The Annals of Statistics
2015, Vol. 43, No. 2, 652–674
DOI: 10.1214/14-AOS1289

MINIMAX-OPTIMAL NONPARAMETRIC REGRESSION IN HIGH DIMENSIONS

BY YUN YANG AND SURYA T. TOKDAR

University of California, Berkeley and Duke University

Minimax L_2 risks for high-dimensional nonparametric regression are derived under two sparsity assumptions: (1) the true regression surface is a sparse function that depends only on $d = O(\log n)$ important predictors among a list of p predictors, with $\log p = o(n)$; (2) the true regression surface depends on $O(n)$ predictors but is an additive function where each additive component is sparse but may contain two or more interacting predictors and may have a smoothness level different from other components. For either modeling assumption, a practicable extension of the widely used Bayesian Gaussian process regression method is shown to adaptively attain the optimal minimax rate (up to $\log n$ terms) asymptotically as both $n, p \to \infty$ with $\log p = o(n)$.

1. Introduction. Rapid advances in technology have empowered researchers to collect data on a large number of explanatory variables to predict many outcomes of interest [5]. Because the relationship between an outcome Y and its predictors $X_1, \ldots, X_p$ may be highly nonlinear and involve interaction, there is a practical need to investigate statistical estimation under multivariate regression models

$$Y = \mu + f(X_1, \ldots, X_p) + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2), \tag{1.1}$$

with minimal assumptions made on f. The quality of estimation that may be achieved under an assumed model can be mathematically quantified by the minimax risk of estimating f from n data points. A classic result due to Charles Stone [24] states that if no assumption is made on how f depends on $X_1, \ldots, X_p$ other than requiring it to be differentiable with a smoothness level $\alpha > 0$ (definition below), then the associated minimax risk decays in n at a rate $n^{-\alpha/(2\alpha+p)}$. This rate is very slow when p is large, which means a very large sample size is needed for high quality statistical estimation—a phenomenon that has been termed the "curse of dimensionality." The curse of dimensionality becomes even more pronounced in the so-called large p small n setting, where the minimax risk decay rate is expressed as a function of both n and p, with p growing faster than n [32].

Received August 2014; revised November 2014.
MSC2010 subject classifications. Primary 62G08, 62C20; secondary 60G15.
Key words and phrases. Adaptive estimation, high-dimensional regression, minimax risk, model selection, nonparametric regression.

Practically motivated modeling assumptions must focus on nonparametric spaces of functions with lower inherent dimensions than the manifest dimension p. An example of such assumptions is

M1. f potentially depends on all elements of $X = (X_1, \ldots, X_p)$, but X itself lies in a low-dimensional manifold $\mathcal{M}^d$ in the ambient space $\mathbb{R}^p$.

It is well known that under M1, the minimax rate is $n^{-2\alpha/(2\alpha+d)}$ which is determined by the smoothness level α of f and the latent manifold dimension d [2, 13, 14, 23, 36, 37], but does not depend on the ambient dimension p. Various nonparametric regression techniques that operate on the ambient space and do not require estimation of the underlying manifold indeed achieve this minimax rate without any prior knowledge of d or α [14, 36].

However, for many high-dimensional applications, such as gene expression studies, a low-dimensional manifold assumption on X may not be tenable or verifiable. In such cases, one often uses the following sparsity inducing assumption:

M2. f depends on a small subset of d predictors with $d \leq \min\{n, p\}$.

M2 has served as the springboard for many widely used regression methods, including high-dimensional linear regression approaches, such as the Lasso [26] and the Dantzig selector [6], and nonparametric regression methods with variable selection, such as the Rodeo [15] and the Gaussian process regression [21]. The latter two allow flexible shape estimation of f and is able to capture interactions among the selected important predictors. However, in light of the classic result due to [24] it is conceivable that when f is allowed to be fully nonparametric, M2 should also suffer from the curse of dimensionality in a large p small n setting, unless d is much much smaller than p, that is, the regression function is assumed to be extremely sparse. A precise result that extends the work of [24] to account for predictor selection is presented in Section 3.

To relax this assumption of extreme sparsity without having to completely give up on nonparametric shape flexibility, we introduce a third modeling assumption:

M3. f may depend on $d \asymp \min\{n^\gamma, p\}$ variables for some $\gamma \in (0, 1)$ but admits an additive structure $f = \sum_{s=1}^k f_s$, where each component function f_s depends on a small d_s number of predictors.

Clearly, M3 subsumes M2 as a special case. In Section 3, we show M2 gives slowest minimax rates within M3. At the opposite extreme is the modeling assumption that f admits a completely additive structure with univariate components $f(X) = f_1(X_{i_1}) + \cdots + f_d(X_{i_d})$ for which scalable algorithms have been devised [11] and attractive minimax risk bounds have been derived albeit under the strong assumption that all component functions f_s have the same smoothness level [12, 17, 20, 22].

Compared to either of these two extremes, M3 provides a much more practically attractive theory of large p nonparametric regression. In Theorem 3.1, we

derive sharp upper and lower bounds on the minimax L_2 estimation risk under M3 as a function of n, p, k, component sizes $d_1, \ldots, d_k$ and smoothness levels of $f_1, \ldots, f_k$ which are allowed to have different levels of smoothness than one another. Minimax rates under M2 and the completely additive structure of [20] follow as corollaries to this general result. Our calculations suggest that M3 offers a minimax risk that decays quickly in n even when p grows almost exponentially in n, f involves nearly $\log p$ many predictors and these predictors interact with each other.

In Section 4, we demonstrate that a conceptually straightforward extension of the widely used Gaussian process regression method (see, e.g., [21], for a review) achieves the minimax rate adaptively across all subclasses of M3 under suitable large p small n asymptotics where p grows almost exponentially in n. In this paper we restrict only to a theoretical study of this new approach, which we name "additive Gaussian process regression." This approach appears entirely practicable with computational demands similar to those of the popular Bayesian additive regression tree method [7]. A full fledged methodological development of the same is underway and will be reported elsewhere.

The rest of the paper is organized as follows. Section 2 introduces the notation and some basic assumptions. Section 3 summarizes our main minimax results for high-dimensional nonparametric regression under M2 and M3. Section 4 proves the adaptive minimax optimality of additive Gaussian process regression. Section 5 provides proofs of our main results in Sections 3 and 4. Supporting technical results and proofs are presented in Section 6.

2. Notation. Let (X^i, Y^i), $i = 1, \ldots, n$ denote the observations on (X, Y). We make a stochastic design assumption that $X^1, \ldots, X^n$ are independent and identically distributed (IID) according to some compactly supported probability measure Q on $\mathbb{R}^p$ and that $f \in L_2(Q)$, the linear space of real valued functions on $\mathbb{R}^p$ equipped with inner product $\langle f, g\rangle_Q = \int f(x)g(x)Q(dx)$ and norm $\|f\|_Q = \langle f, f\rangle_Q^{1/2}$. We do not need to know or estimate Q for the purpose of estimating f, but it is a natural candidate to judge average prediction accuracy at future observations of X drawn from Q, as will be the case under simple exchangeability assumptions. Without loss of generality assume support $(Q) \subset [0, 1]^p$. Let $\|\cdot\|$ stand for the L_2 norm under the Lebesgue measure.

The L_2 minimax risk of estimation associated with any function space $\Sigma \subset L_2(Q)$ is defined as

$$r_n^2(\Sigma, Q, \mu, \sigma) = \inf_{\hat{f} \in \mathcal{A}_n} \sup_{f \in \Sigma} E_{f,Q} \|\hat{f} - f\|_Q^2,$$

where $\mathcal{A}_n$ is the space of all measurable functions of data to $L_2(Q)$ and $E_{f,Q}$ denotes expectation under the model: $X^i \sim Q$, $Y^i | X^i \sim N(\mu + f(X^i), \sigma^2)$, independently across $i = 1, \ldots, n$. When no risk of ambiguity is present, we shorten

the notation $r_n(\Sigma, Q, \mu, \sigma)$ to r_n and call r_n the minimax risk or the minimax *rate*, when viewed as a function of the sample size n.

Let $\mathbb{N}$ denote the set of natural numbers and $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$. For any d-dimensional multiindex $a = (a_1, \ldots, a_d) \in \mathbb{N}_0^d$ define $|a| = a_1 + \cdots + a_d$ and let D^a denote the mixed partial derivative operator $\partial^{|a|}/\partial x_1^{a_1} \cdots \partial x_d^{a_d}$. For any real number b, let $\lfloor b \rfloor$ denote the largest integer strictly smaller than b. Use the notation $C^{\alpha,d}$ to denote the Banach space of Hölder α-smooth functions on $[0,1]^d$ equipped with the norm

$$\|f\|_{C^{\alpha,d}} = \sum_{|k| \le \lfloor \alpha \rfloor} \|D^k f\|_\infty + \max_{x \ne y \in [0,1]^d} |D^{\lfloor \alpha \rfloor} f(x) - D^{\lfloor \alpha \rfloor}(y)| / \|x - y\|^{\alpha - \lfloor \alpha \rfloor}.$$

Let $C_1^{\alpha,d}$ denote the unit ball of $C^{\alpha,d}$.

For any $b \in \{0,1\}^p$ and $x = (x_1, \ldots, x_p) \in \mathbb{R}^p$, let $x_b = (x_j : b_j = 1)$ denote the vector of $|b| = \sum_{j=1}^p b_j$ predictors picked by b and let $T^b : C(\mathbb{R}^{|b|}) \to C(\mathbb{R}^p)$ denote the mapping that takes an $f \in C(\mathbb{R}^{|b|})$ to $T^b f : x \mapsto f(x_b)$. Let $\mathcal{B}^{p,d}$ denote the set of all $b \in \{0,1\}^p$ with $|b| = d$. We formalize the space of centered, p-variate, α-smooth functions of sparsity d and bound λ as

$$\Sigma_S^p(\lambda, \alpha, d) := \left\{ \bigcup_{b \in \mathcal{B}^{p,d}} T^b(\lambda C_1^{\alpha,d}) \right\} \cap \mathcal{Z}_p,$$

where $\mathcal{Z}_p = \{f \in C[0,1]^p : \int f(x)\,dx = 0\}$. The condition that f is centered can be imposed without any loss of generality due to the presence of the overall mean parameter μ in our regression model. The function spaces $\Sigma_S^p(\lambda, \alpha, d)$ make up M2. For M3, we consider additive convolutions of multiple Σ_S^p spaces with an additional restriction on the number of components a single predictor can appear in. For $k, \bar{d} \in \mathbb{N}$, $d \in \mathbb{N}^k$ define

$$\mathcal{B}^{p,k,d,\bar{d}} = \left\{ (b^1, \ldots, b^k) : b^s \in \mathcal{B}^{p,d_s}, b^s \ne b^t, 1 \le s \ne t \le k, \right.$$
$$\left. \sum_{t=1}^k b_j^t \le \bar{d}, 1 \le j \le p \right\}$$

and

$$\Sigma_A^{p,k,\bar{d}}(\lambda, \alpha, d)$$
$$= \left\{ f = \sum_{s=1}^k \lambda_s T^{b^s} f_s : f_s \in C_1^{\alpha_s, d_s}, 1 \le s \le k, (b^1, \ldots, b^k) \in \mathcal{B}^{p,k,d,\bar{d}} \right\} \cap \mathcal{Z}_p.$$

In studying minimax rates for a fixed k, one can set $\bar{d}$ as large as k. But in the more interesting large p small n scenario where k increases with p, the use of a fixed $\bar{d}$ is crucial for interpreting our results.

For a metric space $(\mathcal{S}, \rho)$, the covering ε-entropy of a subset $S \subset \mathcal{S}$ is the logarithm of the minimum number of ρ-balls of radius ε and centers in $\mathcal{S}$ needed to cover S, and is denoted $V(\varepsilon, S, \rho)$. A finite subset $A \subset S$ is called ε-packing in S if any two elements of A have a ρ-distance at least ε. The logarithm of the maximal cardinality of an ε-packing set in S is called the packing ε-entropy of S and is denoted $C(\varepsilon, S, \rho)$.

3. Minimax risks for large *p* small *n* regression. Precise calculations of r_n under M2 and M3 and theoretical results on whether these rates are achieved in practice are known only under additional simplifying assumptions on the shape of f, or, for inference tasks that are simpler than prediction. We provide a brief overview of known results before presenting our main theorem on minimax L_2 risk for regression under M3.

3.1. *A brief overview of existing results.* For linear regression where Σ is taken as the set of functions $f(x) = x^T\beta$ with β in an l_q $(q \le 1)$ ball of $\mathbb{R}^p$ and some additional regularity assumptions are made on the design matrix, [19] shows that

$$r_n^2 \asymp \begin{cases} d\log(p/d)/n, & \text{for } q = 0, \\ (\log d/n)^{1-q/2}, & \text{for } q \in (0, 1], \end{cases}$$

up to some multiplicative constant, where d is the number of important predictors. As shown in [9], these rates are the typical minimax risks associated with variable selection uncertainty. For $q = 0$, the l_q norm precisely encodes the sparsity condition of M_2. See [32, 33] and [34] for additional results and overviews. Many authors have established near minimax performance guarantees of various linear regression methods under the L_2 prediction loss; see, for example, [3, 6, 18] and [38].

As a nonlinear, nonparametric generalization of the linear model, [20] considers the completely additive special case of M3 where all k components are univariate and have the same smoothness $\alpha > 0$ and shows

$$r_n^2 \asymp kn^{-2\alpha/(2\alpha+1)} + \frac{k\log p}{n}.$$

Clearly, the minimax risk decomposes into two terms, where the first term is the sum of minimax risks of estimating each component and the second term is the variable selection uncertainty.

An entirely different generalization of the linear model is the sparse, fully nonparametric regression model M2. To the best of our knowledge, the only minimax rates result in this context is [9], which analyzes minimax risks of support recovery where the objective is to identify the important predictor rather than estimation of f itself. It is shown that if $d\log(p/d)/n$ is lower bounded by some positive constant, then for some constant $c > 0$,

$$\inf_{\hat{J}_n} \sup_{f \in \Sigma} P_f(\hat{J}_n \neq J_f) \ge c,$$

where $\hat{J}_n$ ranges over all variable selection estimators, that is, measurable maps of data to the space of all subsets of $\{1, \ldots, p\}$, Σ is the space of all differentiable functions that depend on only d many predictors and have squared integrable gradients, and $J_f \subset \{1, \ldots, p\}$ is the index set of truly important predictors associated with f. Because of this result, we refer to the term $d \log(p/d)/n$ as the risk associated with variable selection uncertainty. For large p, $d \log(p/d)$ is asymptotically of the same order as the logarithm of $\binom{p}{d}$, the number of ways to select d out of p predictors. Any estimation problem involving high-dimensional variable selection is likely to include a variable selection uncertainty term $d \log(p/d)/n$ in its minimax rate.

3.2. *New results on minimax rates under* M2 *and* M3. We calculate minimax L_2 risks under the following condition on the stochastic design:

ASSUMPTION Q. Q admits a probability density function (p.d.f.) q on $[0, 1]^p$ such that $\bar{q} := \sup_x q(x) < \infty$ and $\inf_{x \in [1/2-\Delta, 1/2+\Delta]^p} q(x) \geq \underline{q}$ for some $\underline{q} > 0$ and $0 < \Delta \leq 1/2$.

The requirement of q being lower bonded on some sub-hypercube inside $[0, 1]^p$ is crucial to obtaining sharp lower bounds on the minimax risk. This requirement is essentially equivalent to asking that X cannot be reduced to a lower dimension without some loss of information, for example, X cannot lie on a lower-dimensional subspace of manifold as assumed under M1.

THEOREM 3.1. *Under Assumption* Q, *there exist* $N_0 \in \mathbb{N}$, $0 < \underline{C} \leq 1 \leq \overline{C}$, *all depending only on* $\bar{d}$, $\max_s d_s$, $\min_s \alpha_s$, $\max_s \alpha_s$, $\min_s \lambda_s$, $\max_s \lambda_s$, *such that for all* $n > N_0$,

$$\underline{C} \underline{\varepsilon}_n^2 \leq r_n^2\big(\Sigma_A^{p,k,\bar{d}}(\lambda, \alpha, d), Q, \mu, \sigma\big) \leq \overline{C} \bar{\varepsilon}_n^2,$$

where

$$\underline{\varepsilon}_n^2 = \sum_{s=1}^{k} \lambda_s^2 \left(\frac{\sqrt{n}\lambda_s}{\sigma} \right)^{-4\alpha_s/(2\alpha_s + d_s)} + \frac{\sigma^2 \sum_s d_s}{n} \log \frac{p}{\sum_s d_s}$$

and

$$\bar{\varepsilon}_n^2 = \sum_{s=1}^{k} \lambda_s^2 \left(\frac{\sqrt{n}\lambda_s}{\sigma} \right)^{-4\alpha_s/(2\alpha_s + d_s)} + \frac{\sigma^2 \sum_s d_s}{n} \log \frac{p}{\min_s d_s}.$$

REMARK 3.2. By choosing $k = 1$ and $\bar{d} = 1$ in Theorem 3.1, we obtain the minimax risk for M2 as a simple corollary,

$$r_n^2\big(\Sigma_S^p(\lambda, \alpha, d), Q, \mu, \sigma\big) \asymp \lambda^2 \left(\frac{\sqrt{n}\lambda}{\sigma} \right)^{-4\alpha/(2\alpha + d)} + \frac{\sigma^2 d}{n} \log \frac{p}{d}. \tag{3.1}$$

REMARK 3.3. One can shed light on the scope and limitations of a model by investigating the conditions needed on the model parameters in order to bound the model's minimax risk by a given margin. From (3.1), the minimax risk of M2 consists of two terms. The second term is the typical risk associated with variable selection uncertainty [9] which remains small as long as $\log p \asymp n^{\beta}$ for some $\beta \in (0, 1)$, which gives the standard large p small n dynamics between sample size and predictor count. The first term in (3.1) is the minimax risk of estimating a d-variate, α-smooth regression function f_0 when there is no variable selection uncertainty. For a fixed smoothness level α, this term remains small as long as $d = o(\log n) = o(\log \log p)$ under standard large p small n dynamics. In other words, meaningful statistical learning is possible under M2 only when the true number of important predictors is much much smaller than the total predictor count.

REMARK 3.4. M3 offers a platform to break away from such extreme sparsity conditions. We consider two special cases for illustration under a standard large p small n dynamic: $\log p = n^{\beta}$ for some $\beta \in (0, 1)$, while allowing k to depend on n. First, suppose all additive components f_s have the same dimension ($d_s \equiv d$), smoothness ($\alpha_s \equiv \alpha$) and magnitude ($\lambda_s \equiv \lambda$), all of which remain fixed as k increases n. This situation includes as a special case the completely additive framework of [20]. From Theorem 3.1, the associated minimax risk $r_n^2 \asymp k n^{-2\alpha/(2\alpha+d)} + kd \log(p/d)/n$ which remains small as long as $k = o(\min\{n^{2\alpha/(2\alpha+d)}, \log p/n\}) \asymp o(n^{\gamma})$ for some $\gamma \in (0, 1)$. Thus, the total number of important predictors, which is of the order kd, could be as large as a fractional power of $\log p$, a number that is much larger than what is allowed under M2.

In the second case, consider an unbalanced case where d_s, α_s vary with s, but remain bounded as k increases with n, and the magnitudes diminish so that the series $\sum_s \lambda_s^{2d_s/(2\alpha_s+d_s)}$ is convergent. Theorem 3.1 suggests that a consistent estimator of f exists in this case as long as $\sum_{s=1}^{k} d_s = o(n)$, that is, the total number of important predictors is $o(n)$.

REMARK 3.5. Consider another unbalanced scenario where k is fixed and one additive component is much more complex than the rest, that is, $d_1/\alpha_1 \gg d_s/\alpha_s$ for $s = 2, \ldots, k$. In this case, Theorem 3.1 gives a minimax risk $r_n^2 \sim n^{-2\alpha_1/(2\alpha_1+d_1)} + \sum_{s=1}^{k} d_s \log(p/d_s)/n$, where the first term is dominated by the largest risk of all additive components, while the second term is still determined by the overall variable selection uncertainty. Therefore, the difficulty of estimating a function with an additive form is determined by the estimation difficulty of its "hardest" component.

4. Adaptive near minimax optimality of Bayesian additive Gaussian process regression. A Gaussian process (GP) on an Euclidean set K is a random element $W = (W_x : x \in \mathcal{X})$ of the supremum-norm Banach space of continuous

functions over $\mathcal{X}$ such that any linear functional of W is univariate Gaussian [29]. The probability law of a GP W is completely determined by the mean and covariance functions $m(x) = EW_x$ and $\mathcal{C}(x, x') = E(W_x - m(x))(W_{x'} - m(x'))$ and is denoted by $\mathrm{GP}(m, \mathcal{C})$. For any function $m : \mathcal{X} \to \mathbb{R}$ and any nonnegative definite function $\mathcal{C} : \mathcal{X} \times \mathcal{X} \to (0, \infty)$, there exist a GP W with law $\mathrm{GP}(m, \mathcal{C})$.

Adaptivity and near minimax optimality of Bayesian Gaussian process regression methods are known for low-dimensional applications [31]. In GP regression, f is assigned a $\mathrm{GP}(m, \mathcal{C})$ prior and inference on f is carried out by summarizing the resulting posterior distribution given data, which also remains a GP law [21]. Theoretical treatments of GP regression have typically focused on $m \equiv 0$ and $\mathcal{C}(x, x') = \mathcal{C}^{\mathrm{SE}}(x, x') = \exp(-\|x - x'\|^2)$, the square exponential covariance function, with additional hyper-parameters inserted inside the covariance function [8, 27, 31]. In particular, in order to achieve adaptation to unknown smoothness, [31] considers as prior distribution the law of a rescaled process W^A defined as $W^A_x = W_{Ax}$ where $W \sim \mathrm{GP}(0, \mathcal{C}^{\mathrm{SE}})$ and A^p follows a gamma distribution, and proves the resulting posterior distribution contracts to the true f at the minimax rate $n^{-\alpha/(2\alpha+p)}$ up to a $\log n$ factor when f is Hölder α-smooth. Extensions to anisotropic function spaces are carried out by [1].

4.1. *Additive Gaussian process regression.* For a stochastic process $W = (W_x : x \in \mathbb{R}^p)$, a scalar $a > 0$ and a binary inclusion vector $b \in \{0, 1\}^p$, define a *selective-rescaled* process $W^{a,b} = (W^{a,b}_x : x \in [0, 1]^p)$ by $W^{a,b}_x = W_{ab \odot x}$ where $\odot$ is the elementwise product operator. Toward a Bayesian estimation of regression functions f described by M3, we consider the following additive Gaussian process (add-GP) prior distribution on f:

$$f = L_1 W_1^{A_1, B_1} + \cdots + L_K W_K^{A_K, B_K}; \qquad K \sim \pi, \tag{4.1}$$

where π is a probability distribution on $\mathbb{N}$, and $L_s W_s^{A_s, B_s}$ are IID copies of the process $LW^{A,B}$ defined as: $W \in C(\mathbb{R}^p)$, $L \in \mathbb{R}_+$ and $(A, B) \in \mathbb{R}_+ \times \{0, 1\}^p$ are mutually independent random elements distributed as

$$\begin{aligned} &W \sim \mathrm{GP}\big(0, \mathcal{C}^{\mathrm{SE}}\big), \qquad L \sim h, \\ &B \sim \left[\bigotimes_{j=1}^{p} \mathrm{Be}\left(\frac{1}{p}\right)\right]\Bigg|_{|B| \le D_0}, \qquad A^{|B|} \big| B \sim \mathrm{Ga}(a_1, a_2), \end{aligned} \tag{4.2}$$

where h is a density function on $(0, \infty)$ and a_1, a_2, D_0 are prespecified, positive valued hyper-parameters.

To complete the add-GP prior specification, we need to specify a prior distribution on (μ, σ). We consider $(\mu, \sigma) \sim \pi_\mu \times \pi_\sigma$ where π_μ is a Gaussian distribution and π_σ admits density function on $\mathbb{R}_+$ with a compact support inside $(0, \infty)$.

4.2. *Posterior contraction rates.* For any $x^{1:\infty} = (x^1, x^2, \ldots) \in ([0,1]^p)^\infty$ and any $\theta = (\mu, f, \sigma)$, let $P_\theta(\cdot|x^{1:\infty})$ denote the conditional distribution of $(Y^i : i \in \mathbb{N})$ given $X^i = x^i$, $i \in \mathbb{N}$, under (1.1). Let $\Pi_n(\cdot|(x^i, y^i), 1 \le i \le n)$ denote the posterior distribution of θ under the add-GP prior given $(X^i, Y^i) = (x^i, y^i)$, $1 \le i \le n$. Following [10, 30], the posterior contraction rate of the add-GP prior at any $\theta^* = (\mu^*, f^*, \sigma^*)$ is said to be at least ε_n if for every $x^{1:\infty}$, other than in a Q^∞-null set,

$$\Pi_n\{\|\mu + f - \mu - f^*\|_n + |\sigma - \sigma^*| \ge M\varepsilon_n|(x^i, Y^i), 1 \le i \le n\} \overset{P_{\theta^*}(\cdot|x^{1:\infty})}{\longrightarrow} 0$$

as $n \to \infty$ for some constant M, where $\|\cdot\|_n$ denotes an empirical version of the $L_2(Q)$ norm: $\|f\|_n^2 = (1/n)\sum_{i=1}^n f^2(x^i)$. It is possible to replace $\|\cdot\|_n$ with $\|\cdot\|_Q$ by appealing to the techniques developed for GP priors in Section 2.4 in [36], but we omit the details.

THEOREM 4.1. *Under Assumption* Q, *for any* $\mu^* \in \mathbb{R}$, $\sigma^* \in \text{support}(\pi_\sigma)$ *and* $f^* \in \Sigma_A^{p,k,\bar{d}}(\lambda^*, \alpha^*, d^*)$ *with* $\max_s d_s^* \le D_0$ *and* $k \le K_0$, *the posterior contraction rate at* $\theta^* = (\mu^*, f^*, \sigma^*)$ *is of the order* $\varepsilon_n(\log n)^{(1+D_0)/2}$ *where*

$$\varepsilon_n^2 = \sum_{s=1}^k \lambda_s^{*2}\left(\frac{\sqrt{n}\lambda_s^*}{\sigma^*}\right)^{-4\alpha_s^*/(2\alpha_s^*+d_s^*)}(\log n)^{2q_s} + \frac{\sigma^{*2}\sum_s d_s^*}{n}\log p$$

with $q_s = (1 + d_s^*)/(2 + d_s^*/\alpha_s^*)$, $1 \le s \le k$, *provided* $K_0 \log p \le n\varepsilon_n^2$.

When p grows with n, add-GP regression essentially employs a sequence of priors changing with n. In this case, it is possible and useful to also let K_0 grow with n and study posterior contraction rate at a sequence of $f^* = f_n^*$ changing with n. Theorem 4.1 remains valid as long as $K_0 \log p \le n\varepsilon_n^2$, the true number of components $k \le K_0$, α_s are bounded from above and below and $\max_s \lambda_s$ is bounded.

REMARK 4.2. Related work on estimation of f under M3 includes [20], where convergence rates are investigated for an M-estimator with a sparsity penalty on the number of additive components and smoothness penalties on each components. However, [20] considers only univariate components. In [25], PAC-Bayesian bounds are derived for general additive regression with additive GP priors. However, [25] assumes that the covariate vector X is pre-divided into M subsets $(X_{(1)}, \ldots, X_{(M)})$ and $f(x) = \sum_{m=1}^M f_m(x_{(m)})$, with sparsity constraints on the component functions. Both these studies assume that important predictors are not shared across components, which makes the studied methods somewhat restricted in application. A lack of overlap comes with the technical advantage that $\|\sum_s f_s\|_Q^2$ decomposes to $\sum_s \|f_s\|_Q^2$ if every f_s has Q-integral 0. In the more general case where components are allowed to share predictors, a naïve application of the Cauchy–Schwarz inequality gives $\|\sum_s f_s\|_Q^2 \le k\sum_s \|f_s^2\|_Q$, but the multiplication by k results in sub-optimal rates unless K_0 grows extremely slowly in n.

Our assumption that any predictor can appear in at most $\bar{d}$ many components, for some fixed $\bar{d}$, overcomes this difficulty with the help of Lemma 6.5.

5. Proofs of the main results.

5.1. *Minimax rates.*

LEMMA 5.1. *For every $\alpha, \lambda, d \in \mathbb{N}$ there exist $N_0 > 0$, $0 < \underline{C} \le 1 \le \overline{C}$, such that for any $n > N_0$ and all $p \in \mathbb{N}$, the ε_n that solves $C(\varepsilon_n, \Sigma_S^p(\lambda, \alpha, d), \|\cdot\|) = n\varepsilon_n^2/\sigma^2$ satisfies*

$$\underline{C} \le \frac{\varepsilon_n^2}{\lambda^2(\sqrt{n}\lambda/\sigma)^{-4\alpha/(2\alpha+d)} + (\sigma^2/n)\log\binom{p}{d}} \le \overline{C}.$$

PROOF. Let ε_1, M_0, M_1 be as in Lemma 6.2. Without loss of generality, $M_0 \le 1 \le M_1$. Let $\delta_n^2 = \lambda^2(\sqrt{n}\lambda/\sigma)^{-4\alpha/(2\alpha+d)} + (\sigma^2/n)\log\binom{p}{d}$ and set N_0 large enough such that $\delta_n < \varepsilon_1$ for all $n > N_0$. For the remainder of this proof, abbreviate $\Sigma_S^p(\lambda, \alpha, d)$ to Σ_S. The arguments below mostly rest on the fact that ε-packing entropy is nonincreasing in ε. Note that

$$C(M_1^{1/2}\delta_n, \Sigma_S, \|\cdot\|) \le C\left(\lambda\left(\frac{\sqrt{n}\lambda}{\sigma}\right)^{-2\alpha/(2\alpha+d)}, \Sigma_S, \|\cdot\|\right) \le M_1 n\delta_n^2/\sigma^2,$$

where the second inequality follows by sticking in $\lambda(\sqrt{n}\lambda/\sigma)^{-2\alpha/(2\alpha+d)}$ as ε in Lemma 6.2. Hence, $\varepsilon_n \le M_1^{1/2}\delta_n$. Also, by Lemma 6.2,

$$\begin{aligned} &C\left(\left(\max\left\{\lambda\left(\frac{\sqrt{n}\lambda}{\sigma}\right)^{-2\alpha/(2\alpha+d)}, \frac{\sigma}{\sqrt{n}}\log^{1/2}\binom{p}{d}\right\}\right), \Sigma_S, \|\cdot\|\right) \\ &\quad \ge M_0 n \max\left\{\lambda\left(\frac{\sqrt{n}\lambda}{\sigma}\right)^{-2\alpha/(2\alpha+d)}, \frac{\sigma}{\sqrt{n}}\log^{1/2}\binom{p}{d}\right\}^2 \Big/ \sigma^2 \end{aligned}$$

and hence $\varepsilon_n^2 \ge M_0 \max\{\lambda(\frac{\sqrt{n}\lambda}{\sigma})^{-2\alpha/(2\alpha+d)}, \frac{\sigma}{\sqrt{n}}\log^{1/2}\binom{p}{d}\} \ge M_0\delta_n^2/2$. This proves the result with $\underline{C} = M_0/2$ and $\overline{C} = M_1$. □

PROOF OF THEOREM 3.1. By Theorem 6 of [35], the minimax risk r_n is the solution to $C(r_n, \Sigma_A^{p,k,\bar{d}}(\lambda, \alpha, d), \|\cdot\|_Q) = nr_n^2/\sigma^2$. For $1 \le s \le k$, let δ_{ns} be the solution to $C(\varepsilon, \Sigma_S^{p_s}(\lambda_s, \alpha_s, d_s), \|\cdot\|) = n\varepsilon^2/\sigma^2$. From Lemma 5.1, there are $N_s > 0, 0 < \underline{C}_s \le 1$, such that for all $n > N_s$, $\delta_{ns}^2 \ge \underline{C}_s\{\lambda_s^2(\sqrt{n}\lambda_s/\sigma)^{-4\alpha_s/(2\alpha_s+d_s)} + (\sigma^2/n)\log\binom{p_s}{d_s}\}$. Denote $\delta_n = (\delta_{n1}, \ldots, \delta_{nk})$, $n > N = \max_s N_s$. Then, by Theorem 6.4, with $b_0 = \underline{q}^{1/2}\Delta^{\max_s \alpha_s + \max_s d_s/2}$,

$$\begin{aligned} C\left(\frac{b_0\|\delta_n\|}{2}, \Sigma_A^{p,k,\bar{d}}(\lambda, \alpha, d), \|\cdot\|_Q\right) &\ge \frac{1}{4}\left\{\frac{3}{4}n\frac{\|\delta_n\|^2}{\sigma^2} - k\log 2\right\} \\ &\ge \frac{1}{16}\frac{n\|\delta_n\|^2}{\sigma^2} \end{aligned} \tag{5.1}$$

provided $k \log 2 \le n\|\delta_n\|^2/(2\sigma^2)$, and hence,

$$r_n^2 \ge \frac{\|\delta_n\|^2}{16} \ge \frac{1}{16}\left\{\sum_{s=1}^{k} \underline{C}_s \lambda_s^2 \left(\frac{\sqrt{n}\lambda_s}{\sigma}\right)^{-4\alpha_s/(2\alpha_s+d_s)} + \frac{\sigma^2}{n}\log\binom{p_s}{d_s}\right\}$$

$$\ge \underline{C}\left\{\sum_{s=1}^{k} \lambda_s^2 \left(\frac{\sqrt{n}\lambda_s}{\sigma}\right)^{-4\alpha_s/(2\alpha_s+d_s)} + \frac{\sigma^2}{n}\sum_s d_s \log p\right\},$$

for some $\underline{C}$.

Next, let $\varepsilon_n = (\varepsilon_{n1}, \ldots, \varepsilon_{nk})$ where ε_{ns} is the solution to $C(\varepsilon_s, \Sigma_S^p(\lambda_s, \alpha_s, d_s), \|\cdot\|) = n\varepsilon_s^2/\sigma^2$, $1 \le s \le k$. By Lemma 5.1, there are $N_s > 0$, $\overline{C}_s \ge 1$, such that for all $n > N_s$, $\varepsilon_{ns}^2 \le \overline{C}_s\{\lambda_s^2(\sqrt{n}\lambda_s/\sigma)^{-4\alpha_s/(2\alpha_s+d_s)} + (\sigma^2/n)\log\binom{p}{d_s}\}$. Set $N = \max_s N_s$. By Theorem 6.4 again, $C(4\bar{q}^{1/2}\sqrt{B}\|\varepsilon_n\|, \Sigma_A^{p,k,\bar{d}}(\lambda, \alpha, d), \|\cdot\|_Q) \le n\|\varepsilon_n\|^2/\sigma^2$, and hence

$$r_n^2 \le 16\bar{q}B\|\varepsilon_n\|^2 \le \overline{C}\left\{\sum_{s=1}^{k} \lambda_s^2 \left(\frac{\sqrt{n}\lambda_s}{\sigma}\right)^{-4\alpha_s/(2\alpha_s+d_s)} + \frac{\sigma^2}{n}\sum_s d_s \log p\right\},$$

completing the proof. □

5.2. *Posterior contraction rates of add-GP.* According to [10], Theorem 1 and Section 7.7, and [29], the conclusion of Theorem 4.1 holds if for Q^∞-almost every $x^{1:\infty}$ there exist $\mathcal{F}_n \subset C(\mathbb{R}^p)$, $n \in \mathbb{N}$, such that

$$\Pi\big(\|\mu + f - \mu^* - f^*\|_n \le \varepsilon_n\big) \ge e^{-n\varepsilon_n^2}, \tag{5.2}$$

$$\Pi(\mu + f \notin \mathcal{F}_n) \le e^{-4n\varepsilon_n^2}, \tag{5.3}$$

$$\log N\big(\bar{\varepsilon}_n, \mathcal{F}_n, \|\cdot\|_\infty\big) \le n\bar{\varepsilon}_n^2, \tag{5.4}$$

where $\bar{\varepsilon}_n = \varepsilon(\log n)^{(1+D_0)/2}$ and Π denotes the add-GP prior on (μ, f, σ). These conditions map to one to one to concentration properties of the selective-rescaled Gaussian processes underlying the add-GP formulation. Without loss of generality, we assume the prior density h on L is a folded Gaussian p.d.f., and that $a_1 = a_2 = 1$.

Two important objects associated with any Gaussian process are its reproducing kernel Hilbert space (RKHS) and concentration function. The RKHS of any GP $W = (W_x : x \in \mathcal{X})$, with $\mathcal{X} \subset \mathbb{R}^d$, is defined to be the set $\mathbb{H}$ of all function $h : \mathcal{X} \to \mathbb{R}$ that can be written as $h(x) = EW_x S$ for some S in the closure of the linear span of the collection of random variables $\{W_x : t \in \mathcal{X}\}$ in L_2 norm. The set $\mathbb{H}$ is a Hilbert space with $\langle EWS_1, EWS_2\rangle_{\mathbb{H}} = ES_1S_2$. With W seen as an element in $C(\mathcal{X})$, its concentration function at any $w \in C(\mathcal{X})$ is defined as

$$\phi_w(\varepsilon) = \inf_{h\in\mathbb{H}:\, \|h-w\|_\infty \le \varepsilon} \|h\|_{\mathbb{H}}^2 - \log \Pi\big(\|W\|_\infty \le \varepsilon\big), \qquad \varepsilon > 0.$$

We make use of the following well-known inequalities involving the RKHS and the concentration function:

$$e^{-\phi_w(2\varepsilon)} \geq \Pi\big(\|W - w\|_\infty \leq 2\varepsilon\big) \geq e^{-\phi_w(\varepsilon)}, \tag{5.5}$$

$$\Pi(W \notin M\mathbb{H}_1 + \varepsilon\mathbb{B}_1) \leq 1 - \Phi\big(\Phi^{-1}\big(e^{-\phi_0(\varepsilon)}\big) + M\big), \tag{5.6}$$

$$V\big(\varepsilon, M\mathbb{H}_1, \|\cdot\|_\infty\big) \leq 1/2 + \phi_0\bigg(\frac{\varepsilon}{2M}\bigg). \tag{5.7}$$

See Lemma 5.3 of [30] for a proof of (5.5). The inequality (5.6) is the well-known Borell's inequality [4], and the right-hand side can be further bounded by $\exp\{-M^2/8\}$ when $M^2/8 \geq \phi_0(\varepsilon)$ since $\Phi^{-1}(u) \geq -\sqrt{2\log(1/u)}$ for all $u \in (0,1)$. Inequality (5.7) holds because the right-hand side gives an upper bound to $C(\varepsilon/(2M), \mathbb{H}_1, \|\cdot\|_\infty)$, since, if $h_1, \ldots, h_N \in \mathbb{H}_1$ are $\varepsilon/(2M)$-separated in $\|\cdot\|_\infty$ then $1 \geq \sum_{j=1}^N \Pi(W \in h_j + \{\varepsilon/(2M)\}\mathbb{B}_1) \geq N\exp\{-1/2 - \phi_0(\varepsilon/(2M))\}$ by (5.5).

For any $b \in \{0,1\}^p$, $a > 0$, let $\mathbb{H}^{a,b}$ and $\phi_w^{a,b}$ denote the RKHS and the concentration function of the selective-rescaled GP $W^{a,b}$ introduced in Section 4.1. By definition, $W^{a,b}$ is isomorphic to a d dimensional, rescaled GP $\tilde{W}^a$ with $\tilde{W} \sim \mathrm{GP}(0, C^{\mathrm{SE}})$ on $\mathbb{R}^d$, whose RKHS and concentration function have been studied extensively in [31]. The following results, which are direct consequences of Lemmas 4.3, 4.6, 4.7 and 4.8 of [31], are of particular interest to us:

$$w \in T^b C^{\alpha,|b|} \quad \Longrightarrow \quad \phi_w^{a,b}(\varepsilon) \leq G_0 a^{|b|}\bigg(\log\frac{a}{\varepsilon}\bigg)^{1+|b|}, \qquad \forall a \geq a_0, \forall \varepsilon < \varepsilon_0 \wedge G_1 a^{-\alpha}, \tag{5.8}$$

$$a_1^{|b|/2}\mathbb{H}_1^{a_1,b} \subset a_2^{|b|/2}\mathbb{H}_1^{a_2,b} \qquad \forall 0 < a_1 < a_2, \tag{5.9}$$

$$h \in \mathbb{H}_1^{a,b} \quad \Longrightarrow \quad |h(0)| \leq 1, \qquad \big\|h - h(0)\big\|_\infty \leq a|b|. \tag{5.10}$$

In (5.8), the constants $\varepsilon_0, a_0, G_0, G_1$ depend only on w and $|b|$.

LEMMA 5.2. *Suppose* $(\varepsilon_n, n \geq 1)$ *satisfies* $n^{-\gamma_1} \leq \varepsilon_n \leq n^{-\gamma_2}$ *for some* $0 < \gamma_1 < \gamma_2 < 1/2$ *and* $K_0 \log p \leq n\varepsilon_n^2$. *Then there exists a sequence of sets* $\mathcal{F}_n \subset C[0,1]^d$ *satisfying*

$$\Pi(\mu + f \notin \mathcal{F}_n) \leq \exp\big(-4n\varepsilon_n^2\big) \tag{5.11}$$

and

$$\log N\big(\bar{\varepsilon}_n, \mathcal{F}_n, \|\cdot\|_\infty\big) \leq n\bar{\varepsilon}_n^2 \tag{5.12}$$

with $\bar{\varepsilon}_n \asymp \varepsilon_n(\log n)^{(1+D_0)/2}$.

PROOF. Let $R_n = K_3 n\varepsilon_n^2$, where K_3 is a large constant to be determined later, and define

$$\bar{L}_n^2 = R_n, \qquad M_n^2 = 8K_4 R_n(\log n)^{1+D_0}, \qquad \delta_n = \frac{\varepsilon_n}{K_0|b|\bar{L}_n M_n},$$

for some constant K_4. By (5.8), and the fact that $\phi_0^{a,b}(\varepsilon)$ is nondecreasing in a, the constant K_4 can be chosen large enough so that

$$M_n^2 \geq 8\phi_0^{a,b}\left(\frac{\varepsilon_n}{K_0\bar{L}_n}\right) \qquad \forall b \in \bigcup_{d\leq D_0} \mathcal{B}^{p,d}, \forall a \leq R_n^{1/|b|}, \tag{5.13}$$

for all large n. Set $N = \lceil D_0 \log\{R_n/\delta_n\}/(\log 4)\rceil$ and take $\Delta_n(b) := \{\delta_n 4^{j/|b|} : 1 \leq j \leq N\}$. For every $r \in \{0\} \cup \Delta_n(b)$, define

$$\mathcal{F}_n^{r,b} = \bigcup_{a\in(0,\delta_n]\cup\{r\}\setminus\{0\}} 2\bar{L}_n M_n \mathbb{H}_1^{a,b} + \frac{\varepsilon_n}{K_0}\mathbb{B}_1.$$

Consider the sieves

$$\mathcal{F}_n := [-\sqrt{n},\sqrt{n}] \oplus \bigcup_{1\leq k\leq K_0} \bigcup_{\substack{b^1,\ldots,b^k\in\bigcup_{d\leq D_0}\mathcal{B}^{p,d},\\ r_s\in\{0\}\cup\Delta_n(b^s),1\leq s\leq k,\\ \sum_s r_s^{|b^s|}\leq 4R_n}} \mathcal{F}_n^{r_1,b^1} \oplus \cdots \oplus \mathcal{F}_n^{r_k,b^k},$$

for $n \in \mathbb{N}$.

Fix any $k \in \{1,\ldots,K_0\}$, $b^1,\ldots,b^k \in \bigcup_{d\leq D_0}\mathcal{B}^{p,d}$ and $a \in \mathbb{R}_+^k$ satisfying $\sum_{s=1}^k a_s^{|b_s|} \leq R_n$. For $1 \leq s \leq k$, if $a_s \leq \delta_n$ set $r_s = 0$, otherwise find $r_s \in \Delta_n(b^s)$ such that $r_s 4^{-1/|b^s|} < a_s \leq r_s$. Then $\sum_{s=1}^k r_s^{|b^s|} \leq 4R_n$ and by (5.9), $\bar{L}_n M_n \mathbb{H}_1^{a_s,b^s} + (\varepsilon_n/K_0)\mathbb{B}_1 \subset \mathcal{F}_n^{r_s,b^s}$ for all $1 \leq s \leq k$. Therefore,

$$\begin{aligned}
&\Pi\{\mu + f \notin \mathcal{F}_n | K = k, (A_s,B^s) = (a_s,b^s), 1\leq s\leq k\}\\
&\quad\leq \Pi(|\mu| > \sqrt{n}) + \sum_{s=1}^k \Pi\left\{L_s W^{a_s,b^s} \notin \bar{L}_n M_n \mathbb{H}_1^{a_s,b^s} + \frac{\varepsilon_n}{K_0}\mathbb{B}_1\right\}\\
&\quad\leq e^{-n/2} + \sum_{s=1}^k\left[\Pi(L_s > \bar{L}_n) + \Pi\left\{W^{a_s,b^s} \notin M_n\mathbb{H}_1^{a_s,b^s} + \frac{\varepsilon_n}{K_0\bar{L}_n}\mathbb{B}_1\right\}\right]\\
&\quad\leq e^{-n/2} + k\{e^{-\bar{L}_n^2} + e^{-M_n^2/8}\}\\
&\quad\leq 3ke^{-R_n}
\end{aligned}$$

for all large n, by (5.6) and (5.13) and the fact $R_n = o(n)$. Consequently,

$$\begin{aligned}
&\Pi(\mu + f \notin \mathcal{F}_n)\\
&\quad\leq \max_{\substack{1\leq k\leq K_0,\\ b^1,\ldots,b^k\in\bigcup_{d\leq D_0}\mathcal{B}^{p,d}}} \left\{\Pi\left(\sum_{s=1}^k A_s^{|b^s|} > R_n \middle| B^s = b^s, s = 1,\ldots,k\right) + 3ke^{-R_n}\right\}\\
&\quad\leq \Pi(G > R_n) + 3K_0e^{-R_n}
\end{aligned}$$

with $G \sim \mathrm{Ga}(K_0, 1)$. Notice $\Pi(G > R_n) \le \exp\{-R_n/2 + K_0 \log 2\}$. Therefore, by the assumption on K_0, $\Pi(\mu + f \notin \mathcal{F}_n)$ is bounded by $\exp(-4n\varepsilon_n^2)$ for all large n, provided K_3 is chosen suitably large.

By (5.10), when $r = 0$, $\mathcal{F}_n^{r,b} \subset 2\bar{L}_n M_n \cdot [-1, 1] + (2\varepsilon_n/K_0)\mathbb{B}_1$, and hence could be covered by $\lceil 4\bar{L}_n M_n K_0/\varepsilon_n \rceil$ many or fewer balls of supremum norm radius $3\varepsilon_n/K_0$. When $r > 0$, by (5.7), at most another $1/2 + \phi_0^{r,b}(\varepsilon_n/(2\bar{L}_n M_n K_0))$ many balls may be needed to maintain $3\varepsilon_n/K_0$ covering. Therefore, by (5.8), $V(3\varepsilon_n/K_0, \mathcal{F}_n^{r,b}, \|\cdot\|_\infty) \le D_1\{r^{|b|}(\log n)^{1+D_0} + \log n\}$ for every $r \in \{0\} \cup \Delta_n(b)$, for some constant D_1 that depends only on D_0, as long as $|b| \le D_0$. Consequently,

$$V\big(4\varepsilon_n, \mathcal{F}_n, \|\cdot\|_\infty\big) \le D_1\big\{R_n(\log n)^{1+D_0} + K_0 \log n\big\} + \log M,$$

where M is the size of the finite set $\{((r_1, b^1), \ldots, (r_k, b^k)) : 1 \le k \le K_0, b^s \in \bigcup_{d \le D_0} \mathcal{B}^{p,d}, r_s \in \{0\} \cup \Delta_n(b^s), 1 \le s \le k\}$. This proves the result because of the assumption on K_0, since $\log M \le \log[K_0\{p^{D_0}(N+1)\}^{K_0}] \le C_6 K_0 \log p$ for some constant C_6 that depends only on D_0. $\square$

LEMMA 5.3. *Under the conditions of Theorem 4.1, for Q-almost every $x^{1:\infty}$, $\Pi(\|\mu + f - \mu^* - f^*\|_n \le \varepsilon_n) \ge \exp(-n\varepsilon_n^2)$ for all large n where*

$$\varepsilon_n^2 \asymp \sum_{s=1}^k \lambda_s^2 \Big(\frac{\sqrt{n}\lambda_s}{\sigma^*}\Big)^{-4\alpha_s/(2\alpha_s + d_s)} (\log n)^{2q_s} + \frac{\sigma^{*2}\sum_s d_s}{n} \log p$$

with $q_s = (1 + d_s)/(2 + d_s/\alpha_s)$, $s = 1, \ldots, k$.

PROOF. By Lemma 6.6, with $\mathcal{F}_n$ as in (5.12), we only need to show $\Pi(\|\mu + f - \mu^* - f^*\|_Q \le \varepsilon_n, \mu + f \in \mathcal{F}_n, \|\mu + f - \mu^* - f^*\|_\infty \le 1) \ge \exp(-n\varepsilon_n^2)$. By inequality (5.11) and the fact $\|\cdot\|_Q \le \bar{q}^{1/2}\|\cdot\|$ it suffices to show that

$$\Pi\big(\|\mu + f - \mu^* - f^*\| \le \varepsilon_n, \|\mu + f - \mu^* - f^*\|_\infty \le 1\big) \ge \exp\big(-n\varepsilon_n^2\big).$$

We can write $f^* = \sum_{s=1}^k \lambda_s T^{b^s} f_s^*$ where $b^s \in \mathcal{B}^{p,d_s}$, $f_s^* \in C_1^{\alpha_s,d_s} \cap \mathcal{Z}_d$, $1 \le s \le k$ and $\max_{1 \le j \le p} \sum_{s=1}^k b_j^s \le \bar{d}$. Let $\delta_{ns} = \lambda_s(\sqrt{n}\lambda_s/\sigma)^{-2\alpha_s/(4\alpha_s + d_s)}(\log n)^{q_s}$, $1 \le s \le k$ and $\delta_n = (\delta_{n1}, \ldots, \delta_{nk})$. Set $B = 1 + \max_s d_s(\bar{d} - 1)$.

For any $a > 0$, $b \in \{0, 1\}^p$ define the Gaussian variable $U^{a,b} = \int W_x^{a,b}\,dx$. Then the Gaussian process $V^{a,b} = W^{a,b} - E(W^{a,b}|U^{a,b})$ satisfies $\int V_x^{a,b}\,dx = 0$, and is independent of the process $E(W^{a,b}|U^{a,b}) = Z\psi^{a,b}$ where $Z \sim N(0, 1)$ and $\psi^{a,b}(x) = \mathrm{cov}(U^{a,b}, W_x^{a,b})/\mathrm{var}^{1/2}(U^{a,b})$, $x \in [0, 1]^p$. By Cauchy–Schwarz inequality, $\|\psi^{a,b}\|_\infty \le 1$. Clearly, $W^{a,b}$ decomposes as $W^{a,b} = V^{a,b} + Z\psi^{a,b}$.

Therefore, for any $\ell, a \in \mathbb{R}_+^k$ and given $K = k$, $(L_s, A_s, B^s) = (\ell_s, a_s, b^s)$, $1 \le s \le k$, we can decompose the additive-GP process f as $f = \sum_{s=1}^k \ell_s Z_s \psi^{a_s,b^s} + \sum_{s=1}^k \bar{f}_s$, where Z_s are independent $N(0, 1)$ variables, $\bar{f}_s$ are mutually independent

with probability laws same as those of $\ell_s V^{a_s,b_s}$, and these two sets of random quantities are independent. Consequently, for large enough n,

$$\begin{aligned}
&\Pi\{\|f-f^*\| \le \sqrt{1+25B}\|\delta_n\|, \|f-f^*\|_\infty \le 1/2 \\
&\qquad |K=k, (L_s, A_s, B^s) = (\ell_s, a_s, b^s), 1 \le s \le k\} \\
&\quad \ge \Pi\Big(\Big\|\sum_s \ell_s Z_s \psi^{a_s,b^s}\Big\| \le \|\delta_n\|, \Big\|\sum_s \ell_s Z_s \psi^{a_s,b^s}\Big\|_\infty \le 1/4\Big) \\
&\qquad \times \Pi\Big\{\Big\|\sum_s (\bar f_s - \lambda_s T^{b^s} f_s^*)\Big\| \le 5\sqrt{B}\|\delta_n\|, \Big\|\sum_s (\bar f_s - \lambda_s T^{b^s} f_s^*)\Big\|_\infty \le 1/4\Big| \\
&\qquad\qquad K=k, (L_s, A_s, B^s) = (\ell_s, a_s, b^s), 1 \le s \le k\Big\} \\
&\quad \ge \Pi\Big(\Big\|\sum_s \ell_s Z_s \psi^{a_s,b^s}\Big\| \le \|\delta_n\|, \Big\|\sum_s \ell_s Z_s \psi^{a_s,b^s}\Big\|_\infty \le 1/4\Big) \\
&\qquad \times \prod_{s=1}^k \Pi(\|\ell_s V^{a_s,b^s} - \lambda_s T^{b^s} f_s^*\| \le 5\delta_{ns}, \|\ell_s V^{a_s,b^s} - \lambda_s T^{b^s} f_s^*\|_\infty \le \delta_{ns}),
\end{aligned}$$

because of Lemma 6.5, since by the assumption on f_s^* and the construction of $\bar f_s$, we have for every $1 \le s \le k$, $\langle \bar f_s - \lambda_s T^{b^s} f_s^*, \bar f_t - \lambda_s T^{b^s} f_s^* \rangle_R \ne 0$ for at most $r_s = 1 + d_s(\bar d - 1)$ many $1 \le t \le k$.

If $\ell_s \in \lambda_s \cdot [1, 1+\delta_{ns}]$, then $\{\|\ell_s V^{a_s,b^s} - \lambda_s T^{b^s} f_s^*\| \le 5\delta_{ns}\} \supset \{\lambda_s \|V^{a_s,b^s} - T^{b^s} f_s^*\|_\infty \le 4\delta_{ns}\} \supset \{\lambda_s \|W^{a_s,b^s} - T^{b^s} f_s^*\|_\infty \le 2\delta_{ns}\}$. When $a_s^{d_s} \in (G_1/\delta_{ns})^{d_s/\alpha_s} \cdot [1,2]$, where G_1 is as in (5.8), the last probability can be lower bounded by $\exp\{-G_2(\lambda_s/\delta_{ns})^{d_s/\alpha_s} \log(1/\delta_{ns})^{1+d_s}\} \ge \exp\{-G_2 n\delta_{ns}^2/\sigma^2\}$ for some constant G_2, for all large n, by (5.5) and (5.8). For the same choices of ℓ_s, a_s, $1 \le s \le k$, $\Pi(\|\sum_s \ell_s Z_s \psi^{a_s,b^s}\| \le \|\delta_n\|, \|\sum_s \ell_s Z_s \psi^{a_s,b^s}\|_\infty \le 1/4) \ge \exp\{-G_3 k \log n\}$ for some constant G_3, for all large n. Therefore, by the assumption on K_0,

$$\begin{aligned}
&\Pi(\|f-f^*\| \le \sqrt{1+25B}\|\delta_n\|, \|f-f^*\|_\infty \le 1/2) \\
&\quad \ge \exp(-G_4 2n\|\delta_n\|^2/\sigma^2)\Pi(K=k) \\
&\qquad \times \prod_{s=1}^k \{\Pi(L_s \in \lambda_s \cdot [1, 1+\delta_{ns}])\Pi(B^s = b^s) \\
&\qquad\qquad \times \Pi(A_s^{d_s} \in (G_1/\delta_{ns})^{d_s/\alpha_s} \cdot [1,2] | |B_s| = d_s)\} \\
&\quad \ge G_5 \exp\Big\{-G_6 n\Big\{\|\delta_n\|^2 + \frac{\sigma^2 \sum_s d_s}{n} \log p\Big\}\Big\}
\end{aligned}$$

for all large n for some constants G_5, G_6 that depend only on $\max_s d_s$, $\min_s \lambda_s$, $\max_s \lambda_s$, $\min_s \alpha_s$ and $\max_s \alpha_s$. This proves the result since μ is independent of f and $\Pi(|\mu - \mu^*| \le \min\{\|\delta_n\|, 1/2\}) \ge \exp\{-G_7 \log n\}$ for some constant G_7. $\square$

PROOF OF THEOREM 4.1. Equations (5.2)–(5.4) are implied by Lemmas 5.3 and 5.2 with the ε_n given in Theorem 4.1. □

6. Auxiliary results. In this section, we provide a number of auxiliary results on packing and covering entropies of regular, sparse and additive Hölder spaces.

LEMMA 6.1. *For every $\alpha > 0$, $d \in \mathbb{N}$ there exist $\varepsilon_0 > 0$, $M_0 > 0$ such that for all $\varepsilon < \varepsilon_0$ there are $N \geq \exp\{M_0(1/\varepsilon)^{d/\alpha}\}$ functions $f_0, \dots, f_N \in C^\infty(\mathbb{R}^d)$ satisfying $f_0 \equiv 0$ and*

$$\text{support}(f_i) \subset [0,1]^d, \qquad f_i|_{[0,1]^d} \in C_1^{\alpha,d}, \qquad 0 \leq i \leq N, \tag{6.1}$$

$$\int_{\mathbb{R}} f_i(u_1, \dots, u_d)\, du_j = 0, \qquad 0 \leq i \leq N, 1 \leq j \leq d, \tag{6.2}$$

$$\|f_i - f_k\| \geq \varepsilon, \qquad 0 \leq i < k \leq N. \tag{6.3}$$

PROOF. Our proof follows the calculations in [28], Section 2.6.2, suitably adapted to handle L_2 norm and condition (6.2). Let $\mathcal{K} \in C^\infty(\mathbb{R}^d)$ such that

$$\text{support}(\mathcal{K}) = [-1,1]^d, \qquad \int \mathcal{K}(u_1, \dots, u_d)\, du_j = 0,\ j = 1, \dots, d. \tag{6.4}$$

For example, one could take $\mathcal{K}(x_1, \dots, x_d) = \prod_{j=1}^d \mathcal{K}_0(x_j)$ where $\mathcal{K}_0(t) = te^{-1/(1-t^2)} I(|t| \leq 1)$, $t \in \mathbb{R}$.

Fix an arbitrary $h \in (0, 1/2)$ and take $m = \lceil 1/(2h) \rceil$, $M = m^d$ and a rectangular grid $\{x^k : k = 1, \dots, M\}$ on $[0,1]^d$ consisting of the M grid points $(\frac{j_1 - 1/2}{m}, \dots, \frac{j_d - 1/2}{m})$, $(j_1, \dots, j_d) \in \{1, \dots, m\}^d$. We assume h is small enough so that $M \geq 8$. For each $1 \leq k \leq M$, the function ϕ_k defined as

$$\phi_k(x) = \frac{1}{\|\mathcal{K}\|_{C^{\alpha,d}}} h^\alpha \mathcal{K}\left(\frac{x - x^k}{h}\right), \qquad x \in [0,1]^d \tag{6.5}$$

has support inside $x^k + [-h, h]^d$ and belongs to $C_1^{\alpha,d}$. Let $\Omega = \{0,1\}^M$ and for each $\omega \in \Omega$ define $f_\omega = \sum_{k=1}^M \omega_k \phi_k$. Clearly, each f_ω is supported on $[0,1]^d$ and $\int f_\omega(u_1, \dots, u_d)\, du_j = 0$ for every $j = 1, \dots, d$. Also, since ϕ_k's are shifted copies of each other with disjoint supports, each $f_\omega \in C_1^{\alpha,d}$ and

$$\begin{aligned} \|f_\omega - f_{\omega'}\| &= \left\{ \sum_{k=1}^M (\omega_k - \omega_k')^2 \int \phi_k^2(x)\, dx \right\}^{1/2} \\ &= h^{\alpha + d/2} \frac{\|\mathcal{K}\|}{\|\mathcal{K}\|_{C^{\alpha,d}}} \rho^{1/2}(\omega, \omega'), \end{aligned} \tag{6.6}$$

where $\rho(\omega, \omega') = \sum_{k=1}^M I(\omega_k \neq \tilde{\omega}_k)$ denotes the Hamming distance.

By the Varshamov–Gilbert bound [28], Lemma 2.9, there are $N \geq 2^{M/8}$ binary strings $\omega^{(0)}, \ldots, \omega^{(N)} \in \Omega$, with $\omega^{(0)} = 0$, satisfying $\rho(\omega^{(k)}, \omega^{(k')}) \geq M/8$, $0 \leq k < k' \leq N$. Then $f_i := f_{\omega^{(i)}}$, $0 \leq i \leq N$, satisfy (6.1)–(6.2) and

$$\|f_i - f_k\| \geq h^{\alpha+d/2} \frac{\|\mathcal{K}\|}{\|\mathcal{K}\|_{C^{\alpha,d}}} \sqrt{\frac{M}{8}} \geq M_1 h^{\alpha}, \qquad 1 \leq i < k \leq N,$$

where $M_1 = \|\mathcal{K}\|/\{2^{(d+3)/2}\|\mathcal{K}\|_{C^{\alpha,d}}\}$ depends on only α and d. This proves the result since with $\varepsilon = M_1 h^{\alpha}$, which could be arbitrarily small, we get $N \geq \exp\{M(\log 2)/8\} \geq \exp\{M_0(1/\varepsilon)^{d/\alpha}\}$ where $M_0 = (M_1^{d/\alpha} \log 2)/2^{d+3}$ depends on only d and α. $\square$

LEMMA 6.2. *For every $\alpha, L > 0$, $d \in \mathbb{N}$ there exist $\varepsilon_1, M_0, M_1 > 0$ such that for any $\varepsilon < \varepsilon_1$ and all $p \in \mathbb{N}$*

$$M_0(L/\varepsilon)^{d/\alpha} + \log \binom{p}{d} \leq C\big(\varepsilon, \Sigma_S^p(\lambda, \alpha, d), \|\cdot\|\big) \leq M_1(L/\varepsilon)^{d/\alpha} + \log \binom{p}{d},$$

and, an ε-packing set satisfying the above lower bound may be obtained entirely with $C^{\infty}(\mathbb{R}^p)$ functions.

PROOF. It suffices to prove for $L = 1$ since $C(\varepsilon, L\Sigma, \|\cdot\|) = C(\varepsilon/L, \Sigma, \|\cdot\|)$ for any set Σ. By Lemma 6.1 there exist ε_0, M_0 such that for any $\varepsilon < \varepsilon_0$ there are functions $f_0 \equiv 0, f_1, \ldots, f_N \in C^{\infty}(\mathbb{R}^d)$ satisfying (6.1)–(6.3) with $\log N \geq M_0(1/\varepsilon)^{d/\alpha}$. Therefore, the set

$$\mathcal{T}^{\alpha,d,p}(\varepsilon) = \bigcup_{\substack{b \in \{0,1\}^p \\ |b| = d}} \{T^b f_i : 1 \leq i \leq N\} \tag{6.7}$$

is a subset of $\Sigma_S^p(1, \alpha, d)$. By (6.2), for any $b \neq b' \in \{0, 1\}^p$, $\langle T^b f_i, T^{b'} f_k \rangle = 0$ for all $1 \leq i, k \leq N$. Hence, $\mathcal{T}^{\alpha,d,p}(\varepsilon)$ is ε-separated in $\|\cdot\|$ since $\|T^b f_i - T^{b'} f_k\| = \|f_i - f_k\| \geq \varepsilon$ by (6.3) if $b = b'$ and $\|T^b f_i - T^{b'} f_k\| = \|f_i\| + \|f_k\| \geq \varepsilon$ by (6.3) and the fact that $f_0 \equiv 0$. This gives the lower bound on $C(\varepsilon, \Sigma_S^p(1, \alpha, d), \|\cdot\|)$ since the cardinality of $\mathcal{T}^{\alpha,d,p}$ is $\binom{p}{d}N$.

It is well known that for every $\alpha > 0$, $d \in \mathbb{N}$ there exist $\varepsilon_0', M_0' > 0$ such that for all $\varepsilon < \varepsilon_0'$, $V(\varepsilon, C_1^{\alpha,d}, \|\cdot\|) \leq M_0'(1/\varepsilon)^{d/\alpha}$ [28], Section 2.6.1, and [16]. Since a union of sets is covered by the union of their covers, it follows that $V(\varepsilon, \Sigma_S^p(L\lambda, \alpha, d), \|\cdot\|) \leq M_0'(1/\varepsilon)^{d/\alpha} + \log \binom{p}{d}$ for all $0 < \varepsilon < \varepsilon_0'$. Consequently, $C(\varepsilon, \Sigma_S^p(\lambda, \alpha, d), \|\cdot\|) \leq V(\varepsilon/2, \Sigma_S^p(\lambda, \alpha, d), \|\cdot\|) \leq M_0' 2^{d/\alpha}(1/\varepsilon)^{d/\alpha} + \log \binom{p}{d}$ for all $\varepsilon < \varepsilon_0'$. This proves the result with $M_1 = M_0' 2^{d/\alpha}$ and $\varepsilon_1 = \min(\varepsilon_0, \varepsilon_0')$. $\square$

LEMMA 6.3. *Let $\mathbb{H}_1, \dots, \mathbb{H}_k$ be mutually orthogonal subsets of a Hilbert space $(\mathbb{H}, \|\cdot\|_{\mathbb{H}})$. Then, for any $\delta \in \mathbb{R}_+^k$ and $c \in (0,1)$*

$$C\left(c\|\delta\|, \bigoplus_{s=1}^{k} \mathbb{H}_s, \|\cdot\|_{\mathbb{H}}\right) \geq \frac{1}{4}\left\{\frac{1-c^2}{C^*}\sum_{s=1}^{k} C(\delta_s, \mathbb{H}_s, \|\cdot\|_{\mathbb{H}}) - k\log 2\right\},$$

where $C^* := \sup_{1\leq s,t\leq k}\{\delta_s^{-2}C(\delta_s, \mathbb{H}_s, \|\cdot\|_{\mathbb{H}})\}/\{\delta_t^{-2}C(\delta_t, \mathbb{H}_t, \|\cdot\|_{\mathbb{H}})\}$.

PROOF. For every $1 \leq s \leq k$, let $\mathcal{H}_s$ denote a maximal δ_s-packing set of $\mathbb{H}_s$ with $C_s := \log|\mathcal{H}_s| = C(\delta_s, \mathbb{H}_s, \|\cdot\|_{\mathbb{H}})$. Take $\Omega = \mathcal{H}_1 \times \cdots \times \mathcal{H}_k$ and let $F = (F_1, \dots, F_k)$ be a random element in Ω with the uniform probability distribution. Fix an $M \in \mathbb{N}$ such that

$$\frac{1}{2}\left\{\frac{1-c^2}{C^*}\sum_s C_s - k\log 2\right\} < 2\log M < \frac{1-c^2}{C^*}\sum_s C_s - k\log 2,$$

and let F^j, $j = 1, \dots, M$, be IID copies of F. If

$$P\left\{\left\|\sum_s F_s^i - \sum_s F_s^j\right\| \geq c\|\delta\|, \forall 1 \leq i < j \leq M\right\} > 0, \tag{6.8}$$

then Ω contains a subset Ω_0 with at least M elements such that for any two $f, f' \in \Omega$, $\|\sum_s f_s - \sum_s f_s'\| > c\|\delta\|$. This would prove the result.

The probability value in (6.8) is at least $1 - M(M-1)/2 \cdot P\{\|\sum_s F_s^1 - \sum_s F_s^2\| < c\|\delta\|\}$, and hence it suffices to show $P\{\|\sum_s F_s^1 - \sum_s F_s^2\| < c\|\delta\|\} \leq 1/M^2$. Define $Z_s = I(F_s^1 \neq F_s^2)$, $s = 1, \dots, k$, which are independent binary variables with $Z_s \sim \text{Bernoulli}(1 - e^{-C_s})$. By orthogonality of $\mathbb{H}_1, \dots, \mathbb{H}_k$,

$$\left\|\sum_s F_s^1 - \sum_s F_s^2\right\|^2 = \sum_{s=1}^{k}\|F_s^1 - F_s^2\|^2 \geq \sum_{s=1}^{k}\delta_s^2 Z_s,$$

and hence it suffices to show

$$P\left(\sum_s \delta_s^2 Z_s < c^2\|\delta\|^2\right) \leq 1/M^2. \tag{6.9}$$

By Markov's inequality, for any $\lambda > 0$,

$$\begin{aligned}
P\left(\sum_s \delta_s^2 Z_s < c^2\|\delta\|^2\right) &\leq P\left\{e^{-\lambda\sum_s \delta_s^2 Z_s} > e^{-\lambda c^2\|\delta\|^2}\right\} \\
&\leq e^{\lambda c^2\|\delta\|^2}\prod_{s=1}^{k} E\left\{e^{-\lambda\delta_s^2 Z_s}\right\} \\
&\leq e^{\lambda c^2\|\delta\|^2}\prod_{s=1}^{k}\left\{e^{-C_s} + e^{-\lambda\delta_s^2}\right\} \\
&= e^{-\lambda(1-c^2)\|\delta\|^2}\prod_{s=1}^{k}\left\{1 + e^{\lambda\delta_s^2 - C_s}\right\}.
\end{aligned}$$

By the assumption on δ, $C_s\delta_t^2/(\delta_s^2 C^*) \le C_t \le C^* C_s \delta_t^2/(\delta_s^2)$ for every $1 \le s, t \le k$, and hence, $\delta_s^2 \le C^* C_s \|\delta\|^2/\sum_t C_t \le C_s/\lambda$ when we set $\lambda = \sum_s C_s/(\|\delta\|^2 C^*)$. Consequently,

$$P\left(\sum_s \delta_s^2 Z_s < c^2\|\delta\|^2\right) \le 2^k e^{-\lambda(1-c^2)\|\delta\|^2} = e^{-(1-c^2)\sum_s C_s/C^* + k\log 2} \le 1/M^2,$$

which completes the proof. □

THEOREM 6.4. *Suppose* $k \max_s d_s \le p$ *and set* $p_s = \lfloor p d_s/\sum_t d_t \rfloor$, $1 \le s \le k$. *Under Assumption* Q, *for any* $\delta \in \mathbb{R}_+^k$,

$$C\big(\underline{q}^{1/2}\Delta^{\max_s(\alpha_s + d_s/2)}\|\delta\|/2,\ \Sigma_A^{p,k,\bar{d}}(\lambda,\alpha,d),\ \|\cdot\|_Q\big) \ge \frac{1}{4}\left\{\frac{3}{4C^*}\sum_{s=1}^k C\big(\delta_s,\ \Sigma_S^{p_s}(\lambda_s,\alpha_s,d_s),\ \|\cdot\|\big) - k\log 2\right\},$$

$$C\big(4\bar{q}^{1/2}\sqrt{B}\|\delta\|,\ \Sigma_A^{p,k,\bar{d}}(\lambda,\alpha,d),\ \|\cdot\|_Q\big) \le \sum_{s=1}^k C\big(\delta_s,\ \Sigma_S^{p}(\lambda_s,\alpha_s,d_s),\ \|\cdot\|\big)$$

with $C^* = \sup_{1\le s,t\le k}\{\delta_s^{-2} C(\delta_s, \Sigma_S^{p_s}(\lambda_s,\alpha_s,d_s), \|\cdot\|)\}/\{\delta_t^{-2} C(\delta_t, \Sigma^{p_t}(L_t,\alpha_t, d_t), \|\cdot\|)\}$ *and* $B = 1 + \max_s d_s(\bar{d} - 1)$.

PROOF. Fix k mutually exclusive subsets $B_1, \ldots, B_k$ of $\{1, \ldots, p\}$ with $|B_s| = p_s$, $1 \le s \le p$. Let Σ^s denote the space of norm λ_s, α_s-smooth regression functions that select d_s predictors from B_s and none from the other subsets, that is, $\Sigma^s = \bigcup_{b\in\{0,1\}^p, |b|=d_s, \text{support}(b)\subset B_s} T^b(\lambda_s C_1^{\alpha_s, d_s})$. These subsets are mutually orthogonal since $f \in \Sigma^s$ and $f' \in \Sigma^t$, $s \ne t$ pick disjoint sets of predictors and $f, f' \in \mathcal{Z}_p$. Clearly, $\bigoplus_{s=1}^k \Sigma^s \subset \Sigma_A^{p,k,\bar{d}}(\lambda,\alpha,d)$. Let $f_i = \sum_{s=1}^k f_{is}$, $i = 1, \ldots, N$, be a $\|\delta\|/2$-packing set of $\bigoplus_{s=1}^k \Sigma^s$ under $\|\cdot\|$. We must have

$$N \ge \frac{1}{4}\left\{\frac{3}{4C^*}\sum_{s=1}^k C\big(\delta_s,\ \Sigma_S^{p_s}(\lambda_s,\alpha_s,d_s),\ \|\cdot\|\big) - k\log 2\right\}, \tag{6.10}$$

by an application of Lemma 6.3 with $c = 1/2$, coupled with the fact that Σ^s is isomorphic with $\Sigma_S^{p_s}(\lambda_s,\alpha_s,d_s)$. Also, by Lemma 6.1 and the packing set construction used in the proof of Lemma 6.3, each f_{si} can be chosen to belong to $\Sigma^s \cup C^\infty(\mathbb{R}^p)$. Define $g_1, \ldots, g_N$ as: $g_i(x) = \Delta^{\bar{\alpha}} f_i(x/\Delta)$ where $\bar{\alpha} = \max_s \alpha_s$. Then each $g_i \in \Sigma_A^{p,k,\bar{d}}(\lambda,\alpha,d)$ and, $\|g_i - g_j\|_Q \ge \underline{q}^{1/2}\Delta^{\bar{\alpha}+\max_s d_s/2}\|f_i - f_j\|$, since every $f_{is} - f_{js}$ involve at most $\max_s d_s$ many variables and they are orthogonal across s. This proves the first assertion of the theorem.

In light of the well-known relation $V(\varepsilon, A, \|\cdot\|) \le C(\varepsilon, A, \|\cdot\|) \le V(\varepsilon/2, A, \|\cdot\|)$ between packing and covering entropies of subsets in a metric space, and the

fact that $\|\cdot\|_Q \le \bar{q}^{1/2}\|\cdot\|$, the second assertion can be established by showing

$$V\big(2\sqrt{B}\|\delta\|, \Sigma_A^{p,k,\bar{d}}(\lambda,\alpha,d), \|\cdot\|\big) \le \sum_{s=1}^{k} V\big(\delta_s, \Sigma_S^p(\lambda_s,\alpha_s,d_s), \|\cdot\|\big).$$

For every $1 \le s \le k$, let $\mathcal{C}^s$ be a minimal δ_s/λ_s-covering set of $C_1^{\alpha_s,d_s}$. For each s, replace every element $f \in \mathcal{C}^s$ by its centered version $\bar{f} = f - \int f(x)\,dx$. The new $\mathcal{C}^s$ remains a $2\delta_s/\lambda_s$-covering set of $C_1^{\alpha_s,d_s} \cap \mathcal{Z}_{d_s}$. Take

$$\mathcal{C}_A = \left\{ f = \sum_{s=1}^{k} \lambda_s T^{b^s} f_s : f_s \in \mathcal{C}^s, 1 \le s \le k, \big(b^1,\dots,b^k\big) \in \mathcal{B}^{p,k,d,\bar{d}} \right\}.$$

Any $f \in \Sigma_A^{p,k,\bar{d}}(\lambda,\alpha,d)$ equals $f = \sum_s \lambda_s T^{b^s} f_s$ for some $f_s \in C_1^{\alpha_s,d_s} \cap \mathcal{Z}_{d_s}$, $1 \le s \le k$ and $(b^1,\dots,b^k) \in \mathcal{B}^{p,k,d,\bar{d}}$. Find $f_s^* \in \mathcal{C}^s$ such that $\|f_s - f_s^*\| \le 2\delta_s/\lambda_s$, $1 \le s \le k$ and set $f^* = \sum_s \lambda_s T^{b^s} f_s^* \in \mathcal{C}_A$. Since every $f_s - f_s^* \in \mathcal{Z}_{d_s}$, we get $\langle T^{b^s}(f_s - f_s^*), T^{b^t}(f_t - f_t^*)\rangle = 0$ whenever $\sum_{j=1}^p b_j^s b_j^t = 0$, that is, b^s, b^t have no shared selection. By assumption on $\mathcal{B}^{p,k,d,\bar{d}}$, for every s, there are at most $d_s(\bar{d}-1)$ many $t \ne s$ with shared selection. Therefore, by Lemma 6.5, $\|f - f^*\|^2 \le B\sum_{s=1}^k \lambda_s^2 \|f_s - f_s^*\|^2 \le 4B\|\delta\|^2$. Consequently, $\mathcal{C}_A$ gives a $(2\sqrt{B}\|\delta\|)$-covering of $\Sigma_A^{p,k,\bar{d}}(\lambda,\alpha,d)$. This completes the proof since $V(\delta_s, \Sigma_S^p(\lambda_s,\alpha_s,d_s), \|\cdot\|) \ge \log|\mathcal{C}^s|$ for every $1 \le s \le k$. □

LEMMA 6.5. *Suppose $f_1,\dots,f_k$ are elements of a Hilbert space $(\mathbb{H}, \|\cdot\|_{\mathbb{H}})$ and for any $1 \le s \le k$, let $r_s = |\{1 \le t \le k : \langle f_s, f_t\rangle_{\mathbb{H}} \ne 0\}|$. Then $\|\sum_{s=1}^k f_s\|_{\mathbb{H}}^2 \le \max_s r_s \sum_{s=1}^k \|f_s\|_{\mathbb{H}}^2$.*

PROOF. Since $2\langle f, g\rangle_{\mathbb{H}} \le \|f\|_{\mathbb{H}}^2 + \|g\|_{\mathbb{H}}^2$, we have

$$\begin{aligned}\Big\|\sum_s f_s\Big\|_{\mathbb{H}}^2 &= \sum_{s,t}\langle f_s, f_t\rangle_{\mathbb{H}} \le \frac{1}{2}\sum_{\langle f_s,f_t\rangle_{\mathbb{H}}\ne 0}\big(\|f_s\|_{\mathbb{H}}^2 + \|f_t\|_{\mathbb{H}}^2\big)\\ &\le \max_s r_s \sum_s \|f_s\|_{\mathbb{H}}^2.\end{aligned}$$

□

LEMMA 6.6. *Suppose $\mathcal{F} \subset C(\mathbb{R}^p)$ satisfies $\sup_{f\in\mathcal{F}}\|f\|_\infty \le 1$. Then, for any sequence δ_n satisfying $n\delta_n^2 \ge 2N(\delta_n,\mathcal{F},\|\cdot\|_\infty)$ and $\sum_{n=1}^\infty e^{-n\delta_n^2} < \infty$,*

$$Q^\infty\Big(\Big\{x^{1:\infty} : \sup_{f\in\mathcal{F}, \|f\|_Q \le \delta_n} \|f\|_n \ge 4\delta_n \text{ infinitely often}\Big\}\Big) = 0.$$

PROOF. Take $f \in \mathcal{F}$ and suppose that $\|f\|_Q \le \delta_n$ and $\|f\|_n \ge 4\delta_n$. Let $\{f_1,\dots,f_N\}$ form an minimal δ_n-covering of $\mathcal{F}$ under the sup-norm with

$2\log N \le n\delta_n^2$. Then there exists some $j_0 \in \{1,\dots,N\}$ such that $\|f - f_{j_0}\|_\infty \le \delta_n$. By the assumptions on f, we have $\|f_{j_0}\|_n \ge 3\delta_n$ and $\|f_{j_0}\|_Q \le 2\delta_n$, implying $|\|f_{j_0}\|_n^2 - \|f_{j_0}\|_Q^2| \ge 5\delta_n^2$. By Bernstein's inequality, we have

$$P\{|\|f_{j_0}\|_n^2 - \|f_{j_0}\|_Q^2| \ge 5\delta_n^2\} \le 2\exp\left[-\frac{5}{8}n\delta_n^2\right].$$

Since there are at most N choices for j_0, we get

$$\begin{aligned} P\Big\{\sup_{f\in\mathcal{F},\|f\|_Q\le\delta_n}\|f\|_n \ge 4\delta_n\Big\} &\le \sum_{j=1}^N P\{|\|f_j\|_n^2 - \|f_j\|_Q^2| \ge 5\delta_n^2\} \\ &\le 2N\exp\left[-\frac{5}{8}n\delta_n^2\right] \le 2\exp\left[-\frac{1}{8}n\delta_n^2\right], \end{aligned}$$

from which the results follows by the Borel–Cantelli lemma. □

Acknowledgments. The authors would like to thank the Associate Editor and two referees for their insightful comments and suggestions.

REFERENCES

[1] BHATTACHARYA, A., PATI, D. and DUNSON, D. (2014). Anisotropic function estimation using multi-bandwidth Gaussian processes. *Ann. Statist.* **42** 352–381. MR3189489

[2] BICKEL, P. J. and LI, B. (2007). Local polynomial regression on unknown manifolds. In *Complex Datasets and Inverse Problems*: *Tomography, Networks and Beyond. Institute of Mathematical Statistics Lecture Notes—Monograph Series* **54** 177–186. IMS, Beachwood, OH. MR2459188

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

[4] BORELL, C. (1975). The Brunn–Minkowski inequality in Gauss space. *Invent. Math.* **30** 207–216. MR0399402

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

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

[7] CHIPMAN, H. A., GEORGE, E. I. and MCCULLOCH, R. E. (2010). BART: Bayesian additive regression trees. *Ann. Appl. Stat.* **4** 266–298. MR2758172

[8] CHOI, T. and SCHERVISH, M. J. (2007). On posterior consistency in nonparametric regression problems. *J. Multivariate Anal.* **98** 1969–1987. MR2396949

[9] COMMINGES, L. and DALALYAN, A. S. (2012). Tight conditions for consistency of variable selection in the context of high dimensionality. *Ann. Statist.* **40** 2667–2696. MR3097616

[10] GHOSAL, S. and VAN DER VAART, A. W. (2007). Convergence rates of posterior distributions for noniid observations. *Ann. Statist.* **35** 192–233.

[11] HASTIE, T. and TIBSHIRANI, R. (1986). Generalized additive models. *Statist. Sci.* **1** 297–318. MR0858512

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

[13] KPOTUFE, S. and DASGUPTA, S. (2012). A tree-based regressor that adapts to intrinsic dimension. *J. Comput. System Sci.* **78** 1496–1515. MR2926146

[14] KULKARNI, S. R. and POSNER, S. E. (1995). Rates of convergence of nearest neighbor estimation under arbitrary sampling. *IEEE Trans. Inform. Theory* **41** 1028–1039. MR1366756

[15] LAFFERTY, J. and WASSERMAN, L. (2008). Rodeo: Sparse, greedy nonparametric regression. *Ann. Statist.* **36** 28–63. MR2387963

[16] LORENTZ, G. G. (1966). Metric entropy and approximation. *Bull. Amer. Math. Soc.* (*N.S.*) **72** 903–937. MR0203320

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

[18] MEINSHAUSEN, N. and YU, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *Ann. Statist.* **37** 246–270. MR2488351

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

[20] 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

[21] RASMUSSEN, C. E. and WILLIAMS, C. K. I. (2006). *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA. MR2514435

[22] 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

[23] SCOTT, C. and NOWAK, R. D. (2006). Minimax-optimal classification with dyadic decision trees. *IEEE Trans. Inform. Theory* **52** 1335–1353. MR2241192

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

[25] SUZUKI, T. (2012). PAC-Bayesian bound for Gaussian process regression and multiple kernel additive model. *JMLR*: *Workshop and Conference Proceedings* **23** 8.1–8.20.

[26] TIBSHIRANI, R. (1996). Regression shrinkage and selection via the Lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

[27] TOKDAR, S. T., ZHU, Y. M. and GHOSH, J. K. (2010). Bayesian density regression with logistic Gaussian process and subspace projection. *Bayesian Anal.* **5** 319–344. MR2719655

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

[29] VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2008). Rates of contraction of posterior distributions based on Gaussian process priors. *Ann. Statist.* **36** 1435–1463. MR2418663

[30] VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2008). Reproducing kernel Hilbert spaces of Gaussian priors. In *Pushing the Limits of Contemporary Statistics*: *Contributions in Honor of Jayanta K. Ghosh*. *Inst. Math. Stat. Collect.* **3** 200–222. IMS, Beachwood, OH. MR2459226

[31] VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2009). Adaptive Bayesian estimation using a Gaussian random field with inverse gamma bandwidth. *Ann. Statist.* **37** 2655–2675. MR2541442

[32] VERZELEN, N. (2012). Minimax risks for sparse regressions: Ultra-high dimensional phenomenons. *Electron. J. Stat.* **6** 38–90. MR2879672

[33] WAINWRIGHT, M. J. (2009). Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting. *IEEE Trans. Inform. Theory* **55** 5728–5741. MR2597190

[34] WAINWRIGHT, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Trans. Inform. Theory* **55** 2183–2202. MR2729873

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

[36] YANG, Y. and DUNSON, B. D. (2013). Bayesian manifold regression. Preprint. Available at arXiv:1305.0617.
[37] YE, G.-B. and ZHOU, D.-X. (2008). Learning and approximation by Gaussians on Riemannian manifolds. *Adv. Comput. Math.* **29** 291–310. MR2438346
[38] ZHANG, C.-H. and HUANG, J. (2008). The sparsity and bias of the LASSO selection in high-dimensional linear regression. *Ann. Statist.* **36** 1567–1594. MR2435448

DEPARTMENT OF EECS
UNIVERSITY OF CALIFORNIA, BERKELEY
BERKELEY, CALIFORNIA 94720
USA
E-MAIL: yy84@berkeley.edu

DEPARTMENT OF STATISTICAL SCIENCE
DUKE UNIVERSITY
BOX 90251
DURHAM, NORTH CAROLINA 27708-0251
USA
E-MAIL: tokdar@stat.duke.edu