The Annals of Statistics
2019, Vol. 47, No. 5, 2567–2600
https://doi.org/10.1214/18-AOS1757

DOUBLY PENALIZED ESTIMATION IN ADDITIVE REGRESSION WITH HIGH-DIMENSIONAL DATA

BY ZHIQIANG TAN[1] AND CUN-HUI ZHANG[2]

Rutgers University

Additive regression provides an extension of linear regression by modeling the signal of a response as a sum of functions of covariates of relatively low complexity. We study penalized estimation in high-dimensional nonparametric additive regression where functional semi-norms are used to induce smoothness of component functions and the empirical L_2 norm is used to induce sparsity. The functional semi-norms can be of Sobolev or bounded variation types and are allowed to be different amongst individual component functions. We establish oracle inequalities for the predictive performance of such methods under three simple technical conditions: a sub-Gaussian condition on the noise, a compatibility condition on the design and the functional classes under consideration and an entropy condition on the functional classes. For random designs, the sample compatibility condition can be replaced by its population version under an additional condition to ensure suitable convergence of empirical norms. In homogeneous settings where the complexities of the component functions are of the same order, our results provide a spectrum of minimax convergence rates, from the so-called slow rate without requiring the compatibility condition to the fast rate under the hard sparsity or certain L_q sparsity to allow many small components in the true regression function. These results significantly broaden and sharpen existing ones in the literature.

1. Introduction. Additive regression is an extension of linear regression where the signal of a response can be written as a sum of functions of covariates of relatively low complexity. Let (Y_i, X_i), $i = 1, \ldots, n$, be a set of n independent (possibly nonidentically distributed) observations, where $Y_i \in \mathbb{R}$ is a response variable and $X_i \in \mathbb{R}^d$ is a covariate (or design) vector. Consider an additive regression model, $Y_i = g^*(X_i) + \varepsilon_i$ with

$$g^*(x) = \sum_{j=1}^{p} g_j^*\big(x^{(j)}\big), \tag{1}$$

Received April 2017; revised July 2018.

[1]Supported in part by PCORI Grant ME-1511-32740.

[2]Supported in part by NSF Grants DMS-1513378, DMS-1721495, IIS-1250985, IIS-1407939 and IIS-1741390.

MSC2010 subject classifications. Primary 62E20, 62F25, 62F35; secondary 62J05, 62J12.

Key words and phrases. Additive model, bounded variation space, ANOVA model, high-dimensional data, metric entropy, penalized estimation, teproducing kernel Hilbert space, Sobolev space, total variation, trend filtering.

where ε_i is a noise with mean 0 given X_i, $x^{(j)}$ is a vector composed of a small subset of the components of $x \in \mathbb{R}^d$ and g_j^* belongs to a certain functional class $\mathcal{G}_j$. That is, $g^*(x)$ lies in the space of additive functions $\mathcal{G} = \{\sum_{j=1}^p g_j(x^{(j)}) : g_j \in \mathcal{G}_j, j = 1, \ldots, p\}$. A function $g \in \mathcal{G}$ may admit the decomposition $g(x) = \sum_{j=1}^p g_j(x^{(j)})$ for multiple choices of $(g_1, \ldots, g_p)$. In what follows, such choices are considered equivalent but a favorite decomposition can be used to evaluate properties (e.g., the L_2 norm) of the components of $g \in \mathcal{G}$.

In a classical setting (e.g., Stone (1985)), each g_j^* is a univariate function and $x^{(j)}$ is the jth component of $x \in [0, 1]^d$, so that $p = d$. We take a broad view of additive regression and our analysis will accommodate the general setting where g_j^* can be multivariate with $X_i^{(j)}$ being a block of covariates, possibly overlapping across different j as in functional ANOVA (e.g., Gu (2002)). However, most concrete examples will be given in the classical setting.

Additive modeling has been well studied in the setting where the number of components p is fixed; see Hastie and Tibshirani (1990) and references therein. Recently, building upon related works in penalized linear regression, there has been considerable progress in the development of theory and methods for sparse additive regression in high-dimensional settings where p can be of greater order than the sample size n but the number of significant components is still smaller than n; see, for example, Lin and Zhang (2006), Meier, van de Geer and Bühlmann (2009), Ravikumar et al. (2009), Huang, Horowitz and Wei (2010), Koltchinskii and Yuan (2010), Raskutti, Wainwright and Yu (2012), Suzuki and Sugiyama (2013), Dalalyan, Ingster and Tsybakov (2014), Petersen, Witten and Simon (2016) and Yuan and Zhou (2016).

In this article, we study a penalized estimator $\hat{g}$ with an associated decomposition $\hat{g} = \sum_{j=1}^p \hat{g}_j$ defined as a minimizer of a penalized loss

$$\|Y - g\|_n^2/2 + \sum_{j=1}^p (\rho_{nj}\|g_j\|_{F,j} + \lambda_{nj}\|g_j\|_n) \tag{2}$$

over $g \in \mathcal{G}$ and decompositions $g = \sum_{j=1}^p g_j$, where $(\lambda_{nj}, \rho_{nj})$ are tuning parameters, $\|\cdot\|_n$ is the empirical L_2 norm based on the data points, for example, $\|Y - g\|_n^2 = n^{-1}\sum_{i=1}^n \{Y_i - g(X_i)\}^2$, and $\|g_j\|_{F,j}$ is a semi-norm describing the complexity of $g_j \in \mathcal{G}_j$. For simplicity, the association of $\|g_j\|_n$ and $\|g_j\|_{F,j}$ with $X_i^{(j)}$ is typically suppressed.

In the penalized loss (2), each component g_j is *doubly penalized* by its empirical norm and functional semi-norm. The empirical norm $\|\cdot\|_n$ is used to induce sparsity, whereas the functional semi-norm $\|\cdot\|_{F,j}$ is used to induce smoothness of the estimated regression function. For example, if $\mathcal{G}_j$ is taken as a Sobolev space $\mathcal{W}_r^m$ on $[0, 1]$, then $\|g_j\|_{F,j} = \{\int_0^1 |g_j^{(m)}|^r \, dz\}^{1/r}$ for $r \geq 1$ and $m \geq 1$, where $g_j^{(m)}$ denotes the mth derivative of g_j. For the special case $r = 2$, the L_2-Sobolev space

$\mathcal{W}_2^m$ is a (reproducing kernel) Hilbert space, used in the construction of smoothing splines (e.g., Gu (2002)). Moreover, if $\mathcal{G}_j$ is a bounded variation space $\mathcal{V}^m$ on $[0, 1]$, then $\|g_j\|_{F,j} = \mathrm{TV}(g_j^{(m-1)})$ for $m \geq 1$, where $\mathrm{TV}(\cdot)$ denotes the total variation. For univariate smoothing, regression splines using total variation penalties have been studied in Mammen and van de Geer (1997); see Section 2 for further discussion.

We consider both fixed and random designs and establish oracle inequalities for the predictive performance of $\hat{g}$ under three simple technical conditions: a sub-Gaussian condition on noises, a compatibility condition on the design and the functional classes $\mathcal{G}_j$ and an entropy condition on $\mathcal{G}_j$. The compatibility condition is similar to the restricted eigenvalue condition used in analysis of the Lasso, and for random designs, the empirical compatibility condition can be replaced by its population version under an additional condition to ensue suitable convergence of empirical norms. For the Sobolev and bounded variation classes, the entropy condition on $\mathcal{G}_j$ follows from standard results in the literature (e.g., Lorentz, Golitschek and Makovoz (1996)).

In the following, we highlight implications of our oracle inequalities and compare our results with existing ones in the classical homogeneous setting where $X_i^{(j)}$ is the jth component of X_i and $\mathcal{G}_j = \mathcal{G}_0$ for all j. Let $\mathcal{G}_0$ be either a Sobolev space $\mathcal{W}_r^m$ or a bounded variation space $\mathcal{V}^m$ on $[0, 1]$. In this setting, it is natural to set $(\lambda_{nj}, \rho_{nj}) = (\lambda_n, \rho_n)$ for all j. Consider random designs, and suppose that (1) holds with some choice of $(g_1^*, \ldots, g_p^*)$ satisfying

$$\sum_{j=1}^{p} \|g_j^*\|_F \leq C_1 M_F, \qquad \sum_{j=1}^{p} \|g_j^*\|_Q^q \leq C_1^q M_q, \tag{3}$$

where $\|f\|_F$ is a semi-norm on $\mathcal{G}_0$, $\|f\|_Q^2 = n^{-1}\sum_{i=1}^n E\{f^2(X_i)\}$, $C_1 > 0$ is a constant depending only on the moments of $(\varepsilon_1, \ldots, \varepsilon_n)$, $0 \leq q \leq 1$, and $M_q > 0$ and $M_F > 0$ are allowed to depend on (n, p). In the case of hard sparsity ($q = 0$), $\#\{j : g_j^* \neq 0\} \leq M_0$. The following self-contained result (Proposition 1) can be deduced from Propositions 3, 5, 7 and 9.

Let $\beta_0 = 1/m$ and define

$$w_n^*(q) = \max\{n^{\frac{-1}{2+\beta_0(1-q)}}, (\log(p)/n)^{\frac{1-q}{2}}\},$$
$$\gamma_n^*(q) = \min\{n^{\frac{-1}{2+\beta_0(1-q)}}, n^{-1/2}(\log(p)/n)^{\frac{-(1-q)\beta_0}{4}}\}.$$

For $0 \leq q < 1$, we assume that the following compatibility condition holds with some constants $C_0^* > 0$, $\kappa_0^* > 0$ and $\xi_0^* > 1$: for any functions $\{f_j \in \mathcal{G}_0 : j = 1, \ldots, p\}$ and $f = \sum_{j=1}^p f_j$, if $w_n^*(q)\sum_{j=1}^p \|f_j\|_{F,j} + \sum_{j\in S^c}\|f_j\|_Q \leq \xi_0^* \sum_{j\in S}\|f_j\|_Q$, then $\kappa_0^{*2}\sum_{j\in S}\|f_j\|_Q^2 \leq \|f\|_Q^2$, where $S = \{1 \leq j \leq p : \|g_j^*\|_Q > C_0^*\lambda_n\}$. This condition is a homogeneous version of Assumption 7 later and can

be relaxed for $q = 0$ according to Assumption 5. For simplicity, we restrict to the case where $1 \le r \le 2$ (including $r = 1$ for $\mathcal{V}^m$). For $rm > 1$, we assume that for $j = 1, \ldots, p$, the average marginal density of $(X_1^{(j)}, \ldots, X_n^{(j)})$ is uniformly bounded away from 0 and, if $q \ne 1$, is also uniformly bounded from above on $[0, 1]$. The assumption of marginal densities bounded from above, as well as the restriction $1 \le r \le 2$, can be relaxed under slightly different technical conditions (see Propositions 3, 4 and 6). For $r = m = 1$, neither the lower nor the upper bound of marginal densities need to be assumed.

PROPOSITION 1. *Let $\mathcal{G}_0$ be a Sobolev space $\mathcal{W}_r^m$ with $1 \le r \le 2$ and $m \ge 1$ or a bounded variation space $\mathcal{V}^m$ with $r = 1$ and $m \ge 1$. Suppose that the noises are sub-Gaussian, and $\log(p) = o(n)$. Let $\tau_0 = 1/(2m + 1 - 2/r)$, $\Gamma_n = 1$ for $rm > 1$ and $\Gamma_n = \sqrt{\log n}$ for $r = m = 1$.*

(i) *Let $q = 1$ and $\lambda_n = \rho_n = A_0\{\log(p)/n\}^{1/2}$ for a sufficiently large constant A_0. If $p \to \infty$, then*

$$\|\hat{g} - g^*\|_Q^2 = O_p(1)C_1^2(M_F^2 + M_1^2)\{n^{-1/2}\Gamma_n + \sqrt{\log(p)/n}\}. \tag{4}$$

(ii) *Let $q = 0$, $\lambda_n = A_0[\gamma_n^*(0) + \{\log(p)/n\}^{1/2}]$ and $\rho_n = \lambda_n w_n^*(0)$. Suppose that*

$$\{w_n^*(0)^{-\tau_0}\sqrt{\log(np)/n}\}(1 + M_F + M_0) = o(1), \tag{5}$$

and the preceding compatibility condition holds. Then for sufficiently large A_0,

$$\|\hat{g} - g^*\|_Q^2 = O_p(1)C_1^2(M_F + M_0)\{n^{\frac{-1}{2+\beta_0}} + \sqrt{\log(p)/n}\}^2. \tag{6}$$

(iii) *Let $0 < q < 1$, $\lambda_n = A_0[\gamma_n^*(q) + \{\log(p)/n\}^{1/2}]$ and $\rho_n = \lambda_n w_n^*(q)$. Suppose that*

$$\{w_n^*(q)^{-\tau_0}(\log(np)/n)^{\frac{1-q}{2}}\}(1 + M_F + M_q) = O(1),$$

and the preceding compatibility condition holds. Then for sufficiently large A_0,

$$\|\hat{g} - g^*\|_Q^2 = O_p(1)C_1^2(M_F + M_q)\{n^{\frac{-1}{2+\beta_0(1-q)}} + \sqrt{\log(p)/n}\}^{2-q}. \tag{7}$$

(iv) *For each of the cases* (i)–(iii), *the convergence rate of $\|\hat{g} - g^*\|_Q^2$ matches the minimax rate over the parameter set* (3) *up to some multiplicative constants depending on (M_F, M_q), except for the extra logarithmic factor $\Gamma_n = \sqrt{\log n}$ in case* (i) *with $r = m = 1$.*

We point out several important features achieved by the foregoing result, distinct from existing results. First, our results are established for additive regression with general L_r-Sobolev spaces and bounded variation spaces. An important innovation in our proofs involves a delicate application of maximal inequalities based on the

metric entropy of a particular choice of bounded subsets of $\mathcal{G}_0$ (see Lemma 1 in the Supplementary Material (Tan and Zhang (2019))). All previous results seem to be limited to the L_2-Sobolev spaces or similar reproducing kernel Hilbert spaces, except for Petersen, Witten and Simon (2016), who studied additive regression with the bounded variation space $\mathcal{V}^1$ and obtained the rate $\{\log(np)/n\}^{1/2}$ for in-sample prediction under assumption (3) with $q = 1$. In contrast, our analysis with $q = 1$ yields the sharper, yet standard, rate $\{\log(p)/n\}^{1/2}$ for in-sample prediction (see Proposition 3), whereas $\{\log(np)/n\}^{1/2}$ for out-of-sample prediction by (4).

Second, the restricted parameter set (3) represents an L_1 ball in $\|\cdot\|_F$ seminorm (inducing smoothness) but an L_q ball in $\|\cdot\|_Q$ norm (inducing sparsity) for the component functions $(g_1^*, \ldots, g_p^*)$. That is, the parameter set (3) decouples conditions for sparsity and smoothness in additive regression: it can encourage sparsity at different levels $0 \le q \le 1$ while enforcing smoothness only to a limited extent. Accordingly, our result leads to a spectrum of convergence rates (6), which are easily seen to slow down as q increases from 0 to 1, corresponding to weaker sparsity assumptions. While most of previous results are obtained under exact sparsity ($q = 0$), Yuan and Zhou (2016) studied additive regression with reproducing kernel Hilbert spaces under an L_q ball in the Hilbert norm $\|\cdot\|_H$: $\sum_{j=1}^p \|g_j^*\|_H^q \le M_q$. This parameter set induces smoothness and sparsity simultaneously and is in general more restrictive than (3). As a result, the minimax rate of estimation obtained by Yuan and Zhou (2016), based on constrained least squares with *known* M_q instead of penalized estimation, is faster than (7), in the form $n^{-2/(2+\beta_0)} + \{\log(p)/n\}^{(2-q)/2}$, unless $q = 0$ or 1.

Third, in the case of $q = 1$, our result (4) shows that the rate $\{\log(p)/n\}^{1/2}$, with an additional $\{\log(n)/n\}^{1/2}$ term for the bounded variation space $\mathcal{V}^1$, can be achieved via penalized estimation without requiring a compatibility condition. This generalizes a slow-rate result for constrained least-squares (instead of penalization) with *known* (M_1, M_F) in additive regression with the Sobolev–Hilbert space in Ravikumar et al. (2009). Both are related to earlier results for linear regression (Greenshtein and Ritov (2004); Bunea, Tsybakov and Wegkamp (2007)).

Fourth, the rate of convergence (6) under exact sparsity ($q = 0$) is known to be in general faster than in Meier, van de Geer and Bühlmann (2009). Compared with previous results giving similar rates of convergence as (6) with $q = 0$ for Hilbert spaces, our results are stronger in requiring much weaker technical conditions. The penalized estimation procedures in Koltchinskii and Yuan (2010) and Raskutti, Wainwright and Yu (2012), while minimizing a similar criterion as (2), involve additional constraints on $(g_1, \ldots, g_p)$: Koltchinskii and Yuan (2010) required that the sup-norm of $\sum_{j=1}^p g_j$ be bounded by a known constant, whereas Raskutti, Wainwright and Yu (2012) required that $\max_j \|g_j\|_H$ be bounded by a known constant. Moreover, Raskutti, Wainwright and Yu (2012) assumed that the covariates $(X_i^{(1)}, \ldots, X_i^{(p)})$ are independent of each other. These restrictions were relaxed in Suzuki and Sugiyama (2013), but only explicitly under the assumption

that the noises ε_i are uniformly bounded by a constant. Moreover, our rate condition (5) about the sizes of (M_0, M_F) is much weaker than in Suzuki and Sugiyama (2013), due to improved analysis of convergence of empirical norms and the more careful choices (λ_n, ρ_n). For example, if (M_0, M_F) are bounded, then condition (5) holds whenever $\log(p)/n = o(1)$ for Sobolev–Hilbert spaces, but the condition previously required amounts to $\log(p)n^{-1/2} = o(1)$. Finally, the seemingly faster rate in Suzuki and Sugiyama (2013) can be deduced from our results when (λ_n, ρ_n) are allowed to depend on (M_0, M_F); see Remarks 10 and 12–14 for relevant discussion.

Finally, minimax rates of convergence in the form (6) have been shown under exact sparsity ($q = 0$) with L_2-Sobolev or similar Hilbert spaces by Raskutti, Wainwright and Yu (2012) and Dalalyan, Ingster and Tsybakov (2014), respectively, in additive regression and white noise models. For additive regression with general L_r-Sobolev or bounded variation spaces, our results provide minimax rates of convergence (achievable by convex programming) under L_q-ball sparsity in $\|\cdot\|_Q$ norm as well as L_1-ball smoothness in $\|\cdot\|_F$ semi-norm. It should be noted that the dependency of our convergence rates of $\hat{g}$ on (M_F, M_q) can be matched with that in the minimax rates when (λ_n, ρ_n) are allowed to depend on (M_0, M_F) (see Remark 18).

The rest of the article is organized as follows. Section 2 gives a review of univariate functional classes and entropies. Section 3 presents general results for fixed designs (Section 3.1) and random designs (Section 3.2). Section 4 provides specific results for Sobolev and bounded variation spaces, and Section 5 studies the convergence of empirical norms. Section 6 concludes the paper with a discussion. For space limitation, all proofs are collected in Section S1 and technical tools are stated in Section S2 of the Supplementary Material.

2. Functional classes and entropies. As a building block of additive regression, we discuss two broad choices for the function space $\mathcal{G}_j$ and the associated semi-norm $\|g_j\|_{F,j}$ in the context of univariate regression. For concreteness, we consider a fixed function space, say $\mathcal{G}_1$, although our discussion is applicable to $\mathcal{G}_j$ for $j = 1, \ldots, p$. For $r \geq 1$, the L_r norm of a function f on $[0, 1]$ is defined as $\|f\|_{L_r} = \{\int_0^1 |f(z)|^r \, dz\}^{1/r}$.

EXAMPLE 1 (Sobolev spaces). For $r \geq 1$ and $m \geq 1$, let $\mathcal{W}_r^m = \mathcal{W}_r^m([0, 1])$ be the Sobolev space of all functions, $g_1 : [0, 1] \to \mathbb{R}$, such that $g_1^{(m-1)}$ is absolutely continuous and the norm $\|g_1\|_{\mathcal{W}_r^m} = \|g_1\|_{L_r} + \|g_1^{(m)}\|_{L_r}$ is finite, where $g_1^{(m)}$ denotes the mth (weak) derivative of g_1. To describe the smoothness, a semi-norm $\|g_1\|_{F,1} = \|g_1^{(m)}\|_{L_r}$ is often used for $g_1 \in \mathcal{W}_r^m$.

In the statistical literature, a major example of Sobolev spaces is $\mathcal{W}_2^m = \{g_1 : \|g_1\|_{L_2} + \|g_1^{(m)}\|_{L_2} < \infty\}$, which is a reproducing kernel Hilbert space (e.g., Gu

(2002)). Consider a univariate regression model

$$Y_i = g_1(X_i^{(1)}) + \varepsilon_i, \qquad i = 1, \ldots, n. \tag{8}$$

The Sobolev space $\mathcal{W}_2^m$ is known to lead to polynomial smoothing splines through penalized estimation: there exists a unique solution, in the form of a spline of order $(2m-1)$, when minimizing over $g_1 \in \mathcal{W}_2^m$ the following criterion:

$$\frac{1}{2n}\sum_{i=1}^{n}\{Y_i - g_1(X_i^{(1)})\}^2 + \rho_{n1}\|g_1\|_{F,1}. \tag{9}$$

This solution can be made equivalent to the standard derivation of smoothing splines (modulus a zero solution), where the penalty in (9) is $\rho'_{n1}\|g_1\|_{F,1}^2$ for a different tuning parameter ρ'_{n1}. Particularly, cubic smoothing splines are obtained with the choice $m = 2$.

EXAMPLE 2 (Bounded variation spaces). For a function f on $[0, 1]$, the total variation (TV) of f is defined as

$$\mathrm{TV}(f) = \sup\left\{\sum_{i=1}^{k}|f(z_i) - f(z_{i-1})| : z_0 < z_1 < \cdots < z_k \text{ is any partition of } [0, 1]\right\}.$$

If f is differentiable, then $\mathrm{TV}(f) = \int_0^1 |f^{(1)}(z)|\,\mathrm{d}z$. For $m \geq 1$, let $\mathcal{V}^m = \mathcal{V}^m([0, 1])$ be the bounded variation space that consists of all functions, $g_1 : [0, 1] \to \mathbb{R}$, such that $g_1^{(m-2)}$, if $m \geq 2$, is absolutely continuous and the norm $\|g_1\|_{\mathcal{V}^m} = \|g_1\|_{L_1} + \mathrm{TV}(g_1^{(m-1)})$ is finite. For $g_1 \in \mathcal{V}^m$, the semi-norm $\|g_1\|_{F,1} = \mathrm{TV}(g_1^{(m-1)})$ is often used to describe smoothness. The bounded variation space $\mathcal{V}^m$ includes as a strict subset the Sobolev space $\mathcal{W}_1^m$, where the semi-norms also agree: $\mathrm{TV}(g_1^{(m-1)}) = \|g_1^{(m)}\|_{L_1}$ for $g_1 \in \mathcal{W}_1^m$.

For univariate regression (8) with bounded variation spaces, TV semi-norms can be used as penalties in (9) for penalized estimation. This leads to a class of TV splines, which are shown to adapt well to spatial inhomogeneous smoothness (Mammen and van de Geer (1997)). For $m = 1$ or 2, a minimizer of (9) over $g_1 \in \mathcal{V}^m$ can always be chosen as a spline of order m, with the knots in the set of design points $\{X_i^{(1)} : i = 1, \ldots, n\}$. But, as a complication, this is in general not true for $m \geq 3$.

Recently, there is another smoothing method related to TV splines, called trend filtering (Kim et al. (2009)), where (9) is minimized over all possible values $\{g_1(X_i^{(1)}) : i = 1, \ldots, n\}$ with $\|g_1\|_{F,1}$ replaced by L_1 norm of mth-order differences of these values. This method is equivalent to TV splines only for $m = 1$ or 2. But when the design points are evenly spaced, it achieves the minimax rate

of convergence over functions of bounded variation for general $m \geq 1$, similarly as TV splines Tibshirani, 2014. Additive models with trend filtering are studied by Sadhanala and Tibshirani (2017) in low-dimensional settings.

The complexity of a functional class can be described by its metric entropy, which plays an important role in the study of empirical processes (van der Vaart and Wellner (1996)). For a subset $\mathcal{F}$ in a metric space $\overline{\mathcal{F}}$ endowed with norm $\|\cdot\|$, the covering number $N(\delta, \mathcal{F}, \|\cdot\|)$ is defined as the smallest number of balls of radius δ in the $\|\cdot\|$-metric needed to cover $\mathcal{F}$, that is, the smallest value of N such that there exist $f_1, \ldots, f_N \in \overline{\mathcal{F}}$, satisfying $\min_{j=1,\ldots,N} \|f - f_j\| \leq \delta$ for any $f \in \mathcal{F}$. The entropy of $(\mathcal{F}, \|\cdot\|)$ is defined as $H(\delta, \mathcal{F}, \|\cdot\|) = \log N(\delta, \mathcal{F}, \|\cdot\|)$.

For analysis of regression models, our approach involves using entropies of functional classes for empirical norms based on design points, for example, $\{X_i^{(1)} : i = 1, \ldots, n\}$ for subsets of $\mathcal{G}_1$. One type of such norms is the empirical L_2 norm, $\|g_1\|_n = \{n^{-1} \sum_{i=1}^n g_1^2(X_i^{(1)})\}^{1/2}$. Another is the empirical supremum norm, $\|g_1\|_{n,\infty} = \max_{i=1,\ldots,n} g_1(X_i^{(1)})$. If $\mathcal{F}$ is the unit ball in the Sobolev space $\mathcal{W}_r^m$ or the bounded variation space $\mathcal{V}^m$ on $[0, 1]$, the general picture is $H(\delta, \mathcal{F}, \|\cdot\|) \lesssim \delta^{-1/m}$ for commonly used norms $\|\cdot\|$; See the Supplementary Material, Section S2.5 (Tan and Zhang (2019)) for more.

3. General results. As in Section 1, consider the estimator

$$\hat{g} = \operatorname*{argmin}_{g \in \mathcal{G}} K_n(g), \qquad K_n(g) = \|Y - g\|_n^2/2 + A_0 R_n(g), \tag{10}$$

where $\mathcal{G} = \{g = \sum_{j=1}^p g_j : g_j \in \mathcal{G}_j\}$, $A_0 > 1$ is a constant, and the penalty is, up to the prefactor A_0 for technical convenience,

$$R_n(g) = \sum_{j=1}^p R_{nj}(g_j) = \sum_{j=1}^p \big(\rho_{nj} \|g_j\|_{F,j} + \lambda_{nj} \|g_j\|_n\big)$$

for any decomposition $g = \sum_{j=1}^p g_j$ with $g_j \in \mathcal{G}_j$. The regularization parameters $(\lambda_{nj}, \rho_{nj})$ are of the form

$$\rho_{nj} = \lambda_{nj} w_{nj}, \qquad \lambda_{nj} = C_1\big\{\gamma_{nj} + \sqrt{\log(p/\epsilon)/n}\big\}, \tag{11}$$

where $C_1 > 0$ is a noise level depending only on parameters in Assumption 1 below, $0 < \epsilon < 1$ is a tail probability for the validity of error bounds, $0 < w_{nj} \leq 1$ is a rate parameter and

$$\gamma_{nj} = n^{-1/2} \psi_{nj}(w_{nj})/w_{nj} \tag{12}$$

for a function $\psi_{nj}(\cdot)$ depending on the entropy of the unit ball of the space $\mathcal{G}_j$ under the associated functional penalty; see Assumption 2 or 4 below.

Before theoretical analysis, we briefly comment on computation of $\hat{g}$. By standard properties of norms and semi-norms, the objective function $K_n(g)$ is convex in g. Moreover, there are at least two situations where the infinite-dimensional problem of minimizing $K_n(g)$ can be reduced to a finite-dimensional one. First, if each class $\mathcal{G}_j$ is a reproducing kernel Hilbert space such as $\mathcal{W}_2^m$, then a solution $\hat{g} = \sum_{j=1}^p \hat{g}_j$ can be obtained such that each $\hat{g}_j$ is a smoothing spline with knots in the design points $\{X_i^{(j)} : i = 1, \ldots, n\}$ (e.g., Meier, van de Geer and Bühlmann (2009)). Second, by the following proposition, the optimization problem can also be reduced to a finite-dimensional one when each class $\mathcal{G}_j$ is the bounded variation space $\mathcal{V}^1$ or $\mathcal{V}^2$.

PROPOSITION 2. *Suppose that $\mathcal{G}_\ell$ is $\mathcal{V}^m$ on $[0, 1]$ for some $1 \le \ell \le p$ and $m \ge 1$. Then a solution $\hat{g} = \sum_{j=1}^p \hat{g}_j$ can be chosen such that $\hat{g}_\ell$ is a spline of order $m - 1$, that is, a piecewise polynomial of degree $m - 1$ and, if $m \ge 2$, an $(m - 2)$th continuously differentiable function. Moreover, $\hat{g}_\ell$ can be defined with knots only in $\{X_i^{(\ell)} : i = 1, \ldots, n\}$ if $m = 1$ or 2.*

The algorithm in Petersen, Witten and Simon (2016), based on the fused Lasso, can be directly used to compute $\hat{g}$ when all classes $(\mathcal{G}_1, \ldots, \mathcal{G}_p)$ are $\mathcal{V}^1$. In general, with a bounded variation class $\mathcal{V}^m$, Yang and Tan (2018) developed a backfitting algorithm for computing $\hat{g}$ as defined above for $m = 1$ or 2, or with the additional restriction that the knots of $\hat{g}$ are contained in the data points for $m \ge 3$. Numerical experiments from these papers showed superior performance of the doubly penalized method, compared with the existing methods as specified in Meier, van de Geer and Bühlmann (2009) and Ravikumar et al. (2009).

3.1. *Fixed designs.* For fixed designs, the covariates $(X_1, \ldots, X_n)$ are fixed as observed, whereas $(\varepsilon_1, \ldots, \varepsilon_n)$ and hence $(Y_1, \ldots, Y_n)$ are independent random variables. The responses are to be predicted when new observations are drawn with covariates from the sample $(X_1, \ldots, X_n)$. The predictive performance of $\hat{g}$ is measured by $\|\hat{g} - g^*\|_n^2$.

Consider the following three assumptions. First, we assume sub-Gaussian tails for the noises. This condition can be relaxed, but with increasing technical complexity and possible modification of the estimators, which we will not pursue here.

ASSUMPTION 1 (Sub-Gaussian noises). Assume that the noises $(\varepsilon_1, \ldots, \varepsilon_n)$ are mutually independent and uniformly sub-Gaussian: For some constants $D_0 > 0$ and $D_1 > 0$,

$$\max_{i=1,\ldots,n} D_0\{E \exp(\varepsilon_i^2/D_0) - 1\} \le D_1.$$

We will also impose this assumption for random designs with the interpretation that the aforementioned independence and expectation are taken conditionally on $(X_1, \ldots, X_n)$.

Second, we impose an entropy condition which describes the relationship between the function $\psi_{nj}(\cdot)$ in the definition of γ_{nj} and the complexity of bounded subsets in $\mathcal{G}_j$. Although entropy conditions are widely used to analyze nonparametric regression (e.g., Section 10.1, van de Geer (2000)), the subset $\mathcal{G}_j(\delta)$ in our entropy condition below is carefully aligned with the penalty $R_{nj}(g_j) = \lambda_{nj}(w_{nj}\|g_j\|_{F,j} + \|g_j\|_n)$. This leads to a delicate use of maximal inequalities so as to relax and in some cases remove restrictions in previous studies of additive models; see Lemma 1 in the Supplementary Material (Tan and Zhang (2019)) and Raskutti, Wainwright and Yu (2012), Lemma 1.

ASSUMPTION 2 (Entropy condition for fixed designs). For $j = 1, \ldots, p$, let $\mathcal{G}_j(\delta) = \{f_j \in \mathcal{G}_j : \|f_j\|_{F,j} + \|f_j\|_n/\delta \le 1\}$ and $\psi_{nj}(\delta)$ be an upper bound of the entropy integral as follows:

$$\psi_{nj}(\delta) \ge \int_0^\delta H^{1/2}\big(u, \mathcal{G}_j(\delta), \|\cdot\|_n\big)\,\mathrm{d}u, \qquad 0 < \delta \le 1. \tag{13}$$

In general, $\mathcal{G}_j(\delta)$ and the entropy $H(\cdot, \mathcal{G}_j(\delta), \|\cdot\|_n)$ may depend on the design points $\{X_i^{(j)}\}$.

Our third assumption is a compatibility condition, which resembles the restricted eigenvalue condition (Bickel, Ritov and Tsybakov (2009)) and the compatibility condition (van de Geer and Bühlmann (2009)) used in high-dimensional analysis of the Lasso in linear regression. We defer to Section 3.2 further discussion about compatibility conditions used in the analysis of additive regression.

ASSUMPTION 3 (Empirical compatibility condition). For certain subset $S \subset \{1, 2, \ldots, p\}$ and constants $\kappa_0 > 0$ and $\xi_0 > 1$, assume that

$$\kappa_0^2\Big(\sum_{j\in S}\lambda_{nj}\|f_j\|_n\Big)^2 \le \Big(\sum_{j\in S}\lambda_{nj}^2\Big)\|f\|_n^2$$

for any functions $\{f_j \in \mathcal{G}_j : j = 1, \ldots, p\}$ and $f = \sum_{j=1}^p f_j \in \mathcal{G}$ satisfying

$$\sum_{j=1}^p \lambda_{nj} w_{nj}\|f_j\|_{F,j} + \sum_{j\in S^c}\lambda_{nj}\|f_j\|_n \le \xi_0 \sum_{j\in S}\lambda_{nj}\|f_j\|_n.$$

REMARK 1. The subset S can be different from $\{1 \le j \le p : g_j^* \ne 0\}$. In fact, S is arbitrary in the sense that a larger S leads to a smaller compatibility coefficient κ_0 which appears as a factor in the denominator of the "noise" term in the prediction error bound below, whereas a smaller S leads to a larger "bias" term. Assumption 3 is automatically satisfied for the choice $S = \varnothing$. In this case, it is possible to take $\xi_0 = \infty$ and any $\kappa_0 > 0$, provided that we treat summation over an empty set as 0 and $\infty \times 0$ as 0.

Our main result for fixed designs is an oracle inequality stated in Theorem 1 below, where $\bar{g} = \sum_{j=1}^{p} \bar{g}_j \in \mathcal{G}$ as an estimation target is an additive function but the true regression function g^* may not be additive. Our oracle inequality (14) is sharp with the coefficient of $\|\hat{g} - g^*\|_n^2$ matching that of $\|\bar{g} - g^*\|_n^2$, similarly as in Koltchinskii, Lounici and Tsybakov (2011). For $\xi \in (0, 1]$, denote as a penalized prediction loss

$$\mathcal{D}_n(\hat{g}, \bar{g}, \xi) = \frac{1}{2}\|\hat{g} - g^*\|_n^2 + \frac{\xi}{2}\|\hat{g} - \bar{g}\|_n^2 + \xi(A_0 - 1)R_n(\hat{g} - \bar{g}).$$

For a subset $S \subset \{1, 2, \ldots, p\}$, write as a bias term for the target $\bar{g}$

$$\Delta_n(\bar{g}, S) = \frac{1}{2}\|\bar{g} - g^*\|_n^2 + 2A_0\left(\sum_{j=1}^{p} \rho_{nj}\|\bar{g}_j\|_{F,j} + \sum_{j \in S^c} \lambda_{nj}\|\bar{g}_j\|_n\right).$$

The bias term is small when $\bar{g}$ is smooth and sparse and predicts g^* well.

THEOREM 1. *Suppose that Assumptions 1, 2 and 3 hold for λ_{nj} and ρ_{nj} in (11). Then for any $A_0 > (\xi_0 + 1)/(\xi_0 - 1)$ we have with probability at least $1 - \epsilon$,*

$$\mathcal{D}_n(\hat{g}, \bar{g}, \xi_1) \le \Delta_n(\bar{g}, S) + \xi_2 A_0 \kappa_0^{-2}\left(\sum_{j \in S} \lambda_{nj}^2\right). \tag{14}$$

where $\xi_1 = 1 - 2A_0/\{(\xi_0 + 1)(A_0 - 1)\} \in (0, 1]$ and $\xi_2 = (\xi_0 + 1)(A_0 - 1)$.

REMARK 2. As seen from our proofs, Theorem 1 and subsequent corollaries are directly applicable to functional ANOVA modeling, where each function g_j may depend on $X_i^{(j)}$, a block of covariates and the variable blocks are allowed to overlap across different j. The entropy associated with the functional class $\mathcal{G}_j$ need to be determined accordingly.

Taking $S = \varnothing$ and $\xi_0 = \infty$ leads to the following corollary, which explicitly does not require the compatibility condition (Assumption 3).

COROLLARY 1. *Suppose that Assumptions 1 and 2 hold. Then for any $A_0 > 1$ we have with probability at least $1 - \epsilon$,*

$$\mathcal{D}_n(\hat{g}, \bar{g}, 1) \le \Delta_n(\bar{g}, \varnothing) = \frac{1}{2}\|\bar{g} - g^*\|_n^2 + 2A_0 R_n(\bar{g}). \tag{15}$$

The following result can be derived from Theorem 1 through the choice $S = \{1 \le j \le p : \|\bar{g}_j\|_n > C_0\lambda_{nj}\}$ for some constant $C_0 > 0$.

COROLLARY 2. *Suppose that Assumptions 1, 2 and 3 hold with $S = \{1 \le j \le p : \|\bar{g}_j\|_n > C_0\lambda_{nj}\}$ for some constant $C_0 > 0$. Then for any $0 \le q \le 1$ and*

$A_0 > (\xi_0 + 1)/(\xi_0 - 1)$ *we have, with probability at least* $1 - \epsilon$,

$$\mathcal{D}_n(\hat{g}, \bar{g}, \xi_1) \le \frac{1}{2}\|\bar{g} - g^*\|_n^2 + O(1)\sum_{j=1}^{p}\bigl(\rho_{nj}\|\bar{g}_j\|_{F,j} + \lambda_{nj}^{2-q}\|\bar{g}_j\|_n^q\bigr),$$

where $O(1)$ *depends only on* $(q, A_0, C_0, \xi_0, \kappa_0)$.

It is instructive to examine the implications of Corollary 2 in a homogenous situation where for some constants $B_0 > 0$ and $0 < \beta_0 < 2$,

$$\max_{j=1,\ldots,p}\int_0^{\delta} H^{1/2}\bigl(u, \mathcal{G}_j(\delta), \|\cdot\|_n\bigr)\,\mathrm{d}u \le B_0\delta^{1-\beta_0/2}, \qquad 0 < \delta \le 1. \tag{16}$$

That is, we assume $\psi_{nj}(\delta) = B_0\delta^{1-\beta_0/2}$ in (13). For $j = 1, \ldots, p$, let

$$w_{nj} = w_n(q) = \{\gamma_n(q)\}^{1-q}, \qquad \gamma_{nj} = \gamma_n(q) = B_0^{\frac{2}{2+\beta_0(1-q)}} n^{\frac{-1}{2+\beta_0(1-q)}}, \tag{17}$$

which are determined by balancing the two rates $\rho_{nj} = \lambda_{nj}^{2-q}$, that is, $w_{nj} = \lambda_{nj}^{1-q}$, along with the definition $\gamma_{nj} = B_0 n^{-1/2} w_{nj}^{-\beta_0/2}$ by (12). For $g = \sum_{j=1}^{p} g_j \in \mathcal{G}$, denote $\|g\|_{F,1} = \sum_{j=1}^{p}\|g_j\|_{F,j}$ and $\|g\|_{n,q} = \sum_{j=1}^{p}\|g_j\|_n^q$. For simplicity, we also assume that g^* is an additive function and set $\bar{g} = g^*$ for Corollary 3.

COROLLARY 3. *Assume that* (1) *holds and* $\|g^*\|_{F,1} \le C_1 M_F$ *and* $\|g^*\|_{n,q} \le C_1^q M_q$ *for* $0 \le q \le 1$, $M_q > 0$, *and* $M_F > 0$, *possibly depending on* (n, p). *In addition, suppose that* (16) *and* (17) *hold, and Assumptions* 1 *and* 3 *are satisfied with* $S = \{1 \le j \le p : \|g_j^*\|_n > C_0\lambda_{nj}\}$ *for some constant* $C_0 > 0$. *If* $0 < w_n(q) \le 1$ *for sufficiently large* n, *then for any* $A_0 > (\xi_0 + 1)/(\xi_0 - 1)$, *we have with probability at least* $1 - \epsilon$,

$$\begin{aligned}\mathcal{D}_n(\hat{g}, g^*, \xi_1) &= \frac{1 + \xi_1}{2}\|\hat{g} - g^*\|_n^2 + \xi_1(A_0 - 1)R_n(\hat{g} - g^*) \\ &\le O(1)C_1^2(M_F + M_q)\bigl\{\gamma_n(q) + \sqrt{\log(p/\epsilon)/n}\bigr\}^{2-q},\end{aligned} \tag{18}$$

where $O(1)$ *depends only on* $(q, A_0, C_0, \xi_0, \kappa_0)$.

REMARK 3. There are several interesting features in the convergence rate (18). First, (18) presents a spectrum of convergence rates in the form

$$\{n^{\frac{-1}{2+\beta_0(1-q)}} + \sqrt{\log(p)/n}\}^{2-q},$$

which are easily shown to become slower as q increases from 0 to 1, that is, the exponent $(2 - q)/\{2 + \beta_0(1 - q)\}$ is decreasing in q for $0 < \beta_0 < 2$. The rate (18) gives the slow rate $\{\log(p)/n\}^{1/2}$ for $q = 1$, or the fast rate $n^{\frac{-2}{2+\beta_0}} + \log(p)/n$

for $q = 0$, as previously obtained for additive regression with reproducing kernel Hilbert spaces. We defer to Section 4 the comparison with existing results in random designs. Second, the rate (18) is in general at least as fast as

$$\{n^{\frac{-1}{2+\beta_0}} + \sqrt{\log(p)/n}\}^{2-q}.$$

Therefore, weaker sparsity (larger q) leads to a slower rate of convergence, but not as slow as the fast rate $\{n^{\frac{-2}{2+\beta_0}} + \log(p)/n\}$ raised to the power of $(2-q)/2$. This is in contrast with previous results on penalized estimation over L_q sparsity balls, for example, the rate $\{k/n + \log(p)/n\}^{(2-q)/2}$ obtained for group Lasso estimation in linear regression (Neghaban et al. (2012)), where k is the group size. Third, the rate (18) is in general not as fast as the following rate (unless $q = 0$ or 1):

$$n^{\frac{-2}{2+\beta_0}} + \{\log(p)/n\}^{(2-q)/2},$$

which was obtained by Yuan and Zhou (2016) using constrained least squares for additive regression with reproducing kernel Hilbert spaces under an L_q ball in the Hilbert norm: $\sum_{j=1}^{p} \|g_j^*\|_H^q \le M_q$. This difference can be explained by the fact that an L_q ball in $\|\cdot\|_H$ norm is more restrictive than in $\|\cdot\|_n$ or $\|\cdot\|_Q$ norm for our results.

3.2. *Random designs*. For random designs, prediction of the responses can be sought when new observations are randomly drawn with covariates from the distributions of $(X_1, \ldots, X_n)$, instead of within the sample $(X_1, \ldots, X_n)$ as in Section 3.1. For such out-of-sample prediction, the performance of $\hat{g}$ is measured by $\|\hat{g} - g^*\|_Q^2$, where $\|\cdot\|_Q$ denotes the theoretical norm: $\|f\|_Q^2 = n^{-1}\sum_{i=1}^{n} E\{f^2(X_i)\}$ for a function $f(x)$.

Consider the following extensions of Assumptions 2 and 3, such that dependency on the empirical norm $\|\cdot\|_n$ and hence on $(X_1, \ldots, X_n)$ are removed.

ASSUMPTION 4 (Entropy condition for random designs). For some constant $0 < \eta_0 < 1$ and $j = 1, \ldots, p$, let $\psi_{nj}(\delta)$ be an upper bound of the entropy integral, independent of the realizations $\{X_i^{(j)} : i = 1, \ldots, n\}$, as follows:

$$\psi_{nj}(\delta) \ge \int_0^{\delta} H^{*1/2}\big((1-\eta_0)u, \mathcal{G}_j^*(\delta), \|\cdot\|_n\big)\,\mathrm{d}u, \qquad 0 < \delta \le 1, \tag{19}$$

where $\mathcal{G}_j^*(\delta) = \{f_j \in \mathcal{G}_j : \|f_j\|_{F,j} + \|f_j\|_Q/\delta \le 1\}$ and

$$H^*\big(u, \mathcal{G}_j^*(\delta), \|\cdot\|_n\big) = \sup_{(X_1^{(j)}, \ldots, X_n^{(j)})} H\big(u, \mathcal{G}_j^*(\delta), \|\cdot\|_n\big).$$

REMARK 4. For $\mathcal{G}_j$ defined as a Sobolev space $\mathcal{W}_r^m$ with $rm > 1$ or bounded variation space $\mathcal{V}^m$ with $m \ge 2$ on $[0, 1]$, the entropy $H(u, \mathcal{G}_j^*(1), \|\cdot\|_n)$ can be

upper-bounded by the standard estimate of $H(u, \mathcal{G}_j^*(1), \|\cdot\|_\infty)$, of order $u^{-1/m}$, independently of the the realizations $\{X_i^{(j)} : i = 1, \ldots, n\}$ (Lorentz, Golitschek and Makovoz (1996)). For the space $\mathcal{W}_1^1$ or $\mathcal{V}^1$, the entropy $H(u, \mathcal{G}_j^*(1), \|\cdot\|_n)$ can be obtained from Mammen (1991), still of order $u^{-1/m} = u^{-1}$, as described in the Supplementary Material, Section S2.5 (Tan and Zhang (2019)). Because $\int_0^\delta (u^{-1/m})^{1/2}\,du = \{2m/(2m-1)\}\delta^{1-1/(2m)}$, the resulting $\psi_{nj}(\delta)$ is of order $\delta^{1-1/(2m)}$. Further discussion is provided in Remarks 20 and 21.

ASSUMPTION 5 (Theoretical compatibility condition). For some subset $S \subset \{1, 2, \ldots, p\}$ and constants $\kappa_0^* > 0$ and $\xi_0^* > 1$, assume that for any functions $\{f_j \in \mathcal{G}_j : j = 1, \ldots, p\}$ and $f = \sum_{j=1}^p f_j \in \mathcal{G}$, if

$$\sum_{j=1}^p \lambda_{nj} w_{nj} \|f_j\|_{F,j} + \sum_{j \in S^c} \lambda_{nj} \|f_j\|_Q \le \xi_0^* \sum_{j \in S} \lambda_{nj} \|f_j\|_Q, \tag{20}$$

then

$$\kappa_0^{*2} \Big(\sum_{j \in S} \lambda_{nj} \|f_j\|_Q \Big)^2 \le \Big(\sum_{j \in S} \lambda_{nj}^2 \Big) \|f\|_Q^2. \tag{21}$$

REMARK 5. Similarly as in Remark 1 about the empirical compatibility condition, Assumption 5 is also automatically satisfied for the choice $S = \varnothing$, in which case it is possible to take $\xi_0^* = \infty$ and any $\kappa_0^* > 0$.

REMARK 6. We discuss the fact that the compatibility assumption involves the tuning parameters (w_{nj}, λ_{nj}). On one hand, in the special case where $(w_{nj}, \gamma_{nj}) \equiv (w_n, \gamma_n)$ for $j = 1, \ldots, p$, Assumption 5 says that if $w_n \times \sum_{j=1}^p \|f_j\|_{F,j} + \sum_{j \in S^c} \|f_j\|_Q \le \xi_0^* \sum_{j \in S} \|f_j\|_Q$, then $\kappa_0^{*2} (\sum_{j \in S} \|f_j\|_Q)^2 \le |S| \|f\|_Q^2$. Because $w_n > 0$, a sufficient condition for this to hold is that

$$\text{if } \sum_{j \in S^c} \|f_j\|_Q \le \xi_0^* \sum_{j \in S} \|f_j\|_Q, \text{ then } \kappa_0^{*2} \Big(\sum_{j \in S} \|f_j\|_Q \Big)^2 \le |S| \cdot \|f\|_Q^2, \tag{22}$$

which, by the Cauchy–Schwartz inequality, is satisfied under the following condition as used in Koltchinskii and Yuan (2010) and Suzuki and Sugiyama (2013):

$$\text{if } \sum_{j \in S^c} \|f_j\|_Q \le \xi_0^* \sum_{j \in S} \|f_j\|_Q, \text{ then } \kappa_0^{*2} \sum_{j \in S} \|f_j\|_Q^2 \le \|f\|_Q^2. \tag{23}$$

Therefore, Assumption 5 is strictly weaker than previous compatibility conditions in the homogeneous setting. On the other hand, there are implications of Assumption 5 in heterogenous settings. If $\lambda_{nj}/(\max_{\ell \in S} \lambda_{n\ell}) \to \infty$ for some $j \in S^c$, then the cone condition (20) essentially restricts $\|f_j\|_Q \approx 0$, which seems harmless to

whether (21) is satisfied. If $\lambda_{nj}/(\max_{\ell\in S}\lambda_{n\ell}) \to 0$ for some $j \in S^c$, then (20) effectively leaves the magnitude of $\|f_j\|_Q$ unrestricted, which roughly imply that $f_j(X_i^{(j)})$ cannot be highly correlated with $\{f_\ell(X_i^{(\ell)}) : \ell \in S\}$ in order for (21) to hold. As compatibility conditions are often invoked with $S = \{1 \le j \le p : \|g_j^*\|_Q > 0\}$ depending on unknown $\|g_j^*\|_Q$, the latter observation suggests that in the presence of function classes of different smoothness, Assumption 5 essentially requires that the correlations between component functions specified with smoother classes and the truly nonzero component functions be bounded away from 1. For example, this restriction is similar to Condition 2.3 in Müller and van de Geer (2015), where smoother components are linear functions of components.

To tackle random designs, our approach relies on establishing appropriate convergence of empirical norms $\|\cdot\|_n$ to $\|\cdot\|_Q$ uniformly over the space of additive functions $\mathcal{G}$, similarly as in Meier, van de Geer and Bühlmann (2009) and Koltchinskii and Yuan (2010). For clarity, we postulate the following assumption on the rate of such convergence to develop general analysis of $\hat{g}$. We will study convergence of empirical norms specifically for Sobolev and bounded variation spaces in Section 5, and then provide corresponding results on the performance of $\hat{g}$ in Section 4. For $g = \sum_{j=1}^p g_j \in \mathcal{G}$, denote

$$R_n^*(g) = \sum_{j=1}^p R_{nj}^*(g_j), \qquad R_{nj}^*(g_j) = \lambda_{nj}\big(w_{nj}\|g_j\|_{F,j} + \|g_j\|_Q\big),$$

as the population version of the penalty $R_n(g)$, with $\|g_j\|_Q$ in place of $\|g_j\|_n$.

ASSUMPTION 6 (Convergence of empirical norms). Assume that

$$P\left\{\sup_{g\in\mathcal{G}} \frac{|\|g\|_n^2 - \|g\|_Q^2|}{R_n^{*2}(g)} > \phi_n\right\} \le \pi, \tag{24}$$

where $0 < \pi < 1$ and $\phi_n > 0$ are such that for sufficiently large n, one or both of the following conditions are valid:

(i) $\phi_n(\max_{j=1,\dots,p}\lambda_{nj}^2) \le \eta_0^2$, where η_0 is from Assumption 4.
(ii) For some constant $0 \le \eta_1 < 1$, we have

$$\phi_n(\xi_0^* + 1)^2\kappa_0^{*-2}\left(\sum_{j\in S}\lambda_{nj}^2\right) \le \eta_1^2, \tag{25}$$

where S is the subset of $\{1, 2, \dots, p\}$ used in Assumption 5.

Our main result, Theorem 2, gives an oracle inequality for random designs, where the predictive performance of $\hat{g}$ is compared with that of an arbitrary additive function $\bar{g} = \sum_{j=1}^p \bar{g}_j \in \mathcal{G}$, but the true regression function g^* may not be additive, similarly as in Theorem 1 for fixed designs. For a subset $S \subset \{1, 2, \dots, p\}$,

denote

$$\Delta_n^*(\bar{g}, S) = \frac{1}{2}\|\bar{g} - g^*\|_n^2 + 2A_0(1-\eta_0)\left(\sum_{j=1}^{p} \rho_{nj}\|\bar{g}_j\|_{F,j} + \sum_{j\in S^c} \lambda_{nj}\|\bar{g}_j\|_Q\right),$$

which, unlike $\Delta_n(\bar{g}, S)$, involves $\|\bar{g}_j\|_Q$ and η_0 from Assumptions 4 and 6(i).

THEOREM 2. *Suppose that Assumptions* 1, 4, 5 *and* 6(i)–(ii) *hold with* $0 < \eta_0 < (\xi_0^* - 1)/(\xi_0^* + 1)$, *for* λ_{nj} *and* ρ_{nj} *in* (11). *Let* $A(\xi_0^*, \eta_0) = \{\xi_0^* + 1 + \eta_0(\xi_0^* + 1)\}/\{\xi_0^* - 1 - \eta_0(\xi_0^* + 1)\} > (1+\eta_0)/(1-\eta_0)$. *Then for any* $A_0 > A(\xi_0^*, \eta_0)$, *we have with probability at least* $1 - \epsilon - \pi$,

$$\frac{1}{2}\|\hat{g} - g^*\|_n^2 + \frac{\xi_1^*}{2}\|\hat{g} - \bar{g}\|_n^2 + \xi_1^* A_1 R_n^*(\hat{g} - \bar{g})$$

$$\le \Delta_n^*(\bar{g}, S) + \xi_2^* A_0 \kappa_0^{*-2}\left(\sum_{j\in S} \lambda_{nj}^2\right), \tag{26}$$

where $A_1 = (A_0 - 1) - \eta_0(A_0 + 1) > 0$, $\xi_1^* = 1 - 2A_0/\{(\xi_0^* + 1)A_1\} \in (0, 1]$ *and* $\xi_2^* = (\xi_0^* + 1)A_1$. *Moreover, we have with probability at least* $1 - \epsilon - \pi$,

$$\mathcal{D}_n^*(\hat{g}, \bar{g}, \xi_1^*, \eta_1) := \frac{1}{2}\|\hat{g} - g^*\|_n^2 + \frac{\xi_3^*}{2}\|\hat{g} - \bar{g}\|_Q^2 + \xi_1^* A_1 R_n^*(\hat{g} - \bar{g})$$

$$\le \Delta_n^*(\bar{g}, S) + \xi_4^* A_0 \kappa_0^{*-2}\left(\sum_{j\in S} \lambda_{nj}^2\right)$$

$$+ \frac{\phi_n}{2A_1^2}\xi_1^{*-2}\Delta_n^{*2}(\bar{g}, S), \tag{27}$$

where $\xi_3^* = \xi_1^*(1 - \eta_1^2)$ *and* $\xi_4^* = \xi_2^*/(1 - \eta_1^2)$.

REMARK 7. Similarly as in Remark 2, we emphasize that Theorem 2 and subsequent corollaries are also applicable to functional ANOVA modeling (e.g., Gu (2002)). For example, consider model (1) studied in Yang and Tokdar (2015), where each g_j^* is assumed to depend only on d_0 of a total of d covariates and lie in a Hölder space with smoothness level α_0. Then $p = \binom{d}{d_0}$, and the entropy condition (31) holds with $\beta_0 = d_0/\alpha_0$. Under certain additional conditions, Corollary 6 with $q = 0$ shows that penalized estimation studied here achieves a convergence rate $M_0 n^{\frac{-2}{2+\beta_0}} + M_0 \log(p)/n$ under exact sparsity of size M_0, where $n^{\frac{-2}{2+\beta_0}}$ is the rate for estimation of a single regression function in the Hölder class in dimension d_0 with smoothness β_0^{-1}, and $\log(p)/n \asymp d_0 \log(d/d_0)/n$ is the term associated with handling p regressors. This result agrees with the minimax rate derived in Yang and Tokdar (2015), but can be applied when more general functional classes are used such as multidimensional Sobolev spaces. In addition, Yang and Tokdar (2015) considered adaptive Bayes estimators which are nearly minimax with some extra logarithmic factor in n.

Taking $S = \varnothing$, $\xi_0^* = \infty$, and $\eta_1 = 0$ leads to the following corollary, which explicitly does not require the theoretical compatibility condition (Assumption 5) or the rate condition, Assumption 6(ii), for convergence of empirical norms.

COROLLARY 4. *Suppose that Assumptions 1, 4 and 6(i) hold. Then for any* $A_0 > (1+\eta_0)/(1-\eta_0)$, *we have with probability at least* $1-\epsilon-\pi$,

$$\frac{1}{2}\|\hat{g}-g^*\|_n^2 + \frac{1}{2}\|\hat{g}-\bar{g}\|_n^2 + A_1 R_n^*(\hat{g}-\bar{g}) \le \Delta_n^*(\bar{g},\varnothing) = \lambda_{n0}^2 + \frac{1}{2}\|\bar{g}-g^*\|_n^2 + 2A_0 R_n^*(\bar{g}). \tag{28}$$

Moreover, we have with probability at least $1-\epsilon-\pi$,

$$\frac{1}{2}\|\hat{g}-g^*\|_n^2 + \frac{1}{2}\|\hat{g}-\bar{g}\|_Q^2 + A_1 R_n^*(\hat{g}-\bar{g}) \le \Delta_n^*(\bar{g},\varnothing) + \frac{\phi_n}{2A_1^2}\Delta_n^{*2}(\bar{g},\varnothing). \tag{29}$$

The preceding results deal with both in-sample and out-of-sample prediction. For space limitation, except in Proposition 3, we hereafter focus on the more challenging out-of-sample prediction. Under some rate condition about ϕ_n in (24), the additional term involving $\phi_n\Delta_n^{*2}(\bar{g},S)$ can be absorbed into the first term, as shown in the following corollary. Two possible scenarios are accommodated. On one hand, taking $\bar{g}=g^*$ directly gives high-probability bounds on the prediction error $\|\hat{g}-g^*\|_Q^2$ provided that g^* is additive, that is, model (1) is correctly specified. On the other hand, the error $\|\hat{g}-g^*\|_Q^2$ can also be bounded, albeit in probability, in terms of an arbitrary additive function $\bar{g}\in\mathcal{G}$, while allowing g^* to be nonadditive.

COROLLARY 5. *Suppose that the conditions of Theorem 2 hold with* $S = \{1 \le j \le p : \|\bar{g}_j\|_Q > C_0^*\lambda_{nj}\}$ *for some constant* $C_0^* > 0$, *and (24) holds with* $\phi_n > 0$ *also satisfying*

$$\phi_n\left(\sum_{j=1}^p \rho_{nj}\|\bar{g}_j\|_{F,j} + \sum_{j\in S^c}\lambda_{nj}\|\bar{g}_j\|_Q\right) \le \eta_2, \tag{30}$$

for some constant $\eta_2 > 0$. *Then for any* $0 \le q \le 1$ *and* $A_0 > A(\xi_0^*,\eta_0)$, *we have with probability at least* $1-\epsilon-\pi$,

$$\mathcal{D}_n^*(\hat{g},\bar{g},\xi_1^*,\eta_1) \le \{O(1)+\phi_n\|\bar{g}-g^*\|_n^2\}\left\{\|\bar{g}-g^*\|_n^2 + \sum_{j=1}^p(\rho_{nj}\|\bar{g}_j\|_{F,j} + \lambda_{nj}^{2-q}\|\bar{g}_j\|_Q^q)\right\},$$

where $O(1)$ *depends only on* $(q, A_0^*, C_0^*, \xi_0^*, \kappa_0^*, \eta_0, \eta_1, \eta_2)$. *In addition, suppose that* $\phi_n \|\bar{g} - g^*\|_Q^2$ *is bounded by a constant and* $\epsilon = \epsilon(n, p)$ *tends to* 0 *in the definition* (11) *of* $(\lambda_{nj}, \rho_{nj})$ *and* $R_n(g)$ *for* $\hat{g}$ *in* (10). *Then for any* $0 \le q \le 1$, *we have*

$$\|\hat{g} - g^*\|_Q^2 \le O_p(1)\left\{\|\bar{g} - g^*\|_Q^2 + \sum_{j=1}^{p}\bigl(\rho_{nj}\|\bar{g}_j\|_{F,j} + \lambda_{nj}^{2-q}\|\bar{g}_j\|_Q^q\bigr)\right\}.$$

Similarly as Corollary 3, it is useful to deduce the following result in a homogeneous situation where we assume $\psi_{nj}(\delta) = B_0^*\delta^{1-\beta_0/2}$ in (19) for some constants $B_0^* > 0$ and $0 < \beta_0 < 2$:

$$\max_{j=1,\dots,p}\int_0^\delta H^{*1/2}\bigl((1-\eta_0)u, \mathcal{G}_j^*(\delta), \|\cdot\|_n\bigr)\,\mathrm{d}u \le B_0^*\delta^{1-\beta_0/2}, \qquad 0 < \delta \le 1. \tag{31}$$

By Remark 4, this assumption is satisfied with $\beta_0 = 1/m$ and B_0^* from existing entropy estimates when each $\mathcal{G}_j$ is a Sobolev space $\mathcal{W}_r^m$ or bounded variation space $\mathcal{V}^m$ on $[0, 1]$ with $r \ge 1$ and $m \ge 1$ under nonvanishing marginal densities of $X_i^{(j)}$. For $j = 1, \dots, p$, let

$$w_{nj} = w_n^*(q) = \max\{\gamma_n(q)^{1-q}, \nu_n^{1-q}\}, \tag{32}$$

$$\gamma_{nj} = \gamma_n^*(q) = \min\{\gamma_n(q), B_0^* n^{-1/2}\nu_n^{-(1-q)\beta_0/2}\}, \tag{33}$$

where $\nu_n = \{\log(p/\epsilon)/n\}^{1/2}$, and $w_n(q) = \gamma_n(q)^{1-q}$ and

$$\gamma_n(q) = B_0^{*\frac{2}{2+\beta_0(1-q)}} n^{\frac{-1}{2+\beta_0(1-q)}} \asymp n^{\frac{-1}{2+\beta_0(1-q)}}$$

are determined from the relationship (12), that is, $\gamma_n(q) = B_0^* n^{-1/2} w_n(q)^{-\beta_0/2}$. The reason for why $(w_n^*(q), \gamma_n^*(q))$ are used instead of the simpler choices $(w_n(q), \gamma_n(q))$ is that the rate condition (34) needed below would become stronger if $\gamma_n^*(q)$ were replaced by $\gamma_n(q)$. The rate of convergence, however, remains the same even if $\gamma_n^*(q)$ is substituted for $\gamma_n(q)$ in (35); see Remark 14 for further discussion. For $g = \sum_{j=1}^p g_j \in \mathcal{G}$, denote $\|g\|_{F,1} = \sum_{j=1}^p \|g_j\|_{F,j}$ and $\|g\|_{Q,q} = \sum_{j=1}^p \|g_j\|_Q^q$.

COROLLARY 6. *Assume that* (1) *holds and* $\|g^*\|_{F,1} \le C_1 M_F$ *and* $\|g^*\|_{Q,q} \le C_1^q M_q$ *for* $0 \le q \le 1$, $M_q > 0$, *and* $M_F > 0$, *possibly depending on* (n, p). *In addition, suppose that* (31), (32), *and* (33) *hold, Assumptions* 1, 5 *and* 6(i) *are satisfied with* $0 < \eta_0 < (\xi_0^* - 1)/(\xi_0^* + 1)$ *and* $S = \{1 \le j \le p : \|g_j^*\|_Q > C_0^*\lambda_{nj}\}$ *for some constant* $C_0^* > 0$, *and* (24) *holds with* $\phi_n > 0$ *satisfying*

$$\phi_n C_1^2(M_F + M_q)\{\gamma_n^*(q) + \sqrt{\log(p/\epsilon)/n}\}^{2-q} = o(1). \tag{34}$$

Then for sufficiently large n, depending on (M_F, M_q) only through the convergence rate in (34), and any $A_0 > A(\xi_0^*, \eta_0)$, we have with probability at least $1 - \epsilon - \pi$,

$$\mathcal{D}_n^*(\hat{g}, g^*, \xi_1^*, 0) \le O(1) C_1^2 (M_F + M_q) \{\gamma_n(q) + \sqrt{\log(p/\epsilon)/n}\}^{2-q}, \tag{35}$$

where $O_P(1)$ depends only on $(q, A_0^*, C_0^*, \xi_0^*, \kappa_0^*, \eta_0)$.

In the case of $q \neq 0$, Corollary 6 can be improved by relaxing the rate condition (34) from $o(1)$ to $O(1)$ but requiring the following compatibility condition.

ASSUMPTION 7 (Monotone compatibility condition). For some subset $S \subset \{1, 2, \ldots, p\}$ and constants $\kappa_0^* > 0$ and $\xi_0^* > 1$, assume that for any functions $\{f_j \in \mathcal{G}_j : j = 1, \ldots, p\}$ and $f = \sum_{j=1}^p f_j \in \mathcal{G}$, if (20) holds then

$$\kappa_0^{*2} \sum_{j \in S} \|f_j\|_Q^2 \le \|f\|_Q^2. \tag{36}$$

REMARK 8. By the Cauchy–Schwartz inequality, (36) implies (21), and hence Assumption 7 is stronger than Assumption 5. However, there is a monotonicity in S for the validity of Assumption 7 with (36) used. In fact, for any subset $S' \subset S$ and any functions $\{f_j' \in \mathcal{G}_j : j = 1, \ldots, p\}$ and $f' = \sum_{j=1}^p f_j' \in \mathcal{G}$, if

$$\sum_{j=1}^p \lambda_{nj} w_{nj} \|f_j'\|_{F,j} + \sum_{j \in S'^c} \lambda_{nj} \|f_j'\|_Q \le \xi_0^* \sum_{j \in S'} \lambda_{nj} \|f_j'\|_Q,$$

then (20) holds with $f_j = f_j'$, $j = 1, \ldots, p$, and hence, via (36), implies

$$\|f'\|_Q^2 \ge \kappa_0^{*2} \sum_{j \in S} \|f_j'\|_Q^2 \ge \kappa_0^{*2} \sum_{j \in S'} \|f_j'\|_Q^2.$$

Therefore, if Assumption 7 holds for a subset S, then it also holds for any subset $S' \subset S$ with the same constants (ξ_0^*, κ^*).

REMARK 9. In the homogeneous setting where $(w_{nj}, \gamma_{nj}) \equiv (w_n, \gamma_n)$, Assumption 7 is implied by condition (23), which is used in Koltchinskii and Yuan (2010) and Suzuki and Sugiyama (2013), whereas Assumption 5 is implied by condition (22) as discussed in Remark 6. Conditions (22) and (23) are extensions from, respectively, the compatibility condition (van de Geer and Bühlmann (2009)) and restricted eigenvalue condition (Bickel, Ritov and Tsybakov (2009)) for the Lasso. The constant κ_0^* from the former condition can be much larger than that from the latter condition, as shown in van de Geer and Bühlmann (2009), Example 10.5). By similar reasoning, condition (22) or Assumption 5 may hold with a larger constant κ_0^* than (23) or Assumption 7.

COROLLARY 7. *Suppose that the conditions of Corollary 6 are satisfied with* $0 < q \le 1$ *(excluding* $q = 0$*), Assumption 7 holds instead of Assumption 5, and the following condition holds instead of* (34):

$$\phi_n C_1^2 (M_F + M_q)\{\gamma_n^*(q) + \sqrt{\log(p/\epsilon)/n}\}^{2-q} \le \eta_3, \tag{37}$$

for some constant $\eta_3 > 0$. *If* $0 < w_n^*(q) \le 1$ *for sufficiently large* n, *then for any* $A_0 > A(\xi_0^*, \eta_0)$, *inequality* (35) *holds with probability at least* $1 - \epsilon - \pi$, *where* $O(1)$ *depends only on* $(q, A_0^*, C_0^*, \xi_0^*, \kappa_0^*, \eta_0, \eta_3)$.

To demonstrate the flexibility of our approach and compare with related results, notably Suzuki and Sugiyama (2013), we provide another result in the context of Corollary 6 with (w_{nj}, γ_{nj}) allowed to depend on (M_F, M_q), in contrast with the choices (32)–(33) independent of (M_F, M_q). For $j = 1, \ldots, p$, let

$$w_{nj} = w_n^\dagger(q) = \max\{w_n'(q), \nu_n^{1-q}(M_q/M_F)\}, \tag{38}$$

$$\gamma_{nj} = \gamma_n^\dagger(q) = \min\{\gamma_n'(q), B_0^* n^{-1/2} \nu_n^{-(1-q)\beta_0/2} (M_q/M_F)^{-\beta_0/2}\}, \tag{39}$$

where $w_n'(q) = \gamma_n'(q)^{1-q}(M_q/M_F)$ and $\gamma_n'(q) = B_0^{*\frac{2}{2+\beta_0(1-q)}} n^{\frac{-1}{2+\beta_0(1-q)}} (M_q/M_F)^{\frac{-\beta_0}{2+\beta_0(1-q)}}$ are determined from the relationship $\gamma_n'(q) = B_0^* n^{-1/2} w_n'(q)^{-\beta_0/2}$ by (12). These choices are picked to balance the two rates: $\lambda_n w_n M_F$ and $\lambda_n^{2-q} M_q$, where w_n and λ_n denote the common values of w_{nj} and λ_{nj} for $j = 1, \ldots, p$.

COROLLARY 8. *Suppose that the conditions of Corollary 6 are satisfied except that* (w_{nj}, γ_{nj}) *are defined by* (38)–(39), *and the following condition holds instead of* (34):

$$\phi_n C_1^2 M_q \{\gamma_n^\dagger(q) + \sqrt{\log(p/\epsilon)/n}\}^{2-q} = o(1). \tag{40}$$

Then for sufficiently large n, *depending on* (M_F, M_q) *only through the convergence rate in* (40), *and any* $A_0 > A(\xi_0^*, \eta_0)$, *we have with probability at least* $1 - \epsilon - \pi$,

$$\mathcal{D}_n^*(\hat{g}, g^*, \xi_1^*, 0) \le O(1) C_1^2 \{M_q^{\frac{2-\beta_0}{2+\beta_0(1-q)}} M_F^{\frac{(2-q)\beta_0}{2+\beta_0(1-q)}} n^{\frac{-(2-q)}{2+\beta_0(1-q)}} + M_q \nu_n^{2-q}\}, \tag{41}$$

where $O(1)$ *depends only on* $(q, B_0^*, A_0^*, C_0^*, \xi_0^*, \kappa_0^*, \eta_0)$.

REMARK 10. In the special case of $q = 0$ (exact sparsity), the convergence rate (41) reduces to $M_0^{\frac{2-\beta_0}{2+\beta_0}} M_F^{\frac{2\beta_0}{2+\beta_0}} n^{\frac{-2}{2+\beta_0}} + M_0 \nu_n^2$. The same rate was obtained in Suzuki and Sugiyama (2013) for additive regression with reproducing kernel Hilbert spaces under

$$\sum_{j=1}^p \|g_j^*\|_Q^0 \le M_0, \qquad \sum_{j=1}^p \|g_j^*\|_H \le M_F \ (\le c M_0),$$

where $\|g_j^*\|_H$ is the Hilbert norm, assumed to satisfy $\|g_j^*\|_H \le c$ for all j. As one of their main points, the rate (41) with $q = 0$ was argued to be faster than the rate $M_0 n^{\frac{-2}{2+\beta_0}} + M_0 \nu_n^2 \asymp (M_F + M_0)(n^{\frac{-2}{2+\beta_0}} + \nu_n^2)$ in (35) with $q = 0$, in the case where $\sum_{j=1}^p \|g_j^*\|_Q^0$ and $\sum_{j=1}^p \|g_j^*\|_H$ are of different orders. Our analysis sheds new light on the relationship between the rates (35) and (41): their difference mainly lies in whether the tuning parameters (w_{nj}, γ_{nj}) are chosen independently of (M_F, M_0) or depending on (M_F, M_0).

4. Minimax rates with Sobolev and bounded variation spaces. For concreteness, consider a fully homogeneous situation where each class $\mathcal{G}_j$ is a Sobolev space $\mathcal{W}_{r_0}^{m_0}$ for some constants $r_0 \ge 1$ and $m_0 \ge 1$ or a bounded variation space $\mathcal{V}^{m_0}$ for $r_0 = 1$ and $m_0 \ge 1$ on $[0, 1]$. Denote $\beta_0 = 1/m_0$. We deduce several explicit rates of convergence for the predictive performance of $\hat{g}$ by combining the results in Section 3.2 and those on convergence of empirical norms involved in Assumption 6, which are deferred to Section 5. We also demonstrate that the obtained rates match minimax lower bounds.

To facilitate justification of conditions related to Assumptions 4 and 6, consider the following assumption on the marginal densities of the covariates, as commonly imposed when handling random designs (e.g., Stone (1982)).

ASSUMPTION 8 (Nonvanishing marginal densities). For $j = 1, \ldots, p$, denote by $q_j(x^{(j)})$ the average marginal density function of $(X_1^{(j)}, \ldots, X_n^{(j)})$, that is, the density function associated with the probability measure $n^{-1} \sum_{i=1}^n Q_{X_i^{(j)}}$, where $Q_{X_i^{(j)}}$ is the marginal distribution of $X_i^{(j)}$. For some constant $0 < \varrho_0 \le 1$, assume that $q_j(x^{(j)})$ is bounded from below by ϱ_0 on $[0, 1]$ simultaneously for $j = 1, \ldots, p$.

We distinguish two cases, $r_0 > \beta_0$ or $r_0 = \beta_0 = 1$. First, under Assumption 8, if $r_0 > \beta_0$, then Assumption 4 (entropy in the empirical norm) and Assumption 10 (entropy in the empirical supremum norm) are satisfied such that $\psi_{nj}(\delta) = B_0^* \delta^{1-\beta_0/2}$ and $\psi_{nj,\infty}(z, \delta) = O(1) B_0^* z^{1-\beta_0/2}$ for $z > 0$ and $0 < \delta \le 1$, where $B_0^* > 0$ is a constant depending on ϱ_0 among others. See Theorem 4 and Remark 20 in Section 5 for the use of Assumption 10 and related discussion. Second, by Remark 21, if $r_0 = \beta_0 = 1$, then Assumptions 4 and 10 are satisfied such that $\psi_{nj}(\delta) = B_0^* \delta^{1/2}$ and $\psi_{nj,\infty}(z, \delta) = O(\log^{1/2}(n)) B_0^* z^{1/2}$ for $z > 0$ and $0 < \delta \le 1$, even when Assumption 8 does not hold. As a result, Γ_n in Section 5 reduces to

$$\Gamma_n = O(1) \quad \text{if } r_0 > \beta_0 \quad \text{or} \quad O\big(\log^{1/2}(n)\big) \quad \text{if } r_0 = \beta_0 = 1. \tag{42}$$

Informally, Γ_n is the ratio of the prefactors in $\psi_{nj,\infty}(z, \delta)$ and $\psi_{nj}(\delta)$.

4.1. *Achieved convergence rates.* We present our results on the convergence rates of $\hat{g}$ in three cases, where the underlying function $g^* = \sum_{j=1}^p g_j^*$ is assumed to satisfy (3) with $q = 1$, $q = 0$, or $0 < q < 1$. As discussed in Section 1, the parameter set (3) decouples sparsity and smoothness, inducing sparsity at different levels through an L_q ball in $\|\cdot\|_Q$ norm for $0 \le q \le 1$, while only enforcing smoothness through an L_1 ball in $\|\cdot\|_F$ norm on the components $(g_1^*, \ldots, g_p^*)$.

The first result deals with the case $q = 1$ for the parameter set (3).

PROPOSITION 3. *Assume that* (1) *holds and* $\|g^*\|_{F,1} \le C_1 M_F$ *and* $\|g^*\|_{Q,1} \le C_1 M_1$ *for* $M_F > 0$ *and* $M_1 > 0$, *possibly depending on* (n, p). *Let* $w_{nj} = 1$ *and* $\gamma_{nj} = \gamma_n^*(1) \asymp n^{-1/2}$ *by* (32)–(33). *Suppose that Assumptions* 1 *and* 8 *hold, and* $\log(p/\epsilon) = o(n)$. *Then for sufficiently large* n, *independently of* (M_F, M_1), *and any* $A_0 > (1+\eta_0)/(1-\eta_0)$, *we have with probability at least* $1 - 2\epsilon$,

$$\|\hat{g} - g^*\|_n^2 + A_1 R_n^*(\hat{g} - g^*) \le O(1) C_1^2 (M_F + M_1) \sqrt{\log(p/\epsilon)/n},$$

where $O(1)$ *depends only on* $(B_0^*, A_0, \eta_0, \varrho_0)$. *Moreover, we have*

$$\frac{1}{2}\|\hat{g} - g^*\|_Q^2 + A_1 R_n^*(\hat{g} - g^*) \le O(1) C_1^2 (M_F^2 + M_1^2)\{n^{-1/2}\Gamma_n + \sqrt{\log(p/\epsilon)/n}\},$$

with probability at least $1 - 2\epsilon$, *where* Γ_n *is from* (42) *and* $O(1)$ *depends only on* $(B_0^*, A_0, \eta_0, \varrho_0)$ *and* (C_2, C_3, C_4) *as in Theorem* 4. *If* $r_0 = \beta_0 = 1$, *then the results are valid even when Assumption* 8, *and hence* ϱ_0 *are removed.*

REMARK 11 (Comparison with existing results). Proposition 3 leads to the slow rate $\{\log(p)/n\}^{1/2}$ under L_1-ball sparsity in $\|\cdot\|_Q$ norm, as previously obtained for additive regression with Sobolev Hilbert spaces in Ravikumar et al. (2009), except in the case where $r_0 = \beta_0 = 1$, that is, each class $\mathcal{G}_j$ is $\mathcal{W}_1^1$ or $\mathcal{V}^1$. In the latter case, Proposition 3 shows that the convergence rate is $\{\log(np)/n\}^{1/2}$ for out-of-sample prediction, but remains $\{\log(p)/n\}^{1/2}$ for in-sample prediction. Previously, only the slower rate, $\{\log(np)/n\}^{1/2}$, was obtained for in-sample prediction in additive regression with the bounded variation space $\mathcal{V}^1$ by Petersen, Witten and Simon (2016).

The second result deals with the case $q = 0$ for the parameter set (3).

PROPOSITION 4. *Assume that* (1) *holds and* $\|g^*\|_{F,1} \le C_1 M_F$ *and* $\|g^*\|_{Q,0} \le M_0$ *for* $M_F > 0$ *and* $M_0 > 0$, *possibly depending on* (n, p). *By* (32)–(33), *let*

$$w_{nj} = w_n^*(0) = \max\left\{ B_0^{*\frac{2}{2+\beta_0}} n^{\frac{-1}{2+\beta_0}}, \left(\frac{\log(p/\epsilon)}{n}\right)^{1/2} \right\},$$

$$\gamma_{nj} = \gamma_n^*(0) = \min\left\{ B_0^{*\frac{2}{2+\beta_0}} n^{\frac{-1}{2+\beta_0}}, B_0^* n^{-1/2} \left(\frac{\log(p/\epsilon)}{n}\right)^{-\frac{\beta_0}{4}} \right\}.$$

Suppose that Assumptions 1, 5 and 8 hold with $0 < \eta_0 < (\xi_0^* - 1)/(\xi_0^* + 1)$ *and* $S = \{1 \le j \le p : \|g_j^*\|_Q > C_0^* \lambda_{nj}\}$ *for some constant* $C_0^* > 0$, *and*

$$\{\Gamma_n w_n^*(0)^{-(1-\beta_0/2)\tau_0} \gamma_n^*(0) + w_n^*(0)^{-\tau_0} \sqrt{\log(p/\epsilon)/n}\}(1 + M_F + M_0)$$

$$= o(1), \tag{43}$$

where $\tau_0 = 1/(2/\beta_0 + 1 - 2/r_0)$. *Then for sufficiently large* n, *depending on* (M_F, M_0) *only through the convergence rate in* (43), *and for any* $A_0 > A(\xi_0^*, \eta_0)$, *we have*

$$\mathcal{D}_n^*(\hat{g}, g^*, \xi_1^*, 0) \le O(1) C_1^2 (M_F + M_0)\{n^{\frac{-1}{2+\beta_0}} + \sqrt{\log(p/\epsilon)/n}\}^2, \tag{44}$$

with probability at least $1 - 2\epsilon$, *where* $O(1)$ *depends only on* $(B_0^*, A_0^*, C_0^*, \xi_0^*, \kappa_0^*, \eta_0, \varrho_0)$. *If* $r_0 = \beta_0 = 1$, *then the results are valid even when Assumption 8, and hence* ϱ_0 *are removed.*

Condition (43) is based on Theorem 4 for convergence of empirical norms. By Remark 23, a weaker condition can be obtained using Theorem 5 when $1 \le r_0 \le 2$ and $\tau_0 < 1$ (i.e., $r_0 > \beta_0$). It is interesting to note that (43) reduces to (45) below in the case $r_0 = \beta_0 = 1$.

PROPOSITION 5. *Proposition 4 is also valid with* (43) *replaced by the weaker condition*

$$\{w_n^*(0)^{-\tau_0} \sqrt{\log(np/\epsilon)/n}\}(1 + M_F + M_0) = o(1), \tag{45}$$

in the case where $1 \le r_0 \le 2$, $r_0 > \beta_0$, *and the average marginal density of* $(X_1^{(j)}, \ldots, X_n^{(j)})$ *is bounded from above for all* j.

REMARK 12 (Comparison with existing results). Propositions 4 and 5 yield the fast rate $n^{\frac{-2}{2+\beta_0}} + \log(p)/n$ under L_0-ball sparsity in $\|\cdot\|_Q$ norm. Previously, the same rate was obtained for high-dimensional additive regression only with reproducing kernel Hilbert spaces (including the Sobolev space $\mathcal{W}_2^m$) by Koltchinskii and Yuan (2010) and Raskutti, Wainwright and Yu (2012), but under more restrictive conditions. They studied hybrid penalized estimation procedures, which involve additional constraints such that the Hilbert norms of $(g_1, \ldots, g_p)$ are bounded by *known* constants when minimizing a penalized criterion. Moreover, Koltchinskii and Yuan (2010) assumed a constant bound on the sup-norm of possible g^*, whereas Raskutti, Wainwright and Yu (2012) assumed the independence of the covariates $(X_i^{(1)}, \ldots, X_i^{(p)})$ for each i. These restrictions were relaxed in subsequent work by Suzuki and Sugiyama (2013), but only explicitly under the assumption that the noises ε_i are uniformly bounded by a constant. Moreover, our condition (43) is much weaker than related ones in Suzuki and Sugiyama (2013),

as discussed in Remarks 13 and 14 below. See also Remark 10 for a discussion about the relationship between our results and the seemingly faster rate in Suzuki and Sugiyama (2013).

REMARK 13. To justify Assumptions 6(i)–(ii) on convergence of empirical norms, our rate condition (43) is much weaker than previous ones used. If each class $\mathcal{G}_j$ is a Sobolev–Hilbert space ($r_0 = 2$), then $\tau_0 = \beta_0/2$ and (43) becomes

$$\{n^{1/2} w_n^*(0)^{\beta_0^2/4} \gamma_n^*(0)^2 + \gamma_n^*(0)\sqrt{\log(p/\epsilon)}\}(1 + M_F + M_0) = o(1). \tag{46}$$

Moreover, by Proposition 5, condition (43) can be weakened to (45), that is,

$$\gamma_n^*(0)\sqrt{\log(np/\epsilon)}(1 + M_F + M_0) = o(1), \tag{47}$$

under an additional condition that the average marginal density of $(X_1^{(j)}, \ldots, X_n^{(j)})$ is bounded from above for all j. Either condition (46) or (47) is much weaker than those in related analysis with reproducing kernel Hilbert spaces. In fact, techniques based on the contraction inequality (Ledoux and Talagrand (1991) as used in Meier, van de Geer and Bühlmann (2009) and Koltchinskii and Yuan (2010)), lead to a rate condition such as

$$n^{1/2}\{\gamma_n^2(0) + \nu_n^2\}(1 + M_F + M_0) = o(1), \tag{48}$$

where $\gamma_n(0) = B_0^{*\frac{2}{2+\beta_0}} n^{\frac{-1}{2+\beta_0}}$ and $\nu_n = \{\log(p/\epsilon)/n\}^{1/2}$. This amounts to condition (6) assumed in Suzuki and Sugiyama (2013), in addition to the requirement $n^{-1/2}(\log p) \le 1$. But condition (48) is even stronger than the following condition:

$$n^{1/2}\{\gamma_n(0)^{2+\beta_0^2/4} + \gamma_n(0)\nu_n\}(1 + M_F + M_0) = o(1), \tag{49}$$

because $\Gamma_n \gamma_n(0)^{2+\beta_0^2/4} + \gamma_n(0)\nu_n \ll \gamma_n^2(0) + \nu_n^2$ if either $\gamma_n(0) \gg \nu_n$ or $\gamma_n(0) \ll \nu_n$. Condition (49) implies (46) and (47), as we explain in the next remark.

REMARK 14. Our rate condition (43) is in general weaker than the corresponding condition with $(w_n^*(0), \gamma_n^*(0))$ replaced by $(w_n(0), \gamma_n(0))$, that is,

$$\{\Gamma_n \gamma_n(0)^{1-(1-\beta_0/2)\tau_0} + \gamma_n(0)^{-\tau_0}\nu_n\}(1 + M_F + M_0) = o(1). \tag{50}$$

This demonstrates the advantage of using the more careful choices $(w_n^*(0), \gamma_n^*(0))$ and also explains why (49) implies (46) in Remark 13. In fact, if $\gamma_n(0) \ge \nu_n$ then (43) and (50) are identical to each other. On the other hand, if $\gamma_n(0) < \nu_n$, then $w_n^*(0) = \nu_n > \gamma_n(0)$ and $w_n^*(0)^{-(1-\beta_0/2)\tau_0}\gamma_n^*(0) = B_0^* n^{-1/2} \times w_n^*(0)^{-(1-\beta_0/2)\tau_0-\beta_0/2} < \gamma_n(0)^{1-(1-\beta_0/2)\tau_0}$. This also shows that if $\gamma_n(0) \ll \nu_n$, then (43) is much weaker than (50). For illustration, if $r_0 = 2$ and hence $\tau_0 = \beta_0/2$, then (50) or equivalently (49) requires at least $\gamma_n(0)^{-\beta_0/2}\nu_n = o(1)$, that is, $(\log p)n^{\frac{-2}{2+\beta_0}} = o(1)$, and (48) requires at least $n^{1/2}\nu_n^2 = o(1)$, that

is, $\log(p)n^{-1/2} = o(1)$. In contrast, the corresponding requirement for (43), $w_n^*(0)^{-\beta_0/2}\nu_n = o(1)$, is automatically valid as long as $\nu_n = o(1)$, that is, $\log(p)n^{-1} = o(1)$.

The following result deals with the case $0 < q < 1$ for the parameter set (3).

PROPOSITION 6. *Assume that* (1) *holds and* $\|g^*\|_{F,1} \le C_1 M_F$ *and* $\|g^*\|_{Q,q} \le C_1^q M_q$ *for* $0 < q < 1$, $M_q > 0$, *and* $M_F > 0$, *possibly depending on* (n, p). *Let* $w_{nj} = w_n^*(q)$ *and* $\gamma_{nj} = \gamma_n^*(q)$ *by* (32)–(33). *Suppose that Assumptions* 1, 7 *and* 8 *hold with* $0 < \eta_0 < (\xi_0^* - 1)/(\xi_0^* + 1)$ *and* $S = \{1 \le j \le p : \|g_j^*\|_Q > C_0^* \lambda_{nj}\}$ *for some constant* $C_0^* > 0$, $\log(p/\epsilon) = o(n)$, *and*

$$\{\Gamma_n w_n^*(q)^{-(1-\beta_0/2)\tau_0}\gamma_n^*(q)^{1-q} + w_n^*(q)^{-\tau_0}\nu_n^{1-q}\}(1 + M_F + M_q) \le \eta_4, \tag{51}$$

for some constant $\eta_4 > 0$, *where* $\nu_n = \{\log(p/\epsilon)/n\}^{1/2}$. *Then for sufficiently large* n, *independently of* (M_F, M_q), *and any* $A_0 > A(\xi_0^*, \eta_0)$, *we have*

$$\mathcal{D}_n^*(\hat{g}, g^*, \xi_1^*, 0) \le O(1)C_1^2(M_F + M_q)\{n^{\frac{-1}{2+\beta_0(1-q)}} + \sqrt{\log(p/\epsilon)/n}\}^{2-q},$$

with probability at least $1 - 2\epsilon$, *where* $O(1)$ *depends only on* $(q, B_0^*, A_0^*, C_0^*, \xi_0^*, \kappa_0^*, \eta_0, \varrho_0, \eta_4)$ *and* (C_2, C_3, C_4) *as in Theorem* 4. *If* $r_0 = \beta_0 = 1$, *then the results are valid even when Assumption* 8, *and hence* ϱ_0 *are removed.*

Similarly as in Propositions 4 and 5, condition (51) can be weakened as follows when $1 \le r_0 \le 2$ and $\tau_0 < 1$ (i.e., $r_0 > \beta_0$). It should also be noted that (51) is equivalent to (52) below (with different η_4 in the two equations) in the case $r_0 = \beta_0 = 1$, because $\gamma_n^*(q)$ with $q < 1$ is of a slower polynomial order than $n^{-1/2}$ and hence $\{\log(n)/n\}^{1/2}\gamma_n^*(q)^{-1} = o(1)$.

PROPOSITION 7. *Proposition* 6 *is also valid with* (43) *replaced by the weaker condition*

$$\{w_n^*(q)^{-\tau_0}(\log(np/\epsilon)/n)^{(1-q)/2}\}(1 + M_F + M_0) \le \eta_4, \tag{52}$$

for some constant $\eta_4 > 0$, *in the case where* $1 \le r_0 \le 2$, $r_0 > \beta_0$ *and the average marginal density of* $(X_1^{(j)}, \ldots, X_n^{(j)})$ *is bounded from above for all* j.

REMARK 15. Propositions 6 and 7 yield, under L_q-ball sparsity in $\|\cdot\|_Q$ norm, a convergence rate interpolating the slow and fast rates smoothly from $q = 1$ to $q = 0$, similarly as in fixed designs (Section 3.1). However, the rate condition (51) involved does not always exhibit a smooth transition to those for the slow and fast rates. In the extreme case $q = 1$, condition (51) with $q = 1$ cannot be satisfied when M_1 is unbounded or when M_1 is bounded but Γ_n is unbounded with $r_0 = \beta_0 = 1$. In contrast, Proposition 3 allows for unbounded M_1 and the case

TABLE 1
Convergence rates for out-of-sample prediction under parameter set (3) with (M_F, M_q) bounded from above

		$r_0 = \beta_0 = 1$			
	$r_0 > \beta_0$			$q = 0$	
	$0 \le q \le 1$	$q = 1$	$0 < q < 1$	$\nu_n = o(\gamma_n(0))$	otherwise
Scale adaptive	yes	yes	yes	yes	no
Rate	$\{\gamma_n(q) + \nu_n\}^{2-q}$	$\sqrt{\log(n)/n} + \nu_n$	$\{\gamma_n(q) + \nu_n\}^{2-q}$		

Note: $\gamma_n(q) \asymp n^{\frac{-1}{2+\beta_0(1-q)}}$ and $\nu_n = \{\log(p/\epsilon)/n\}^{1/2}$. Scale-adaptiveness means the convergence rate is achieved with (w_{nj}, γ_{nj}) chosen independently of (M_F, M_q).

$r_0 = \beta_0 = 1$. This difference is caused by the need to justify Assumption 6(ii) with $q \neq 1$. In the extreme case $q = 0$, condition (51) with $q = 0$ also differ drastically from (43) in Proposition 4. As seen from the proof of Corollary 7, this difference arises because Assumption 6(ii) can be justified by exploiting the fact that $z^q \to \infty$ as $z \to \infty$ for $q > 0$ (but not $q = 0$).

For illustration, Table 1 gives the convergence rates from Propositions 3–6 in the simple situation where (M_F, M_q) are bounded from above, independently of (n, p). The rate conditions (43) and (51) are easily seen to hold in all cases except that (43) is not satisfied for $q = 0$ when $r_0 = \beta_0 = 1$ but $\nu_n \neq o(\gamma_n(0))$. In this case, we show in the following result that the convergence rate $\{\gamma_n(0) + \nu_n\}^2$ can still be achieved, but with the tuning parameters (w_{nj}, γ_{nj}) chosen suitably depending on the upper bound of (M_F, M_q). This is in contrast with the other cases in Table 1 where the convergence rates are achieved by our penalized estimators in a scale-adaptive manner: $(w_{nj}, \gamma_{nj}) = (w_n^*(q), \gamma_n^*(q))$ are chosen independently of (M_F, M_q) or their upper bounds.

PROPOSITION 8. *Assume that $r_0 = \beta_0 = 1$, and M_F and M_0 are bounded from above by a constant $\overline{M} > 0$. Suppose that the conditions of Proposition 4 are satisfied except with (43) and Assumption 8 removed, and Assumption 7 holds instead of Assumption 5. Let $\hat{g}'$ be the estimator with (w_{nj}, γ_{nj}) replaced by $w'_{nj} = K_0 w_n^*(0)$ and $\gamma'_{nj} = K_0^{-\beta_0/2} \gamma_n^*(0)$ for $K_0 > 0$. Then K_0 can be chosen, depending on $\overline{M}$ but independently of (n, p), such that for sufficiently large n, depending on $\overline{M}$, and any $A_0 > A(\xi_0^*, \eta_0)$, we have*

$$\mathcal{D}_n^*(\hat{g}', g^*, \xi_1^*, 0) \le O(1) C_1^2 (M_F + M_0) \{n^{\frac{-1}{2+\beta_0}} + \sqrt{\log(p/\epsilon)/n}\}^2,$$

with probability at least $1 - 2\epsilon$, where $O(1)$ depends only on $(\overline{M}, B_0^*, A_0^*, C_0^*, \xi_0^*, \kappa_0^*, \eta_0)$ and (C_2, C_3, C_4) as in Theorem 4.

4.2. *Minimax lower bounds.* We demonstrate minimiax optimality of the rates achieved by the doubly penalized estimator $\hat{g}$. To clarify main ideas, we first provide a general result on minimax lower bounds for estimation in additive model (1) under the following conditions. Assume that each noise ϵ_i is distributed as $\mathrm{N}(0, \sigma^2)$, independently of $(X_i^{(1)}, \ldots, X_i^{(p)})$, and the vectors $\{(\varepsilon_i, X_i^{(1)}, \ldots, X_i^{(p)}) : i = 1, \ldots, n\}$ are independent and identically distributed. In addition, assume that all functions $g_j \in \mathcal{G}_j$ are centered: $\int g_j(z)\,\mathrm{d}z = 0$. Suppose that for $1 \le j \le p$, there exist basis functions $\{u_{j\ell}(\cdot) : \ell \ge 1\}$ in $\mathcal{G}_j$ such that for all integers $k \ge 1$ and real numbers a_ℓ,

$$c_0 \sum_{\ell=1}^{k} a_\ell^2 \le \Bigg\| \sum_{\ell=1}^{k} a_\ell u_{j\ell} \Bigg\|_Q^2 \le \sum_{\ell=1}^{k} a_\ell^2, \tag{53}$$

and for all signs $e_{j\ell} \in \{-1, 1\}$,

$$\Bigg\| \sum_{\ell=1}^{k} e_{j\ell} u_{j\ell} \Bigg\|_{F,j} \le C_F k^{1/\beta_0 + 1/2}, \tag{54}$$

where $c_0 \in (0, 1]$, $\beta_0 \in (0, 2)$, and $C_F > 0$ are constants. Denote the parameter set as

$$\mathcal{G}(M_F, M_q) = \Bigg\{ g(x) = \sum_{j=1}^{p} g_j\big(x^{(j)}\big) : \sum_{j=1}^{p} \|g_j\|_{F,j} \le \sigma M_F, \sum_{j=1}^{p} \|g_j\|_Q^q \le \sigma^q M_q \Bigg\},$$

where $0 \le q \le 1$, $M_q > 0$ and $M_F > 0$ are known.

THEOREM 3. (i) *Suppose that* (53) *and* (54) *hold. Let integers* $1 \le s \le p$ *and* $k \ge 1$ *be determined such that*

$$M_F = C_F s n^{-1/2} k^{1/\beta_0 + 1/2}, \qquad M_q = s n^{-q/2} k^{q/2}. \tag{55}$$

Then

$$\inf_{(\tilde{g}_1, \ldots, \tilde{g}_p)} \sup_{g^* \in \mathcal{G}(M_F, M_q)} E\Bigg\{ \sum_{j=1}^{p} \|\tilde{g}_j - g_j^*\|_Q^2 \Bigg\}$$

$$\ge c_0 c_1 s k \sigma^2 / n = c_0 c_1 \sigma^2 \bigg(\frac{M_F}{C_F} \bigg)^{1-q_1} M_q^{q_1} n^{\frac{-(2-q)}{2+(1-q)\beta_0}},$$

where $q_1 = (2 - \beta_0)/\{2 + (1 - q)\beta_0\}$ *and* $c_1 = \sqrt{2/\pi} \int_1^\infty e^{-z^2/2}\,\mathrm{d}z$.

(ii) *Suppose that* (54) *hold with* $k = 1$ *and* C_F *set to* $C_{F,1}$, *for a basis function* $u_{j1}(\cdot)$ *with* $\|u_j\|_Q = 1$. *Let integer* $s_0 \ge 1$ *and* λ_0 *be determined such that* $\lambda_0 = \sqrt{(2/n)\log(ep/s_0)}$ *and* $s_0 \le \min(M_q/\lambda_0^q, M_F/(C_{F,1}\lambda_0), p)$. *Then*

$$\inf_{(\tilde{g}_1, \ldots, \tilde{g}_p)} \sup_{g^* \in \mathcal{G}(M_F, M_q)} E\Bigg\{ \sum_{j=1}^{p} \|\tilde{g}_j - g_j^*\|_Q^2 \Bigg\} \ge \sigma^2 \frac{(1 - s_0/p)/16}{1 - s_0/p + 1/e} s_0 \lambda_0^2.$$

REMARK 16. Suppose $M_q/\lambda_0^q \le \min(M_F/(C_F\lambda_0), p/2)$ in Theorem 3(ii). The largest possible s_0 and the corresponding λ_0 must satisfy $s_0 \le M_q/\lambda_0^q < s_0 + 1 \le p/2 + 1$. This implies $n\lambda_0^2 = 2\log(ep/s_0) \asymp 2\log(ep/M_q) + q\log\lambda_0^2 = 2\log(ep/(M_q n^q)) + q\log(2\log(ep/s_0)) \asymp 2\log(ep/(M_q n^q))$. The last step above is valid due to $2\log(ep/s_0) \ge 2(1+\log 2)$. Thus, $\lambda_0 \asymp \{(2/n)\log(ep/(M_q n^q))\}^{1/2}$ and

$$s_0\lambda_0^2 \asymp M_q\lambda_0^{2-q} \asymp M_q\{(2/n)\log(ep/(M_q n^q))\}^{1-q/2}.$$

REMARK 17. Return to the setting where $\mathcal{G}_j$ is a Sobolev class $\mathcal{W}_{r_0}^{m_0}$ or a bounded variation class $\mathcal{V}^{m_0}$ on $[0,1]$ for $j = 1,\dots,p$. Condition (53) is satisfied for any L_2-orthogonal bases, properly scaled, provided that the marginal density of $X_i^{(j)}$ is bounded from below and above on $[0,1]$ for all j. Let $u_{j\ell}(z) = \sqrt{k}u_0(kz - (\ell-1))$, $\ell = 1,\dots,k$, with $\operatorname{supp}(u_0) \subset [0,1]$ and $\int_0^1 u_0^2(z)\,dz = 1$. Then (54) holds with $\beta_0 = 1/m_0$ and $C_F = \|u_0^{(m_0)}\|_{L_{r_0}}$ or $\mathrm{TV}(u_0^{(m_0-1)})$. In Theorem 3(ii), we take $u_{j1}(z) = b_j(z - 1/2)$ for some coefficient b_j. Then (54) holds with $k = 1$ and $C_{F,1} = \max_{1\le j\le p}\|u_{j1}\|_{F,j}$, which is bounded from above provided that $E\{(X_i^{(j)} - 1/2)^2\}$ is bounded from below for all j. In particular, $C_{F,1} = 0$ in the case of $m_0 \ge 2$.

The prediction error $\|\tilde{g} - g^*\|_Q^2$ can be bounded from below by

$$\|\tilde{g} - g^*\|_Q^2 \ge (c_3/c_2)\sum_{j=1}^p \|\tilde{g}_j - g_j^*\|_Q^2,$$

under the following assumption, which is qualitatively similar to the assumption that the vector $(X_i^{(1)},\dots,X_i^{(p)})$ is uniformly distributed on $[0,1]^p$ for establishing minimax lowers bounds in Raskutti, Wainwright and Yu (2012) and Suzuki and Sugiyama (2013).

ASSUMPTION 9. Assume that the marginal density of $X_i^{(j)}$ is bounded from above by $c_2 > 0$ on $[0,1]$ for all $1 \le j \le p$, and the joint density of $(X_i^{(1)},\dots,X_i^{(p)})$ is bounded from below by $c_3 > 0$ on $[0,1]^p$.

With these remarks, Theorem 3 then leads to the following minimax lower bound.

PROPOSITION 9. *Let each $\mathcal{G}_j$, $j = 1,\dots,p$, be a Sobolev class $\mathcal{W}_{r_0}^{m_0}$ or a bounded variation class $\mathcal{V}^{m_0}$ on $[0,1]$ for $r_0 \ge 1$ and $m_0 \ge 1$. Suppose that*

Assumption 9 holds, and $M_q/\lambda_0^q \le \min(M_F/(C_{F,1}\lambda_0), p/2)$, *with* λ_0 *as in Remark 16 and* $C_{F,1}$ *as in Remark 17. Then*

$$\inf_{\tilde g} \sup_{g^* \in \mathcal{G}(M_F, M_q)} E\{\|\tilde g - g^*\|_Q^2\}$$

$$\ge O(1)\sigma^2\{(M_F/C_F)^{1-q_1} M_q^{q_1} n^{-(2-q)/\{2+(1-q)\beta_0\}} + M_q \lambda_0^{2-q}\}, \tag{56}$$

where q_1 *is as in Theorem 3(i),* $C_F = \|u_0^{(m_0)}\|_{L_{r_0}}$ *or* $\mathrm{TV}(u_0^{(m_0-1)})$ *as in Remark 17, and* $O(1)$ *depends only on* (c_1, c_2, c_3).

REMARK 18. The lower bound (56) is matched by the convergence rate (41) for the doubly penalized estimator $\hat g$, where the tuning parameters (λ_n, ρ_n) are allowed to depend on (M_F, M_q). Moreover, (56) is matched by the convergence rate (35) in Corollaries 6 and 7 (as well as the rates in Propositions 3–8), up to multiplicative constants depending on (M_F, M_q).

REMARK 19. The lower bound (56) with $q = 0$ is similar to those obtained by Raskutti, Wainwright and Yu (2012) and Suzuki and Sugiyama (2013) in additive regression and Dalalyan, Ingster and Tsybakov (2014) in white noise models, all with L_2-Sobolev or similar Hilbert spaces. The extension involved in our results is to handle $0 < q \le 1$ as well as L_r-Sobolev and bounded variation spaces.

5. Convergence of empirical norms. We provide two explicit results on the convergence of empirical norms as needed for Assumption 6. These results can also be useful for other applications.

Our first result, Theorem 4, is applicable (but not limited) to Sobolev and bounded variation spaces in general. For clarity, we postulate another entropy condition, similar to Assumption 4 but with the empirical supremum norms.

ASSUMPTION 10 (Entropy condition in supremum norms). For $j = 1, \dots, p$, let $\psi_{nj,\infty}(\cdot, \delta)$ be an upper envelope of the entropy integral, independent of the realizations $\{X_i^{(j)} : i = 1, \dots, n\}$, as follows:

$$\psi_{nj,\infty}(z, \delta) \ge \int_0^z H^{*1/2}\big(u/2, \mathcal{G}_j^*(\delta), \|\cdot\|_{n,\infty}\big)\,\mathrm{d}u, \qquad z > 0, 0 < \delta \le 1,$$

where $\mathcal{G}_j^*(\delta) = \{f_j \in \mathcal{G}_j : \|f_j\|_{F,j} + \|f_j\|_Q/\delta \le 1\}$ as in Assumption 4 and

$$H^*\big(u, \mathcal{G}_j^*(\delta), \|\cdot\|_{n,\infty}\big) = \sup_{(X_1^{(j)}, \dots, X_n^{(j)})} H\big(u, \mathcal{G}_j^*(\delta), \|\cdot\|_{n,\infty}\big).$$

We also make use of the following two conditions about metric entropies and sup-norms. Suppose that for $j = 1, \dots, p$, $\psi_{nj}(\delta)$ and $\psi_{nj,\infty}(z, \delta)$ in Assumptions

4 and 10 are in the polynomial forms

$$\psi_{nj}(\delta) = B_{nj}\delta^{1-\beta_j/2}, \qquad 0 < \delta \le 1, \tag{57}$$

$$\psi_{nj,\infty}(z,\delta) = B_{nj,\infty}z^{1-\beta_j/2}, \qquad z > 0, 0 < \delta \le 1, \tag{58}$$

where $0 < \beta_j < 2$ is a constant, and $B_{nj} > 0$ and $B_{nj,\infty} > 0$ are constants, possibly depending on n. Denote $\Gamma_n = \max_{j=1,\dots,p}(B_{nj,\infty}/B_{nj})$. In addition, suppose that for $j = 1, \dots, p$,

$$\|g_j\|_\infty \le (C_{4,j}/2)\big(\|g_j\|_{F,j} + \|g_j\|_Q\big)^{\tau_j}\|g_j\|_Q^{1-\tau_j}, \qquad g_j \in \mathcal{G}_j, \tag{59}$$

where $C_{4,j} \ge 1$ and $0 < \tau_j \le (2/\beta_j - 1)^{-1}$ are constants. Let $\gamma_{nj} = n^{-1/2} \times \psi_{nj}(w_{nj})/w_{nj} = n^{-1/2}B_{nj}w_{nj}^{-\beta_j/2}$ by (12) and $\tilde{\gamma}_{nj} = n^{-1/2}w_{nj}^{-\tau_j}$ for $j = 1, \dots, p$. As a function of w_{nj}, the quantity $\tilde{\gamma}_{nj}$ in general differs from γ_{nj} even up to a multiplicative constant unless $\tau_j = \beta_j/2$ as in the case where $\mathcal{G}_j$ is an L_2-Sobolev space; see (61) below.

THEOREM 4. *Suppose that Assumptions 4 and 10 hold with $\psi_{nj}(\delta)$ and $\psi_{nj,\infty}(z,\delta)$ in the forms (57) and (58), and condition (59) holds. In addition, suppose that for sufficiently large n, $\gamma_{nj} \le w_{nj} \le 1$ and $\Gamma_n\gamma_{nj}^{1-\beta_j/2} \le 1$ for $j = 1, \dots, p$. Then for any $0 < \epsilon' < 1$ (e.g., $\epsilon' = \epsilon$), inequality (24) holds with $\pi = \epsilon'^2$ and $\phi_n > 0$ such that*

$$\begin{aligned}\phi_n = O(1)\Big\{ & n^{1/2}\Gamma_n \max_j \frac{\gamma_{nj}}{\lambda_{nj}} \max_j \frac{\tilde{\gamma}_{nj}w_{nj}^{\beta_{p+1}\tau_j/2}}{\lambda_{nj}} \\ & + \max_j \frac{\tilde{\gamma}_{nj}}{\lambda_{nj}} \max_j \frac{\sqrt{\log(p/\epsilon')}}{\lambda_{nj}} + \max_j \frac{\tilde{\gamma}_{nj}^2\log(p/\epsilon')}{\lambda_{nj}^2}\Big\},\end{aligned} \tag{60}$$

where $\beta_{p+1} = \min_{j=1,\dots,p}\beta_j$, and $O(1)$ depends only on (C_2, C_3) from Lemmas 13 and 14 in the Supplementary Material (Tan and Zhang (2019)) and $C_4 = \max_{j=1,\dots,p} C_{4,j}$ from condition (59).

REMARK 20. Conditions (57), (58) and (59) are satisfied under Assumption 8, when each $\mathcal{G}_j$ is a Sobolev space $\mathcal{W}_{r_j}^{m_j}$ for $r_j \ge 1$ and $m_j \ge 1$, or a bounded variation space $\mathcal{V}^{m_j}$ for $r_j = 1$ and $m_j \ge 1$, on $[0,1]$. Let $\beta_j = 1/m_j$. First, (59) is implied by the interpolation inequalities for Sobolev spaces Nirenberg, 1966 with

$$\tau_j = (2/\beta_j + 1 - 2/r_j)^{-1} \tag{61}$$

and $C_{4,j} = \varrho_0^{-1}C_4(m_j, r_j)$ depending on $C_4(m_j, r_j)$ in Lemma 21 of the Supplementary Material (Tan and Zhang (2019)). Moreover, if $f_j \in \mathcal{G}_j^*(\delta)$ with $0 < \delta \le 1$, then $\|f_j\|_{F,j} \le 1$ and $\|f_j\|_Q \le \delta$, and hence $\|f_j\|_{L_{r_j}} \le \|f_j\|_\infty \le C_{4,j}$ by (59). By rescaling the entropy estimates for Sobolev and bounded variation spaces (Lorentz,

Golitschek and Makovoz (1996)) as in Lemmas 19 and 20 in the Supplementary Material (Tan and Zhang (2019)), Assumptions 4 and 10 are satisfied such that (57) and (58) hold with B_{nj} independent of n, and $B_{nj,\infty} = O(1)B_{nj}$ if $r_j > \beta_j$ or $B_{nj,\infty} = O(\log^{1/2}(n))B_{nj}$ if $r_j = \beta_j = 1$.

REMARK 21. Assumption 8 is not needed for justification of (57), (58) and (59), when each class $\mathcal{G}_j$ is $\mathcal{W}_1^1$ or $\mathcal{V}^1$ on $[0, 1]$, that is, $r_j = m_j = 1$. In this case, condition (59) directly holds with $\tau_j = 1$, because $\|g_j\|_\infty \le \mathrm{TV}(g_j) + \|g_j\|_Q$. Then (57) and (58) easily follow from the entropy estimates in Lemmas 19 and 20 in the Supplementary Material (Tan and Zhang (2019)).

Our second result provides a sharper rate than in Theorem 4, applicable (but not limited) to Sobolev and bounded variation spaces, provided that the following conditions hold. For $g_j \in \mathcal{G}_j$, assume that $g_j(\cdot)$ can be written as $\sum_{\ell=1}^\infty \theta_{j\ell} u_{j\ell}(\cdot)$ for certain coefficients $\theta_{j\ell}$ and basis functions $u_{j\ell}(\cdot)$ on a set Ω. In addition, for certain positive constants $C_{5,1}$, $C_{5,2}$, $C_{5,3}$, $0 < \tau_j < 1$ and $0 < w_{nj} \le 1$, assume that for all $1 \le j \le p$,

$$\sup\left\{\sum_{\ell=1}^k u_{j\ell}^2(x)/k : x \in \Omega, k \ge \ell_{j0}\right\} \le C_{5,1}, \tag{62}$$

$$\max_{k\ge 1} \sum_{\ell_{j,k-1}<\ell\le\ell_{jk}} \theta_{j\ell}^2 \ell_{jk}^{1/\tau_j} \le C_{5,2}\big(\|g_j\|_{F,j} + w_{nj}^{-1}\|g_j\|_Q\big)^2, \tag{63}$$

$$\sum_{\ell=1}^{\ell_{j0}} \theta_{j\ell}^2 w_{nj}^{-2} \le C_{5,2}\big(\|g_j\|_{F,j} + w_{nj}^{-1}\|g_j\|_Q\big)^2, \tag{64}$$

with $\ell_{jk} = \lceil (2^k/w_{nj})^{2\tau_j} \rceil$ for $k \ge 0$ and $\ell_{j,-1} = 0$, and for all $1 \le j \le p$ and $k \ge 0$,

$$\sup\left\{\Big\|\sum_{\ell_{j,k-1}<\ell\le\ell_{jk}} \theta_{j\ell} u_{j\ell}\Big\|_Q^2 : \sum_{\ell_{j,k-1}<\ell\le\ell_{jk}} \theta_{j\ell}^2 = 1\right\} \le C_{5,3}. \tag{65}$$

THEOREM 5. *Suppose that* (62), (63), (64) *and* (65) *hold as above, and* $\max_{j=1,\dots,p}\{e^{2/(1-\tau_j)} + 2w_{nj}^{-\tau_j}\} \le n$. *Then for any* $0 < \epsilon' < 1$ (*e.g.*, $\epsilon' = \epsilon$), *inequality* (24) *holds with* $\pi = \epsilon'^2$ *and* $\phi_n > 0$ *such that*

$$\phi_n = O(1)\left\{\max_j \frac{\tilde\gamma_{nj}}{(1-\tau_j)\lambda_{nj}} \max_j \frac{\sqrt{\log(np/\epsilon')}}{\lambda_{nj}} + \max_j \frac{\tilde\gamma_{nj}^2 \log(np/\epsilon')}{(1-\tau_j)^2\lambda_{nj}^2}\right\},$$

where $\tilde\gamma_{nj} = n^{-1/2} w_{nj}^{-\tau_j}$ *and* $O(1)$ *depends only on* $\{C_{5,1}, C_{5,2}, C_{5,3}\}$.

REMARK 22. Let $\mathcal{G}_j$ be a Sobolev space $\mathcal{W}_{r_j}^{m_j}$ with $r_j \ge 1$, $m_j \ge 1$, and $(r_j \wedge 2)m_j > 1$ or a bounded variation space $\mathcal{V}^{m_j}$ with $r_j = 1$ and $m_j > 1$ (excluding

$m_j = 1$) on $[0, 1]$. Condition (62) holds for commonly used Fourier, wavelet and spline bases. For any L_2-orthonormal bases $\{u_{j\ell}, \ell \geq 1\}$, condition (64) follows from Assumption 8 with $C_{5,2} \geq \varrho_0^{-1}$, and condition (65) is also satisfied under an additional assumption that the average marginal density of $\{X_i^{(j)} : i = 1, \ldots, n\}$ is bounded from above on $[0, 1]$ by $C_{5,3}$ for all j. In the proof of Proposition 5, we verify (62) and (63) for suitable wavelet bases with $\tau_j = 1/\{2m_j + 1 - 2/(r_j \wedge 2)\}$, which satisfies $\tau_j < 1$ because $(r_j \wedge 2)m_j > 1$. In fact, $\mathcal{G}_j$ is allowed to be a Besov space $\mathcal{B}_{r_j,\infty}^{m_j}$, which contains $\mathcal{W}_{r_j}^{m_j}$ for $r_j \geq 1$ and $\mathcal{V}^{m_j}$ for $r_j = 1$ (e.g., DeVore and Lorentz (1993)).

REMARK 23. The convergence rate of ϕ_n in Theorem 5 is no slower than (60) in Theorem 4 if $1 \leq r_j \leq 2$ and $(1 - \tau_j)^{-1}\{\log(n)/n\}^{1/2} = O(\tilde{\gamma}_{nj})$, the latter of which is valid whenever τ_j is bounded away from 1 and $\tilde{\gamma}_{nj} = n^{-1/2} w_{nj}^{-\tau_j/2}$ is of a slower polynomial order than $n^{-1/2}$. However, Theorem 5 requires an additional side condition (65) along with the requirement of $\tau_j < 1$, which excludes for example the bounded variation space $\mathcal{V}^1$ on $[0, 1]$; See equations (46) and (47) for implications of these rates when used in Assumption 6.

6. Discussion. For additive regression with high-dimensional data, we have established new convergence results on the predictive performance of doubly penalized estimation when each component function can be a Sobolev space $\mathcal{W}_r^m$ or a bounded variation space $\mathcal{V}^m$. There remain various open problems to be fully investigated. First, the doubly penalized estimators are shown under certain conditions to be adaptive to the sizes of $L_1(\|\cdot\|_F)$ and $L_q(\|\cdot\|_Q)$ balls with fixed sparsity index q and smoothness index m. For $q = 0$ and in white noise models with unknown component functions in L_2-Sobolev spaces, Dalalyan, Ingster and Tsybakov (2014) developed adaptive estimation with respect to smoothness m. It is desirable to study how adaptive estimation can be achieved over such balls with varying q and m. Moreover, it is interesting to study variable selection and inference about component functions for high-dimensional additive regression, in addition to predictive performance studied here.

Acknowledgment. The authors thank a referee for extensive comments that led to improvement of the paper.

SUPPLEMENTARY MATERIAL

Supplement to "Doubly penalized estimation in additive regression with high-dimensional data" (DOI: 10.1214/18-AOS1757SUPP; .pdf). We provide proofs and technical tools.

REFERENCES

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

BUNEA, F., TSYBAKOV, A. and WEGKAMP, M. (2007). Sparsity oracle inequalities for the Lasso. *Electron. J. Stat.* **1** 169–194. MR2312149

DALALYAN, A., INGSTER, Y. and TSYBAKOV, A. B. (2014). Statistical inference in compound functional models. *Probab. Theory Related Fields* **158** 513–532. MR3176357

DEVORE, R. A. and LORENTZ, G. G. (1993). *Constructive Approximation. Grundlehren der Mathematischen Wissenschaften* [*Fundamental Principles of Mathematical Sciences*] **303**. Springer, Berlin. MR1261635

GREENSHTEIN, E. and RITOV, Y. (2004). Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli* **10** 971–988. MR2108039

GU, C. (2002). *Smoothing Spline ANOVA Models. Springer Series in Statistics*. Springer, New York. MR1876599

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

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

KIM, S.-J., KOH, K., BOYD, S. and GORINEVSKY, D. (2009). l_1 trend filtering. *SIAM Rev.* **51** 339–360. MR2505584

KOLTCHINSKII, V., LOUNICI, K. and TSYBAKOV, A. B. (2011). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Ann. Statist.* **39** 2302–2329. MR2906869

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

LEDOUX, M. and TALAGRAND, M. (1991). *Probability in Banach Spaces*: *Isoperimetry and Processes. Ergebnisse der Mathematik und Ihrer Grenzgebiete* (3) [*Results in Mathematics and Related Areas* (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

LORENTZ, G. G., GOLITSCHEK, M. V. and MAKOVOZ, Y. (1996). *Constructive Approximation*: *Advanced Problems. Grundlehren der Mathematischen Wissenschaften* [*Fundamental Principles of Mathematical Sciences*] **304**. Springer, Berlin. MR1393437

MAMMEN, E. (1991). Nonparametric regression under qualitative smoothness assumptions. *Ann. Statist.* **19** 741–759. MR1105842

MAMMEN, E. and VAN DE GEER, S. (1997). Locally adaptive regression splines. *Ann. Statist.* **25** 387–413. MR1429931

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

MÜLLER, P. and VAN DE GEER, S. (2015). The partial linear model in high dimensions. *Scand. J. Stat.* **42** 580–608. MR3345123

NIRENBERG, L. (1966). An extended interpolation inequality. *Ann. Sc. Norm. Super. Pisa Cl. Sci.* (3) **20** 733–737. MR0208360

PETERSEN, A., WITTEN, D. and SIMON, N. (2016). Fused lasso additive model. *J. Comput. Graph. Statist.* **25** 1005–1025. MR3572026

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

SADHANALA, V. and TIBSHIRANI, R. J. (2017). Additive models with trend filtering. Preprint. Available at arXiv:1702.05037.

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

TAN, Z. and ZHANG, C.-H. (2019). Supplement to "Doubly penalized estimation in additive regression with high-dimensional data." DOI:10.1214/18-AOS1757SUPP.

TIBSHIRANI, R. J. (2014). Adaptive piecewise polynomial estimation via trend filtering. *Ann. Statist.* **42** 285–323. MR3189487

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

VAN DE GEER, S. A. and BÜHLMANN, P. (2009). On the conditions used to prove oracle results for the Lasso. *Electron. J. Stat.* **3** 1360–1392. MR2576316

VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes*: *With Applications to Statistics*. *Springer Series in Statistics*. Springer, New York. MR1385671

YANG, T. and TAN, Z. (2018). Backfitting algorithms for total-variation and empirical-norm penalized additive modelling with high-dimensional data. *Stat* **7** e198. MR3905854

YANG, Y. and TOKDAR, S. T. (2015). Minimax-optimal nonparametric regression in high dimensions. *Ann. Statist.* **43** 652–674. MR3319139

YUAN, M. and ZHOU, D.-X. (2016). Minimax optimal rates of estimation in high dimensional additive models. *Ann. Statist.* **44** 2564–2593. MR3576554

DEPARTMENT OF STATISTICS
RUTGERS UNIVERSITY
110 FRELINGHUYSEN ROAD
PISCATAWAY, NEW JERSEY 08854
USA
E-MAIL: ztan@stat.rutgers.edu
czhang@stat.rutgers.edu