1/(2\alpha)}\\ &\quad= n^{-1/2}\sum_{j:\|\Delta_j\|_{\mathcal{H}_1}>n^{-1/2}} \|\Delta_j\|_{\mathcal{H}_1}^{1/(2\alpha)}\|\Delta_j\|_{L_2(\Pi_{jn})}^{1-1/(2\alpha)}\\ &\qquad+ n^{-1/2}\sum_{j:\|\Delta_j\|_{\mathcal{H}_1}\le n^{-1/2}} \|\Delta_j\|_{\mathcal{H}_1}^{1/(2\alpha)}\|\Delta_j\|_{L_2(\Pi_{jn})}^{1-1/(2\alpha)}.\end{aligned}$$

For the first term on the right-hand side, by a similar argument as before, we have

$$\begin{aligned}&n^{-1/2}\sum_{j:\|\Delta_j\|_{\mathcal{H}_1}>n^{-1/2}} \|\Delta_j\|_{\mathcal{H}_1}^{1/2\alpha}\|\Delta_j\|_{L_2(\Pi_{jn})}^{1-1/(2\alpha)}\\ &\quad\le \zeta^{-4\alpha/(2\alpha-1)}\sum_{j:\|\Delta_j\|_{\mathcal{H}_1}>n^{-1/2}} \|\Delta_j\|_{L_2(\Pi_{jn})}^2\\ &\qquad+ \zeta^{4\alpha/(2\alpha+1)}n^{-2\alpha/(2\alpha+1)}\sum_{j:\|\Delta_j\|_{\mathcal{H}_1}>n^{-1/2}} \|\Delta_j\|_{\mathcal{H}_1}^{2/(2\alpha+1)}\\ &\quad\le \zeta^{-4\alpha/(2\alpha-1)}\sum_{j:\|\Delta_j\|_{\mathcal{H}_1}>n^{-1/2}} \|\Delta_j\|_{L_2(\Pi_{jn})}^2\\ &\qquad+ \zeta^{4\alpha/(2\alpha+1)}n^{-(1-q/2)}\sum_{j:\|\Delta_j\|_{\mathcal{H}_1}>n^{-1/2}} \|\Delta_j\|_{\mathcal{H}_1}^{q}\\ &\quad\le \zeta^{-4\alpha/(2\alpha-1)}\sum_{j:\|\Delta_j\|_{\mathcal{H}_1}>n^{-1/2}} \|\Delta_j\|_{L_2(\Pi_{jn})}^2 + 4\zeta^{4\alpha/(2\alpha+1)}n^{-(1-q/2)},\end{aligned}$$

where in the last inequality we used the fact that

$$\sum_{j:\|\Delta_j\|_{\mathcal{H}_1}>n^{-1/2}} \|\Delta_j\|_{\mathcal{H}_1}^{q} \le \sum_{j=1}^{d}\|\Delta_j\|_{\mathcal{H}_1}^{q} \le 2\sum_{j=1}^{d}\big(\|\widehat{f}_j\|_{\mathcal{H}_1}^{q} + \|f_j\|_{\mathcal{H}_1}^{q}\big) \le 4.$$

On the other hand, because

$$\|\Delta_j\|_{L_2(\Pi_{jn})} \le \|\Delta_j\|_{L_\infty} \le \|\Delta_j\|_{\mathcal{H}_1},$$

we get

$$\begin{aligned}
n^{-1/2} & \sum_{j:\|\Delta_j\|_{\mathcal{H}_1}\le n^{-1/2}} \|\Delta_j\|_{\mathcal{H}_1}^{1/2\alpha} \|\Delta_j\|_{L_2(\Pi_{jn})}^{1-1/(2\alpha)} \\
&\le n^{-1/2} \sum_{j:\|\Delta_j\|_{\mathcal{H}_1}\le n^{-1/2}} \|\Delta_j\|_{\mathcal{H}_1} \\
&\le n^{-(1-q/2)} \sum_{j:\|\Delta_j\|_{\mathcal{H}_1}\le n^{-1/2}} \|\Delta_j\|_{\mathcal{H}_1}^q \\
&\le n^{-(1-q/2)} \sum_{j=1}^d \|\Delta_j\|_{\mathcal{H}_1}^q \\
&\le 4n^{-(1-q/2)}.
\end{aligned}$$

Thus,

$$\begin{aligned}
n^{-1/2} & \sum_{j=1}^d \|\Delta_j\|_{\mathcal{H}_1}^{1/2\alpha} \|\Delta_j\|_{L_2(\Pi_{jn})}^{1-1/(2\alpha)} \\
&\le \zeta^{-4\alpha/(2\alpha-1)} \sum_{j=1}^d \|\Delta_j\|_{L_2(\Pi_{jn})}^2 + 8\zeta^{4\alpha/(2\alpha+1)} n^{-(1-q/2)}.
\end{aligned} \tag{4.7}$$

Combing (4.6) and (4.7), we get

$$\begin{aligned}
n^{-1/2} & \sum_{j=1}^d \|\Delta_j\|_{\mathcal{H}_1}^{1/2\alpha} \|\Delta_j\|_{L_2(\Pi_{jn})}^{1-1/(2\alpha)} \\
&\le \zeta^{-4\alpha/(2\alpha-1)} \sum_{j=1}^d \|\Delta_j\|_{L_2(\Pi_{jn})}^2 + 8\zeta^{4\alpha/(2\alpha+1)} n^{-(1-\max\{q/2,1/(2\alpha+1)\})}.
\end{aligned} \tag{4.8}$$

By Theorem 4 of Koltchinskii and Yuan (2010), there exists a numerical constant $C_2 > 1$ such that with probability at least $1 - d^{-\beta}$ for all $h \in \mathcal{H}_1$, and $j = 1, \dots, d$,

$$\|h\|_{L_2(\Pi_j)} \le C_2\left[\|h\|_{L_2(\Pi_{jn})} + \left(n^{-\alpha/(2\alpha+1)} + \sqrt{\frac{(\beta+1)\log d}{n}}\right)\|h\|_{\mathcal{H}_1}\right] \tag{4.9}$$

and

$$\|h\|_{L_2(\Pi_{jn})} \le C_2\left[\|h\|_{L_2(\Pi_j)} + \left(n^{-\alpha/(2\alpha+1)} + \sqrt{\frac{(\beta+1)\log d}{n}}\right)\|h\|_{\mathcal{H}_1}\right]. \tag{4.10}$$

Denote by $\mathcal{E}_2$ the event that both (4.9) and (4.10) hold. Under $\mathcal{E}_2$,

$$\begin{aligned}
&\sum_{j=1}^{d} \|\Delta_j\|_{L_2(\Pi_{jn})}^2 \\
&\quad \le 2C_2^2 \sum_{j=1}^{d} \left[\|\Delta_j\|_{L_2(\Pi_j)}^2 + \left(n^{-2\alpha/(2\alpha+1)} + \frac{(\beta+1)\log d}{n} \right) \|\Delta_j\|_{\mathcal{H}_1}^2 \right] \\
&\quad \le 2C_2^2 \sum_{j=1}^{d} \|\Delta_j\|_{L_2(\Pi_j)}^2 + 8C_2^2 \left(n^{-2\alpha/(2\alpha+1)} + \frac{(\beta+1)\log d}{n} \right),
\end{aligned}$$

where the second inequality follows from the fact that

$$\sum_{j=1}^{d} \|\Delta_j\|_{\mathcal{H}_1}^2 \le 4.$$

By (3.1), this implies that

$$\sum_{j=1}^{d} \|\Delta_j\|_{L_2(\Pi_{jn})}^2 \le 2C_2^2 \eta_q \|\Delta\|_{L_2(\Pi)}^2 + 8C_2^2 \left(n^{-2\alpha/(2\alpha+1)} + \frac{(\beta+1)\log d}{n} \right).$$

Together with (4.8), we get

$$\begin{aligned}
&n^{-1/2} \sum_{j=1}^{d} \|\Delta_j\|_{\mathcal{H}_1}^{1/2\alpha} \|\Delta_j\|_{L_2(\Pi_{jn})}^{1-1/(2\alpha)} \\
&\quad \le 2C_2^2 \eta_q \zeta^{-4\alpha/(2\alpha-1)} \|\Delta\|_{L_2(\Pi)}^2 \\
&\qquad + 8C_2^2 \zeta^{-4\alpha/(2\alpha-1)} \left(n^{-2\alpha/(2\alpha+1)} + \frac{(\beta+1)\log d}{n} \right) \\
&\qquad + 8\zeta^{4\alpha/(2\alpha+1)} n^{-(1-\max\{q/2,1/(2\alpha+1)\})}.
\end{aligned} \tag{4.11}$$

The second term on the rightmost-hand side of (4.5) can also be bounded under event $\mathcal{E}_2$. By (4.10),

$$\begin{aligned}
&\sum_{j=1}^{d} \|\Delta_j\|_{L_2(\Pi_{jn})} \\
&\quad \le C_2 \sum_{j=1}^{d} \|\Delta_j\|_{L_2(\Pi_j)} + C_2 \left(n^{-\alpha/(2\alpha+1)} + \sqrt{\frac{(\beta+1)\log d}{n}} \right) \sum_{j=1}^{d} \|\Delta_j\|_{\mathcal{H}_1} \qquad (4.12) \\
&\quad \le C_2 \sum_{j=1}^{d} \|\Delta_j\|_{L_2(\Pi_j)} + 4C_2 \left(n^{-\alpha/(2\alpha+1)} + \sqrt{\frac{(\beta+1)\log d}{n}} \right),
\end{aligned}$$

where in the second inequality we used the fact that

$$\sum_{j=1}^{d} \|\Delta_j\|_{\mathcal{H}_1} \le \sum_{j=1}^{d} \|\Delta_j\|_{\mathcal{H}_1}^q \le 4.$$

Write

$$\sum_{j=1}^{d} \|\Delta_j\|_{L_2(\Pi_j)} \le \sum_{j:\|\Delta_j\|_{L_2(\Pi_j)} > \sqrt{(\log d)/n}} \|\Delta_j\|_{L_2(\Pi_j)} + \sum_{j:\|\Delta_j\|_{L_2(\Pi_j)} \le \sqrt{(\log d)/n}} \|\Delta_j\|_{L_2(\Pi_j)}.$$

The first term can be bounded by the Cauchy–Schwarz inequality

$$\begin{aligned}\sum_{j:\|\Delta_j\|_{L_2(\Pi_j)} > \sqrt{(\log d)/n}} \|\Delta_j\|_{L_2(\Pi_j)} &\le \left(\operatorname{card}\left\{ j : \|\Delta_j\|_{L_2(\Pi_j)} > \sqrt{\frac{\log d}{n}} \right\}\right)^{1/2} \\ &\quad \times \left(\sum_{j:\|\Delta_j\|_{L_2(\Pi_j)} > \sqrt{(\log d)/n}} \|\Delta_j\|_{L_2(\Pi_j)}^2\right)^{1/2}.\end{aligned}$$

Observe that

$$\operatorname{card}\left\{ j : \|\Delta_j\|_{L_2(\Pi_j)} > \sqrt{\frac{\log d}{n}} \right\} \le \left(\frac{\log d}{n}\right)^{-q/2} \sum_{j=1}^{d} \|\Delta_j\|_{\mathcal{H}_1}^q \le 4\left(\frac{\log d}{n}\right)^{-q/2}.$$

Thus,

$$\begin{aligned}\sum_{j:\|\Delta_j\|_{L_2(\Pi_j)} > \sqrt{(\log d)/n}} \|\Delta_j\|_{L_2(\Pi_j)} &\le 4\left(\frac{\log d}{n}\right)^{-q/4} \left(\sum_{j=1}^{d} \|\Delta_j\|_{L_2(\Pi_j)}^2\right)^{1/2} \\ &\le 4\eta_q^{1/2}\left(\frac{\log d}{n}\right)^{-q/4} \|\Delta\|_{L_2(\Pi)}.\end{aligned}$$

Together with the fact that

$$\begin{aligned}\sum_{j:\|\Delta_j\|_{L_2(\Pi_j)} \le \sqrt{(\log d)/n}} \|\Delta_j\|_{L_2(\Pi_j)} \\ \le \sum_{j:\|\Delta_j\|_{L_2(\Pi_j)} \le \sqrt{(\log d)/n}} \|\Delta_j\|_{L_2(\Pi_j)}^q \left(\frac{\log d}{n}\right)^{(1-q)/2}\end{aligned}$$

$$\le \left(\frac{\log d}{n}\right)^{(1-q)/2} \sum_{j=1}^{d} \|\Delta_j\|_{L_2(\Pi_j)}^q$$

$$\le 4\left(\frac{\log d}{n}\right)^{(1-q)/2},$$

we get

$$\sum_{j=1}^{d} \|\Delta_j\|_{L_2(\Pi_j)} \le 4\eta_q^{1/2}\left(\frac{\log d}{n}\right)^{-q/4} \|\Delta\|_{L_2(\Pi)} + 4\left(\frac{\log d}{n}\right)^{(1-q)/2}. \tag{4.13}$$

In light of (4.12), we have

$$\begin{aligned} &\sqrt{\frac{\log d}{n}} \sum_{j=1}^{d} \|\Delta_j\|_{L_2(\Pi_{jn})} \\ &\quad \le 4C_2\eta_q^{1/2}\left(\frac{\log d}{n}\right)^{1/2-q/4} \|\Delta\|_{L_2(\Pi)} \\ &\qquad + 4C_2 n^{-\alpha/(2\alpha+1)}\sqrt{\frac{\log d}{n}} + 8C_2\sqrt{\beta+1}\left(\frac{\log d}{n}\right)^{1-q/2}, \end{aligned} \tag{4.14}$$

where we used the fact that $\log d < n$ and $C_2 > 1$.

Combing (4.5), (4.11), (4.14) and the fact that

$$\sum_{j=1}^{d} \|\Delta_j\|_{\mathcal{H}_1} \le 4,$$

we get

$$\begin{aligned} &\frac{2}{n}\sum_{i=1}^{n} \varepsilon_i[\widehat{f}(X_i) - f(X_i)] \\ &\quad \le C_3\eta_q\zeta^{-4\alpha/(2\alpha-1)}\|\Delta\|_{L_2(\Pi)}^2 \\ &\qquad + C_3\zeta^{-4\alpha/(2\alpha-1)}\left(n^{-2\alpha/(2\alpha+1)} + \frac{(\beta+1)\log d}{n}\right) \\ &\qquad + C_3\zeta^{4\alpha/(2\alpha+1)} n^{-(1-\max\{q/2, 1/(2\alpha+1)\})} \\ &\qquad + C_3\sqrt{\beta+1}\,\eta_q^{1/2}\left(\frac{\log d}{n}\right)^{1/2-q/4} \|\Delta\|_{L_2(\Pi)} \\ &\qquad + C_3\sqrt{\beta+1}\,n^{-\alpha/(2\alpha+1)}\sqrt{\frac{\log d}{n}} \\ &\qquad + C_3\sqrt{\beta+1}\left(\frac{\log d}{n}\right)^{1-q/2} + C_3 n^{-1/2} e^{-d}, \end{aligned} \tag{4.15}$$

for some constant $C_3 > 0$, under the event $\mathcal{E}_1 \cap \mathcal{E}_2$.

Step 2. *Bounding the left-hand side of* (4.4). To bound the left-hand side of (4.4), first observe that

$$\begin{aligned}\|\Delta\|^2_{L_2(\Pi)} - \|\Delta\|^2_{L_2(\Pi_n)} \\ \leq \sup_{\substack{g\in\mathcal{B}_4(\ell_q(\mathcal{H}_d))\\ \|g\|_{L_2(\Pi)}\leq\|\Delta\|_{L_2(\Pi)}}} \big(\|g\|^2_{L_2(\Pi)} - \|g\|^2_{L_2(\Pi_n)}\big).\end{aligned} \tag{4.16}$$

Note that for any $g \in \mathcal{B}_4(\ell_q(\mathcal{H}_d))$,

$$\|g\|^2_{L_\infty} \leq \|g\|^2_{\ell_1(\mathcal{H}_d)} \leq \big(\|g\|^q_{\ell_q(\mathcal{H}_d)}\big)^2 \leq 16$$

and

$$\|g\|^4_{L_2(\Pi)} \leq \|g\|^2_{L_\infty}\|g\|^2_{L_2(\Pi)} \leq 16\|g\|^2_{L_2(\Pi)}.$$

By Talagrand's concentration inequality, for any fixed $u \in [0, 1]$,

$$\begin{aligned}&\sup_{\substack{g\in\mathcal{B}_4(\ell_q(\mathcal{H}_d))\\ \|g\|_{L_2(\Pi)}\leq u}} \big(\|g\|^2_{L_2(\Pi)} - \|g\|^2_{L_2(\Pi_n)}\big) \\ &\quad\leq 2\left(\mathbb{E}\sup_{\substack{g\in\mathcal{B}_4(\ell_q(\mathcal{H}_d))\\ \|g\|_{L_2(\Pi)}\leq u}} \big(\|g\|^2_{L_2(\Pi)} - \|g\|^2_{L_2(\Pi_n)}\big) + 4u\sqrt{\frac{t}{n}} + \frac{16t}{n}\right),\end{aligned}$$

with probability at least $1 - e^{-t}$. By the symmetrization inequality,

$$\begin{aligned}&\mathbb{E}\sup_{\substack{g\in\mathcal{B}_4(\ell_q(\mathcal{H}_d))\\ \|g\|_{L_2(\Pi)}\leq u}} \big(\|g\|^2_{L_2(\Pi)} - \|g\|^2_{L_2(\Pi_n)}\big) \\ &\quad\leq 2\mathbb{E}\sup_{\substack{g\in\mathcal{B}_4(\ell_q(\mathcal{H}_d))\\ \|g\|_{L_2(\Pi)}\leq u}} \left(\frac{1}{n}\sum_{i=1}^n \sigma_i g^2(X_i)\right).\end{aligned}$$

Note that g^2 is 8-Lipschitz function on $\mathcal{B}_4(\ell_q(\mathcal{H}_d))$. By the contraction inequality,

$$\begin{aligned}&\mathbb{E}\sup_{\substack{g\in\mathcal{B}_4(\ell_q(\mathcal{H}_d))\\ \|g\|_{L_2(\Pi)}\leq u}} \left(\frac{1}{n}\sum_{i=1}^n \sigma_i g^2(X_i)\right) \\ &\quad\leq 8\mathbb{E}\sup_{\substack{g\in\mathcal{B}_4(\ell_q(\mathcal{H}_d))\\ \|g\|_{L_2(\Pi)}\leq u}} \left(\frac{1}{n}\sum_{i=1}^n \sigma_i g(X_i)\right).\end{aligned}$$

Again by Talagrand's concentration inequality, there exists a numerical constant $C_4 > 0$ such that with probability at least $1 - e^{-t}$,

$$\mathbb{E} \sup_{\substack{g \in \mathcal{B}_4(\ell_q(\mathcal{H}_d)) \\ \|g\|_{L_2(\Pi)} \le u}} \left(\frac{1}{n} \sum_{i=1}^n \sigma_i g(X_i) \right)$$

$$\le C_4 \left(\sup_{\substack{g \in \mathcal{B}_4(\ell_q(\mathcal{H}_d)) \\ \|g\|_{L_2(\Pi)} \le u}} \left(\frac{1}{n} \sum_{i=1}^n \sigma_i g(X_i) \right) + u\sqrt{\frac{t}{n}} + \frac{t}{n} \right)$$

$$\le C_4 \left(\sup_{\substack{\sum_{j=1}^d \|g_j\|_{\mathcal{H}_1}^q \le 4 \\ \|\sum_{j=1}^d g_j\|_{L_2(\Pi)} \le u}} \sum_{j=1}^d \left(\frac{1}{n} \sum_{i=1}^n \sigma_i g_j(x_{ij}) \right) + u\sqrt{\frac{t}{n}} + \frac{t}{n} \right).$$

In other words,

$$\begin{aligned} &\sup_{\substack{g \in \mathcal{B}_4(\ell_q(\mathcal{H}_d)) \\ \|g\|_{L_2(\Pi)} \le u}} \left(\|g\|_{L_2(\Pi)}^2 - \|g\|_{L_2(\Pi_n)}^2 \right) \\ &\quad \le 16 C_4 \left(\sup_{\substack{\sum_{j=1}^d \|g_j\|_{\mathcal{H}_1}^q \le 4 \\ \|\sum_{j=1}^d g_j\|_{L_2(\Pi)} \le u}} \sum_{j=1}^d \left(\frac{1}{n} \sum_{i=1}^n \sigma_i g_j(x_{ij}) \right) + u\sqrt{\frac{t}{n}} + \frac{t}{n} \right), \end{aligned} \tag{4.17}$$

with probability at least $1 - 2e^{-t}$.

Note that

$$\frac{1}{n} \sum_{i=1}^n \sigma_i g_j(x_{ij}) \le \|g_j\|_{\mathcal{H}_1} \sup_{\substack{\|h\|_{\mathcal{H}_1} = 1 \\ \|h\|_{L_2(\Pi_j)} \le \|g_j\|_{L_2(\Pi_j)} / \|g_j\|_{\mathcal{H}_1}}} \left(\frac{1}{n} \sum_{i=1}^n \sigma_i h(x_{ij}) \right).$$

By Lemma 2.2 and union bound, there exists a constant $C_5 > 0$ such that

$$\sup_{\substack{\|h\|_{\mathcal{H}_1} = 1 \\ \|h\|_{L_2(\Pi_j)} \le u}} \left(\frac{1}{n} \sum_{i=1}^n \sigma_i h(x_{ij}) \right) \le C_5 n^{-1/2} \left(u^{1-1/(2\alpha)} + u\sqrt{(\beta+1)\log d} + e^{-d} \right),$$

uniformly over $u \in [0, 1]$ and $j = 1, \ldots, d$ with probability at least $1 - d^{-\beta}$. Denote this event by $\mathcal{E}_3$, and we shall now proceed conditional on $\mathcal{E}_3$.

It is not hard to see that, under $\mathcal{E}_3$,

$$\begin{aligned} \sum_{j=1}^d \left(\frac{1}{n} \sum_{i=1}^n \sigma_i g_j(x_{ij}) \right) &\le C_5 n^{-1/2} \sum_{j=1}^d \left(\|g_j\|_{\mathcal{H}_1}^{1/2\alpha} \|g_j\|_{L_2(\Pi_j)}^{1-1/(2\alpha)} \right. \\ &\quad \left. + \|g_j\|_{L_2(\Pi_j)} \sqrt{(\beta+1)\log d} + e^{-d} \|g_j\|_{\mathcal{H}_1} \right). \end{aligned} \tag{4.18}$$

Following the same argument as that for (4.8), it can derived

$$
\begin{aligned}
n^{-1/2} & \sup_{\substack{\sum_{j=1}^d \|g_j\|_{\mathcal{H}_1}^q \le 4 \\ \|\sum_{j=1}^d g_j\|_{L_2(\Pi)} \le u}} \sum_{j=1}^d \|g_j\|_{\mathcal{H}_1}^{1/2\alpha} \|g_j\|_{L_2(\Pi_j)}^{1-1/(2\alpha)} \\
& \le \zeta^{-4\alpha/(2\alpha-1)} \sup_{\substack{\sum_{j=1}^d \|g_j\|_{\mathcal{H}_1}^q \le 4 \\ \|\sum_{j=1}^d g_j\|_{L_2(\Pi)} \le u}} \sum_{j=1}^d \|g_j\|_{L_2(\Pi_j)}^2 \\
& \quad + 8\zeta^{4\alpha/(2\alpha+1)} n^{-(1-\max\{q/2, 1/(2\alpha+1)\})} \\
& \le \zeta^{-4\alpha/(2\alpha-1)} \eta_q u^2 + 8\zeta^{4\alpha/(2\alpha+1)} n^{-(1-\max\{q/2, 1/(2\alpha+1)\})}.
\end{aligned} \tag{4.19}
$$

Similar to (4.13), it can also be shown that for any $g_1, \dots, g_d$ such that

$$
\sum_{j=1}^d \|g_j\|_{\mathcal{H}_1}^q \le 4 \quad \text{and} \quad \sum_{j=1}^d \|g_j\|_{L_2(\Pi_j)} \le u,
$$

we have

$$
\sum_{j=1}^d \|g_j\|_{L_2(\Pi_j)} \le 4\eta_q^{1/2} \left(\frac{\log d}{n}\right)^{-q/4} u + 4\left(\frac{\log d}{n}\right)^{1-q/2}. \tag{4.20}
$$

Combining (4.18), (4.19) and (4.20), we have

$$
\begin{aligned}
& \sup_{\substack{\sum_{j=1}^d \|g_j\|_{\mathcal{H}_1}^q \le 4 \\ \|\sum_{j=1}^d g_j\|_{L_2(\Pi)} \le u}} \sum_{j=1}^d \left(\frac{1}{n}\sum_{i=1}^n \sigma_i g_j(x_{ij})\right) \\
& \quad \le C_5 \zeta^{-4\alpha/(2\alpha-1)} \eta_q u^2 \\
& \qquad + 8C_5 \zeta^{4\alpha/(2\alpha+1)} n^{-(1-\max\{q/2, 1/(2\alpha+1)\})} \\
& \qquad + 4C_5 \sqrt{\frac{(\beta+1)\log d}{n}} \left(\eta_q^{1/2}\left(\frac{\log d}{n}\right)^{-q/4} u + \left(\frac{\log d}{n}\right)^{(1-q)/2}\right) \\
& \qquad + C_5 n^{-1/2} e^{-d}.
\end{aligned}
$$

Together with (4.17), conditional on $\mathcal{E}_3$,

$$
\begin{aligned}
& \sup_{\substack{g \in \mathcal{B}_4(\ell_q(\mathcal{H}_d)) \\ \|g\|_{L_2(\Pi)} \le u}} \left(\|g\|_{L_2(\Pi)}^2 - \|g\|_{L_2(\Pi_n)}^2\right) \\
& \quad \le C_6 \zeta^{-4\alpha/(2\alpha-1)} \eta_q u^2 + C_6 \zeta^{4\alpha/(2\alpha+1)} n^{-(1-\max\{q/2, 1/(2\alpha+1)\})}
\end{aligned}
$$

$$+ C_6\sqrt{\frac{(\beta+1)\log d}{n}}\left(\eta_q^{1/2}\left(\frac{\log d}{n}\right)^{-q/4} u + \left(\frac{\log d}{n}\right)^{(1-q)/2}\right)$$
$$+ C_6 n^{-1/2} e^{-d} + C_6\left(u\sqrt{\frac{t}{n}} + \frac{t}{n}\right)$$

holds for some constant $C_6 > 0$, with probability at least $1-2e^{-t}$. Using a peeling argument similar to that for Lemma 2.1, we can make this bound uniformly over $u \in [0, 1]$. More specifically, it can be shown that there exist constants $C_7 > 0$ such that, conditional on $\mathcal{E}_3$,

$$\begin{aligned}
&\sup_{\substack{g\in\mathcal{B}_4(\ell_q(\mathcal{H}_d))\\ \|g\|_{L_2(\Pi)}\le u}} \left(\|g\|^2_{L_2(\Pi)} - \|g\|^2_{L_2(\Pi_n)}\right) \\
&\quad\le C_6\zeta^{-4\alpha/(2\alpha-1)}\eta_q u^2 + C_6\zeta^{4\alpha/(2\alpha+1)} n^{-(1-\max\{q/2,1/(2\alpha+1)\})} \\
&\qquad + C_6\sqrt{\frac{(\beta+1)\log d}{n}}\left(\eta_q^{1/2}\left(\frac{\log d}{n}\right)^{-q/4} u + \left(\frac{\log d}{n}\right)^{(1-q)/2}\right) \\
&\qquad + C_6 n^{-1/2} e^{-d} \\
&\qquad + C_7\left(u\sqrt{\frac{(\beta+1)\log d}{n}} + \frac{(\beta+1)\log d}{n}\right),
\end{aligned} \tag{4.21}$$

uniformly over all $u \in [0, 1]$ with probability at least $1 - d^{-\beta}$. Denote by $\mathcal{E}_4$ the event that inequality (4.21) holds. Then

$$\mathbb{P}\{\mathcal{E}_4\} \ge \mathbb{P}\{\mathcal{E}_4|\mathcal{E}_3\}\mathbb{P}(\mathcal{E}_3) \ge \left(1-d^{-\beta}\right)^2 \ge 1 - 2d^{-\beta}.$$

Together with (4.16), we get, under event $\mathcal{E}_4$,

$$\begin{aligned}
&\|\Delta\|^2_{L_2(\Pi)} \\
&\quad\le \|\Delta\|^2_{L_2(\Pi_n)} + C_8\zeta^{-4\alpha/(2\alpha-1)}\eta_q u^2 \\
&\qquad + C_8\zeta^{4\alpha/(2\alpha+1)} n^{-(1-\max\{q/2,1/(2\alpha+1)\})} \\
&\qquad + C_8\sqrt{\frac{(\beta+1)\log d}{n}}\left(\eta_q^{1/2}\left(\frac{\log d}{n}\right)^{-q/4} \|\Delta\|_{L_2(\Pi)} + \left(\frac{\log d}{n}\right)^{1-q/2}\right) \\
&\qquad + C_8 n^{-1/2} e^{-d} \\
&\qquad + C_8\left(u\sqrt{\frac{(\beta+1)\log d}{n}} + \frac{(\beta+1)\log d}{n}\right),
\end{aligned} \tag{4.22}$$

for some constant $C_8 > 0$.

Step 3. *Putting it together*. Combining (4.15) and (4.22), we get

$$\begin{aligned}\|\Delta\|_{L_2(\Pi)}^2 \le{}& C_9\eta_q\zeta^{-4\alpha/(2\alpha-1)}\|\Delta\|_{L_2(\Pi)}^2\\ &+ C_9\zeta^{-4\alpha/(2\alpha-1)}\frac{(\beta+1)\log d}{n}\\ &+ C_9\zeta^{4\alpha/(2\alpha+1)}n^{-(1-\max\{q/2,1/(2\alpha+1)\})}\\ &+ C_9\sqrt{\beta+1}\left(\frac{\log d}{n}\right)^{1/2-q/4}\|\Delta\|_{L_2(\Pi)}\\ &+ C_9\sqrt{\beta+1}n^{-\alpha/(2\alpha+1)}\sqrt{\frac{\log d}{n}}\\ &+ C_9(\beta+1)\left(\frac{\log d}{n}\right)^{1-q/2}\\ &+ C_9 n^{-1/2}e^{-d},\end{aligned}$$

for some constant $C_9 > 0$, under the event $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_4$.

Take ζ large enough so that

$$C_9\eta_q\zeta^{-4\alpha/(2\alpha-1)} \le 1/2.$$

Then

$$\begin{aligned}\|\Delta\|_{L_2(\Pi)}^2 \le{}& 2C_9\zeta^{-4\alpha/(2\alpha-1)}\frac{(\beta+1)\log d}{n}\\ &+ 2C_9\zeta^{4\alpha/(2\alpha+1)}n^{-(1-\max\{q/2,1/(2\alpha+1)\})}\\ &+ 2C_9\sqrt{\beta+1}\left(\frac{\log d}{n}\right)^{1/2-q/4}\|\Delta\|_{L_2(\Pi)}\\ &+ 2C_9\sqrt{\beta+1}n^{-\alpha/(2\alpha+1)}\sqrt{\frac{\log d}{n}}\\ &+ 2C_9\sqrt{\beta+1}\left(\frac{\log d}{n}\right)^{1-q/2}\\ &+ 2C_9 n^{-1/2}e^{-d}.\end{aligned}$$

Therefore, there exists a constant $C_{10} > 0$ such that, under the event $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_4$,

$$\begin{aligned}&\|\Delta\|_{L_2(\Pi)}^2\\ &\quad\le C_{10}(\beta+1)\left(n^{-2\alpha/(2\alpha+1)} + \left(\frac{\log d}{n}\right)^{1-q/2} + \left(\frac{\log d}{n}\right)^{1/2-q/4}\|\Delta\|_{L_2(\Pi)}\right),\end{aligned}$$

which implies (3.4). Statement (3.4) now follows from the fact that

$$\mathbb{P}\{\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_4\} \ge 1 - \mathbb{P}\{\mathcal{E}_1^c\} - \mathbb{P}\{\mathcal{E}_2^c\} - \mathbb{P}\{\mathcal{E}_4^c\} \ge 1 - 4d^{-\beta},$$

and appropriate rescaling of the constants.

To show (3.5), we first derive, via an identical argument to Step 2, that

$$
\begin{aligned}
\|\Delta\|^2_{L_2(\Pi_n)} \le {}& \|\Delta\|^2_{L_2(\Pi)} + C_{11}\zeta^{-4\alpha/(2\alpha-1)}\eta_q u^2 \\
&+ C_{11}\zeta^{4\alpha/(2\alpha+1)} n^{-(1-\max\{q/2,1/(2\alpha+1)\})} \\
&+ C_{11}\sqrt{\frac{(\beta+1)\log d}{n}}\biggl(\biggl(\frac{\log d}{n}\biggr)^{-q/4} u + 2\biggl(\frac{\log d}{n}\biggr)^{1-q/2}\biggr) \\
&+ C_{11} n^{-1/2} e^{-d} \\
&+ C_{11}\biggl(u\sqrt{\frac{(\beta+1)\log d}{n}} + \frac{(\beta+1)\log d}{n}\biggr),
\end{aligned}
\tag{4.23}
$$

for some constant $C_{11} > 0$. Together with (3.4), this implies (3.5).

4.2. *Proof of auxiliary results.* We now present the proofs of Lemmas 2.1 and 2.2.

PROOF OF LEMMA 2.1. An application of Talagrand's concentration inequality yields, with probability at least $1 - e^{-t}$

$$
R_{jn}(u) \le 2\biggl(\mathbb{E}R_{jn}(u) + u\sqrt{\frac{t}{n}} + \frac{t}{n}\biggr).
$$

It is well known that there exists a numerical constant $C_1 > 0$

$$
\mathbb{E}R_{jn}(u) \le \{\mathbb{E}[R_{jn}(u)]^2\}^{1/2} \le C_1 n^{-1/2} u^{1-1/(2\alpha)}.
$$

See, for example, Mendelson (2002) or Koltchinskii (2011). In other words, with probability at least $1 - e^{-t}$,

$$
R_{jn}(u) \le C_2\biggl(n^{-1/2}u^{1-1/(2\alpha)} + u\sqrt{\frac{t}{n}} + \frac{t}{n}\biggr)
$$

for some numerical constant $C_2 > 0$. We now make this inequality uniform over $u \in [0, 1]$ via a peeling argument.

In particular, with probability at least $1 - \exp(-\beta \log d - 2\log j)$ for some constant $\beta > 0$,

$$
\begin{aligned}
&\sup_{\substack{\|h\|_{\mathcal{H}_1}\le 1 \\ 2^{-j}\le \|h\|_{L_2(\Pi_j)} \le 2^{-j+1}}} \biggl|\frac{1}{n}\sum_{i=1}^n \sigma_i h(x_{ij})\biggr| \\
&\quad \le R_{jn}(2^{-j+1}) \\
&\quad \le C_2 n^{-1/2}\bigl[(2^{-j+1})^{1-1/(2\alpha)} + 2^{-j+1}(\beta\log d + 2\log j)^{1/2} \\
&\qquad + n^{-1/2}(\beta \log d + 2\log j)\bigr].
\end{aligned}
$$

By union bound, there exists a constant $C_3 > 0$ such that

$$R_{jn}(u) \leq C_3 n^{-1/2}\left(u^{1-1/(2\alpha)} + u\sqrt{\beta \log d} + \frac{\beta \log d}{\sqrt{n}}\right),$$

holds for any $u \in (e^{-d(2\alpha/(2\alpha-1))}, 1]$, with probability at least

$$1 - \sum_{j=1}^{\lceil 2\alpha d \log_2 e/(2\alpha-1)\rceil} \exp(-\beta \log d - 2\log j) \geq 1 - 2d^{-\beta}.$$

On the other hand, when $u \leq e^{-d(2\alpha/(2\alpha-1))}$,

$$\begin{aligned}
R_{jn}(u) &\leq R_{jn}\big(e^{-d(2\alpha/(2\alpha-1))}\big) \\
&\leq C_2 n^{-1/2}\left(e^{-d} + e^{-d(2\alpha/(2\alpha-1))}\sqrt{\beta \log d} + \frac{\beta \log d}{\sqrt{n}}\right) \\
&\leq 2C_2 n^{-1/2}\left(e^{-d} + \frac{\beta \log d}{\sqrt{n}}\right),
\end{aligned}$$

with probability at least $1 - d^{-\beta}$, for sufficiently large d. In summary, there exists a constant $C_4 > 0$ such that

$$R_{jn}(u) \leq C_4 n^{-1/2}\left(u^{1-1/(2\alpha)} + u\sqrt{\beta \log d} + \frac{\beta \log d}{\sqrt{n}} + e^{-d}\right),$$

uniformly over all $u \in [0, 1]$ with probability at least $1 - 3d^{-\beta}$. □

PROOF OF LEMMA 2.2. Note that

$$\int_0^u [\log \mathcal{N}(\mathcal{B}_1(\mathcal{H}_1), \delta, \|\cdot\|_{L_\infty})]^{1/2}\, du \leq c_\alpha \delta^{1-1/(2\alpha)}.$$

Therefore, there exist constants $C_1, C_2 > 0$ such that for any fixed $u \in [0, 1]$

$$\mathbb{P}\{\widehat{Z}_{jn}(u) \leq C_1 n^{-1/2}(u^{1-1/(2\alpha)} + ut^{1/2})\} \leq C_2 \exp[-(u^{-1/\alpha} + t)].$$

See, for example, an de Geer [(2000); Corollary 8.3]. The rest of the proof follows a similar peeling argument as that for Lemma 2.1 and is omitted for brevity. □

REFERENCES

ARONSZAJN, N. (1950). Theory of reproducing kernels. *Trans. Amer. Math. Soc.* **68** 337–404. MR0051437

BACH, F. R. (2008). Consistency of the group lasso and multiple kernel learning. *J. Mach. Learn. Res.* **9** 1179–1225. MR2417268

BARTLETT, P. L. and MENDELSON, S. (2002). Rademacher and Gaussian complexities: Risk bounds and structural results. *J. Mach. Learn. Res.* **3** 463–482. MR1984026

BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. MR2533469

BOUSQUET, O. (2002). A Bennett concentration inequality and its application to suprema of empirical processes. *C. R. Math. Acad. Sci. Paris* **334** 495–500. MR1890640

BOUSQUET, O. and HERRMANN, D. (2003). On the complexity of learning the kernel matrix. In *Advances in Neural Information Processing Systems* **15** 415–422.

BREIMAN, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics* **37** 373–384. MR1365720

BÜHLMANN, P. and VAN DE GEER, S. (2011). *Statistics for High-Dimensional Data*: *Methods*, *Theory and Applications*. Springer, Heidelberg. MR2807761

CANDES, E. and TAO, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.* **35** 2313–2351. MR2382644

COVER, T. M. and THOMAS, J. A. (1991). *Elements of Information Theory*. Wiley, New York. MR1122806

CRAMMER, K., KESHET, J. and SINGER, Y. (2003). Kernel design using boosting. In *Advances in Neural Information Processing Systems* **15** 553–560. MIT Press, Cambridge, MA.

CUI, X., PENG, H., WEN, S. and ZHU, L. (2013). Component selection in the additive regression model. *Scand. J. Stat.* **40** 491–510. MR3091694

FAN, J., FENG, Y. and SONG, R. (2011). Nonparametric independence screening in sparse ultra-high-dimensional additive models. *J. Amer. Statist. Assoc.* **106** 544–557. MR2847969

FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581

HASTIE, T. J. and TIBSHIRANI, R. J. (1990). *Generalized Additive Models*. *Monographs on Statistics and Applied Probability* **43**. Chapman & Hall, London. MR1082147

HUANG, J., HOROWITZ, J. L. and WEI, F. (2010). Variable selection in nonparametric additive models. *Ann. Statist.* **38** 2282–2313. MR2676890

KOLTCHINSKII, V. (2011). *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*. *Lecture Notes in Math.* **2033**. Springer, Heidelberg. MR2829871

KOLTCHINSKII, V. and YUAN, M. (2008). Sparse recovery in large ensembles of kernel machines. In *Proceedings of* 19*th Annual Conference on Learning Theory* (*COLT* 2008) 229–238. Omnipress, Madison, WI.

KOLTCHINSKII, V. and YUAN, M. (2010). Sparsity in multiple kernel learning. *Ann. Statist.* **38** 3660–3695. MR2766864

LANCKRIET, G. R. G., CRISTIANINI, N., BARTLETT, P., EL GHAOUI, L. and JORDAN, M. I. (2004). Learning the kernel matrix with semidefinite programming. *J. Mach. Learn. Res.* **5** 27–72. MR2247973

LEDOUX, M. and TALAGRAND, M. (1991). *Probability in Banach Spaces*: *Isoperimetry and Processes*. *Ergebnisse der Mathematik und Ihrer Grenzgebiete* (3) **23**. Springer, Berlin. MR1102015

LIN, Y. and ZHANG, H. H. (2006). Component selection and smoothing in multivariate nonparametric regression. *Ann. Statist.* **34** 2272–2297. MR2291500

MASSART, P. (2007). *Concentration Inequalities and Model Selection*. *Lecture Notes in Math.* **1896**. Springer, Berlin. MR2319879

MEIER, L., VAN DE GEER, S. and BÜHLMANN, P. (2009). High-dimensional additive modeling. *Ann. Statist.* **37** 3779–3821. MR2572443

MENDELSON, S. (2002). Geometric parameters of kernel machines. In *Computational Learning Theory* (*Sydney*, 2002). *Lecture Notes in Computer Science* **2375** 29–43. Springer, Berlin. MR2040403

MICCHELLI, C. A. and PONTIL, M. (2005). Learning the kernel function via regularization. *J. Mach. Learn. Res.* **6** 1099–1125. MR2249850

RASKUTTI, G., WAINWRIGHT, M. J. and YU, B. (2011). Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Trans. Inform. Theory* **57** 6976–6994. MR2882274

RASKUTTI, G., WAINWRIGHT, M. J. and YU, B. (2012). Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *J. Mach. Learn. Res.* **13** 389–427. MR2913704

RAVIKUMAR, P., LAFFERTY, J., LIU, H. and WASSERMAN, L. (2009). Sparse additive models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **71** 1009–1030. MR2750255

SREBRO, N. and BEN-DAVID, S. (2006). Learning bounds for support vector machines with learned kernels. In *Learning Theory. Lecture Notes in Computer Science* **4005** 169–183. Springer, Berlin. MR2280605

STONE, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *Ann. Statist.* **8** 1348–1360. MR0594650

STONE, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10** 1040–1053. MR0673642

STONE, C. J. (1985). Additive regression and other nonparametric models. *Ann. Statist.* **13** 689–705. MR0790566

SUZUKI, T. and SUGIYAMA, M. (2013). Fast learning rate of multiple kernel learning: Trade-off between sparsity and smoothness. *Ann. Statist.* **41** 1381–1405. MR3113815

TALAGRAND, M. (1996). New concentration inequalities in product spaces. *Invent. Math.* **126** 505–563. MR1419006

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **58** 267–288. MR1379242

TSYBAKOV, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer, New York. MR2724359

VAN DE GEER, S. (2000). *Empirical Processes in M-Estimation*. Cambridge Univ. Press, Cambridge.

VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer, New York. MR1385671

WAHBA, G. (1990). *Spline Models for Observational Data*. *CBMS-NSF Regional Conference Series in Applied Mathematics* **59**. SIAM, Philadelphia, PA. MR1045442

YANG, Y. and BARRON, A. (1999). Information-theoretic determination of minimax rates of convergence. *Ann. Statist.* **27** 1564–1599. MR1742500

YE, F. and ZHANG, C. (2010). Rate minimaxity of the Lasso and Dantzig selector for the ℓ_q loss in ℓ_r balls. *J. Mach. Learn. Res.* **11** 3519–3540. MR2756192

YUAN, M. (2007). Nonnegative garrote component selection in functional ANOVA models. In *Conference on Artificial Intelligence and Statistics* 660–666.

DEPARTMENT OF STATISTICS
UNIVERSITY OF WISCONSIN-MADISON
1300 UNIVERSITY AVENUE
MADISON, WISCONSIN 53706
USA
E-MAIL: myuan@stat.wisc.edu

DEPARTMENT OF MATHEMATICS
CITY UNIVERSITY OF HONG KONG
TAT CHEE AVENUE
KOWLOON
HONG KONG