Minimax-optimal rates for sparse additive models over kernel classes via convex programming

Garvesh Raskutti[1]
garveshr@stat.berkeley.edu

Martin J. Wainwright[1,2]
wainwrig@stat.berkeley.edu

Bin Yu[1,2]
binyu@stat.berkeley.edu

Departments of Statistics[1], and EECS[2]
UC Berkeley, Berkeley, CA

Revised version of August 2010 paper

Abstract

Sparse additive models are families of d-variate functions that have the additive decomposition $f^* = \sum_{j \in S} f_j^*$, where S is an unknown subset of cardinality $s \ll d$. In this paper, we consider the case where each univariate component function f_j^* lies in a reproducing kernel Hilbert space (RKHS), and analyze a method for estimating the unknown function f^* based on kernels combined with ℓ_1-type convex regularization. Working within a high-dimensional framework that allows both the dimension d and sparsity s to increase with n, we derive convergence rates (upper bounds) in the $L^2(\mathbb{P})$ and $L^2(\mathbb{P}_n)$ norms over the class $\mathcal{F}_{d,s,\mathcal{H}}$ of sparse additive models with each univariate function f_j^* in the unit ball of a univariate RKHS with bounded kernel function. We complement our upper bounds by deriving minimax lower bounds on the $L^2(\mathbb{P})$ error, thereby showing the optimality of our method. Thus, we obtain optimal minimax rates for many interesting classes of sparse additive models, including polynomials, splines, and Sobolev classes. We also show that if, in contrast to our univariate conditions, the multivariate function class is assumed to be globally bounded, then much faster estimation rates are possible for any sparsity $s = \Omega(\sqrt{n})$, showing that global boundedness is a significant restriction in the high-dimensional setting.

1 Introduction

The past decade has witnessed a flurry of research on sparsity constraints in statistical models. Sparsity is an attractive assumption for both practical and theoretical reasons: it leads to more interpretable models, reduces computational cost, and allows for model identifiability even under high-dimensional scaling, where the dimension d exceeds the sample size n. While a large body of work has focused on sparse linear models, many applications call for the additional flexibility provided by non-parametric models. In the general setting, a non-parametric regression model takes the form $y = f^*(x_1, \ldots, x_d) + w$, where $f^* : \mathbb{R}^d \to \mathbb{R}$ is the unknown regression function, and w is scalar observation noise. Unfortunately, this general non-parametric model is known to suffer severely from the so-called "curse of dimensionality", in that for most natural function classes (e.g., twice differentiable functions), the sample size n required to achieve any given error grows exponentially in the dimension d. Given this curse of dimensionality, it is essential to further constrain the complexity of possible functions f^*. One attractive candidate is the class of *additive non-parametric models* [17], in which the function f^* has an additive decomposition of the form

$$f^*(x_1, x_2, \dots, x_d) = \sum_{j=1}^{d} f_j^*(x_j), \tag{1}$$

where each component function f_j^* is univariate. Given this additive form, this function class no longer suffers from the exponential explosion in sample size of the general non-parametric model. Nonetheless, one still requires a sample size $n \gg d$ for consistent estimation; note that this is true even for the linear model, which is a special case of equation (1).

A natural extension of sparse linear models is the class of *sparse additive models*, in which the unknown regression function is assumed to have a decomposition of the form

$$f^*(x_1, x_2 \dots, x_d) = \sum_{j \in S} f_j^*(x_j), \tag{2}$$

where $S \subseteq \{1, 2, \dots, d\}$ is some unknown subset of cardinality $|S| = s$. Of primary interest is the case when the decomposition is genuinely sparse, so that $s \ll d$. To the best of our knowledge, this model class was first introduced in Lin and Zhang [22], and has since been studied by various researchers (e.g., [20, 24, 31, 40]). Note that the sparse additive model (2) is a natural generalization of the sparse linear model, to which it reduces when each univariate function is constrained to be linear.

In past work, several groups have proposed computationally efficient methods for estimating sparse additive models (2). Just as ℓ_1-based relaxations such as the Lasso have desirable properties for sparse parametric models, more general ℓ_1-based approaches have proven to be successful in this setting. Lin and Zhang [22] proposed the COSSO method, which extends the Lasso to cases where the component functions f_j^* lie in a reproducing kernel Hilbert space (RKHS); see also Yuan [40] for a similar extension of the non-negative garrote [8]. Bach [3] analyzes a closely related method for the RKHS setting, in which least-squares loss is penalized by an ℓ_1-sum of Hilbert norms, and establishes consistency results in the classical (fixed d) setting. Other related ℓ_1-based methods have been proposed in independent work by Koltchinskii and Yuan [19], Ravikumar et al. [31] and Meier et al. [24], and analyzed under high-dimensional scaling ($n \ll d$). As we describe in more detail in Section 3.4, each of the above papers establish consistency and convergence rates for the prediction error under certain conditions on the covariates as well as the sparsity s and dimension d. However, it is not clear whether the rates obtained in these papers are sharp for the given methods, nor whether the rates are minimax-optimal. Past work by Koltchinskii and Yuan [20] establishes rates for sparse additive models with an additional global boundedness condition, but as will be discussed at more length in the sequel, these rates are not minimax optimal in general.

This paper makes three main contributions to this line of research. Our first contribution is to analyze a simple polynomial-time method for estimating sparse additive models and provide upper bounds on the error in the $L^2(\mathbb{P})$ and $L^2(\mathbb{P}_n)$ norms. The estimator[1] we analyze is based on a combination of least-squares loss with two ℓ_1-based sparsity penalty terms, one corresponding to an $\ell_1/L^2(\mathbb{P}_n)$ norm and the other an $\ell_1/\|\cdot\|_{\mathcal{H}}$ norm. Our first main result (Theorem 1) shows

[1] The estimator is the same as the estimator analyzed in Koltchinskii and Yuan [20]. We proposed the estimator concurrently (see our earlier preprint [29]) and as we discuss later analyze the same estimator under less restrictive conditions than those imposed in Koltchinskii and Yuan [20].

that with high probability, if we assume the univariate functions are bounded and independent, the error of our procedure in the squared $L^2(\mathbb{P}_n)$ and $L^2(\mathbb{P})$ norms is bounded by $\mathcal{O}\big(\frac{s \log d}{n} + s\nu_n^2\big)$, where the quantity ν_n^2 corresponds to the optimal rate for estimating a single univariate function. Importantly, our analysis does *not* require a global boundedness condition on the class $\mathcal{F}_{d,s,\mathcal{H}}$ of all s-sparse models, an assumption that is often imposed in classical non-parametric analysis. Indeed, as we discuss below, when such a condition is imposed, then significantly faster rates of estimation are possible. The proof of Theorem 1 involves a combination of techniques for analyzing M-estimators with decomposable regularizers [27] and techniques in empirical process theory for analyzing kernel classes [4, 25, 35].

Our second contribution is complementary in nature, in that it establishes algorithm-independent minimax lower bounds on $L^2(\mathbb{P})$ error. These minimax lower bounds, stated in Theorem 2, are specified in terms of the metric entropy of the underlying univariate function classes. For both finite-rank kernel classes and Sobolev-type classes, these lower bounds match our achievable results, as stated in Corollaries 1 and 2, up to constant factors in the regime of sub-linear sparsity ($s = o(d)$). Thus, for these function classes, we have a sharp characterization of the associated minimax rates. The lower bounds derived in this paper initially appeared in the Proceedings of the NIPS Conference (December 2009). The proofs of Theorem 2 is based on characterizing the packing entropies of the class of sparse additive models, combined with classical information theoretic techniques involving Fano's inequality and variants (see, e.g. the papers [16, 38, 39]).

Our third contribution is to determine upper bounds on minimax $L^2(\mathbb{P})$ and $L^2(\mathbb{P}_n)$ error when we impose a global boundedness assumption on the class $\mathcal{F}_{d,s,\mathcal{H}}$, meaning that the quantity $B(\mathcal{F}_{d,s,\mathcal{H}}) = \sup_{f \in \mathcal{F}_{d,s,\mathcal{H}}} \sup_x |\sum_{j=1}^d f_j(x_j)|$ is assumed to be bounded independently of (s, d). As mentioned earlier, our upper bound in Theorem 1 does *not* impose a global boundedness condition, whereas in contrast, past work by Koltchinskii and Yuan [20] did impose such a global boundedness condition in their analysis of the same ℓ_1-kernel-based estimator. Under global boundedness, their work provides rates on the $L^2(\mathbb{P})$ and $L^2(\mathbb{P}_n)$ norm that are of the same order as the results presented here. It is natural to wonder whether or not this difference is actually significant—that is, do the minimax rates for the class of sparse additive models depend on whether or not global boundedness is imposed? In Section 3.5, we define the class $\mathcal{F}^*_{d,s,\mathcal{H}}(B)$ of sparse additive models with the additional assumption that $B(\mathcal{F}_{d,s,\mathcal{H}}) \leq B$. Theorem 3 and Corollary 3 in this paper provide upper bounds on the minimax rate in $L^2(\mathbb{P})$ and $L^2(\mathbb{P}_n)$ error over the class $\mathcal{F}^*_{d,s,\mathcal{H}}(B)$. These rates are faster than those of Theorem 3 in the paper [20] (KY), showing that the KY rates are not minimax optimal for problems with $s = \Omega(\sqrt{n})$. In this way, we see that the assumption of global boundedness, though relatively innocuous for classical (low-dimensional) non-parametric problems, can be quite limiting in high dimensions.

The remainder of the paper is organized as follows. In Section 2, we provide background on kernel spaces and the class of sparse additive models considered in this paper. Section 3 is devoted to the statement of our main results and discussion of their consequences; it includes description of our method, the upper bounds on the convergence rate that it achieves, and a matching set of minimax lower bounds. Section 3.5 emphasizes the restrictiveness of the global uniform boundedness assumption and in particular Theorem 3 and Corollary 3 show that there are classes of Sobolev spaces where under the scaling $s = \Omega(\sqrt{n})$, optimal rates of convergence are faster than rates proven in Theorem 2. Section 4 is devoted to the proofs of our three main theorems, with the more technical details deferred to the Appendices. We conclude with a discussion in Section 5.

2 Background and problem set-up

We begin with some background on reproducing kernel Hilbert spaces, before providing a precise definition of the class of sparse additive models studied in this paper.

2.1 Reproducing kernel Hilbert spaces

Given a subset $\mathcal{X} \subset \mathbb{R}$ and a probability measure $\mathbb{Q}$ on $\mathcal{X}$, we consider a Hilbert space $\mathcal{H} \subset L^2(\mathbb{Q})$, meaning a family of functions $g : \mathcal{X} \to \mathbb{R}$, with $\|g\|_{L^2(\mathbb{Q})} < \infty$, and an associated inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ under which $\mathcal{H}$ is complete. The space $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS) if there exists a symmetric function $\mathbb{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ such that: (a) for each $x \in \mathcal{X}$, the function $\mathbb{K}(\cdot, x)$ belongs to the Hilbert space $\mathcal{H}$, and (b) we have the reproducing relation $f(x) = \langle f, \mathbb{K}(\cdot, x) \rangle_{\mathcal{H}}$ for all $f \in \mathcal{H}$. Any such kernel function must be positive semidefinite; under suitable regularity conditions, Mercer's theorem [26] guarantees that the kernel has an eigen-expansion of the form

$$\mathbb{K}(x, x') = \sum_{k=1}^{\infty} \mu_k \phi_k(x) \phi_\ell(x'), \tag{3}$$

where $\mu_1 \geq \mu_2 \geq \mu_3 \geq \ldots \geq 0$ are a non-negative sequence of eigenvalues, and $\{\phi_k\}_{k=1}^{\infty}$ are the associated eigenfunctions, taken to be orthonormal in $L^2(\mathbb{Q})$. The decay rate of these eigenvalues will play a crucial role in our analysis, since they ultimately determine the rate ν_n for the univariate RKHS's in our function classes.

Since the eigenfunctions $\{\phi_k\}_{k=1}^{\infty}$ form an orthonormal basis, any function $f \in \mathcal{H}$ has an expansion of the $f(x) = \sum_{k=1}^{\infty} a_k \phi_k(x)$, where $a_k = \langle f, \phi_k \rangle_{L^2(\mathbb{Q})} = \int_{\mathcal{X}} f(x) \phi_k(x) \, d\,\mathbb{Q}(x)$ are (generalized) Fourier coefficients. Associated with any two functions in $\mathcal{H}$—say $f = \sum_{k=1}^{\infty} a_k \phi_k$ and $g = \sum_{k=1}^{\infty} b_k \phi_k$—are two distinct inner products. The first is the usual inner product in the space $L^2(\mathbb{Q})$—namely, $\langle f, g \rangle_{L^2(\mathbb{Q})} := \int_{\mathcal{X}} f(x) g(x) \, d\,\mathbb{Q}(x)$. By Parseval's theorem, it has an equivalent representation in terms of the expansion coefficients—namely

$$\langle f, g \rangle_{L^2(\mathbb{Q})} = \sum_{k=1}^{\infty} a_k b_k.$$

The second inner product, denoted $\langle f, g \rangle_{\mathcal{H}}$, is the one that defines the Hilbert space; it can be written in terms of the kernel eigenvalues and generalized Fourier coefficients as

$$\langle f, g \rangle_{\mathcal{H}} = \sum_{k=1}^{\infty} \frac{a_k b_k}{\mu_k}.$$

Using this definition, the Hilbert ball of radius 1 for the Hilbert space $\mathcal{H}$ with eigenvalues μ_k and eigenfunctions $\phi_k(\cdot)$, is:

$$\mathbb{B}_{\mathcal{H}}(1) = \{f \in \mathcal{H};\ f(\cdot) = \sum_{k=1}^{\infty} a_k \phi_k(\cdot) \mid \sum_{k=1}^{\infty} \frac{a_k^2}{\mu_k} \leq 1\}. \tag{4}$$

For more background on reproducing kernel Hilbert spaces, we refer the reader to various standard references [2, 14, 32, 33, 37].

2.2 Sparse additive models over RKHS

For each $j = 1, \ldots, d$, let $\mathcal{H}_j \subset L^2(\mathbb{Q})$ be a reproducing kernel Hilbert space of univariate functions on the domain $\mathcal{X} \subset \mathbb{R}$. We assume that

$$\mathbb{E}[f_j(x)] = \int_{\mathcal{X}} f_j(x) d\,\mathbb{Q}(x) \;=\; 0 \qquad \text{for all } f_j \in \mathcal{H}_j, \text{ and for each } j = 1, 2, \ldots, d.$$

As will be clarified momentarily, our observation model (7) allows for the possibility of a non-zero mean μ, so that there is no loss of generality in this assumption. For a given subset $S \subset \{1, 2, \ldots, d\}$, we define

$$\mathcal{H}(S) := \big\{ f = \sum_{j \in S} f_j \;\mid\; f_j \in \mathcal{H}_j, \text{ and } f_j \in \mathbb{B}_{\mathcal{H}_j}(1) \;\forall\, j \in S \big\}, \tag{5}$$

corresponding to the class of functions $f : \mathcal{X}^d \to \mathbb{R}$ that decompose as sums of univariate functions on co-ordinates lying within the set S. Note that $\mathcal{H}(S)$ is also (a subset of) a reproducing kernel Hilbert space, in particular with the norm

$$\|f\|^2_{\mathcal{H}(S)} = \sum_{j \in S} \|f_j\|^2_{\mathcal{H}_j},$$

where $\|\cdot\|_{\mathcal{H}_j}$ denotes the norm on the univariate Hilbert space $\mathcal{H}_j$. Finally, for a cardinality $s \in \{1, 2, \ldots, \lfloor d/2 \rfloor\}$, we define the function class

$$\mathcal{F}_{d,s,\mathcal{H}} := \bigcup_{\substack{S \subset \{1,2,\ldots,d\} \\ |S| = s}} \mathcal{H}(S). \tag{6}$$

To ease notation, we frequently adopt the shorthand $\mathcal{F} = \mathcal{F}_{d,s,\mathcal{H}}$, but the reader should recall that $\mathcal{F}$ depends on the choice of Hilbert spaces $\{\mathcal{H}_j\}_{j=1}^d$, and moreover, that we are actually studying a *sequence of function classes* indexed by (d, s).

Now let $\mathbb{P} = \mathbb{Q}^d$ denote the product measure on the space $\mathcal{X}^d \subseteq \mathbb{R}^d$. Given an arbitrary $f^* \in \mathcal{F}$, we consider the observation model

$$y_i = \mu + f^*(x_i) + w_i, \quad \text{for } i = 1, 2, \ldots, n, \tag{7}$$

where $\{w_i\}_{i=1}^n$ is an i.i.d. sequence of standard normal variates, and $\{x_i\}_{i=1}^n$ is a sequence of design points in $\mathbb{R}^d$, sampled in an i.i.d. manner from $\mathbb{P}$.

Given an estimate $\widehat{f}$, our goal is to bound the error $\widehat{f} - f^*$ under two norms. The first is the *usual* $L^2(\mathbb{P})$ *norm* on the space $\mathcal{F}$; given the product structure of $\mathbb{P}$ and the additive nature of any $f \in \mathcal{F}$, it has the additive decomposition $\|f\|^2_{L^2(\mathbb{P})} = \sum_{j=1}^d \|f_j\|^2_{L^2(\mathbb{Q})}$. In addition, we consider the error in the *empirical* $L^2(\mathbb{P}_n)$*-norm* defined by the sample $\{x_i\}_{i=1}^n$, defined as

$$\|f\|^2_{L^2(\mathbb{P}_n)} := \frac{1}{n} \sum_{i=1}^n f^2(x_i).$$

Unlike the $L^2(\mathbb{P})$ norm, this norm does not decouple across the dimensions, but part of our analysis will establish an approximate form of such decoupling. For shorthand, we frequently use the notation $\|f\|_2 = \|f\|_{L^2(\mathbb{P})}$ and $\|f\|_n = \|f\|_{L^2(\mathbb{P}_n)}$ for a d-variate function $f \in \mathcal{F}$. With a minor abuse of notation, for a univariate function $f_j \in \mathcal{H}_j$, we also use the shorthands $\|f_j\|_2 = \|f_j\|_{L^2(\mathbb{Q})}$ and $\|f_j\|_n = \|f_j\|_{L^2(\mathbb{Q}_n)}$.

3 Main results and their consequences

This section is devoted to the statement of our three main results, and discussion of some of their consequences. We begin in Section 3.1 by describing a regularized M-estimator for sparse additive models, and we state our upper bounds for this estimator in Section 3.2. This estimator is essentially equivalent to that analyzed in the paper KY [20], except that we allow for a non-zero mean for the function, and estimate it as well. We illustrate our upper bounds for various concrete instances of kernel classes. In Section 3.3, we state minimax lower bounds on the $L^2(\mathbb{P})$ error over the class $\mathcal{F}_{d,s,\mathcal{H}}$, which establish the optimality of our procedure. In Section 3.4, we provide a detailed comparison between our results to past work, and in Section 3.5 we discuss the effect of global boundedness conditions on optimal rates.

3.1 A regularized M-estimator for sparse additive models

For any function of the form $f = \sum_{j=1}^d f_j$, the $(L^2(\mathbb{Q}_n), 1)$ and $(\mathcal{H}, 1)$-norms are given by

$$\|f\|_{n,1} := \sum_{j=1}^d \|f_j\|_n, \quad \text{and} \quad \|f\|_{\mathcal{H},1} := \sum_{j=1}^d \|f_j\|_{\mathcal{H}}, \tag{8}$$

respectively. Using this notation and defining the sample mean $\bar{y}_n = \frac{1}{n}\sum_{i=1}^n y_i$, we define the cost functional

$$\mathcal{L}(f) = \frac{1}{2n}\sum_{i=1}^n \big(y_i - \bar{y}_n - f(x_i)\big)^2 + \lambda_n \|f\|_{n,1} + \rho_n \|f\|_{\mathcal{H},1}. \tag{9}$$

The cost functional $\mathcal{L}(f)$ is least-squares loss with a sparsity penalty $\|f\|_{n,1}$ and a smoothness penalty $\|f\|_{\mathcal{H},1}$. Here (λ_n, ρ_n) are a pair of positive regularization parameters whose choice will be specified by our theory. Given this cost functional, we then consider the M-estimator

$$\widehat{f} \in \arg\min_f \mathcal{L}(f) \quad \text{subject to } f = \textstyle\sum_{j=1}^d f_j \text{ and } \|f_j\|_{\mathcal{H}} \le 1 \text{ for all } j = 1, 2, \ldots, d. \tag{10}$$

In this formulation (10), the problem is infinite-dimensional in nature, since it involves optimization over Hilbert spaces. However, an attractive feature of this M-estimator is that, as a straightforward consequence of the representer theorem [18, 33], it can be reduced to an equivalent convex program in $\mathbb{R}^n \times \mathbb{R}^d$. In particular, for each $j = 1, 2, \ldots, d$, let $\mathbb{K}^j$ denote the kernel function for co-ordinate j. Using the notation $x_i = (x_{i1}, x_{i2}, \ldots, x_{id})$ for the i^{th} sample, we define the collection of empirical kernel matrices $K^j \in \mathbb{R}^{n \times n}$, with entries $K^j_{i\ell} = \mathbb{K}^j(x_{ij}, x_{\ell j})$. By the representer theorem, any solution $\widehat{f}$ to the variational problem (10) can be written in the form

$$\widehat{f}(z_1, \ldots, z_d) = \sum_{i=1}^n \sum_{j=1}^d \widehat{\alpha}_{ij} \mathbb{K}^j(z_j, x_{ij}),$$

for a collection of weights $\big\{\widehat{\alpha}_j \in \mathbb{R}^n,\ j = 1, \ldots, d\big\}$. The optimal weights are obtained by solving the convex program

$$(\widehat{\alpha}_1, \ldots, \widehat{\alpha}_d) \in \arg \min_{\substack{\alpha_j \in \mathbb{R}^n \\ \alpha_j^T K^j \alpha_j \le 1}} \left\{ \frac{1}{2n}\|y - \bar{y}_n - \sum_{j=1}^d K^j \alpha_j\|_2^2 + \lambda_n \sum_{j=1}^d \sqrt{\frac{1}{n}\|K^j\alpha_j\|_2^2} + \rho_n \sum_{j=1}^d \sqrt{\alpha_j^T K^j \alpha_j} \right\}. \tag{11}$$

This problem is a second-order cone program (SOCP), and there are various algorithms for finding a solution to arbitrary accuracy in time polynomial in (n, d), among them interior point methods (e.g., see §11 in the book [7]).

Various combinations of sparsity and smoothness penalties have been used in past work on sparse additive models. For instance, the method of Ravikumar et. al [31] is based on least-squares loss regularized with single sparsity constraint, and separate smoothness constraints for each univariate function. They solve the resulting optimization problem using a back-fitting procedure. Koltchinskii and Yuan [19] develop a method based on least-squares loss combined with a single penalty term $\sum_{j=1}^{d} \|f_j\|_{\mathcal{H}}$. Their method also leads to an SOCP if $\mathcal{H}$ is a reproducing kernel Hilbert space, but differs from the program (11) in lacking the additional sparsity penalties. Meier et. al [24] analyzed least-squares regularized with a penalty term of the form $\sum_{j=1}^{d} \sqrt{\lambda_1 \|f_j\|_n^2 + \lambda_2 \|f_j\|_{\mathcal{H}}^2}$, where λ_1 and λ_2 are a pair of regularization parameters. In their method, λ_1 controls the sparsity while λ_2 controls the smoothness. If $\mathcal{H}$ is an RKHS, the method in Meier et. al [24] reduces to an ordinary group Lasso problem on a different set of variables, which can be cast as a quadratic program. The more recent work of Koltchinskii and Yuan [20] is based on essentially the same estimator as problem (10), but they impose stronger assumptions in their analysis. We provide a more in-depth comparison of our analysis and results with the past work listed above in Sections 3.4 and 3.5.

3.2 Upper bound

We now state a result that provides upper bounds on the estimation error achieved by the estimator (10), or equivalently (11). To simplify presentation, we state our result in the special case that the univariate Hilbert space $\mathcal{H}_j, j = 1, \ldots, d$ are all identical, denoted by $\mathcal{H}$. However, the analysis and results extend in a straightforward manner to the general setting of distinct univariate Hilbert spaces, as we discuss following the statement of Theorem 1.

Let $\mu_1 \geq \mu_2 \geq \ldots \geq 0$ denote the non-negative eigenvalues of the kernel operator defining the univariate Hilbert space $\mathcal{H}$, as defined in equation (3), and define the function

$$\mathcal{Q}_{\sigma,n}(t) := \frac{1}{\sqrt{n}} \big[\sum_{\ell=1}^{\infty} \min\{t^2, \mu_\ell\} \big]^{1/2}. \tag{12}$$

Let $\nu_n > 0$ be the smallest positive solution to the inequality

$$40 \nu_n^2 \geq \mathcal{Q}_{\sigma,n}(\nu_n), \tag{13}$$

where the 40 is simply used for technical convenience. We refer to ν_n as the *critical univariate rate*, as it is the minimax-optimal rate for $L^2(\mathbb{P})$-estimation of a single univariate function in the Hilbert space $\mathcal{H}$ (e.g., [25, 35]). This quantity will be referred to throughout the remainder of the paper.

Our choices of regularization parameters are specified in terms of the quantity

$$\gamma_n := \kappa \max\big\{\nu_n, \sqrt{\frac{\log d}{n}}\big\}, \tag{14}$$

where κ is a fixed constant that we choose later. We assume that each function within the unit ball of the univariate Hilbert space is uniformly bounded by a constant multiple of its Hilbert norm—that is, for each $j = 1, \ldots, d$ and each $f_j \in \mathcal{H}$,

$$\|f_j\|_\infty := \sup_{x_j} |f_j(x_j)| \leq c\, \|f_j\|_{\mathcal{H}}. \tag{15}$$

This condition is satisfied for many kernel classes including Sobolev spaces, and any univariate kernel function[2] bounded uniformly by c. Such a condition is routinely imposed for proving upper bounds on rates of convergence for non-parametric least squares in the univariate case $d = 1$ (see e.g. [34, 35]). Note that this univariate boundedness does not imply that the multivariate functions $f = \sum_{j \in S} f_j$ in $\mathcal{F}$ are uniformly bounded independently of (d, s); rather, they can take on values of the order $\sqrt{s}$.

The following result applies to any class $\mathcal{F}_{d,s,\mathcal{H}}$ of sparse additive models based on the univariate Hilbert space satisfying condition (15), and to the estimator (10) based on n i.i.d. samples $(x_i, y_i)_{i=1}^n$ from the observation model (7).

Theorem 1. *Let $\widehat{f}$ be any minimizer of the convex program (10) with regularization parameters $\lambda_n \geq 16\gamma_n$ and $\rho_n \geq 16\gamma_n^2$. Then provided that $n\gamma_n^2 = \Omega(\log(1/\gamma_n))$, there are universal constants (C, c_1, c_2) such that*

$$\mathbb{P}\Big[\max\{\|\widehat{f} - f^*\|_2^2,\ \|\widehat{f} - f^*\|_n^2\} \geq C\big\{s\lambda_n^2 + s\rho_n\big\}\Big] \leq c_1 \exp(-c_2 n\gamma_n^2). \tag{16}$$

We provide the proof of Theorem 1 in Section 4.1.

Remarks: First, the technical condition $n\gamma_n^2 = \Omega(\log(1/\gamma_n))$ is quite mild, and satisfied in most cases of interest, among them the kernels considered below in Corollaries 1 and 2.

Second, note that setting $\lambda_n = c\gamma_n$ and $\rho_n = c\gamma_n^2$ for some constant $c \in (16, \infty)$ yields the rate $\Theta(s\gamma_n^2 + s\rho_n) = \Theta(\frac{s \log d}{n} + s\nu_n^2)$. This rate may be interpreted as the sum of a subset selection term $(\frac{s \log d}{n})$ and an s-dimensional estimation term $(s\nu_n^2)$. Note that the subset selection term $(\frac{s \log d}{n})$ is independent of the choice of Hilbert space $\mathcal{H}$ whereas the s-dimensional estimation term is independent of the ambient dimension d. Depending on the scaling of the triple (n, d, s) and the smoothness of the univariate RKHS $\mathcal{H}$, either the subset selection term or function estimation term may dominate. In general, if $\frac{\log d}{n} = o(\nu_n^2)$, the s-dimensional estimation term dominates, and vice versa otherwise. At the boundary, the scalings of the two terms are equivalent.

Finally, for clarity, we have stated our result in the case where the univariate Hilbert space $\mathcal{H}$ is identical across all co-ordinates. However, our proof extends with only notational changes to the general setting, in which each co-ordinate j is endowed with a (possibly distinct) Hilbert space $\mathcal{H}_j$. In this case, the M-estimator returns a function $\widehat{f}$ such that (with high probability)

$$\max\big\{\|\widehat{f} - f^*\|_n^2,\ \|\widehat{f} - f^*\|_2^2\big\} \;\leq\; C\Big\{\frac{s \log d}{n} + \sum_{j \in S} \nu_{n,j}^2\Big\},$$

[2] Indeed, we have

$$\sup_{x_j} |f_j(x_j)| = \sup_{x_j} |\langle f_j(.), \mathbb{K}(., x_j)\rangle_{\mathcal{H}}| \;\leq\; \sup_{x_j} \sqrt{\mathbb{K}(x_j, x_j)}\|f_j\|_{\mathcal{H}}.$$

where $\nu_{n,j}$ is the critical univariate rate associated with the Hilbert space $\mathcal{H}_j$, and S is the subset on which f^* is supported.

Theorem 1 has a number of corollaries, obtained by specifying particular choices of kernels. First, we discuss m-rank operators, meaning that the kernel function $\mathbb{K}$ can be expanded in terms of m eigenfunctions. This class includes linear functions, polynomial functions, as well as any function class based on finite dictionary expansions. First we present a corollary for finite-rank kernel classes.

Corollary 1. *Under the same conditions as Theorem 1, consider an univariate kernel with finite rank m. Then any solution $\widehat{f}$ to the problem (10) with $\lambda_n = c\gamma_n$ and $\rho_n = c\gamma_n^2$ with $16 \leq c < \infty$ satisfies*

$$\mathbb{P}\Big[\max\big\{\|\widehat{f}-f^*\|_n^2, \|\widehat{f}-f^*\|_2^2\big\} \geq C\big\{\frac{s\log d}{n} + s\frac{m}{n}\big\}\Big] \leq c_1 \exp\big(-c_2(m+\log d)\big). \tag{17}$$

Proof. It suffices to show that the critical univariate rate (13) satisfies the scaling $\nu_n^2 = \mathcal{O}(m/n)$. For a finite-rank kernel and any $t > 0$, we have

$$\mathcal{Q}_{\sigma,n}(t) = \frac{1}{\sqrt{n}}\sqrt{\sum_{j=1}^m \min\{t^2, \mu_j\}} \;\leq\; t\sqrt{\frac{m}{n}},$$

from which the claim follows by the definition (13). $\square$

Next, we present a result for the RKHS's with infinitely many eigenvalues, but whose eigenvalues decay at a rate $\mu_k \simeq (1/k)^{2\alpha}$ for some parameter $\alpha > 1/2$. Among other examples, this type of scaling covers the case of Sobolev spaces, say consisting of functions with α derivatives (e.g., [6, 15]).

Corollary 2. *Under the same conditions as Theorem 1, consider an univariate kernel with eigenvalue decay $\mu_k \simeq (1/k)^{2\alpha}$ for some $\alpha > 1/2$. Then the kernel estimator defined in (10) with $\lambda_n = c\gamma_n$ and $\rho_n = c\gamma_n^2$ with $16 \leq c < \infty$ satisfies*

$$\mathbb{P}\Big[\max\big\{\|\widehat{f}-f^*\|_n^2, \|\widehat{f}-f^*\|_2^2\big\} \geq C\big\{\frac{s\log d}{n} + s\big(\frac{1}{n}\big)^{\frac{2\alpha}{2\alpha+1}}\big\}\Big] \leq c_1 \exp\big(-c_2(n^{\frac{1}{2\alpha+1}}+\log d)\big). \tag{18}$$

Proof. As in the previous corollary, we need to compute the critical univariate rate ν_n. Given the assumption of polynomial eigenvalue decay, a truncation argument shows that $\mathcal{Q}_{\sigma,n}(t) = \mathcal{O}\big(\frac{t^{1-\frac{1}{2\alpha}}}{\sqrt{n}}\big)$. Consequently, the critical univariate rate (13) satisfies the scaling $\nu_n^2 \asymp \nu_n^{1-\frac{1}{2\alpha}}/\sqrt{n}$, or equivalently, $\nu_n^2 \asymp n^{-\frac{2\alpha}{2\alpha+1}}$. $\square$

3.3 Minimax lower bounds

In this section, we provide minimax lower bounds in $L^2(\mathbb{P})$ error so as to complement the achievability results derived in Theorem 1. Given the function class $\mathcal{F}$, the minimax $L^2(\mathbb{P})$-error is given by

$$\mathfrak{M}_{\mathbb{P}}(\mathcal{F}_{d,s,\mathcal{H}}) := \inf_{\widehat{f}_n} \sup_{f^* \in \mathcal{F}} \|\widehat{f}_n - f^*\|_2^2, \tag{19}$$

where the infimum is taken over all measurable functions of n samples $\{(x_i, y_i)\}_{i=1}^n$. As defined, this minimax error is a random variable, and our goal is to obtain a lower bound in probability.

Central to our proof of the lower bounds is the metric entropy structure of the univariate reproducing kernel Hilbert spaces. More precisely, our lower bounds depend on the *packing entropy,* defined as follows. Let $(\mathcal{G}, \rho)$ be a totally bounded metric space, consisting of a set $\mathcal{G}$ and a metric $\rho : \mathcal{G} \times \mathcal{G} \to \mathbb{R}_+$. An ϵ-packing of $\mathcal{G}$ is a collection $\{f^1, \ldots, f^M\} \subset \mathcal{G}$ such that $\rho(f^i, f^j) \geq \epsilon$ for all $i \neq j$. The ϵ-packing number $M(\epsilon; \mathcal{G}, \rho)$ is the cardinality of the largest ϵ-packing. The packing entropy is the simply the logarithm of the packing number, namely the quantity $\log M(\epsilon; \mathcal{G}, \rho)$, to which we also refer as the metric entropy. In this paper, we derive explicit minimax lower bounds for two different scalings of the univariate metric entropy.

Logarithmic metric entropy: There exists some $m > 0$ such that

$$\log M(\epsilon; \mathbb{B}_{\mathcal{H}}(1), L^2(\mathbb{P})) \simeq m \log(1/\epsilon) \qquad \text{for all } \epsilon \in (0,1). \tag{20}$$

Function classes with metric entropy of this type include linear functions (for which $m = k$), univariate polynomials of degree k (for which $m = k+1$), and more generally, any function space with finite VC-dimension [36]. This type of scaling also holds for any RKHS based on a kernel with rank m (e.g., see [11]), and these finite-rank kernels include both linear and polynomial functions as special cases.

Polynomial metric entropy There exists some $\alpha > 0$ such that

$$\log M(\epsilon; \mathbb{B}_{\mathcal{H}}(1), L^2(\mathbb{P})) \simeq (1/\epsilon)^{1/\alpha} \qquad \text{for all } \epsilon \in (0,1). \tag{21}$$

Various types of Sobolev/Besov classes exhibit this type of metric entropy decay [6, 15]. In fact, any RKHS in which the kernel eigenvalues decay at a rate $k^{-2\alpha}$ have a metric entropy with this scaling [10, 11].

We are now equipped to state our lower bounds on the minimax risk (19):

Theorem 2. *Given n i.i.d. samples from the sparse additive model (7) with sparsity $s \leq d/4$, there is an universal constant $C > 0$ such that:*

(a) For a univariate class $\mathcal{H}$ with logarithmic metric entropy (20) indexed by parameter m, we have

$$\mathfrak{M}_{\mathbb{P}}(\mathcal{F}_{d,s,\mathcal{H}}) \;\geq\; C \Big\{ \frac{s \log(d/s)}{n} + s\, \frac{m}{n} \Big\} \tag{22}$$

with probability greater than $1/2$.

(b) For a univariate class $\mathcal{H}$ with polynomial metric entropy (21) indexed by α, we have

$$\mathfrak{M}_{\mathbb{P}}(\mathcal{F}_{d,s,\mathcal{H}}) \;\geq\; C \left\{ \frac{s \log(d/s)}{n} + s \left(\frac{1}{n}\right)^{\frac{2\alpha}{2\alpha+1}} \right\} \tag{23}$$

with probability greater than $1/2$.

The proof of Theorem 2 is provided in Section 4.2. Our choice of stating bounds that hold with probability $1/2$ is simply a convention often used in information-theoretic approaches (see, for instance, the papers [16, 38, 39]). We note that analogous lower bounds can established with probabilities arbitrarily close to one, albeit at the expense of worse constants. The most important consequence of Theorem 2 is in establishing the minimax-optimality of the results given in Corollary 1 and 2; in particular, in the regime of sub-linear sparsity (i.e., for which $\log d = \mathcal{O}(\log(d/s))$), the combination of Theorem 2 with these corollaries identifies the minimax rates up to constant factors.

3.4 Comparison with other estimators

It is interesting to compare these convergence rates in $L^2(\mathbb{P}_n)$ error with those established in past work [19, 24, 31] using different estimators. Ravikumar et. al [31] show that any solution to their back-fitting method is consistent in terms of mean-squared error risk (see Theorem 3 in their paper) but they don't have rates in the regime where $s \to \infty$. An earlier method of Koltchinskii and Yuan [19] is based regularizing the least-squares loss with the $(\mathcal{H}, 1)$-norm penalty—that is, $\sum_{j=1}^{d} \|f_j\|_{\mathcal{H}}$ but no $(\|.\|_n, 1)$-norm penalty; Theorem 2 in their paper presents a rate that captures the decomposition into two terms, a subset selection and s-dimensional estimation term. In quantitative terms however, their rates are looser than those given here; in particular, their bound includes a term of the order $\frac{s^3 \log d}{n}$, which is larger than the bound in Theorem 1. Meier et al. [24] analyze a different M-estimator to the one we analyze in this paper. Rather than adding two separate $(\mathcal{H}, 1)$-norm and an $(\|.\|_n, 1)$-norm penalties, they combine the two terms into a single sparsity and smoothness penalty. For their estimator, Meier et al. [24] establish a convergence rate of the form $\mathcal{O}(s(\frac{\log d}{n})^{\frac{2\alpha}{2\alpha+1}})$ in the case of α-smooth Sobolev spaces (see Theorem 1 in their paper). This result is sub-optimal compared to the optimal rate proven in Theorem 2(b). More precisely, we either have $\frac{\log d}{n} < (\frac{\log d}{n})^{\frac{2\alpha}{2\alpha+1}}$, when subset selection term dominates, or $(\frac{1}{n})^{\frac{2\alpha}{2\alpha+1}} < (\frac{\log d}{n})^{\frac{2\alpha}{2\alpha+1}}$, when the s-dimensional estimation term dominates. In all of the above-mentioned methods, it is unclear whether or not sharper analysis would yield better rates. Koltchinskii and Yuan [20] analyzes the same estimator as the M-estimator (10) and achieve the same rates as in Theorem 1, under a global boundedness condition. In the following section, we show that rates in their paper are not minimax optimal for Sobolev spaces when $s = \Omega(\sqrt{n})$.

3.5 Upper bounds under a global boundedness assumption

As discussed previously in the introduction, the past work of Koltchinski and Yuan [20], referred to as KY for short, is based on the M-estimator (10). In terms of rates obtained, they establish a convergence rate based on two terms as in Theorem 1, but with a pre-factor that depends on the global quantity

$$B = \sup_{f \in \mathcal{F}_{d,s,\mathcal{H}}} \|f\|_\infty = \sup_{f \in \mathcal{F}_{d,s,\mathcal{H}}} \sup_x |f(x)|, \tag{24}$$

assumed to be bounded independently of dimension and sparsity. Such types of global boundedness conditions are fairly standard in classical non-parametric estimation and is equivalent to univariate boundedness, and they have no effect on minimax rates. In sharp contrast, the analysis of this section shows that for sparse additive models in the regime $s = \Omega(\sqrt{n})$, such global boundedness can *substantially speed up* minimax rates, showing that the rates proven in KY are not minimax optimal for these classes. The underlying insight is as follows: when the sparsity grows, imposing global boundedness over s-variate functions substantially reduces the effective dimension from its original size s to a lower dimensional quantity, which we denote by $sK_B(s,n)$, and moreover, the quantity $K_B(s,n) \to 0$ when $s = \Omega(\sqrt{n})$ as described below.

Recall the definition (6) of the function class $\mathcal{F}_{d,s,\mathcal{H}}$. The model considered in the KY paper is the smaller function class

$$\mathcal{F}^*_{d,s,\mathcal{H}}(B) := \bigcup_{\substack{S \subset \{1,2,\ldots,d\} \\ |S|=s}} \mathcal{H}(S,B),$$

where $\mathcal{H}(S,B) := \big\{ f = \sum_{j \in S} f_j \;\mid\; f_j \in \mathcal{H}, \text{ and } f_j \in \mathbb{B}_{\mathcal{H}}(1) \; \forall\, j \in S \text{ and } \|f\|_\infty \leq B \big\}$.

The following theorem provides sharper rates for the Sobolev case, in which each univariate Hilbert space has eigenvalues decaying as $\mu_k \simeq k^{-2\alpha}$ for some smoothness parameter $\alpha > 1/2$. Our probabilistic bounds involve the quantity

$$\delta_n := \max\Big(\sqrt{\frac{s\log(d/s)}{n}}, B^{1/2}\big(\frac{s^{\frac{1}{\alpha}}\log s}{n}\big)^{1/4}\Big), \tag{25}$$

and our rates are stated in terms of the function

$$K_B(s,n) := B\sqrt{\log s}\,(s^{-1/2\alpha} n^{1/(4\alpha+2)})^{2\alpha-1}. \tag{26}$$

Note that $K_B(s,n) \to 0$ if $s = \Omega(\sqrt{n})$. With this notation, we have the following *upper bound* on the minimax risk over the function class $\mathcal{F}^*_{d,s,\mathcal{H}}(B)$.

Theorem 3. *Consider a Sobolev RKHS $\mathcal{H}$ with eigenvalue decay $k^{-2\alpha}$ and eigenfunctions such that $\|\phi_k\|_\infty \leq C < \infty$. Then there are universal constants (c_1, c_2, κ) such that with probability greater than $1 - 2\exp\big(-c_1 n\delta_n^2\big)$, we have*

$$\underbrace{\min_{\hat{f}} \max_{f^* \in \mathcal{F}^*_{d,s,\mathcal{H}}(B)} \|\widehat{f} - f^*\|_2^2}_{\mathfrak{M}_{\mathbb{P}}(\mathcal{F}^*_{d,s,\mathcal{H}}(B))} \leq \kappa^2(1+B)Csn^{-\frac{2\alpha}{2\alpha+1}}\Bigg(K_B(s,n) + n^{-1/(2\alpha+1)}\log(d/s)\Bigg), \tag{27}$$

as long as $n\delta_n^2 = \Omega(\log(1/\delta_n))$.

We provide the proof of Theorem 3 in Section 4.3; it is based on analyzing directly the least-squares estimator over $\mathcal{F}^*_{d,s,\mathcal{H}}(B)$. The assumption that $\|\phi_k\|_\infty \leq C < \infty$ for all k includes the usual Sobolev spaces in which ϕ_k are (rescaled) Fourier basis functions. An immediate consequence of Theorem 3 is that the minimax rates over the function class $\mathcal{F}^*_{d,s,\mathcal{H}}(B)$ can be strictly faster than minimax rates for the class $\mathcal{F}_{d,s,\mathcal{H}}$ (which does not assume global boundedness). Recall that the minimax lower bound from Theorem 2 (b) takes the form:

$$\mathfrak{M}_{\mathbb{P}}(\mathcal{F}_{d,s,\mathcal{H}}) \;:=\; \min_{\widehat{f}} \max_{f^* \in \mathcal{F}_{d,s,\mathcal{H}}} \|\widehat{f} - f^*\|_2^2 \geq C_1 sn^{-\frac{2\alpha}{2\alpha+1}}\Bigg(1 + n^{-1/(2\alpha+1)}\log(d/s)\Bigg),$$

for a universal constant C_1. Note that up to constant factors, the achievable rate (27) from Theorem 3 is the same except that the term 1 is replaced by the function $K_B(s,n)$. Consequently, for scalings of (s,n) such that $K_B(s,n) \to 0$, global boundedness conditions lead to strictly faster rates.

Corollary 3. *Under the conditions of Theorem 3, we have*

$$\frac{\mathfrak{M}_{\mathbb{P}}(\mathcal{F}_{d,s,\mathcal{H}})}{\mathfrak{M}_{\mathbb{P}}(\mathcal{F}^*_{d,s,\mathcal{H}}(B))} \geq \frac{C_1(1+n^{-1/(2\alpha+1)}\log(d/s))}{C\,\kappa^2(1+B)\,(K_B(s,n)+n^{-1/(2\alpha+1)}\log(d/s))} \to +\infty$$

with probability at least $1/2$*, whenever* $B = \mathcal{O}(1)$ *and* $K_B(s,n) \to 0$.

Remarks: The quantity $K_B(s,n)$ is guaranteed to decay to zero as long as the sparsity index s grows in a non-trivial way with the sample size. For instance, if we have $s = \Omega(\sqrt{n})$ for a problem of dimension $d = \mathcal{O}(n^\beta)$ for any $\beta \geq 1/2$, then it can be verified that $f_B(s,n) = o(1)$. As an alternative view of the differences, it can be noted that there are scalings of (n,s,d) for which the minimax rate $\mathfrak{M}_{\mathbb{P}}(\mathcal{F}_{d,s,\mathcal{H}})$ over $\mathcal{F}_{d,s,\mathcal{H}}$ is constant—that is, does not vanish as $n \to +\infty$—while the minimax rate $\mathfrak{M}_{\mathbb{P}}(\mathcal{F}^*_{d,s,\mathcal{H}}(B))$ does vanish. As an example, consider the Sobolev class with smoothness $\alpha = 2$, corresponding to twice-differentiable functions. For a sparsity index $s = \mathcal{O}(n^{4/5})$, then Theorem 2(b) implies that $\mathfrak{M}_{\mathbb{P}}(\mathcal{F}_{d,s,\mathcal{H}}) = \Omega(1)$, so that the minimax rate over $\mathcal{F}_{d,s,\mathcal{H}}$ is strictly bounded away from zero for all sample sizes. In contrast, under a global boundedness condition, Theorem 3 shows that the minimax rate is upper bounded as $\mathfrak{M}_{\mathbb{P}}(\mathcal{F}^*_{d,s,\mathcal{H}}(B)) = \mathcal{O}\big(n^{-1/5}\sqrt{\log n}\big)$, which tends to zero.

In summary, Theorem 3 and Theorem 2(b) together show that the minimax rates over $\mathcal{F}_{d,s,\mathcal{H}}$ and $\mathcal{F}^*_{d,s,\mathcal{H}}(B)$ can be drastically different. Thus, global boundedness is a stringent condition in the high-dimensional setting; in particular, the rates given in Theorem 3 of Koltchinskii and Yuan [20] are not minimax optimal when $s = \Omega(\sqrt{n})$.

4 Proofs

In this section, we provide the proofs of our three main theorems. For clarity in presentation, we split the proofs up into a series of lemmas, with the bulk of the more technical proofs deferred to the appendices. This splitting allows our presentation in Section 4 to be relatively streamlined.

4.1 Proof of Theorem 1

At a high-level, Theorem 1 is based on an appropriate adaptation to the non-parametric setting of various techniques that have been developed for bounding the error to those in sparse linear regression (e.g., [5, 27]). In contrast to the parametric setting where classical tail bounds are sufficient, controlling the error terms in the non-parametric case requires more advanced techniques from empirical process theory. In particular, we make use of concentration theorems for Gaussian and empirical processes (e.g., [12, 21, 23, 28, 35]) as well as results on the Rademacher complexity of kernel classes [4, 25].

At the core of the proof are three technical lemmas. First, Lemma 1 provides an upper bound on the Gaussian complexity of any function of the form $f = \sum_{j=1}^d f_j$ in terms of the norms $\|\cdot\|_{\mathcal{H},1}$ and $\|\cdot\|_{n,1}$ previously defined. Lemma 2 exploits the notion of decomposability [27], as applied to these norms, in order to show that the error function belongs to a particular cone-shaped set. Finally, Lemma 3 establishes an upper bound on the $L^2(\mathbb{P})$ error of our estimator in terms of the $L^2(\mathbb{P}_n)$ error which completes the proof. The latter lemma can be interpreted as proving that our problem satisfies non-parametric analog of a restricted strong convexity [27] or restricted eigenvalue condition [5]. The proof of Lemma 3 involves a new approach that combines the Sudakov minoration [28] with a one-sided concentration bound for non-negative random variables [12].

Throughout the proof, we use C and c_i, $i = 1, 2, 3, 4$ to denote universal constants, independent of (n, d, s). Note that the precise numerical values of these constants may change from line to line. The reader should recall the definitions of ν_n and γ_n from equations (13) and (14) respectively. For a subset $A \subseteq \{1, 2, \ldots, d\}$ and a function of the form $f = \sum_{j=1}^d f_j$, we adopt the convenient notation

$$\|f_A\|_{n,1} := \sum_{j \in A} \|f_j\|_n, \quad \text{and} \quad \|f_A\|_{\mathcal{H},1} := \sum_{j \in A} \|f_j\|_{\mathcal{H}}. \tag{28}$$

We begin by establishing an inequality on the error function $\widehat{\Delta} := \widehat{f} - f^*$. Since $\widehat{f}$ and f^* are, respectively, optimal and feasible for the problem (10), we are guaranteed that $\mathcal{L}(\widehat{f}) \leq \mathcal{L}(f^*)$, and hence that the error function $\widehat{\Delta}$ satisfies the bound

$$\frac{1}{2n} \sum_{i=1}^n (w_i + \mu - \bar{y}_n - \widehat{\Delta}(x_i))^2 + \lambda_n \|\widehat{f}\|_{n,1} + \rho_n \|\widehat{f}\|_{\mathcal{H},1} \leq \frac{1}{2n} \sum_{i=1}^n (w_i + \mu - \bar{y}_n)^2 + \lambda_n \|f^*\|_{n,1} + \rho_n \|f^*\|_{\mathcal{H},1}.$$

Some simple algebra yields the bound

$$\frac{1}{2} \|\widehat{\Delta}\|_n^2 \leq \big|\frac{1}{n} \sum_{i=1}^n w_i \widehat{\Delta}(x_i)\big| + |\bar{y}_n - \mu| \big|\frac{1}{n} \sum_{i=1}^n \widehat{\Delta}(x_i)\big| + \lambda_n \|\widehat{\Delta}\|_{n,1} + \rho_n \|\widehat{\Delta}\|_{\mathcal{H},1}. \tag{29}$$

Following the terminology of van de Geer [35], we refer to this bound as our *basic inequality*.

4.1.1 Controlling deviation from the mean

Our next step is to control the error due to estimating the mean $|\bar{y}_n - \mu|$. We begin by observing that this error term can be written as $\bar{y}_n - \mu = \frac{1}{n} \sum_{i=1}^n (y_i - \mu)$, Now consider the random variable $y_i - \mu = \sum_{j \in S} f_j^*(x_{ij}) + w_i$. It is the sum of the s independent random variables $f_j^*(x_{ij})$, each bounded in absolute value with one, along with the independent sub-Gaussian noise term w_i; consequently, the variable $y_i - \mu$ is sub-Gaussian with parameter at most $\sqrt{s+1}$ (see, e.g., Lemma 1.4 in Buldygin and Kozachenko [9]). By applying standard sub-Gaussian tail bounds, we have $\mathbb{P}(|\bar{y}_n - \mu| > t) \leq 2 \exp(-\frac{nt^2}{2(s+1)})$, and hence, if we define the event $\mathcal{C}(\gamma_n) = \{|\bar{y}_n - \mu| \leq \sqrt{s} \gamma_n\}$, we are guaranteed

$$\mathbb{P}[\mathcal{C}(\gamma_n)] \geq 1 - 2 \exp(-\frac{n \gamma_n^2}{4}).$$

For the remainder of the proof, we condition on the event $\mathcal{C}(\gamma_n)$. Under this conditioning, the bound (29) simplifies to:

$$\frac{1}{2} \|\widehat{\Delta}\|_n^2 \leq \big|\frac{1}{n} \sum_{i=1}^n w_i \widehat{\Delta}(x_i)\big| + \sqrt{s} \gamma_n \|\widehat{\Delta}\|_n + \lambda_n \|\widehat{\Delta}\|_{n,1} + \rho_n \|\widehat{\Delta}\|_{\mathcal{H},1},$$

where we have applied the Cauchy-Schwarz inequality to write $\big|\frac{1}{n} \sum_{i=1}^n \widehat{\Delta}(x_i)\big| \leq \|\widehat{\Delta}\|_n$.

4.1.2 Controlling the Gaussian complexity term

The following lemma provides control the Gaussian complexity term on the right-hand side of inequality (29) by bounding the Gaussian complexity for the univariate functions $\widehat{\Delta}_j$, $j = 1, 2, \ldots, d$ in terms of their $\|\cdot\|_n$ and $\|\cdot\|_{\mathcal{H}}$ norms. In particular, recalling that $\gamma_n = \kappa \max\{\sqrt{\frac{\log d}{n}},\ \nu_n\}$, we have the following lemma.

Lemma 1. *Define the event*

$$\mathcal{T}(\gamma_n) := \Big\{ \forall\ \ j = 1, 2, \ldots, d,\ \Big| \frac{1}{n} \sum_{i=1}^n w_i \widehat{\Delta}_j(x_{ij}) \Big| \leq 8\gamma_n^2\ \|\widehat{\Delta}_j\|_{\mathcal{H}} + 8\gamma_n\ \|\widehat{\Delta}_j\|_n \Big\}. \tag{30}$$

Then under the condition $n\gamma_n^2 = \Omega(\log(1/\gamma_n))$, we have

$$\mathbb{P}(\mathcal{T}(\gamma_n)) \geq 1 - c_1 \exp(-c_2 n \gamma_n^2). \tag{31}$$

The proof of this lemma, provided in Appendix B, uses concentration of measure for Lipschitz functions over Gaussian random variables [21] combined with peeling and weighting arguments [1, 35]. In particular, the subset selection term $(\frac{s \log d}{n})$ in Theorem 1 arises from taking the maximum over all d components.

The remainder of our analysis involves conditioning on the event $\mathcal{T}(\gamma_n) \cap \mathcal{C}(\gamma_n)$. Using Lemma 1, when conditioned on the event $\mathcal{T}(\gamma_n) \cap \mathcal{C}(\gamma_n)$ we have:

$$\|\widehat{\Delta}\|_n^2 \quad \leq \quad 2\sqrt{s}\gamma_n \|\widehat{\Delta}\|_n + (16\gamma_n + 2\lambda_n)\|\widehat{\Delta}\|_{n,1} + (16\gamma_n^2 + 2\rho_n)\|\widehat{\Delta}\|_{\mathcal{H},1}. \tag{32}$$

4.1.3 Exploiting decomposability

Recall that S denotes the true support of the unknown function f^*. By the definition (28), we can write $\|\widehat{\Delta}\|_{n,1} = \|\widehat{\Delta}_S\|_{n,1} + \|\widehat{\Delta}_{S^c}\|_{n,1}$, where $\widehat{\Delta}_S := \sum_{j \in S} \widehat{\Delta}_j$ and $\widehat{\Delta}_{S^c} := \sum_{j \in S^c} \widehat{\Delta}_j$. Similarly, we have an analogous representation for $\|\widehat{\Delta}\|_{\mathcal{H},1}$. The next lemma shows that conditioned on the event $\mathcal{T}(\gamma_n)$, the quantities $\|\widehat{\Delta}\|_{\mathcal{H},1}$ and $\|\widehat{\Delta}\|_{n,1}$ are not significantly larger than the corresponding norms as applied to the function $\widehat{\Delta}_S$.

Lemma 2. *Conditioned on the events $\mathcal{T}(\gamma_n)$ and $\mathcal{C}(\gamma_n)$, and with the choices $\lambda_n \geq 16\gamma_n$ and $\rho_n \geq 16\gamma_n^2$, we have*

$$\lambda_n \|\widehat{\Delta}\|_{n,1} + \rho_n \|\widehat{\Delta}\|_{\mathcal{H},1} \leq 4\lambda_n \|\widehat{\Delta}_S\|_{n,1} + 4\rho_n \|\widehat{\Delta}_S\|_{\mathcal{H},1} + \frac{1}{2} s \gamma_n^2. \tag{33}$$

The proof of this lemma, provided in Appendix C, is based on the decomposability [27] of the $\|\cdot\|_{\mathcal{H},1}$ and $\|\cdot\|_{n,1}$ norms. This lemma allows us to exploit the sparsity assumption, since in conjunction with Lemma 1, we have now bounded the right-hand side of the inequality (32) by terms involving only $\widehat{\Delta}_S$.

For the remainder of the proof of Theorem 1, we assume $\lambda_n \geq 16\gamma_n$ and $\rho_n \geq 16\gamma_n^2$. In particular, still conditioning on $\mathcal{C}(\gamma_n) \cap \mathcal{T}(\gamma_n)$ and applying Lemma 2 to inequality (32), we obtain

$$\begin{aligned}
\|\widehat{\Delta}\|_n^2 \quad &\leq \quad 2\sqrt{s}\gamma_n \|\widehat{\Delta}\|_n + 3\lambda_n \|\widehat{\Delta}\|_{n,1} + 3\rho_n \|\widehat{\Delta}\|_{\mathcal{H},1} \\
&\leq \quad 2\sqrt{s}\lambda_n \|\widehat{\Delta}\|_n + 12\lambda_n \|\widehat{\Delta}_S\|_{n,1} + 12\rho_n \|\widehat{\Delta}_S\|_{\mathcal{H},1} + \frac{3}{32} s \rho_n,
\end{aligned}$$

Finally, since both $\widehat{f}_j$ and f_j^* belong to $\mathbb{B}_{\mathcal{H}}(1)$, we have $\|\widehat{\Delta}_j\|_{\mathcal{H}} \leq \|\widehat{f}_j\|_{\mathcal{H}} + \|f_j^*\|_{\mathcal{H}} \leq 2$, which implies that $\|\widehat{\Delta}_S\|_{\mathcal{H},1} \leq 2s$, and hence

$$\|\widehat{\Delta}\|_n^2 \leq 2\sqrt{s}\lambda_n \|\widehat{\Delta}\|_n + 12\lambda_n \|\widehat{\Delta}_S\|_{n,1} + 25 s\rho_n. \tag{34}$$

4.1.4 Upper bounding $\|\widehat{\Delta}_S\|_{n,1}$

The final step is to control the term $\|\widehat{\Delta}_S\|_{n,1} = \sum_{j\in S} \|\widehat{\Delta}_j\|_n$ that appears in the upper bound (34). Ideally, we would like to upper bound it by a quantity of the order $\sqrt{s}\|\widehat{\Delta}_S\|_2$. Such an upper bound would follow immediately if it were phrased in terms of the population $\|\cdot\|_2$-norm rather than the empirical-$\|\cdot\|_n$ norm, but there are additional cross-terms with the empirical norm. Accordingly, a somewhat more delicate argument is required, which we provide here. First define the events

$$\mathcal{A}_j(\lambda_n) := \{\|\widehat{\Delta}_j\|_n \leq 2\|\widehat{\Delta}_j\|_2 + \lambda_n\},$$

and $\mathcal{A}(\lambda_n) = \cap_{j=1}^d \mathcal{A}_j(\lambda_n)$. By applying Lemma 7 from Appendix A with $t = \lambda_n \geq 16\gamma_n$ and $b = 2$, we conclude that $\|f_j\|_n \leq 2\|f_j\|_2 + \lambda_n$. with probability greater than $1 - c_1 \exp(-c_2 n\lambda_n^2)$. Consequently, if we define the event $\mathcal{A}(\lambda_n) = \cap_{j\in S}\mathcal{A}_j(\lambda_n)$, then this tail bound together with the union bound implies that

$$\mathbb{P}[\mathcal{A}^c(\lambda_n)] \leq s\, c_1 \exp(-c_2 n \lambda_n^2) \;\leq\; c_1 \exp(-c_2' n\lambda_n^2), \tag{35}$$

where we have used the fact that $\lambda_n = \Omega(\sqrt{\frac{\log s}{n}})$. Now, conditioned on the event $\mathcal{A}(\lambda_n)$, we have

$$\|\widehat{\Delta}_S\|_{n,1} \;\leq\; \sum_{j\in S} \|\widehat{\Delta}_j\|_n \;\leq\; 2\sum_{j\in S}\|\widehat{\Delta}_j\|_2 + s\lambda_n \;\leq\; 2\sqrt{s}\|\widehat{\Delta}_S\|_2 + s\lambda_n \;\leq\; 2\sqrt{s}\|\widehat{\Delta}\|_2 + s\lambda_n. \tag{36}$$

Substituting this upper bound (36) on $\|\widehat{\Delta}_S\|_{n,1}$ into our earlier inequality (34) yields

$$\|\widehat{\Delta}\|_n^2 \;\leq\; 2\sqrt{s}\lambda_n\|\widehat{\Delta}\|_n + 24\sqrt{s}\lambda_n\|\widehat{\Delta}\|_2 + 12 s\lambda_n^2 + 25 s\rho_n. \tag{37}$$

At this point, we encounter a challenge due to the unbounded nature of our function class. In particular, if $\|\widehat{\Delta}\|_2$ were upper bounded by $C\max(\|\widehat{\Delta}\|_n, \sqrt{s}\lambda_n, \sqrt{s\rho_n})$, then the upper bound (37) would immediately imply the claim of Theorem 1. If one were to assume global boundedness of the multivariate functions $\widehat{f}$ and f^*, as done in past work [20], then an upper bound on $\|\widehat{\Delta}\|_2$ of this form would directly follow from known results (e.g., Theorem 2.1 in Bartlett et al. [4].) However, since we do not impose global boundedness, we need to develop a novel approach to obtaining suitable control $\|\widehat{\Delta}\|_2$, the task to which we now turn.

4.1.5 Controlling $\|\widehat{\Delta}\|_2$ for unbounded classes

For the remainder of the proof, we condition on the event $\mathcal{A}(\lambda_n) \cap \mathcal{T}(\gamma_n) \cap \mathcal{C}(\gamma_n)$. We split our analysis into three cases. Throughout the proof, we make use of the quantity

$$\tilde{\delta}_n := B\max(\sqrt{s}\lambda_n, \sqrt{s\rho_n}), \tag{38}$$

where $B \in (1, \infty)$ is a constant to be chosen later in the argument.

Case 1: If $\|\widehat{\Delta}\|_2 < \|\widehat{\Delta}\|_n$, then combined with inequality (37), we conclude that

$$\|\widehat{\Delta}\|_n^2 \leq 2\sqrt{s}\lambda_n\|\widehat{\Delta}\|_n + 24\sqrt{s}\lambda_n\|\widehat{\Delta}\|_n + 12s\lambda_n^2 + 25s\rho_n.$$

This is a quadratic inequality in terms of the quantity $\|\widehat{\Delta}\|_n$, and some algebra shows that it implies the bound $\|\widehat{\Delta}\|_n \leq 15\max(\sqrt{s}\lambda_n, \sqrt{s\rho_n})$. By assumption, we then have $\|\widehat{\Delta}\|_2 \leq 15\max(\sqrt{s}\lambda_n, \sqrt{s\rho_n})$ as well, thereby completing the proof of Theorem 1.

Case 2: If $\|\widehat{\Delta}\|_2 < \tilde{\delta}_n$, then together with the bound (37), we conclude that

$$\|\widehat{\Delta}\|_n^2 \leq 2\sqrt{s}\lambda_n\|\widehat{\Delta}\|_n + 24\sqrt{s}\lambda_n\tilde{\delta}_n + 12s\lambda_n^2 + 25s\rho_n.$$

This inequality is again a quadratic in $\|\widehat{\Delta}\|_n$; moreover, note that by definition (38) of $\tilde{\delta}_n$, we have $s\lambda_n^2 + s\rho_n = \mathcal{O}(\tilde{\delta}_n^2)$. Consequently, this inequality implies that $\|\widehat{\Delta}\|_n \leq C\tilde{\delta}_n$ for some constant C. Our starting assumption implies that $\|\widehat{\Delta}\|_2 \leq \tilde{\delta}_n$, so that the claim of Theorem 1 follows in this case.

Case 3: Otherwise, we may assume that $\|\widehat{\Delta}\|_2 \geq \tilde{\delta}_n$ and $\|\widehat{\Delta}\|_2 \geq \|\widehat{\Delta}\|_n$. In this case, the inequality (37) together with the bound $\|\widehat{\Delta}\|_2 \geq \|\widehat{\Delta}\|_n$ implies that

$$\|\widehat{\Delta}\|_n^2 \leq 2\sqrt{s}\lambda_n\|\widehat{\Delta}\|_2 + 24\sqrt{s}\lambda_n\|\widehat{\Delta}\|_2 + 12s\lambda_n^2 + 25s\rho_n. \tag{39}$$

Our goal is to establish a lower bound on the left-hand-side—namely, the quantity $\|\widehat{\Delta}\|_n^2$—in terms of $\|\widehat{\Delta}\|_2^2$. In order to do so, we consider the function class $\mathcal{G}(\lambda_n, \rho_n)$ defined by functions of the form $g = \sum_{j=1}^d g_j$, and such that

$$\lambda_n\|g\|_{n,1} + \rho_n\|g\|_{\mathcal{H},1} \leq 4\lambda_n\|g_S\|_{n,1} + 4\rho_n\|g_S\|_{\mathcal{H},1} + \frac{1}{32}s\rho_n, \tag{40a}$$

$$\|g_S\|_{1,n} \leq 2\sqrt{s}\|g_S\|_2 + s\lambda_n \qquad \text{and} \tag{40b}$$

$$\|g\|_n \leq \|g\|_2. \tag{40c}$$

Conditioned on the events $\mathcal{A}(\gamma_n)$, $\mathcal{T}(\gamma_n)$ and $\mathcal{C}(\gamma_n)$, and with our choices of regularization parameter, we are guaranteed that the error function $\widehat{\Delta}$ satisfies all three of these constraints, and hence that $\widehat{\Delta} \in \mathcal{G}(\lambda_n, \rho_n)$. Consequently, it suffices to establish a lower bound on $\|g\|_n$ that holds uniformly over the class $\mathcal{G}(\lambda_n, \rho_n)$. In particular, define the event

$$\mathcal{B}(\lambda_n, \rho_n) := \Big\{\|g\|_n^2 \geq \|g\|_2^2/2 \quad \text{for all } g \in \mathcal{G}(\lambda_n, \rho_n) \text{ where } \|g\|_2 \geq \tilde{\delta}_n\Big\}. \tag{41}$$

The following lemma shows that this event holds with high probability.

Lemma 3. *Under the conditions of Theorem 1, there are universal constants c_i such that*

$$\mathbb{P}[\mathcal{B}(\lambda_n, \rho_n)] \geq 1 - c_1\exp(-c_2 n\gamma_n^2). \tag{42}$$

We note that this lemma can be interpreted as guaranteeing a version of restricted strong convexity [27] for the least-squares loss function, suitably adapted to the non-parametric setting. Since we do not assume global boundedness, the proof of this lemma requires a novel technical argument, one which combines a one-sided concentration bound for non-negative random variables (Theorem 3.5 in Chung and Lu [12]) with the Sudakov minoration [28] for Gaussian complexity. We refer the reader to Appendix D for all the details of the proof.

Using Lemma 3 and conditioning $\mathcal{B}(\lambda_n, \rho_n)$, we are guaranteed that $\|\widehat{\Delta}\|_n^2 \geq \|\widehat{\Delta}\|_2^2/2$, and hence, combined with our earlier bound (39), we conclude that

$$\|\widehat{\Delta}\|_2^2 \leq 4\sqrt{s}\lambda_n \|\widehat{\Delta}\|_2 + 48\sqrt{s}\lambda_n \|\widehat{\Delta}\|_2 + 24 s\lambda_n^2 + 50 s\rho_n.$$

Hence $\|\widehat{\Delta}\|_n \leq \|\widehat{\Delta}\|_2 \leq C \max(\sqrt{s}\lambda_n, \sqrt{s\rho_n})$, completing the proof of the claim in the third case.

In summary, the entire proof is based on conditioning on the three events $\mathcal{T}(\gamma_n)$, $\mathcal{A}(\lambda_n)$ and $\mathcal{B}(\lambda_n, \rho_n)$. From the bound (35) as well as Lemmas 1 and 3, we have

$$\mathbb{P}\big[\mathcal{T}(\gamma_n) \cap \mathcal{A}(\lambda_n) \cap \mathcal{B}(\lambda_n, \rho_n) \cap \mathcal{C}(\gamma_n)\big] \geq 1 - c_1 \exp\big(-c_2 n \gamma_n^2\big),$$

thereby showing that $\max\{\|\widehat{f} - f^*\|_n^2, \|\widehat{f} - f^*\|_2^2\} \leq C \max(s\lambda_n^2, s\rho_n)$ with the claimed probability. This completes the proof of Theorem 1.

4.2 Proof of Theorem 2

We now turn to the proof of the minimax lower bounds stated in Theorem 2. For both parts (a) and (b), the first step is to follow a standard reduction to testing (e.g., [16, 38, 39]) so as to obtain a lower bound on the minimax error $\mathfrak{M}_{\mathbb{P}}(\mathcal{F}_{d,s,\mathcal{H}})$ in terms of the probability of error in a multi-way hypothesis testing. We then apply different forms of the Fano inequality [38, 39] in order to lower bound the probability of error in this testing problem. Obtaining useful bounds requires a precise characterization of the metric entropy structure of $\mathcal{F}_{d,s,\mathcal{H}}$, as stated in Lemma 4.

4.2.1 Reduction to testing

We begin with the reduction to a testing problem. Let $\{f^1, \ldots, f^M\}$ be a δ_n-packing of $\mathcal{F}$ in the $\|\cdot\|_2$-norm, and let Θ be a random variable uniformly distributed over the index set $[M] := \{1, 2, \ldots, M\}$. Note that we are using M as a shorthand for the packing number $M(\delta_n; \mathcal{F}, \|\cdot\|_2)$. A standard argument (e.g., [16, 38, 39]) then yields the lower bound

$$\inf_{\widehat{f}} \sup_{f^* \in \mathcal{F}} \mathbb{P}\big[\|\widehat{f} - f^*\|_2^2 \geq \delta_n^2/2\big] \geq \inf_{\widehat{\Theta}} \mathbb{P}[\widehat{\Theta} \neq \Theta], \tag{43}$$

where the infimum on the right-hand side is taken over all estimators $\widehat{\Theta}$ that are measurable functions of the data, and take values in the index set $[M]$.

Note that $\mathbb{P}[\widehat{\Theta} \neq \Theta]$ corresponds to the error probability in a multi-way hypothesis test, where the probability is taken over the random choice of Θ, the randomness of the design points $X_1^n := \{x_i\}_{i=1}^n$, and the randomness of the observations $Y_1^n := \{y_i\}_{i=1}^n$. Our initial analysis is performed conditionally on the design points, so that the only remaining randomness in the observations Y_1^n comes from the observation noise $\{w_i\}_{i=1}^n$. From Fano's inequality [13], for any estimator $\widehat{\Theta}$, we have $\mathbb{P}\big[\widehat{\Theta} \neq \Theta \mid X_1^n\big] \geq 1 - \frac{I_{X_1^n}(\Theta; Y_1^n) + \log 2}{\log M}$, where $I_{X_1^n}(\Theta; Y_1^n)$ denotes the mutual information between Θ and Y_1^n with X_1^n fixed. Taking expectations over X_1^n, we obtain the lower bound

$$\mathbb{P}\big[\widehat{\Theta} \neq \Theta\big] \geq 1 - \frac{\mathbb{E}_{X_1^n}\big[I_{X_1^n}(\Theta; Y_1^n)\big] + \log 2}{\log M}. \tag{44}$$

The remainder of the proof consists of constructing appropriate packing sets of $\mathcal{F}$, and obtaining good upper bounds on the mutual information term in the lower bound (44).

4.2.2 Constructing appropriate packings

We begin with results on packing numbers. Recall that $\log M(\delta; \mathcal{F}, \|\cdot\|_2)$ denotes the δ-packing entropy of $\mathcal{F}$ in the $\|\cdot\|_2$ norm.

Lemma 4. *(a) For all $\delta \in (0,1)$ and $s \leq d/4$, we have*

$$\log M(\delta; \mathcal{F}, \|\cdot\|_2) = \mathcal{O}\big(s \;\log M(\frac{\delta}{\sqrt{s}}; \mathbb{B}_{\mathcal{H}}(1), \|\cdot\|_2) + s \log \frac{d}{s}\big). \tag{45}$$

(b) For a Hilbert class with logarithmic metric entropy (20) and such that $\|f\|_2 \leq \|f\|_{\mathcal{H}}$, there exists set $\{f^1, \ldots, f^M\}$ with $\log M \;\geq\; C\,\big\{s\log(d/s) + sm\big\}$, and

$$\delta \;\leq\; \|f^k - f^m\|_2 \;\leq\; 8\delta \qquad \text{for all } k \neq m \in \{1, 2, \ldots, M\}. \tag{46}$$

The proof, provided in Appendix E, is combinatorial in nature. We now turn to the proofs of parts (a) and (b) of Theorem 2.

4.2.3 Proof of Theorem 2(a)

In order to prove this claim, it remains to exploit Lemma 4 in an appropriate way, and to upper bound the resulting mutual information. For the latter step, we make use of the generalized Fano approach (e.g., [39]).

From Lemma 4, we can find a set $\{f^1, \ldots, f^M\}$ that is a δ-packing of $\mathcal{F}$ in ℓ_2-norm, and such that $\|f^k - f^\ell\|_2 \leq 8\delta$ for all $k, \ell \in [M]$. For $k = 1, \ldots, M$, let $\mathbb{Q}^k$ denote the conditional distribution of Y_1^n conditioned on X_1^n and the event $\{\Theta = k\}$, and let $D(\mathbb{Q}^k \,\|\, \mathbb{Q}^\ell)$ denote the Kullback-Leibler divergence. From the convexity of mutual information [13], we have the upper bound $I_{X_1^n}(\Theta; Y_1^n) \leq \frac{1}{\binom{M}{2}} \sum_{k,\ell=1}^M D(\mathbb{Q}^k \,\|\, \mathbb{Q}^\ell)$. Given our linear observation model (7), we have

$$D(\mathbb{Q}^k \,\|\, \mathbb{Q}^\ell) = \frac{1}{2\sigma^2} \sum_{i=1}^n \big(f^k(x_i) - f^\ell(x_i)\big)^2 \;=\; \frac{n\,\|f^k - f^\ell\|_n^2}{2},$$

and hence

$$\mathbb{E}_{X_1^n}\big[I_{X_1^n}(Y_1^n; \Theta)\big] \leq \frac{n}{2} \frac{1}{\binom{M}{2}} \sum_{k,\ell=1}^M \mathbb{E}_{X_1^n}[\|f^k - f^\ell\|_n^2] \;=\; \frac{n}{2} \frac{1}{\binom{M}{2}} \sum_{k,\ell=1}^M \|f^k - f^\ell\|_2^2.$$

Since our packing satisfies $\|f^k - f^\ell\|_2^2 \leq 64\delta^2$, we conclude that

$$\mathbb{E}_{X_1^n}\big[I_{X_1^n}(Y_1^n; \Theta)\big] \leq 32 n \delta^2.$$

From the Fano bound (44), for any $\delta > 0$ such that $\frac{32n\delta^2 + \log 2}{\log M} < \frac{1}{4}$, then we are guaranteed that $\mathbb{P}[\widehat{\Theta} \neq \Theta] \geq \frac{3}{4}$. From Lemma 4(b), our packing set satisfies $\log M \geq C\big\{sm + s\log(d/s)\big\}$, so that so that the choice $\delta^2 = C'\left\{\frac{sm}{n} + \frac{s\log(d/s)}{n}\right\}$, for a suitably small $C' > 0$, can be used to guarantee the error bound $\mathbb{P}[\widehat{\Theta} \neq \Theta] \geq \frac{3}{4}$.

4.2.4 Proof of Theorem 2(b)

In this case, we use an upper bounding technique due to Yang and Barron [38] in order to upper bound the mutual information. Although the argument is essentially the same, it does not follow verbatim from their claims—in particular, there are some slight differences due to our initial conditioning—so that we provide the details here. By definition of the mutual information, we have

$$I_{X_1^n}(\Theta; Y_1^n) = \frac{1}{M}\sum_{k=1}^{M} D(\mathbb{Q}^k \,\|\, \mathbb{P}_Y),$$

where $\mathbb{Q}^k$ denotes the conditional distribution of Y_1^n given $\Theta = k$ and still with X_1^n fixed, whereas $\mathbb{P}_Y$ denotes the marginal distribution of $\mathbb{P}_Y$.

First we define *covering numbers*. Let $(\mathcal{G}, \rho)$ be a totally bounded metric space, consisting of a set $\mathcal{G}$ and a metric $\rho : \mathcal{G} \times \mathcal{G} \to \mathbb{R}_+$. An ϵ-covering set of $\mathcal{G}$ is a collection $\{f^1, \ldots, f^N\}$ of functions such that for all $f \in \mathcal{G}$ there exists $k \in \{1, 2, ..., N\}$ such that $\rho(f, f^k) \leq \epsilon$. The ϵ-covering number $N(\epsilon; \mathcal{G}, \rho)$ is the cardinality of the smallest ϵ-covering set.

Now let $\{g^1, \ldots, g^N\}$ be an ϵ-cover of $\mathcal{F}$ in the $\|\cdot\|_2$ norm, for a tolerance ϵ to be chosen. As argued in Yang and Barron [38], we have

$$I_{X_1^n}(\Theta; Y_1^n) \;=\; \frac{1}{M}\sum_{j=1}^{M} D(\mathbb{Q}^j \,\|\, \mathbb{P}_Y) \leq D(\mathbb{Q}^k \,\|\, \frac{1}{N}\sum_{k=1}^{N} \mathbb{P}^k),$$

where $\mathbb{P}^\ell$ denotes the conditional distribution of Y_1^n given g^ℓ and X_1^n. For each ℓ, let us choose $\ell^*(k) \in \arg\min_{\ell=1,\ldots,N} \|g^\ell - f^k\|_2$. We then have the upper bound

$$I_{X_1^n}(\Theta; Y_1^n) \leq \frac{1}{M}\sum_{k=1}^{M} \big\{ \log N + \frac{n}{2}\|g^{\ell^*(k)} - f^k\|_n^2 \big\}.$$

Taking expectations over X_1^n, we obtain

$$\begin{aligned}
\mathbb{E}_{X_1^n}[I_{X_1^n}(\Theta; Y_1^n)] &\leq \frac{1}{M}\sum_{k=1}^{M} \big\{ \log N + \frac{n}{2}\mathbb{E}_{X_1^n}[\|g^{\ell^*(k)} - f^k\|_n^2] \big\} \\
&\leq \log N + \frac{n}{2}\,\epsilon^2,
\end{aligned}$$

where the final inequality follows from the choice of our covering set.

From this point, we can follow the same steps as Yang and Barron [38]. The polynomial scaling (21) of the metric entropy guarantees that their conditions are satisfied, and we conclude that the minimax error is lower bounded any $\delta_n > 0$ such that $n\delta_n^2 \;\geq\; C\log N(\delta_n; \mathcal{F}, \|\cdot\|_2)$. From Lemma 4 and the assumed scaling (21), it is equivalent to solve the equation

$$n\delta_n^2 \;\geq\; C\Big\{ s\log(d/s) + s(\sqrt{s}/\delta_n)^{1/\alpha} \Big\},$$

from which some algebra yields $\delta_n^2 = C\big\{\frac{s\log(d/s)}{n} + s\big(\frac{1}{n}\big)^{\frac{2\alpha}{2\alpha+1}}\big\}$ as a suitable choice.

4.3 Proof of Theorem 3

Recall the definition of $\mathcal{F}^*_{d,s,\mathcal{H}}(B)$ and $\mathcal{H}(S,B)$ from Section 3.5 which ensures f^* is uniformly bounded by B. In order to establish upper bounds on the minimax rate in $L^2(\mathbb{P})$-error over $\mathcal{F}^*_{d,s,\mathcal{H}}(B)$, we we analyze a least-squares estimator (not the same as the original M-estimator (10)) constrained to $\mathcal{F}^*_{d,s,\mathcal{H}}(B)$:

$$\widehat{f} \in \arg\min_{f \in \mathcal{F}^*_{d,s,\mathcal{H}}(B)} \sum_{i=1}^n (y_i - \bar{y}_n - f(x_i))^2. \tag{47}$$

Since our goal is to upper bound the minimax rate in $L^2(\mathbb{P})$ error, it is sufficient to upper bound the $L^2(\mathbb{P})$-norm of $\widehat{f} - f^*$ where $\widehat{f}$ is any solution to (47). The proof shares many steps with the proof of Theorem 1. First, the same reasoning shows that the error $\widehat{\Delta} := \widehat{f} - f^*$ satisfies the basic inequality

$$\frac{1}{n}\sum_{i=1}^n \widehat{\Delta}^2(x_i) \leq \frac{2}{n}\Big|\sum_{i=1}^n w_i \widehat{\Delta}(x_i)\Big| + |\bar{y}_n - \mu|\Big|\frac{1}{n}\sum_{i=1}^n \widehat{\Delta}(x_i)\Big|. \tag{48}$$

Recall the definition (25) of the critical rate δ_n. Once again, we first control the term error due to estimating the mean $|\bar{y}_n - \mu|$. Noting that $\bar{y}_n - \mu = \frac{1}{n}\sum_{i=1}^n (y_i - \mu)$ and the random variable $y_i - \mu = f^*(x_i) + w_i$ is sub-Gaussian with parameter $\sqrt{B+1}$. This follows since f^* is bounded by B and using standard results on sums of independent sub-Gaussian random variables (see e.g. Lemma 1.4 in Buldygin and Kozachenko [9]). Therefore

$$\mathbb{P}(|\bar{y}_n - \mu| > t) \leq 2\exp\Big(-\frac{nt^2}{2(B+1)}\Big).$$

Setting $\mathcal{A}(\delta_n) = \{|\bar{y}_n - \mu| \leq \sqrt{B}\delta_n\}$, it is clear that

$$\mathbb{P}[\mathcal{A}(\delta_n)] \geq 1 - 2\exp\Big(-\frac{n\delta_n^2}{4}\Big).$$

For the remainder of the proof, we condition on the event $\mathcal{A}(\delta_n)$, in which case equation (29) simplifies to

$$\frac{1}{2}\|\widehat{\Delta}\|_n^2 \leq \Big|\frac{1}{n}\sum_{i=1}^n w_i \widehat{\Delta}(x_i)\Big| + \sqrt{B}\delta_n \|\widehat{\Delta}\|_n. \tag{49}$$

Here we have used the fact that $\big|\frac{1}{n}\sum_{i=1}^n \widehat{\Delta}(x_i)\big| \leq \|\widehat{\Delta}\|_n$, by the Cauchy-Schwartz inequality.

Now we control the Gaussian complexity term $\big|\frac{1}{n}\sum_{i=1}^n w_i\widehat{\Delta}(x_i)\big|$. For any fixed subset S, define the random variable

$$\widehat{Z}_n(w,t;\mathcal{H}(S,2B)) := \sup_{\substack{\Delta \in \mathcal{H}(S,2B) \\ \|\Delta\|_n \leq t}} \Big|\frac{1}{n}\sum_{i=1}^n w_i \Delta(x_i)\Big|. \tag{50}$$

We first bound this random variable for a fixed subset S of size $2s$, and then take the union bound over all $\binom{d}{2s}$ possible subsets.

Lemma 5. *Assume that the RKHS $\mathcal{H}$ has eigenvalues $(\mu_k)_{k=1}^{\infty}$ that satisfy $\mu_k \simeq k^{-2\alpha}$ and eigenfunctions such that $\|\phi_k\|_\infty \leq C$. Then we have*

$$\mathbb{P}\big[\exists t > 0 \text{ such that } \widehat{Z}_n(w, t; \mathcal{H}(S, 2B)) \geq 16B\, C \sqrt{\frac{s^{1/\alpha}\log s}{n}} + 3t\delta_n\big] \leq c_1 \exp(-9n\delta_n^2). \tag{51}$$

The proof of Lemma 5 is provided Appendix F.1. Returning to inequality (49), we note that by definition,

$$\frac{2}{n}|\sum_{i=1}^{n} w_i \widehat{\Delta}(x_i)| \leq \max_{|S|=2s} \widehat{Z}_n(w, \|\widehat{\Delta}\|_n; \mathcal{H}(S, 2B)).$$

Lemma 5 combined with the union bound implies that

$$\max_{|S|=2s} \widehat{Z}_n(w, \|\widehat{\Delta}\|_n; \mathcal{H}(S, 2B)) \leq 16B\, C \sqrt{\frac{s^{1/\alpha}\log s}{n}} + 3\delta_n \|\widehat{\Delta}\|_n$$

with probability at least $1 - c_1 \binom{d}{2s} \exp(-3n\delta_n^2)$. Our choice (25) of δ_n ensures that this probability is at least $1 - c_1 \exp(-c_2 n\delta_n^2)$. Combined with the basic inequality (49), we conclude that

$$\|\widehat{\Delta}\|_n^2 \leq 32B\, C \sqrt{\frac{s^{1/\alpha}\log s}{n}} + 7B\delta_n \,\|\widehat{\Delta}\|_n \tag{52}$$

with probability $1 - c_1 \exp(-c_2 n\delta_n^2)$.

By definition (25) of δ_n, the bound (52) implies that $\|\widehat{\Delta}\|_n = \mathcal{O}(\delta_n)$ with high probability. In order to translate this claim into a bound on $\|\widehat{\Delta}\|_2$, we require the following result:

Lemma 6. *There exist universal constants (c, c_1, c_2) such that for all $t \geq c\delta_n$, we have*

$$\frac{\|g\|_2}{2} \;\leq\; \|g\|_n \;\leq\; \frac{3}{2}\|g\|_2 \qquad \text{for all } g \in \mathcal{H}(S, 2B) \text{ with } \|g\|_2 \geq t \tag{53}$$

with probability at least $1 - c_1 \exp(-c_2 n t^2)$.

Proof. The bound (53) follows by applying Lemma 7 in Appendix A with $\mathcal{G} = \mathcal{H}(S, 2B)$ and $b = 2B$. The critical radius from equation (55) needs to satisfy the relation $\mathcal{Q}_{w,n}(\epsilon_n; \mathcal{H}(S, 2B)) \leq \frac{\epsilon_n^2}{40}$. From Lemma 11, the choice $\epsilon_n^2 = 320B\, C\sqrt{\frac{s^{1/\alpha}\log s}{n}}$ satisfies this relation. By definition (25) of δ_n, we have $\delta_n \geq c\epsilon_n$ for some universal constant c, which completes the proof. $\square$

This lemma implies that with probability at least $1 - c_1 \exp(-c_2 B n\delta_n^2)$, we have $\|\widehat{\Delta}\|_2 \leq 2\|\widehat{\Delta}\|_n + C\delta_n$. Combined with our earlier upper bound on $\|\widehat{\Delta}\|_n$, this completes the proof of Theorem 3.

5 Discussion

In this paper, we have studied estimation in the class of sparse additive models defined by univariate reproducing kernel Hilbert spaces. In conjunction, Theorems 1 and 2 provide a precise characterization of the minimax-optimal rates for estimating f^* in the $L^2(\mathbb{P})$-norm for various kernel classes with bounded univariate functions. These classes include finite-rank kernels (with logarithmic metric entropy), as well as kernels with polynomially decaying eigenvalues (and hence polynomial metric entropy). In order to establish achievable rates, we analyzed a simple M-estimator based on regularizing the least-squares loss with two kinds of ℓ_1-based norms, one defined by the univariate Hilbert norm and the other by the univariate empirical norm. On the other hand, we obtained our lower bounds by a combination of approximation-theoretic and information-theoretic techniques.

An important feature of our analysis is we assume only that each univariate function is bounded, but do not assume that the multivariate function class is bounded. As discussed in Section 3.5, imposing a global boundedness condition in the high-dimensional setting could lead to a substantially smaller function classes; for instance, for Sobolev classes and sparsity $s = \Omega(\sqrt{n})$, Theorem 3 shows that it is possible to obtain much faster rates than the optimal rates for the class of sparse additive models with univariate functions bounded. Theorem 3 in our paper shows that the rates obtained under global boundedness conditions are not minimax optimal for Sobolev spaces in the regime $s = \Omega(\sqrt{n})$.

There are a number of ways in which this work could be extended. For instance, although our analysis was based on assuming independence of the covariates x_j, $j = 1, 2, \ldots d$, it would be interesting to investigate the case when the random variables are endowed with some correlation structure. One might expect some changes in the optimal rates, particularly if many of the variables are strongly dependent. This work considered only the function class consisting of sums of univariate functions; a natural extension would be to consider nested non-parametric classes formed of sums over hierarchies of subsets of variables. Analysis in this case would require dealing with dependencies between the different functions and is left for future research.

Acknowledgements

This work was partially supported by NSF grants DMS-0605165 and DMS-0907632 to MJW and BY. In addition, BY was partially supported by the NSF grant SES-0835531 (CDI), the SRO grant (INSERT NUMBER) and the Purdue grant (INSERT NUMBER). MJW was also partially supported AFOSR Grant FA9550-09-1-0466. During this work, GR was financially supported by a Berkeley Graduate Fellowship.

A A general result on equivalence of $L^2(\mathbb{P})$ and $L^2(\mathbb{P}_n)$ norms

Since it is required in a number of our proofs, we begin by stating and proving a general result that provides uniform control on the difference between the empirical $\|\cdot\|_n$ and population $\|\cdot\|_2$ norms over a uniformly bounded function class $\mathcal{G}$. We impose two conditions on this class:

(a) it is uniformly bounded, meaning that there is some $b \geq 1$ such that $\|g\|_\infty \leq b$ for all $g \in \mathcal{G}$.

(b) it is star-shaped, meaning that if $g \in \mathcal{G}$, then $\lambda g \in \mathcal{G}$ for all $\lambda \in [0, 1]$.

For each co-ordinate, the Hilbert ball $\mathbb{B}_{\mathcal{H}}(2)$ satisfies both of these conditions; we use $\mathcal{G} = \mathbb{B}_{\mathcal{H}}(2)$. (To be clear, we cannot apply this result to the multivariate function class $\mathcal{F}_{d,s,\mathcal{H}}$, since it is not uniformly bounded.)

Let $\{\sigma_i\}_{i=1}^n$ be an i.i.d. sequence of Rademacher variables, and let $\{x_i\}_{i=1}^n$ be an i.i.d. sequence of variables from $\mathcal{X}$, drawn according to some distribution $\mathbb{Q}$. For each $t > 0$, we define the local

Rademacher complexity

$$\mathcal{Q}_{\sigma,n}(t,\mathcal{G}) \; := \mathbb{E}_{x,\sigma}\big[\sup_{\substack{\|g\|_2\le t\\ g\in\mathcal{G}}} \frac{1}{n}\sum_{i=1}^n \sigma_i g(x_i)\big] \tag{54}$$

We let ϵ_n denote the smallest solution (of size at least $1/\sqrt{n}$) to the inequality

$$\mathcal{Q}_{\sigma,n}(\epsilon_n,\mathcal{G}) = \frac{\epsilon_n^2}{40}, \tag{55}$$

where our scaling by the constant 40 is for later theoretical convenience. Such an ϵ_n exists, because the star-shaped property implies that the function $\mathcal{Q}_{\sigma,n}(t,\mathcal{G})/t$ is non-increasing in t. This quantity corresponds to the critical rate associated with the population Rademacher complexity. For any $t \ge \epsilon_n$, we define the event $\mathcal{E}(t) := \big\{ \sup_{\substack{g\in\mathcal{G}\\ \|g\|_2\le t}} \big|\|g\|_n - \|g\|_2\big| \ge \frac{bt}{2}\big\}$.

Lemma 7. *Suppose that $\|g\|_\infty \le b$ for all $g\in\mathcal{G}$. Then there exist universal constants (c_1,c_2) such that for any $t\ge\epsilon_n$,*

$$\mathbb{P}\big[\mathcal{E}(t)\big] \le c_1\exp(-c_2 n t^2). \tag{56}$$

In addition, for any $g\in\mathcal{G}$ with $\|g\|_2\ge t$, we have $\|g\|_n \;\le\; \|g\|_2(1+\frac{b}{2})$, and moreover, for all $g\in\mathcal{G}$ with $\|g\|_2 \ge bt$, we have

$$\frac{1}{2}\|g\|_2 \;\le\; \|g\|_n \;\le\; \frac{3}{2}\|g\|_2, \tag{57}$$

both with probability at least $1-c_1\exp(-c_2nt^2)$.

Lemma 7 follows from a relatively straightforward adaptation of known results (e.g., Lemma 5.16 in van de Geer [35] and Theorem 2.1 in Bartlett et al. [4]), so we omit the proof details here.

B Proof of Lemma 1

The proof of this lemma is based on peeling and weighting techniques from empirical theory [1, 35] combined with results on the local Rademacher and Gaussian complexities [4, 25]. For each univariate Hilbert space $\mathcal{H}_j = \mathcal{H}$, let us introduce the random variables

$$\widehat{Z}_n(w,t;\mathcal{H}) := \sup_{\substack{\|g_j\|_{\mathcal{H}}\le 1\\ \|g_j\|_n\le t}} \big|\frac{1}{n}\sum_{i=1}^n w_i g_j(x_{ij})\big|, \quad \text{and} \quad Z_n(w,t;\mathcal{H}) := \mathbb{E}_x\Big[\sup_{\substack{\|g_j\|_{\mathcal{H}}\le 1\\ \|g_j\|_2\le t}} \big|\frac{1}{n}\sum_{i=1}^n w_i g_j(x_{ij})\big|\Big], \tag{58}$$

where $w_i \sim N(0,1)$ are i.i.d. standard normal. The empirical and population Gaussian complexities are given by

$$\widehat{\mathcal{Q}}_{w,n}(t,\mathcal{H}) \; := \mathbb{E}_w\big[\widehat{Z}_n(w;t,\mathcal{H})\big] \quad \text{and} \quad \mathcal{Q}_{w,n}(t,\mathcal{H}) \; := \mathbb{E}_w\big[Z_n(w;t,\mathcal{H})\big]. \tag{59}$$

For future reference, we note that in the case of a univariate Hilbert space $\mathcal{H}$ with eigenvalues $\{\mu_k\}_{k=1}^\infty$, results in Mendelson [25] imply that there are universal constants $c_\ell \le c_u$ such that for all $t^2 \ge 1/n$, we have

$$\frac{c_\ell}{\sqrt{n}}\big[\sum_{k=1}^\infty \min\{t^2,\mu_k\}\big]^{1/2} \;\le\; \mathcal{Q}_{w,n}(t,\mathcal{H}) \;\le\; \frac{c_u}{\sqrt{n}}\big[\sum_{k=1}^\infty \min\{t^2,\mu_k\}\big]^{1/2}, \tag{60}$$

for all j The same bounds also hold for the local Rademacher complexities for Reproducing kernel Hilbert spaces.

Let $\widehat{\nu}_{n,j} > 0$ denote the smallest positive solution r of the inequality

$$\widehat{\mathcal{Q}}_{w,n}(r, \mathcal{H}) \leq 4\, r^2. \tag{61}$$

The function $\widehat{\mathcal{Q}}_{w,n}(r, \mathcal{H})$ defines the local Gaussian complexity of the kernel class in co-ordinate j. Recall the bounds (60) that apply to both the empirical and population Gaussian complexities. Recall that the critical univariate rate ν_n is defined in terms of the population Gaussian complexity (see equation (13)).

B.1 Some auxiliary results

In order to prove Lemma 1, we also need some auxiliary results, stated below as Lemmas 8 and 9.

Lemma 8. *For any function class $\mathcal{G}$ and all $\delta \geq 0$, we have*

$$\mathbb{P}\big[|\widehat{Z}_n(w, t, \mathcal{G}) - \widehat{\mathcal{Q}}_{w,n}(t, \mathcal{G})| \geq \delta t\big] \leq 2 \exp\big(-\frac{n\delta^2}{2}\big), \quad \textit{and} \tag{62a}$$

$$\mathbb{P}\big[|Z_n(w, t, \mathcal{G}) - \mathcal{Q}_{w,n}(t, \mathcal{G})| \geq \delta t\big] \leq 2 \exp\Big(-\frac{n\delta^2}{2}\Big). \tag{62b}$$

Proof. We have

$$|\widehat{Z}_n(w, t, \mathcal{G}) - \widehat{Z}_n(w', t, \mathcal{G})| \;\leq\; \sup_{\substack{g \in \mathcal{G} \\ \|g\|_n \leq t}} \frac{1}{n} \Big|\sum_{i=1}^n (w_i - w_i') g(x_i)\Big| \;\leq\; \frac{t}{\sqrt{n}} \|w - w'\|_2,$$

showing that $\widehat{Z}_n(w, t, \mathcal{G})$ is $\frac{t}{\sqrt{n}}$-Lipschitz with respect to the ℓ_2 norm. Consequently, concentration for Lipschitz functions of Gaussian random variables [21] yields the tail bound (62a). Turning to the quantity $Z_n(w, t, \mathcal{H})$, a similar argument yields that

$$\begin{aligned}
|Z_n(w, t, \mathcal{G}) - Z_n(w', t, \mathcal{G})| &\leq \mathbb{E}_x\Big[\sup_{\substack{g \in \mathcal{G} \\ \|g\|_2 \leq t}} \frac{1}{n} \Big|\sum_{i=1}^n (w_i - w_i') g(x_i)\Big|\Big] \\
&\leq \sup_{\substack{g \in \mathcal{G} \\ \|g\|_2 \leq t}} \mathbb{E}_x\Big[\Big(\frac{1}{n}\sum_{i=1}^n g^2(x_i)\Big)^{1/2}\Big] \, \|w - w'\|_2 \;\leq\; \frac{t}{\sqrt{n}} \, \|w - w'\|_2,
\end{aligned}$$

where the final step uses Jensen's inequality and the fact that $\mathbb{E}_x[g^2(x_i)] \leq t^2$ for all $i = 1, \ldots, n$. The same reasoning then yields the tail bound (62b). ☐

Our second lemma involves the event $\mathcal{D}(\gamma_n) := \big\{\widehat{\nu}_{n,j} \leq \gamma_n, \quad \text{for all } j = 1, 2, \ldots, d\big\}$, where we recall the definition (61) of $\widehat{\nu}_{n,j}$, and that $\gamma_n := \kappa \max\big\{\nu_n, \sqrt{\frac{\log d}{n}}\big\}$.

Lemma 9. *For all $1 \leq j \leq d$, we have*

$$\mathbb{P}\big[\widehat{\nu}_{n,j} \leq \gamma_n\big] \geq 1 - c_1 \exp(-c_2 n \gamma_n^2). \tag{63}$$

Proof. We first bound the probability of the event $\{\widehat{\nu}_{n,j} > \gamma_n\}$ for a fixed $\mathcal{H}_j$. Let $g \in \mathbb{B}_{\mathcal{H}_j}(1)$ be any function such that $\|g\|_2 > t \geq \nu_n$. Then conditioned on the sandwich relation (57) with $b = 1$, we are guaranteed that $\|g\|_n > \frac{t}{2}$. Taking the contrapositive, we conclude that $\|g\|_n \leq \frac{t}{2}$ implies $\|g\|_2 \leq t$, and hence that $\widehat{Z}_n(w, t/2, \mathcal{H}) \leq Z_n(w, t, \mathcal{H})$ for all $t \geq \nu_n$, under the stated conditioning.

For any $t \geq \nu_n$, the inequalities (57), (62a) and (62b) hold with probability at least $1 - c_1 \exp(-c_2 n t^2)$. Conditioning on these inequalities, we can set $t = \gamma_n > \nu_n$, and thereby obtain

$$
\begin{aligned}
\widehat{\mathcal{Q}}_{w,n}(\gamma_n, \mathcal{H}) &\overset{(a)}{\leq} \widehat{Z}_n(w, \gamma_n, \mathcal{H}) + \gamma_n^2 \\
&\overset{(b)}{\leq} Z_n(w, 2\gamma_n, \mathcal{H}) + \gamma_n^2 \\
&\overset{(c)}{\leq} \mathcal{Q}_{w,n}(2\gamma_n, \mathcal{H}) + 2\gamma_n^2 \\
&\overset{(d)}{\leq} 4\gamma_n^2,
\end{aligned}
$$

where inequality (a) follows from the bound (62a), inequality (b) follows the initial argument, inequality (c) follows from the bound (62b), and inequality (d) follows since $2\gamma_n > \epsilon_n$ and the definition of ϵ_n.

By the definition of $\widehat{\nu}_{n,j}$ as the minimal t such that $\widehat{\mathcal{Q}}_{w,n}(t, \mathcal{H}) \leq 4t^2$, we conclude that for each fixed $j = 1, \ldots, n$, we have $\widehat{\nu}_{n,j} \leq \gamma_n$ with probability at least $1 - c_1 \exp(-c_2 n \gamma_n^2)$. Finally, the uniformity over $j = 1, 2, \ldots, d$ follows from the union bound and our choice of $\gamma_n \geq \kappa \sqrt{\frac{\log d}{n}}$. $\square$

B.2 Main argument to prove Lemma 1

We can now proceed with the proof of Lemma 1. Combining Lemma 9 with the union bound over $j = 1, 2, \ldots, d$, we conclude that that

$$
\mathbb{P}[\mathcal{D}(\gamma_n)] \geq 1 - c_1 \exp(-c_2 n \gamma_n^2),
$$

as long as $c_2 \geq 1$. For the remainder of our proofs, we condition on the event $\mathcal{D}(\gamma_n)$. In particular, our goal is to prove that

$$
\Big|\frac{1}{n} \sum_{i=1}^{n} w_i f_j(x_{ij})\Big| \leq C \,\{\gamma_n^2 \,\|f_j\|_{\mathcal{H}} + \gamma_n \,\|f_j\|_n\} \qquad \text{for all } f_j \in \mathcal{H} \tag{64}
$$

with probability greater than $1 - c_1 \exp(-c_2 n \gamma_n^2)$. By combining this result with our choice of γ_n and the union bound, the claimed bound then follows on $\mathbb{P}[\mathcal{T}(\gamma_n)]$.

If $f_j = 0$, then the claim (64) is trivial. Otherwise we renormalize f_j by defining $g_j := f_j / \|f_j\|_{\mathcal{H}}$, and we write

$$
\frac{1}{n} \sum_{i=1}^{n} w_i f_j(x_{ij}) = \|f_j\|_{\mathcal{H}} \, \frac{1}{n} \sum_{i=1}^{n} w_i g_j(x_{ij}) \;\leq\; \|f_j\|_{\mathcal{H}} \, \widehat{Z}_n\big(w; \|g_j\|_n, \mathcal{H}\big),
$$

where the final inequality uses the definition (58), and the fact that $\|g_j\|_{\mathcal{H}} = 1$. We now split the analysis into two cases: (1) $\|g_j\|_n \leq \gamma_n$, and (2) $\|g_j\|_n > \gamma_n$.

Case 1: $\|g_j\|_n \le \gamma_n$**.** In this case, it suffices to upper bound the quantity $\widehat{Z}_n(w; \gamma_n, \mathcal{H})$. Note that $\|g_j\|_{\mathcal{H}} = 1$ and recall definition (58) of the random variable $\widehat{Z}_n$. On one hand, since $\gamma_n \ge \widehat{\nu}_{n,j}$ by Lemma 9, the definition of $\widehat{\nu}_{n,j}$ implies that $\widehat{\mathcal{Q}}_{w,n}(\gamma_n, \mathcal{H}) \le 4\gamma_n^2$, and hence

$$\mathbb{E}[\widehat{Z}_n(w; \gamma_n; \mathcal{H})] = \widehat{\mathcal{Q}}_{w,n}(\gamma_n; \mathcal{H}) \le 4\gamma_n^2.$$

Applying the bound (62a) from Lemma 8 with $\delta = \gamma_n = t$, we conclude that $\widehat{Z}_n(w; \gamma_n; \mathcal{H}) \le C\,\gamma_n^2$ with probability at least $1 - c_1 \exp\big\{-c_2 n\gamma_n^2\big\}$, which completes the proof in the case where $\|g\|_n \le \gamma_n$.

Case 2: $\|g_j\|_n > \gamma_n$**.** In this case, we study the random variable $\widehat{Z}_n(w; r_j; \mathcal{H})$ for some $r_j > \gamma_n$. Our intermediate goal is to prove the bound

$$\mathbb{P}\Big[\widehat{Z}_n(w; r_j; \mathcal{H}) \ge C\, r_j\, \gamma_n\Big] \le c_1 \exp\big\{-c_2 n\gamma_n^2\big\}. \tag{65}$$

Applying the bound (62a) with $t = r_j$ and $\delta = \gamma_n$, we are guaranteed an upper bound of the form $\widehat{Z}_n(w; r_j; \mathcal{H}) \le \widehat{\mathcal{Q}}_{w,n}(r_j, \mathcal{H}) + r_j\,\gamma_n$ with probability at least $1 - c_1 \exp\big(-c_2 n\gamma_n^2\big)$. In order to complete the proof, we need to show that $\widehat{\mathcal{Q}}_{w,n}(r_j, \mathcal{H}) \le r_j\,\gamma_n$. Since $r_j > \gamma_n > \widehat{\nu}_{n,j}$, we have

$$\widehat{\mathcal{Q}}_{w,n}(r_j, \mathcal{H}) = \frac{r_j}{\widehat{\nu}_{n,j}} \mathbb{E}_w\big[\sup_{\substack{\|g_j\|_n \le \widehat{\nu}_{n,j} \\ \|g_j\|_{\mathcal{H}} \le \frac{\widehat{\nu}_{n,j}}{r_j}}} \big|\frac{1}{n}\sum_{i=1}^n w_i g_j(x_{ij})\big|\big] \;\le\; \frac{r_j}{\widehat{\nu}_{n,j}} \widehat{\mathcal{Q}}_{w,n}(\widehat{\nu}_{n,j}, \mathcal{H}) \;\le 4\, r_j \widehat{\nu}_{n,j},$$

where the final inequality uses the fact that $\widehat{\mathcal{Q}}_{w,n}(\widehat{\nu}_{n,j}, \mathcal{H}) \le 4\,\widehat{\nu}_{n,j}^2$. On the event $\mathcal{D}(\gamma_n)$ from Lemma 9, we have $\widehat{\nu}_{n,j} \le \gamma_n$, from which the claim (65) follows.

We now use the bound (65) to prove the bound (64), in particular via a "peeling" operation over all choices of $r_j = \|f_j\|_n / \|f_j\|_{\mathcal{H}}$. (See van de Geer [35] for more details on these peeling arguments.) We claim that it suffices to consider $r_j \le 1$. It is equivalent to show that $\|g_j\|_n \le 1$ for any $g_j \in \mathbb{B}_{\mathcal{H}}(1)$. Since $\|g_j\|_\infty \le \|g_j\|_{\mathcal{H}} \le 1$, we have $\|g_j\|_n^2 = \frac{1}{n}\sum_{i=1}^n g_j^2(x_{ij}) \le 1$, as required. Now define the event

$$\mathcal{T}_j(\gamma_n) := \Big\{\exists f_j \in \mathbb{B}_{\mathcal{H}}(1) \;\mid\; \big|\frac{1}{n}\sum_{i=1}^n w_i f_j(x_{ij})\big| > 8\,\|f_j\|_{\mathcal{H}}\,\gamma_n\,\frac{\|f_j\|_n}{\|f_j\|_{\mathcal{H}}},\ \text{and}\ \frac{\|f_j\|_n}{\|f_j\|_{\mathcal{H}}} \in (\gamma_n, 1]\Big\}.$$

and the sets $S_m := \big\{2^{m-1}\gamma_n \le \frac{\|f_j\|_n}{\|f_j\|_{\mathcal{H}}} \le 2^m \gamma_n\big\}$ for $m = 1, 2, \ldots, M$. By choosing $M = 2\log_2(1/\gamma_n)$, we ensure that $2^M \gamma_n \ge 1$, and hence that if the event $\mathcal{T}_j(\gamma_n)$ occurs, then it must occur for function f_j belonging to some S_m, so that we have a function f_j such that $\frac{\|f_j\|_n}{\|f_j\|_{\mathcal{H}}} \le t_m := 2^m \gamma_n$, and

$$\big|\frac{1}{n}\sum_{i=1}^n w_i f_j(x_{ij})\big| > \;8\,\|f_j\|_{\mathcal{H}}\,\gamma_n\,\frac{\|f_j\|_n}{\|f_j\|_{\mathcal{H}}}\; \ge C\,\|f_j\|_{\mathcal{H}}\,t_m,$$

which implies that $\widehat{Z}_n(w; t_m, \mathcal{H}) \ge 4t_m$. Consequently, by union bound and the tail bound (65), we have

$$\mathbb{P}[\mathcal{T}_j(\gamma_n)] \le M\; c_1 \exp\big\{-c_2 n\gamma_n^2\big\} \;\le\; c_1 \exp\big\{-c_2' n\gamma_n^2\big\}$$

by the condition $n\gamma_n^2 = \Omega(\log(1/\gamma_n))$, which completes the proof.

C Proof of Lemma 2

Define the function

$$\widetilde{\mathcal{L}}(\Delta) := \frac{1}{2n}\sum_{i=1}^{n}\big(w_i + \mu + \bar{y}_n - \Delta(x_i)\big)^2 + \lambda_n \|f^* + \Delta\|_{n,1} + \rho_n \|f^* + \Delta\|_{\mathcal{H},1}$$

and note that by definition of our M-estimator, the error function $\widehat{\Delta} := \widehat{f} - f^*$ minimizes $\widetilde{\mathcal{L}}$. From the inequality $\widetilde{\mathcal{L}}(\widehat{\Delta}) \leq \widetilde{\mathcal{L}}(0)$, we obtain the upper bound $\frac{1}{2}\|\widehat{\Delta}\|_n^2 \leq T_1 + T_2$, where

$$\begin{aligned}
T_1 &:= \Big|\frac{1}{n}\sum_{i=1}^{n} w_i \widehat{\Delta}(x_i)\Big| + |\bar{y}_n - \mu|\Big|\frac{1}{n}\sum_{i=1}^{n}\widehat{\Delta}(x_i)\Big|, \quad \text{and} \\
T_2 &:= \lambda_n \sum_{j=1}^{d}\big\{\|f_j^*\|_n - \|f_j^* + \widehat{\Delta}_j\|_n\big\} + \rho_n \sum_{j=1}^{d}\big\{\|f_j^*\|_{\mathcal{H}} - \|f_j^* + \widehat{\Delta}_j\|_{\mathcal{H}}\big\}.
\end{aligned}$$

Conditioned on the event $\mathcal{C}(\gamma_n)$, we have the bound $|\bar{y}_n - \mu|\big|\frac{1}{n}\sum_{i=1}^{n}\widehat{\Delta}(x_i)\big| \leq \sqrt{s}\gamma_n\|\widehat{\Delta}\|_n$, and hence $\frac{1}{2}\|\widehat{\Delta}\|_n^2 \leq T_2 + \big|\frac{1}{n}\sum_{i=1}^{n} w_i\widehat{\Delta}(x_i)\big| + \sqrt{s}\gamma_n\|\widehat{\Delta}\|_n$, or equivalently

$$0 \;\leq\; \frac{1}{2}\big(\|\widehat{\Delta}\|_n - \sqrt{s}\gamma_n\big)^2 \leq T_2 + \Big|\frac{1}{n}\sum_{i=1}^{n} w_i\widehat{\Delta}(x_i)\Big| + \frac{1}{2}s\gamma_n^2. \tag{66}$$

It remains to control the term T_2. On one hand, for any $j \in S^c$, we have

$$\|f_j^*\|_n - \|f_j^* + \widehat{\Delta}_j\|_n \;=\; -\|\widehat{\Delta}_j\|_n, \quad \text{and} \quad \|f_j^*\|_{\mathcal{H}} - \|f_j^* + \widehat{\Delta}_j\|_{\mathcal{H}} \;=\; -\|\widehat{\Delta}_j\|_{\mathcal{H}}.$$

On the other hand, for any $j \in S$, the triangle inequality yields $\|f_j^*\|_n - \|f_j^* + \widehat{\Delta}_j\|_n \leq \|\widehat{\Delta}_j\|_n$, with a similar inequality for the terms involving $\|\cdot\|_{\mathcal{H}}$. Combined with the bound (66), we conclude that

$$0 \leq \frac{1}{n}\sum_{i=1}^{n} w_i\widehat{\Delta}(x_i) + \lambda_n\big\{\|\widehat{\Delta}_S\|_{n,1} - \|\widehat{\Delta}_{S^c}\|_{n,1}\big\} + \rho_n\big\{\|\widehat{\Delta}_S\|_{\mathcal{H},1} - \|\widehat{\Delta}_{S^c}\|_{\mathcal{H},1}\big\} + \frac{1}{2}s\gamma_n^2. \tag{67}$$

Recalling our conditioning on the event $\mathcal{T}(\gamma_n)$, by Lemma 1, we have the upper bound

$$\Big|\frac{1}{n}\sum_{i=1}^{n} w_i\widehat{\Delta}(x_i)| \leq 8\,\big\{\gamma_n\|\widehat{\Delta}\|_{n,1} + \gamma_n^2\|\widehat{\Delta}\|_{\mathcal{H},1}\big\}.$$

Combining with the inequality (67) yields

$$\begin{aligned}
0 &\leq 8\,\big\{\gamma_n\|\widehat{\Delta}\|_{n,1} + \gamma_n^2\|\widehat{\Delta}\|_{\mathcal{H},1}\big\} + \lambda_n\big\{\|\widehat{\Delta}_S\|_{n,1} - \|\widehat{\Delta}_{S^c}\|_{n,1}\big\} + \rho_n\big\{\|\widehat{\Delta}_S\|_{\mathcal{H},1} - \|\widehat{\Delta}_{S^c}\|_{\mathcal{H},1}\big\} + \frac{1}{2}s\gamma_n^2 \\
&\leq \frac{\lambda_n}{2}\|\widehat{\Delta}\|_{n,1} + \frac{\rho_n}{2}\|\widehat{\Delta}\|_{\mathcal{H},1} + \lambda_n\big\{\|\widehat{\Delta}_S\|_{n,1} - \|\widehat{\Delta}_{S^c}\|_{n,1}\big\} + \rho_n\big\{\|\widehat{\Delta}_S\|_{\mathcal{H},1} - \|\widehat{\Delta}_{S^c}\|_{\mathcal{H},1}\big\} + \frac{1}{2}s\gamma_n^2,
\end{aligned}$$

where we have recalled our choices of (λ_n, ρ_n). Finally, re-arranging terms yields the claim (33).

D Proof of Lemma 3

Recalling the definition (40) of the function class $\mathcal{G}(\lambda_n, \rho_n)$ and the critical radius $\tilde{\delta}_n$ from equation (38), we define the function class $\mathcal{G}'(\lambda_n, \rho_n, \tilde{\delta}_n) := \{h \in \mathcal{G}(\lambda_n, \rho_n) \mid \|h\|_2 = \tilde{\delta}_n\}$, and the alternative event

$$\mathcal{B}'(\lambda_n, \rho_n) := \big\{\{\|h\|_n^2 \geq \tilde{\delta}_n^2/2 \quad \text{for all } h \in \mathcal{G}'(\lambda_n, \rho_n, \tilde{\delta}_n)\big\}.$$

We claim that it suffices to show that $\mathcal{B}'(\lambda_n, \rho_n)$ holds with probability at least $1 - c_1 \exp(-c_2 n \gamma_n^2)$. Indeed, given an arbitrary non-zero function $g \in \mathcal{G}(\lambda_n, \rho_n)$, consider the rescaled function $h = \frac{\tilde{\delta}_n}{\|g\|_2} g$. Since $g \in \mathcal{G}(\lambda_n, \rho_n)$ and $\mathcal{G}(\lambda_n, \rho_n)$ is star-shaped, we have $h \in \mathcal{G}(\lambda_n, \rho_n)$, and also $\|h\|_2 = \tilde{\delta}_n$ by construction. Consequently, when the event $\mathcal{B}'(\lambda_n, \rho_n)$ holds, we have $\|h\|_n^2 \geq \tilde{\delta}_n^2/2$, or equivalently $\|g\|_n^2 \geq \|g\|_2^2/2$, showing that $\mathcal{B}(\lambda_n, \rho_n)$ holds. Accordingly, the remainder of the proof is devoted to showing that $\mathcal{B}'(\lambda_n, \rho_n)$ holds with probability greater than $1 - c_1 \exp(-c_2 n \gamma_n^2)$. Alternatively, if we define the random variable $Z_n(\mathcal{G}') := \sup_{f \in \mathcal{G}'} \big\{\tilde{\delta}_n^2 - \frac{1}{n}\sum_{i=1}^n f^2(x_i)\big\}$, then it suffices to show that $Z_n(\mathcal{G}') \leq \tilde{\delta}_n^2/2$ with high probability.

Recall from Section 4.2.4 the definition of a covering set; here we use the notion of a proper covering, which restricts the covering to use only members of the set $\mathcal{G}$. Letting $N_{\text{pr}}(\epsilon; \mathcal{G}, \rho)$ denote the propert covering number, it can be shown that $N_{\text{pr}}(\epsilon; \mathcal{G}, \rho) \leq N(\epsilon; \mathcal{G}, \rho) \leq N_{\text{pr}}(\epsilon/2; \mathcal{G}, \rho)$. Now let $g^1, \ldots, g^N$ be a minimal $\tilde{\delta}_n/8$-proper covering of $\mathcal{G}'$ in the $L^2(\mathbb{P}_n)$-norm, so that for all $f \in \mathcal{G}'$, there exists $g = g^k \in \mathcal{G}'$ such that $\|f - g\|_n \leq \tilde{\delta}_n/8$. We can then write

$$\tilde{\delta}_n^2 - \frac{1}{n}\sum_{i=1}^n f^2(x_i) = \big\{\tilde{\delta}_n^2 - \frac{1}{n}\sum_{i=1}^n g^2(x_i)\big\} + \big\{\frac{1}{n}\sum_{i=1}^n (g^2(x_i) - f^2(x_i))\big\}.$$

By the Cauchy-Schwartz inequality, we have

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n (g^2(x_i) - f^2(x_i)) &= \frac{1}{n}\sum_{i=1}^n (g(x_i) - f(x_i))(g(x_i) + f(x_i)) \\
&\leq \sqrt{\frac{1}{n}\sum_{i=1}^n (g(x_i) - f(x_i))^2}\sqrt{\frac{1}{n}\sum_{i=1}^n (f(x_i) + g(x_i))^2} \\
&= \|g - f\|_n \sqrt{\frac{1}{n}\sum_{i=1}^n (f(x_i) + g(x_i))^2}.
\end{aligned}$$

By our choice of the covering, we have $\|g - f\|_n \leq \tilde{\delta}_n/8$. On the other hand, we have

$$\sqrt{\frac{1}{n}\sum_{i=1}^n (f(x_i) + g(x_i))^2} \leq \sqrt{2\|f\|_n^2 + 2\|g\|_n^2} \;\leq\; \sqrt{4\tilde{\delta}_n^2} \;=\; 2\tilde{\delta}_n,$$

where the final inequality follows since $\|f\|_n = \|g\|_n = \tilde{\delta}_n$. Overall, we have established the upper bound $\frac{1}{n}\sum_{i=1}^n (g^2(x_i) - f^2(x_i)) \leq \frac{\tilde{\delta}_n^2}{4}$, and hence shown that

$$Z_n(\mathcal{G}') \leq \max_{g^1, g^2, \ldots, g^N} \big\{\tilde{\delta}_n^2 - \frac{1}{n}\sum_{i=1}^n (g^k(x_i))\big\} + \frac{\tilde{\delta}_n^2}{4},$$

where $N = N_{\rm pr}(\tilde{\delta}_n/8, \mathcal{G}', \|\cdot\|_n)$. For any g in our covering set, since $g^2(x_i) \geq 0$, we may apply Theorem 3.5 from Chung and Lu [12] with $t = \tilde{\delta}_n^2/4$ to obtain the one-sided tail bound

$$\mathbb{P}[\tilde{\delta}_n^2 - \frac{1}{n}\sum_{i=1}^n g^2(x_i) \geq \frac{\tilde{\delta}_n^2}{4}] \leq \exp\big(- \frac{n\tilde{\delta}_n^4}{32\mathbb{E}[g^4(x)]}\big), \tag{68}$$

where we used the upper bound $\mathrm{var}(g^2(x)) \leq \mathbb{E}[g^4(x)]$. Next using the fact that the variables $\{g_j(x_j)\}_{j=1}^d$ are independent and zero-mean, we have

$$\begin{aligned}
\mathbb{E}[g^4(x)] &= \sum_{j=1}^d \mathbb{E}[g_j^4(x_j)] + \binom{4}{2}\sum_{j\neq k}\mathbb{E}[[g_j^2(x_j)]\mathbb{E}[g_k^2(x_k)] \\
&\leq 4\sum_{j=1}^d \mathbb{E}[g_j^2(x_j)] + 6\sum_{j=1}^d \mathbb{E}[g_j^2(x_j)]\sum_{k=1}^d \mathbb{E}[g_k^2(x_k)] \\
&\leq 4\tilde{\delta}_n^2 + 6\tilde{\delta}_n^4 \\
&\leq 10\tilde{\delta}_n^2,
\end{aligned}$$

where the second inequality follows since $\|g_j\|_\infty \leq \|g_j\|_{\mathcal{H}} \leq 2$ for each j. Combining this upper bound on $\mathbb{E}[g^4(x)]$ with the earlier tail bound (68) and applying union bound yields

$$\mathbb{P}[\max_{k=1,2,\ldots,N}\big\{\tilde{\delta}_n^2 - \frac{1}{n}\sum_{i=1}^n g^2(x_i)\big\} \geq \frac{\tilde{\delta}_n^2}{4}] \leq \exp\big(\log N_{\rm pr}(\tilde{\delta}_n/8, \mathcal{G}', \|\cdot\|_n) - \frac{n\tilde{\delta}_n^2}{320}\big). \tag{69}$$

It remains to bound the covering entropy $\log N_{\rm pr}(\tilde{\delta}_n/8, \mathcal{G}', \|\cdot\|_n)$. Since the proper covering entropy $\log N_{\rm pr}(\tilde{\delta}_n/8, \mathcal{G}', \|\cdot\|_n)$ is at most $\log N(\tilde{\delta}_n/16, \mathcal{G}', \|\cdot\|_n)$, it suffices to upper bound the usual covering entropy. Viewing the samples $(x_1, x_2, ..., x_n)$ as fixed, let us define the zero-mean Gaussian process $\{W_g, g \in \mathcal{G}'\}$ via $W_g := \frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i g(x_i)$, where the variables $\{\varepsilon_i\}_{i=1}^n$ are i.i.d. standard Gaussian variates. By construction, we have $\mathrm{var}[(W_g - W_f))] = \|g-f\|_n^2$. Consequently, by the Sudakov minoration [28], for all $\epsilon > 0$, we have $\epsilon\sqrt{\log N(\epsilon; \mathcal{G}', \|\cdot\|_n)} \leq 4\mathbb{E}_\varepsilon[\sup_{g\in\mathcal{G}'} W_g]$. Setting $\epsilon = \tilde{\delta}_n/16$ and performing some algebra, we obtain the upper bound

$$\frac{1}{\sqrt{n}}\sqrt{\log N(\tilde{\delta}_n/16; \mathcal{G}', \|\cdot\|_n)} \leq \frac{64}{\tilde{\delta}_n}\mathbb{E}_\varepsilon[\sup_{g\in\mathcal{G}'}\frac{1}{n}\sum_{i=1}^n \varepsilon_i g(x_i)]. \tag{70}$$

The final step is to upper bound the Gaussian complexity $\mathbb{E}_\varepsilon[\sup\limits_{g\in\mathcal{G}'}\frac{1}{n}\sum_{i=1}^n \varepsilon_i g(x_i)]$. In the proof of Lemma 1, we showed that for any co-ordinate $j \in \{1, 2, \ldots, d\}$, the univariate Gaussian complexity is upper bounded as

$$\mathbb{E}\big[\sup_{\substack{\|g_j\|_n \leq r_j \\ \|g_j\|_{\mathcal{H}} \leq R_j}} \frac{1}{n}\sum_{i=1}^n \varepsilon_i g_j(x_{ij})\big] \;\leq\; C\,\big\{\gamma_n\, r_j + \gamma_n^2 R_j\big\}.$$

Summing across co-ordinates and recalling the fact that the constant C may change from line to line, we obtain the upper bound

$$\begin{aligned}
\mathbb{E}_\varepsilon[\sup_{g\in\mathcal{G}'}\frac{1}{n}\sum_{i=1}^n \varepsilon_i g(x_i)] &\le C\sup_{g\in\mathcal{G}'}\{\gamma_n\,\|g\|_{1,n}+\gamma_n^2\|g\|_{1,\mathcal{H}}\}\\
&\stackrel{(a)}{\le} C\sup_{g\in\mathcal{G}'}\{4\gamma_n\,\|g_S\|_{1,n}+4\gamma_n^2\|g_S\|_{1,\mathcal{H}}+\frac{1}{32}s\rho_n\}\\
&\stackrel{(b)}{\le} C\sup_{g\in\mathcal{G}'}\{\gamma_n\,\|g_S\|_{1,n}+s\rho_n\}\\
&\stackrel{(c)}{\le} C\sup_{g\in\mathcal{G}'}\Big\{\gamma_n\,[2\sqrt{s}\|g\|_2+s\gamma_n]+s\rho_n\Big\},
\end{aligned}$$

where step (a) uses inequality (40a) in the definition of $\mathcal{G}'$; step (b) uses the inequality $\|g_j\|_{\mathcal{H}}\le 2$ for each co-ordinate and hence $\|g_S\|_{1,\mathcal{H}}\le 2s$, and our choice of regularization parameter $\rho_n\ge\gamma_n^2$; and step (c) uses inequality (40b) in the definition of $\mathcal{G}'$. Since $\|g\|_2=\tilde{\delta}_n$ for all $g\in\mathcal{G}'$, we have shown that

$$\mathbb{E}_\varepsilon[\sup_{g\in\mathcal{G}'}\frac{1}{n}\sum_{i=1}^n \varepsilon_i g(x_i)] \le C\big\{s\gamma_n^2+\sqrt{s}\gamma_n\tilde{\delta}_n+s\rho_n\big\}\;\stackrel{(d)}{\le}\; C\big\{\frac{\tilde{\delta}_n^2}{B^2}+\frac{\tilde{\delta}_n^2}{B}\big\}, \tag{71}$$

where inequality (d) follows from our choice (38) of $\tilde{\delta}_n$, and the constant B can be chosen as large as we please. In particular, by choosing B sufficiently large, and combining the bound (71) with the Sudakov bound (70), we can ensure that

$$\frac{1}{n}\,\log N(\tilde{\delta}_n/16;\mathcal{G}',\|\cdot\|_n)\le\frac{\tilde{\delta}_n^2}{640}.$$

Combined with the earlier tail bound (69), we conclude that

$$\mathbb{P}[\max_{k=1,2,\ldots,N}\{\tilde{\delta}_n^2-\frac{1}{n}\sum_{i=1}^n g^2(x_i)\}\ge\frac{\tilde{\delta}_n^2}{4}]\le\exp\big(-\frac{n\tilde{\delta}_n^2}{640}\big),$$

which completes the proof of Lemma 3.

E Proof of Lemma 4

Proof of part (a): Let $N=M(\frac{\delta}{\sqrt{s}};\mathbb{B}_{\mathcal{H}}(1),\|\cdot\|_2)-1$, and define $\mathcal{I}=\{0,1,\ldots,N\}$. Using $\|u\|_0=\sum_{j=1}^d\mathbb{I}[u_j\neq 0]$ to denote the number of non-zero components in a vector, consider the set

$$\mathfrak{S}:=\big\{u\in\mathcal{I}^d\;\mid\;\|u\|_0=s\big\}. \tag{72}$$

Note that this set has cardinality $|\mathfrak{S}|=\binom{d}{s}N^s$, since any element is defined by first choosing s co-ordinates are non-zero, and then for each co-ordinate, choosing non-zero entry from a total of N possible symbols.

For each $j = 1, \ldots, d$, let $\{0, f_j^1, f_j^2, \ldots, f_j^N\}$ be a $\delta/\sqrt{s}$-packing of $\mathbb{B}_{\mathcal{H}}(1)$. Based on these packings of the univariate function classes, we can use $\mathfrak{S}$ to index a collection of functions contained inside $\mathcal{F}$. In particular, any $u \in \mathfrak{S}$ uniquely defines a function $g^u = \sum_{j=1}^d g_j^{u_j} \in \mathcal{F}$, with elements

$$g_j^{u_j} = \begin{cases} f_j^{u_j} & \text{if } u_j \neq 0 \\ 0 & \text{otherwise.} \end{cases} \tag{73}$$

Since $\|u\|_0 = s$, we are guaranteed that at most s co-ordinates of g are non-zero, so that $g \in \mathcal{F}$.

Now consider two functions g^u and h^v contained within the class $\{g^u, u \in \mathfrak{S}\}$. By definition, we have

$$\|g^u - h^v\|_2^2 = \sum_{j=1}^d \|f_j^{u_j} - f_j^{v_j}\|_2^2 \;\geq\; \frac{\delta^2}{s} \sum_{j=1}^d \mathbb{I}[u_j \neq v_j], \tag{74}$$

Consequently, it suffices to establish the existence of a "large" subset $\mathcal{A} \subset \mathfrak{S}$ such that the Hamming metric $\rho_H(u,v) := \sum_{j=1}^d \mathbb{I}[u_j \neq v_j]$ is at least $s/2$ for all pairs $u, v \in \mathcal{A}$, in which case we are guaranteed that $\|g - h\|_2^2 \geq \delta^2$. For any $u \in \mathfrak{S}$, we observe that

$$\Big|\big\{v \in \mathfrak{S} \;\mid\; \rho_H(u,v) \leq \frac{s}{2}\big\}\Big| \leq \binom{d}{\frac{s}{2}} \, (N+1)^{\frac{s}{2}}.$$

This bound follows because we simply need to choose a subset of size $s/2$ where u and v agree, and the remaining $s/2$ co-ordinates can be chosen arbitrarily in $(N+1)^{\frac{s}{2}}$ ways. For a given set $\mathcal{A}$, we write $\rho_H(u, \mathcal{A}) \leq \frac{s}{2}$ if there exists some $v \in \mathcal{A}$ such that $\rho_H(u,v) \leq \frac{s}{2}$. Using this notation, we have

$$\Big|\big\{u \in \mathfrak{S} \;\mid\; \rho_H(u,\mathcal{A}) \leq \frac{s}{2}\big\}\Big| \leq |\mathcal{A}| \binom{d}{\frac{s}{2}} \, (N+1)^{\frac{s}{2}} \stackrel{(a)}{<} |\mathfrak{S}|,$$

where inequality (a) follows as long as

$$|\mathcal{A}| \leq N^* \;:= \frac{1}{2} \frac{\binom{d}{s}}{\binom{d}{\frac{s}{2}}} \; \frac{N^s}{(N+1)^{s/2}}.$$

Thus, as long as $|\mathcal{A}| \leq N^*$, there must exist some element $u \in \mathfrak{S}$ such that $\rho_H(u, \mathcal{A}) > \frac{s}{2}$, in which case we can form the augmented set $\mathcal{A} \cup \{u\}$. Iterating this procedure, we can form a set with N^* elements such that $\rho_H(u,v) \geq \frac{s}{2}$ for all $u, v \in \mathcal{A}$.

Finally, we lower bound N^*. We have

$$\begin{aligned} N^* &\stackrel{(i)}{\geq} \frac{1}{2} \big(\frac{d-s}{s/2}\big)^{\frac{s}{2}} \; \frac{(N)^s}{(N+1)^{s/2}} \\ &= \frac{1}{2} \big(\frac{d-s}{s/2}\big)^{\frac{s}{2}} \; N^{s/2} \big(\frac{N}{N+1}\big)^{s/2} \\ &\geq \frac{1}{2} \big(\frac{d-s}{s/2}\big)^{\frac{s}{2}} \; N^{s/2}, \end{aligned}$$

where inequality (i) follows by elementary combinatorics (see Lemma 5 in the paper [30] for details). We conclude that for $s \leq d/4$, we have

$$\log N^* = \Omega\big(s \log \frac{d}{s} + s \log M(\frac{\delta}{\sqrt{s}}; \mathbb{B}_{\mathcal{H}}(1), \|\cdot\|_2)\big),$$

thereby completing the proof of Lemma 4(a).

Proof of part (b): In order to prove part (b), we instead let $N = M(\frac{1}{2}; \mathbb{B}_{\mathcal{H}}(1), \|\cdot\|_2) - 1$, and then follow the same steps. Since $\log N = \Omega(m)$, we have the modified lower bound

$$\log N^* = \Omega\big(s \log \frac{d}{s} + sm\big),$$

Moreover, instead of the lower bound (74), we have

$$\|g^u - h^v\|_2^2 = \sum_{j=1}^{d} \|f_j^{u_j} - f_j^{v_j}\|_2^2 \;\geq\; \frac{1}{4} \sum_{j=1}^{d} \mathbb{I}[u_j \neq v_j] \;\geq\; \frac{s}{8}, \tag{75}$$

using our previous result on the Hamming separation. Furthermore, since $\|f_j\|_2 \leq \|f_j\|_{\mathcal{H}}$ for any univariate function, we have the upper bound

$$\|g^u - h^v\|_2^2 = \sum_{j=1}^{d} \|f_j^{u_j} - f_j^{v_j}\|_2^2 \;\leq\; \sum_{j=1}^{d} \|f_j^{u_j} - f_j^{v_j}\|_{\mathcal{H}}^2.$$

By the definition (72) of $\mathfrak{S}$, at most $2s$ of the terms $f_j^{u_j} - f_j^{v_j}$ can be non-zero. Moreover, by construction we have $\|f_j^{u_j} - f_j^{v_j}\|_{\mathcal{H}} \leq 2$, and hence

$$\|g^u - h^v\|_2^2 \leq 8s.$$

Finally, by rescaling the functions by $\sqrt{8}\,\delta/\sqrt{s}$, we obtain a class of N^* rescaled functions $\{\widetilde{g}^u, u \in \mathcal{I}\}$ such that

$$\|\widetilde{g}^u - \widetilde{h}^v\|_2^2 \geq \delta^2, \quad \text{and} \quad \|\widetilde{g}^u - \widetilde{h}^v\|_2^2 \leq 64\delta^2,$$

as claimed.

F Results for proof of Theorem 3

The reader should recall from Section 3.5 the definitions of the function classes $\mathcal{F}^*_{d,s,\mathcal{H}}(B)$ and $\mathcal{H}(S, B)$. The function class $\mathcal{H}(S, B)$ can be parameterized by the two-dimensional sequence $(a_{j,k})_{j \in S\,, k \in \mathbb{N}}$ of co-efficients, and expressed in terms of two-dimesnional sequence of basis functions $(\phi_{j,k})_{j \in S\,, k \in \mathbb{N}}$ and the sequence of eigenvalues $(\mu_k)_{k \in \mathbb{N}}$ for the univariate RKHS $\mathcal{H}$ as follows:

$$\mathcal{H}(S, B) := \big\{ f = \sum_{j \in S} \sum_{k=1}^{\infty} a_{j,k} \phi_{j,k} \mid \sum_{k=1}^{\infty} \frac{a_{j,k}^2}{\mu_k} \leq 1 \;\forall\, j \in S \text{ and } \|f\|_\infty \leq B \big\}.$$

For any integer $M \geq 1$, we also consider the truncated function class

$$\mathcal{H}(S, B, M) := \big\{ f = \sum_{j \in S} \sum_{k=1}^{M} a_{j,k} \phi_{j,k} \mid \sum_{k=1}^{\infty} \frac{a_{j,k}^2}{\mu_k} \leq 1 \;\forall\, j \in S \text{ and } \|f\|_\infty \leq B \big\}.$$

Lemma 10. *We have the inclusion* $\mathcal{H}(S, B, M) \subseteq \big\{ f \in \mathcal{H}(S) \mid \sum_{j \in S} \sum_{k=1}^{M} |a_{j,k}| \leq B\sqrt{M} \big\}$.

Proof. Without loss of generality, let us assume that $S = \{1, 2, ..., s\}$, and consider a function $f = \sum_{j=1}^{s} f_j \in \mathcal{H}(S, B, M)$. Since each f_j acts on a different co-ordinate, we are guaranteed that $\|f\|_\infty = \sum_{j=1}^{s} \|f_j\|_\infty$. Consider any univariate function $f_j = \sum_{k=1}^{M} a_{j,k}\phi_{j,k}$. We have

$$\sum_{k=1}^{M} |a_{j,k}| \leq \sqrt{M} \left(\sum_{k=1}^{M} a_{j,k}^2 \right)^{1/2} \overset{(a)}{\leq} \sqrt{M} \left[\mathbb{E}[f_j^2(X_j)]\right]^{1/2} \leq \sqrt{M}\|f_j\|_\infty,$$

where step (a) uses the fact that $\mathbb{E}[f_j^2(X_j)] = \sum_{k=1}^{\infty} a_{j,k}^2 \geq \sum_{k=1}^{M} a_{j,k}^2$ for any $M \geq 1$. Adding up the bounds over all co-ordinates, we obtain

$$\|a\|_1 \;=\; \sum_{j=1}^{s}\sum_{k=1}^{M} |a_{j,k}| \leq \sqrt{M} \sum_{j=1}^{s} \|f_j\|_\infty \;=\; \sqrt{M}\|f\|_\infty \;\leq\; \sqrt{M}B,$$

where the final step uses the uniform boundedness condition. □

F.1 Proof of Lemma 5

Recalling the definition of $\widehat{Z}_n(w; t, \mathcal{H}(S, 2B))$ stated from (50), let us view it as a function of the standard Gaussian random vector $(w_1, \ldots, w_n)$. It is straightforward to verify that this variable is Lipschitz (with respect to the Euclidean norm) with parameter at most $t/\sqrt{n}$. Consequently, by concentration for Lipschitz functions [21], we have

$$\mathbb{P}\big[\widehat{Z}_n(w; t, \mathcal{H}(S, 2B)) \geq \mathbb{E}[\widehat{Z}_n(w; t, \mathcal{H}(S, 2B))] + 3t\delta_n\big] \leq \exp\big(-\frac{9n\delta_n^2}{2}\big).$$

Next we prove an upper bound on the expectations

$$\widehat{\mathcal{Q}}_{w,n}(t; \mathcal{H}(S, 2B)) := \; \mathbb{E}_w\big[\sup_{\substack{g\in\mathcal{H}(S,2B)\\ \|g\|_n\leq t}} \frac{1}{n}\sum_{i=1}^{n} w_i g(x_i)\big], \quad \text{and} \tag{76a}$$

$$\mathcal{Q}_{w,n}(t; \mathcal{H}(S, 2B)) := \; \mathbb{E}_{x,w}\big[\sup_{\substack{g\in\mathcal{H}(S,2B)\\ \|g\|_2\leq t}} \frac{1}{n}\sum_{i=1}^{n} w_i g(x_i)\big]. \tag{76b}$$

Lemma 11. *Under the conditions of Theorem 3, we have*

$$\max\big\{\widehat{\mathcal{Q}}_{w,n}(t; \mathcal{H}(S, 2B)),\; \mathcal{Q}_{w,n}(t; \mathcal{H}(S, 2B))\big\} \;\;\leq\;\; 8B\,C\sqrt{\frac{s^{1/\alpha}\log s}{n}}.$$

Proof. By definition, any function $g \in \mathcal{H}(S, 2B)$ has support at most $2s$, and without loss of generality (re-indexing as necessary), we assume that $S = \{1, 2, ..., 2s\}$. We can thus view functions in $\mathcal{H}(S, 2B)$ as having domain $\mathbb{R}^{2s}$, and we can an operator Φ that maps from $\mathbb{R}^{2s}$ to $[\ell^2(\mathbb{N})]^{2s}$, via

$$x \mapsto \Phi_{j,k}(x) = \phi_{j,k}(x_j), \qquad \text{for } j = 1, \ldots, 2s, \text{ and } k \in \mathbb{N}.$$

Any function in $g \in \mathcal{H}(S, 2B)$ can be expressed in terms of two-dimensional sequence $(a_{j,k})$ and the functions $(\Phi_{j,k})$ as $g(x) = g(x_1, x_2, \ldots, x_{2s}) = \sum_{j=1}^{2s}\sum_{k=1}^{\infty} \Phi_{j,k}(x)a_{j,k} \;=\; \langle\!\langle \Phi(x),\; a\rangle\!\rangle$, where $\langle\!\langle\cdot,\;\cdot\rangle\!\rangle$

is a convenient shorthand for the inner product between the two arrays.

For any function $g \in \mathcal{H}(S, 2B)$, triangle inequality yields the upper bound

$$\sup_{g\in 2\mathcal{H}(S,2B)} \frac{1}{n}\Big|\sum_{i=1}^{n} w_i \langle\langle \Phi(x_i),\ a\rangle\rangle\Big| \le \underbrace{\sup_{g\in 2\mathcal{H}(S,2B)} \frac{1}{n}\Big|\sum_{i=1}^{n} w_i \langle\langle \Phi_{\cdot,1:M}(x_i),\ a_{\cdot,1:M}\rangle\rangle}_{A_1} + A_2 \tag{77}$$

where $A_2 := \sup_{g\in 2\mathcal{H}(S,2B)} \frac{1}{n}|\sum_{i=1}^{n} w_i \langle\langle \Phi_{\cdot,M+1:\infty}(x_i),\ a_{\cdot,M+1:\infty}\rangle\rangle|$.

Bounding term $\mathbb{E}_{x,w}[A_1]$ and $\mathbb{E}_w[A_1]$: By Hölder's inequality and Lemma 10, we have

$$A_1 \le \frac{1}{\sqrt{n}} \sup_{g\in 2\mathcal{H}(S,2B)} \|a_{\cdot,1:M}\|_{1,1} \max_{j,k} \Big|\sum_{i=1}^{n} \frac{w_i}{\sqrt{n}} \Phi_{j,k}(x_i)\Big| \ \le\ \frac{2\ B\ \sqrt{M}}{\sqrt{n}} \max_{j,k} \Big|\sum_{i=1}^{n} \frac{w_i}{\sqrt{n}} \Phi_{j,k}(x_i)\Big|.$$

By assumption, we have $|\Phi_{j,k}(x_i)| \le C$ for all indices (i,j,k), implying that $\sum_{i=1}^{n} \frac{w_i}{\sqrt{n}} \Phi_{j,k}(x_i)$ is zero-mean with sub-Gaussian parameter bounded by C and we are taking the maximum of $2s \times M$ such terms. Consequently, we conclude that

$$\mathbb{E}_w[A_1] \le 8BC\sqrt{\frac{M \log(Ms)}{n}}. \tag{78}$$

The same bound holds for $\mathbb{E}_{x,w}[A_1]$.

Bounding term $\mathbb{E}_{x,w}[A_2]$ and $\mathbb{E}_w[A_2]$: In order to control this term, we simply recognize that it corresponds to the usual Gaussian complexity of the sum of $2s$ univariate Hilbert spaces, each of which is an RKHS truncated to the eigenfunctions $\{\mu_k\}_{k\ge M+1}$.

$$\begin{aligned}
\frac{1}{n}\Big|\sum_{i=1}^{n} w_i \langle\langle \Phi_{\cdot,M+1:\infty}(x_i),\ a_{\cdot,M+1:\infty}\rangle\rangle\Big| \ &\le\ \frac{1}{\sqrt{n}} \sum_{j=1}^{2s} \Big|\sum_{k\ge M+1} a_{j,k}\Phi_{j,k}(x) \sum_{i=1}^{n} \frac{w_i}{\sqrt{n}}\Big| \\
&\le\ \frac{C}{\sqrt{n}} \sum_{j=1}^{2s} \Big|\sum_{k\ge M+1} \frac{a_{j,k}}{\sqrt{\mu_k}} \sqrt{\mu_k} \sum_{i=1}^{n} \frac{w_i}{\sqrt{n}}\Big| \\
&\le\ \frac{C}{\sqrt{n}} \sum_{j=1}^{2s} \sqrt{\sum_{k\ge M+1} \frac{a_{j,k}^2}{\mu_k}} \sqrt{\sum_{k\ge M+1} \mu_k \Big(\sum_{i=1}^{n} \frac{w_i}{\sqrt{n}}\Big)^2},
\end{aligned}$$

where the final inequality follows from Cauchy-Schwartz. Exploiting the fact that $\sum_{k\ge M+1} \frac{a_{j,k}^2}{\mu_k} \le 1$ for all j, we have the bound

$$\mathbb{E}_w[A_2] \le 4Cs\, \frac{\sqrt{\sum_{k\ge M+1} \mu_k}}{\sqrt{n}}. \tag{79}$$

One again a similar bound holds for $\mathbb{E}_{x,w}[A_2]$.

Substituting the bound (78) and (79) into the upper bound (77), we conclude that

$$\begin{aligned}\mathcal{Q}_{w,n}(2\mathcal{H}(S,2B)) &\leq 4BC\sqrt{\frac{M\log(Ms)}{n}} + 4Cs\sqrt{\frac{\sum_{k\geq M+1}\mu_k}{n}} \\ &\leq 4BC\sqrt{\frac{M\log(Ms)}{n}} + 4Cs\sqrt{\frac{M^{1-2\alpha}}{n}},\end{aligned}$$

where the second inequality follows from the relation $\mu_k \simeq k^{-2\alpha}$. Finally, setting $M = s^{\frac{1}{\alpha}}$ yields the claim.

Note that the same argument works for the Rademacher complexity, since we only exploited the sub-Gaussianity of the variables w_i. $\square$

Returning to the proof of Lemma 5, combining Lemma 11 with the bound (62a) in Lemma 8:

$$\mathbb{P}\big[\widehat{Z}_n(w;t,\mathcal{H}(S,2B)) \geq 8B\,C\,\sqrt{\frac{s^{1/\alpha}\log s}{n}} + 3t\delta_n\big] \leq \exp\big(-\frac{9n\delta_n^2}{2}\big).$$

Since $\|g\|_n \leq 2B$ for any function $g \in \mathcal{H}(S,2B)$, the proof Lemma 5 is completed using a peeling argument over the radius, analogous to the proof of Lemma 1 (see Appendix B).

References

[1] K. S. Alexander. Rates of growth and sample moduli for weighted empirical processes indexed by sets. *Probability Theory and Related Fields*, 75:379–423, 1987.

[2] N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68:337–404, 1950.

[3] F. Bach. Consistency of the group Lasso and multiple kernel learning. *Journal of Machine Learning Research*, 9:1179–1225, 2008.

[4] P. Bartlett, O. Bousquet, and S. Mendelson. Local Rademacher complexities. *Annals of Statistics*, 33:1497–1537, 2005.

[5] P. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. Submitted to Annals of Statistics, 2008.

[6] M. S. Birman and M. Z. Solomjak. Piecewise-polynomial approximations of functions of the classes W_p^α. *Math. USSR-Sbornik*, 2(3):295–317, 1967.

[7] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, Cambridge, UK, 2004.

[8] L. Breiman. Better subset regression using the nonnegative garrote. *Technometrics*, 37:373–384, 1995.

[9] V. V. Buldygin and Y. V. Kozachenko. *Metric characterization of random variables and random processes*. American Mathematical Society, Providence, RI, 2000.

[10] B. Carl and I. Stephani. *Entropy, compactness and the approximation of operators.* Cambridge Tracts in Mathematics. Cambridge University Press, Cambridge, UK, 1990.

[11] B. Carl and H. Triebel. Inequalities between eigenvalues, entropy numbers and related quantities of compact operators in banach spaces. *Annals of Mathematics*, 251:129–133, 1980.

[12] F. Chung and L. Lu. Concentration inequalities and martingale inequalities. *Internet Mathematics*, 3:79–127, 2006.

[13] T.M. Cover and J.A. Thomas. *Elements of Information Theory.* John Wiley and Sons, New York, 1991.

[14] Howard L. Weinert (ed.), editor. *Reproducing Kernel Hilbert Spaces : Applications in Statistical Signal Processing.* Hutchinson Ross Publishing Co., Stroudsburg, PA, 1982.

[15] C. Gu. *Smoothing spline ANOVA models.* Springer Series in Statistics. Springer, New York, NY, 2002.

[16] R. Z. Has'minskii. A lower bound on the risks of nonparametric estimates of densities in the uniform metric. *Theory Prob. Appl.*, 23:794–798, 1978.

[17] T. Hastie and R. Tibshirani. Generalized additive models. *Statistical Science*, 1(3):297–310, 1986.

[18] G. Kimeldorf and G. Wahba. Some results on Tchebycheffian spline functions. *Jour. Math. Anal. Appl.*, 33:82–95, 1971.

[19] V. Koltchinskii and M. Yuan. Sparse recovery in large ensembles of kernel machines. In *Proceedings of COLT*, 2008.

[20] V. Koltchinskii and M. Yuan. Sparsity in multiple kernel learning. *Annals of Statistics*, 38:3660–3695, 2010.

[21] M. Ledoux. *The Concentration of Measure Phenomenon.* Mathematical Surveys and Monographs. American Mathematical Society, Providence, RI, 2001.

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

[23] P. Massart. About the constants in talagrand's concentration inequalities for empirical processes. *Annals of Probability*, 28(2):863–884, 2000.

[24] L. Meier, S. van de Geer, and P. Buhlmann. High-dimensional additive modeling. *Annals of Statistics*, 37:3779–3821, 2009.

[25] S. Mendelson. Geometric parameters of kernel machines. In *Proceedings of COLT*, pages 29–43, 2002.

[26] J. Mercer. Functions of positive and negative type and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society A*, 209:415–446, 1909.

[27] S. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. In *NIPS Conference*, 2009.

[28] G. Pisier. *The Volume of Convex Bodies and Banach Space Geometry*, volume 94 of *Cambridge Tracts in Mathematics*. Cambridge University Press, Cambridge, UK, 1989.

[29] G. Raskutti, M. J. Wainwright, and B. Yu. Minimax-optimal rates for sparse additive models over kernel classes via convex programming. Technical Report arXiv:0910.2042, UC Berkeley, Department of Statistics, 2010.

[30] G. Raskutti, M. J. Wainwright, and B. Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Trans. Information Theory*, 57(10):6976—6994, October 2011.

[31] P. Ravikumar, H. Liu, J. Lafferty, and L. Wasserman. SpAM: sparse additive models. *Journal of the Royal Statistical Society, Series B*, 2010. To appear.

[32] S. Saitoh. *Theory of Reproducing Kernels and its Applications*. Longman Scientific & Technical, Harlow, UK, 1988.

[33] B. Schölkopf and A. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.

[34] C. J. Stone. Additive regression and other nonparametric models. *Annals of Statistics*, 13(2):689–705, 1985.

[35] S. van de Geer. *Empirical Processes in M-Estimation*. Cambridge University Press, 2000.

[36] A. W. van der Vaart and J. Wellner. *Weak Convergence and Empirical Processes*. Springer-Verlag, New York, NY, 1996.

[37] G. Wahba. *Spline models for observational data*. CBMS-NSF Regional Conference Series in Applied Mathematics. SIAM, Philadelphia, PN, 1990.

[38] Y. Yang and A. Barron. Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, 27(5):1564–1599, 1999.

[39] B. Yu. Assouad, Fano and Le Cam. *Research Papers in Probability and Statistics: Festschrift in Honor of Lucien Le Cam*, pages 423–435, 1996.

[40] M. Yuan. Nonnegative garrote component selection in functional anova models. In *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*, pages 660–666, 2007.