Statistical inference in compound functional models

Arnak S. Dalalyan, Yuri Ingster, Alexandre Tsybakov

Statistical inference in compound functional models

Arnak Dalalyan[1] · Yuri Ingster · Alexandre B. Tsybakov[1]

January 2, 2013

Key words. Compound functional model, minimax estimation, sparse additive structure, dimension reduction, structure adaptation

Abstract. We consider a general nonparametric regression model called the compound model. It includes, as special cases, sparse additive regression and nonparametric (or linear) regression with many covariates but possibly a small number of relevant covariates. The compound model is characterized by three main parameters: the structure parameter describing the macroscopic form of the compound function, the microscopic sparsity parameter indicating the maximal number of relevant covariates in each component and the usual smoothness parameter corresponding to the complexity of the members of the compound. We find non-asymptotic minimax rate of convergence of estimators in such a model as a function of these three parameters. We also show that this rate can be attained in an adaptive way.

1. Introduction

High dimensional statistical inference has known a tremendous development over the past ten years motivated by applications in various fields such as bioinformatics, computer vision, financial engineering. The most intensively investigated models in the context of high-dimensionality are the (generalized) linear models, for which efficient procedures are well known and the theoretical properties are well understood (cf., for instance, [3, 10, 11, 29]). More recently, increasing interest is demonstrated for studying nonlinear models in high-dimensional setting [16, 12, 6, 21, 27] under various types of sparsity assumption. The present paper introduces a general framework that unifies these studies and describes the theoretical limits of statistical procedures in high-dimensional nonlinear problems.

In order to reduce the technicalities and focus on the main ideas, we consider the Gaussian white noise model, which is known to be asymptotically equivalent, under some natural conditions, to the model of regression [5, 22], as well as to other nonparametric models [9, 13]. Thus, we assume that we observe a real-valued Gaussian process $\boldsymbol{Y} = \{Y(\phi) : \phi \in L^2([0,1]^d)\}$ such that

$$\mathbf{E}_f[Y(\phi)] = \int_{[0,1]^d} f(\boldsymbol{x})\,\phi(\boldsymbol{x})\,d\boldsymbol{x}, \qquad \mathbf{Cov}_f(Y(\phi), Y(\phi')) = \varepsilon^2 \int_{[0,1]^d} \phi(\boldsymbol{x})\phi'(\boldsymbol{x})\,d\boldsymbol{x},$$

for all $\phi, \phi' \in L^2([0,1]^d)$, where f is an unknown function in $L^2([0,1]^d)$, $\mathbf{E}_f$ and $\mathbf{Cov}_f$ are the expectation and covariance signs, and ε is some positive number. It is well known that these two properties uniquely characterize the probability distribution of a Gaussian process that we will further denote by $\mathbf{P}_f$ (respectively, by $\mathbf{P}_0$ if $f \equiv 0$). Alternatively, $\boldsymbol{Y}$ can be considered as a

ENSAE-CREST-GENES
3, avenue Pierre Larousse
92245 MALAKOFF Cedex, FRANCE
e-mail: arnak.dalalyan,alexandre.tsybakov@ensae.fr

St. Petersburg State Electrotechnical University

trajectory of the process

$$dY(\boldsymbol{x}) = f(\boldsymbol{x})\,d\boldsymbol{x} + \varepsilon dW(\boldsymbol{x}), \qquad \boldsymbol{x} \in [0,1]^d,$$

where $W(\boldsymbol{x})$ is a d-parameter Brownian sheet. The parameter ε is assumed known; in the model of regression it corresponds to the quantity $\sigma^2 n^{-1/2}$, where σ^2 is the variance of noise. Without loss of generality, we assume in what follows that $0 < \varepsilon < 1$.

1.1. Notation

First, we introduce some notation. Vectors in finite-dimensional spaces and infinite sequences will be denoted by boldface letters, vector norms will be denoted by $|\cdot|$ while function norms will be denoted by $\|\cdot\|$. Thus, for $\mathbf{v} = (v_1, \dots, v_d) \in \mathbb{R}^d$ we set

$$|\mathbf{v}|_0 = \sum\nolimits_{j=1}^{d} \mathbf{1}(v_j \neq 0), \quad |\mathbf{v}|_\infty = \max_{j=1,\dots,d} |v_j|, \quad |\mathbf{v}|_q^q = \sum\nolimits_{j=1}^{d} |v_j|^q,\ 1 \le q < \infty,$$

whereas for a function $f : [0,1]^d \to \mathbb{R}$ we set

$$\|f\|_\infty = \sup_{\boldsymbol{x}\in[0,1]^d} |f(\boldsymbol{x})|, \qquad \|f\|_q^q = \int_{[0,1]^d} |f(\boldsymbol{x})|^q\,d\boldsymbol{x},\ 1 \le q < \infty.$$

We denote by $L_0^2([0,1]^d)$ the subspace of $L^2([0,1]^d)$ containing all the functions f such that $\int_{[0,1]^d} f(\boldsymbol{x})\,d\boldsymbol{x} = 0$. The notation $\langle\cdot,\cdot\rangle$ will be used for the inner product in $L^2([0,1]^d)$, that is $\langle h, \tilde{h}\rangle = \int_{[0,1]^d} h(\boldsymbol{x})\tilde{h}(\boldsymbol{x})\,d\boldsymbol{x}$ for any $h, \tilde{h} \in L^2([0,1]^d)$. For two integers a and a', we denote by $[\![a, a']\!]$ the set of all integers belonging to the interval $[a, a']$. We denote by $[t]$ the integer part of a real number t. For a finite set V, we denote by $|V|$ its cardinality. For a vector $\boldsymbol{x} \in \mathbb{R}^d$ and a set of indices $V \subseteq \{1, \dots, d\}$, the vector $\boldsymbol{x}_V \in \mathbb{R}^{|V|}$ is defined as the restriction of $\boldsymbol{x}$ to the coordinates with indices belonging to V. For every $s \in \{1, \dots, d\}$ and $m \in \mathbb{N}$, we define $\mathcal{V}_s^d = \big\{V \subseteq \{1, \dots, d\} : |V| \le s\big\}$ and the set of binary vectors $\mathcal{B}_{s,m}^d = \big\{\boldsymbol{\eta} \in \{0,1\}^{\mathcal{V}_s^d} : |\boldsymbol{\eta}|_0 = m\big\}$. We also use the notation $M_{d,s} \triangleq |\mathcal{V}_s^d|$. We extend these definitions to $s = 0$ by setting $\mathcal{V}_0^d = \{\varnothing\}$, $M_{d,0} = 1$, $|\mathcal{B}_{0,1}^d| = 1$, and $|\mathcal{B}_{0,m}^d| = 0$ for $m > 1$. For a vector $\boldsymbol{a}$, we denote by $\mathrm{supp}(\boldsymbol{a})$ the set of indices of its non-zero coordinates. In particular, the support $\mathrm{supp}(\boldsymbol{\eta})$ of a binary vector $\boldsymbol{\eta} = \{\eta_V\}_{V\in\mathcal{V}_s^d} \in \mathcal{B}_{s,m}^d$ is the set of V's such that $\eta_V = 1$.

1.2. Compound functional model

In this paper we impose the following assumption on the unknown function f.

Compound functional model: *There exists an integer $s \in \{1, \dots, d\}$, a binary sequence $\boldsymbol{\eta} \in \mathcal{B}_{s,m}^d$, a set of functions $\{f_V \in L_0^2([0,1]^{|V|})\}_{V\in\mathcal{V}_s^d}$ and a constant $\bar{f}$ such that*

$$f(\boldsymbol{x}) = \bar{f} + \sum_{V\in\mathcal{V}_s^d} f_V(\boldsymbol{x}_V)\eta_V = \bar{f} + \sum_{V\in\mathrm{supp}(\boldsymbol{\eta})} f_V(\boldsymbol{x}_V), \qquad \forall \boldsymbol{x} \in \mathbb{R}^d. \tag{1}$$

The functions f_V are called the atoms of the compound model.

Note that, under the compound model, $\bar{f} = \int_{[0,1]^d} f(\boldsymbol{x})\,d\boldsymbol{x}$.

The atoms f_V are assumed to be sufficiently regular, namely, each f_V is an element of a suitable functional class Σ_V. In particular, one can consider a smoothness class Σ_V and more specifically the Sobolev ball of functions of s variables[1]. In what follows, we will mainly deal with this example.

[1] Note that every function of less than s variables can also be considered as a function of s variables.

Given a collection $\boldsymbol{\Sigma} = \{\Sigma_V\}_{V \in \mathcal{V}_s^d}$ of subsets of $L_0^2([0,1]^s)$ and a subset $\tilde{\mathcal{B}}$ of $\mathcal{B}_{s,m}^d$, we define the classes

$$\mathcal{F}_{s,m}(\boldsymbol{\Sigma}) = \bigcup\nolimits_{\boldsymbol{\eta} \in \tilde{\mathcal{B}}} \mathcal{F}_{\boldsymbol{\eta}}(\boldsymbol{\Sigma}),$$

where

$$\mathcal{F}_{\boldsymbol{\eta}}(\boldsymbol{\Sigma}) = \Big\{ f : \mathbb{R}^d \to \mathbb{R} : \exists \bar{f} \in \mathbb{R}, \{f_V\}_{V \in \mathrm{supp}(\boldsymbol{\eta})}, f_V \in \Sigma_V, \text{ such that } f = \bar{f} + \sum_{V \in \mathrm{supp}(\boldsymbol{\eta})} f_V \Big\}.$$

The class $\mathcal{F}_{s,m}(\boldsymbol{\Sigma})$ is defined for any $s \in \{0, \dots, d\}$ and any $m \in \{0, \dots, M_{d,s}\}$. In what follows, we assume that $\tilde{\mathcal{B}}$ is fixed and for this reason we do not include it in the notation. Examples of $\tilde{\mathcal{B}}$ can be the set of all $\boldsymbol{\eta} \in \mathcal{B}_{s,m}^d$ such that $V \in \mathrm{supp}(\boldsymbol{\eta})$ are pairwise disjoint or of all $\boldsymbol{\eta} \in \mathcal{B}_{s,m}^d$ such that every set V from $\mathrm{supp}(\boldsymbol{\eta})$ has a non-empty intersection with at most one other set from $\mathrm{supp}(\boldsymbol{\eta})$.

It is clear from the definition that the parameters $\big(\boldsymbol{\eta}, \{f_V\}_{V \in \mathrm{supp}(\boldsymbol{\eta})}\big)$ are not identifiable. Indeed, two different collections $\big(\boldsymbol{\eta}, \{f_V\}_{V \in \mathrm{supp}(\boldsymbol{\eta})}\big)$ and $\big(\bar{\boldsymbol{\eta}}, \{\bar{f}_V\}_{V \in \mathrm{supp}(\bar{\boldsymbol{\eta}})}\big)$ may lead to the same compound function f. Of course, this is not necessarily an issue as long as only the problem of estimating f is considered.

We now define the Sobolev classes of functions of many variables that will play the role of Σ_V. Consider an orthonormal system of functions $\{\varphi_{\boldsymbol{j}}\}_{\boldsymbol{j} \in \mathbb{Z}^d}$ in $L^2([0,1]^d)$ such that $\varphi_{\boldsymbol{0}}(\boldsymbol{x}) \equiv 1$. We assume that the system $\{\varphi_{\boldsymbol{j}}\}$ and the set $\tilde{\mathcal{B}}$ are such that

$$\Big\| \sum_{V \in \mathrm{supp}(\boldsymbol{\eta})} \sum_{\substack{\boldsymbol{j}: \boldsymbol{j} \neq \boldsymbol{0} \\ \mathrm{supp}(\boldsymbol{j}) \subseteq V}} \theta_{\boldsymbol{j},V} \varphi_{\boldsymbol{j}} \Big\|_2^2 \leq C_* \sum_{V \in \mathrm{supp}(\boldsymbol{\eta})} \sum_{\substack{\boldsymbol{j}: \boldsymbol{j} \neq \boldsymbol{0} \\ \mathrm{supp}(\boldsymbol{j}) \subseteq V}} \theta_{\boldsymbol{j},V}^2, \tag{2}$$

for all $\boldsymbol{\eta} \in \tilde{\mathcal{B}}$ and all square-summable arrays $(\theta_{\boldsymbol{j},V}, (\boldsymbol{j}, V) \in \mathbb{Z}^d \times \mathcal{V}_s^d)$, where $C_* > 0$ is a constant independent of s, m and d. For example, this condition holds with $C_* = 1$ if $\tilde{\mathcal{B}}$ is the set of all $\boldsymbol{\eta} \in \mathcal{B}_{s,m}^d$ such that $V \in \mathrm{supp}(\boldsymbol{\eta})$ are pairwise disjoint and with $C^* = 3/2$ if $\tilde{\mathcal{B}}$ is the set of all $\boldsymbol{\eta} \in \mathcal{B}_{s,m}^d$ such that every set V from $\mathrm{supp}(\boldsymbol{\eta})$ has a non-empty intersection with at most one other set from $\mathrm{supp}(\boldsymbol{\eta})$.

One example of $\{\varphi_{\boldsymbol{j}}\}_{\boldsymbol{j} \in \mathbb{Z}^d}$ is a tensor product orthonormal basis:

$$\varphi_{\boldsymbol{j}}(\boldsymbol{x}) = \bigotimes\nolimits_{\ell=1}^d \varphi_{j_\ell}(x_\ell), \tag{3}$$

where $\boldsymbol{j} = (j_1, \dots, j_d) \in \mathbb{Z}^d$ is a multi-index and $\{\varphi_k\}$, $k \in \mathbb{Z}$, is an orthonormal basis in $L^2([0,1])$. Specifically, we can take the trigonometric basis with $\varphi_0(u) \equiv 1$ on $[0,1]$, $\varphi_k(u) = \sqrt{2}\cos(2\pi\, ku)$ for $k > 0$ and $\varphi_k(u) = \sqrt{2}\sin(2\pi\, ku)$ for $k < 0$. To ease notation, we set $\theta_{\boldsymbol{j}}[f] = \langle f, \varphi_{\boldsymbol{j}} \rangle$ for $\boldsymbol{j} \in \mathbb{Z}^d$.

For any set of indices $V \subseteq \{1, \dots, d\}$ and any $\beta > 0$, $L > 0$, we define the Sobolev class of functions

$$W_V(\beta, L) = \Big\{ g \in L_0^2([0,1]^d) : \quad g = \sum_{\boldsymbol{j} \in \mathbb{Z}^d : \mathrm{supp}(\boldsymbol{j}) \subseteq V} \theta_{\boldsymbol{j}}[g] \varphi_{\boldsymbol{j}} \quad \text{and} \quad \sum_{\boldsymbol{j} \in \mathbb{Z}^d} |\boldsymbol{j}|_\infty^{2\beta} \theta_{\boldsymbol{j}}[g]^2 \leq L \Big\}. \tag{4}$$

Assuming that $\{\varphi_{\boldsymbol{j}}\}$ is the trigonometric basis and f is periodic with period one in each coordinate, *i.e.*, $f(\boldsymbol{x} + \boldsymbol{j}) = f(\boldsymbol{x})$ for every $\boldsymbol{x} \in \mathbb{R}^d$ and every $\boldsymbol{j} \in \mathbb{Z}^d$, the condition $f_V \in W_V(\beta, L)$ can be interpreted as the square integrability of all partial derivatives of f_V up to the order β.

Let us give some examples of compound models.

– *Additive models* are the special case $s = 1$ of compound models. Here, additive models are understood in a wider sense than originally defined by Stone [26]. Namely, for $s = 1$ we have the model
$$f(\boldsymbol{x}) = \bar{f} + \sum_{j \in J} f_j(x_j), \qquad \boldsymbol{x} = (x_1, \dots, x_d) \in \mathbb{R}^d,$$
where J is any (unknown) subset of indices and not necessarily $J = \{1, \dots, d\}$. Estimation and testing problems in this model when the atoms belong to some smoothness classes have been studied in Ingster and Lepski [14], Meier et al. [19], Koltchinskii and Yuan [16], Raskutti et al. [21], Gayraud and Ingster [12], Suzuki [27].
– *Single atom models* are the special case $m = 1$ of compound models. If $m = 1$ we have $f(\boldsymbol{x}) = f_V(\boldsymbol{x}_V)$ for some unknown $V \subseteq \{1, \dots, d\}$, *i.e.*, there exists only one set V for which $\eta_V = 1$, and $|V| \le s$. Estimation and variable selection in this model were considered by Bertin and Lecué [2], Comminges and Dalalyan [7], Rosasco et al. [25]. The case of small s and large d is particularly interesting in the context of sparsity. In a parametric model, when f_V is a linear function, we are back to the sparse high-dimensional linear regression setting, which has been extensively studied, see, e.g., van de Geer and Bühlmann [29].
– *Tensor product models.* Let $\mathcal{A}$ be a given finite subset of $\mathbb{Z}$, and assume that $\varphi_{\boldsymbol{j}}$ is a tensor product basis defined by (3). Consider the following parametric class of functions
$$\boldsymbol{T}_{\boldsymbol{\eta}}(\mathcal{A}) = \Big\{ f : \mathbb{R}^d \to \mathbb{R} : \exists \bar{f}, \{\theta_{\boldsymbol{j},V}\}, \text{ such that } f = \bar{f} + \sum_{V \in \mathrm{supp}(\boldsymbol{\eta})} \sum_{\boldsymbol{j} \in \mathcal{J}_{V,\mathcal{A}}} \theta_{\boldsymbol{j},V} \varphi_{\boldsymbol{j}} \Big\}, \tag{5}$$
where
$$\mathcal{J}_{V,\mathcal{A}} = \Big\{ \boldsymbol{j} \in \mathcal{A}^d : \mathrm{supp}(\boldsymbol{j}) \subseteq V \Big\}. \tag{6}$$
We say that function f satisfies the tensor product model if it belongs to the set $\boldsymbol{T}_{\boldsymbol{\eta}}(\mathcal{A})$ for some $\boldsymbol{\eta} \in \tilde{\mathcal{B}}$. We define
$$\mathcal{F}_{s,m}(\boldsymbol{T}_{\mathcal{A}}) = \bigcup\nolimits_{\boldsymbol{\eta} \in \tilde{\mathcal{B}}} \boldsymbol{T}_{\boldsymbol{\eta}}(\mathcal{A}).$$
Important examples are sparse high-dimensional multilinear/polynomial systems. Motivated respectively by applications in genetics and signal processing, they have been recently studied by Nazer and Nowak [20] in the context of compressed sensing without noise and by Kekatos and Giannakis [15] in the case where the observations are corrupted by a Gaussian noise. With our notation, the models they considered are the tensor product models with $\mathcal{A} = \{0, 1\}$ (linear basis functions φ_j) in the multilinear model of [20] and $\mathcal{A} = \{-1, 0, 1\}$ in the Volterra filtering problem of [15] (second-order Volterra systems with $\varphi_0(x) \equiv 1$, $\varphi_1(x) \propto (x - 1/2)$ and $\varphi_{-1}(x) \propto x^2 - x + 1/6$). More generally, the set $\mathcal{A}$ should be of small cardinality to guarantee efficient dimension reduction. Another approach is to introduce hierarchical structures on the coefficients of tensor product representation [4, 1].

In what follows, we assume that f belongs to the functional class $\mathcal{F}_{s,m}(\boldsymbol{\Sigma})$ where either $\boldsymbol{\Sigma} = \{W_V(\beta, L)\}_{V \in \mathcal{V}_s^d} \triangleq \boldsymbol{W}(\beta, L)$ or $\boldsymbol{\Sigma} = \boldsymbol{T}_{\mathcal{A}}$.

The compound model is described by three main parameters, which are the dimension m that we call the *macroscopic* parameter and that characterizes the complexity of possible structure vectors $\boldsymbol{\eta}$, the dimension s of atoms in the compound that we call the *microscopic* parameter, and the complexity of functional class $\boldsymbol{\Sigma}$. The latter can be described by entropy numbers of $\boldsymbol{\Sigma}$ in convenient norms, and in the particular case of Sobolev classes, it is naturally characterized by the smoothness parameter β. The integers m and s are "effective dimension" parameters. As soon as they grow, the structure becomes less pronounced and the compound model approaches the global nonparametric regression in dimension d, which is known to suffer from the curse of dimensionality already for moderate d. Therefore, an interesting case is the sparsity scenario where s and/or m are small.

2. Overview of the results and relation to the previous work

Several statistical problems arise naturally in the context of compound functional model.

Estimation of f. This is the subject of the present paper. We measure the risk of arbitrary estimator $\widetilde{f}_\varepsilon$ by its mean integrated squared error $\mathbf{E}_f[\|\widetilde{f}_\varepsilon - f\|_2^2]$ and we study the minimax risk

$$\inf_{\widetilde{f}_\varepsilon} \sup_{f\in\mathcal{F}_{s,m}(\boldsymbol{\Sigma})} \mathbf{E}_f[\|\widetilde{f}_\varepsilon - f\|_2^2],$$

where $\inf_{\widetilde{f}_\varepsilon}$ denotes the minimum over all estimators[2]. A first general question is to establish the minimax rates of estimation, *i.e.*, to find values $\psi_{s,m,\varepsilon}(\boldsymbol{\Sigma})$ such that

$$\inf_{\widetilde{f}_\varepsilon} \sup_{f\in\mathcal{F}_{s,m}(\boldsymbol{\Sigma})} \mathbf{E}_f[\|\widetilde{f}_\varepsilon - f\|_2^2] \asymp \psi_{s,m,\varepsilon}(\boldsymbol{\Sigma}),$$

when $\boldsymbol{\Sigma}$ is a Sobolev, Hölder or other class of functions. A second question is to construct optimal estimators in a minimax sense, *i.e.*, estimators $\widehat{f}_\varepsilon$ such that

$$\sup_{f\in\mathcal{F}_{s,m}(\boldsymbol{\Sigma})} \mathbf{E}_f[\|\widehat{f}_\varepsilon - f\|_2^2] \le C\psi_{s,m,\varepsilon}(\boldsymbol{\Sigma}), \tag{7}$$

for some constant C independent of s, m, ε and $\boldsymbol{\Sigma}$. Some results on minimax rates of estimation of f are available only for the case $s = 1$ (cf. the discussion below). Finally, a third question that we address here is whether the optimal rate can be attained adaptively, *i.e.*, whether one can construct an estimator $\widehat{f}_\varepsilon$ that satisfies (7) simultaneously for all s, m, β and L when $\boldsymbol{\Sigma} = \boldsymbol{W}(\beta, L)$. We will show that the answer to this question is positive.

Variable selection. Assume that $m = 1$. This means that $f(\boldsymbol{x}) = f_V(\boldsymbol{x}_V)$ for some unknown $V \subseteq \{1, \dots, d\}$, *i.e.*, there exists only one set V for which $\eta_V = 1$ (a single atom model). Then it is of interest to identify V under the constraint $|V| \le s$. In particular, d can be very large while s can be small. This corresponds to estimating the relevant covariates and generalizes the problem of selection of sparsity pattern in linear regression. An estimator $\widehat{V}_n \subseteq \{1, \dots, d\}$ of V is considered as good, if the probability $\mathbf{P}(\widehat{V}_n = V)$ is close to one.

Hypotheses testing (detection): The problem is to test the hypothesis $H_0 : f \equiv 0$ (no signal) against the alternative $H_1 : f \in \mathcal{A}$, where $\mathcal{A} = \big\{f \in \mathcal{F}_{s,m}(\boldsymbol{\Sigma}) : \|f\|_2 \ge r\big\}$. Here, it is interesting to characterize the minimax rates of separation $r > 0$ in terms of s, m and $\boldsymbol{\Sigma}$.

Some of the above three problems have been studied in the literature for special cases $s = 1$ (additive model) and $m = 1$ (single atom model). Ingster and Lepski [14] studied the problem of testing in additive model and provided asymptotic minimax rates of separation. Sharp asymptotic optimality under additional assumptions in the same problem was obtained by Gayraud and Ingster [12]. Recently, Comminges and Dalalyan [7] established tight conditions for variable selection in the single atom model. We also mention an earlier work of Bertin and Lecué [2] dealing with variable selection.

The problem of estimation has been also considered for additive model and class $\boldsymbol{\Sigma}$ defined as a reproducing kernel Hilbert space, cf. Koltchinskii and Yuan [16], Raskutti et al. [21]. In particular, these papers showed that if $s = 1$ and $\boldsymbol{\Sigma} = \boldsymbol{W}(\beta, L)$ is a Sobolev class, then there is an estimator of f for which the mean integrated squared error converges to zero at the rate

$$\max\Big(m\varepsilon^{4\beta/(2\beta+1)},\ m\varepsilon^2 \log d\Big). \tag{8}$$

[2] We focus our attention on the behavior of the expected error of estimation. Alternatively, one can be interested in establishing similar type of upper bounds on the error of estimation that hold true with large probability [8, 17].

Furthermore, Raskutti et al. [21, Thm. 2] provided the following lower bound on the minimax risk:

$$\max\Big(m\varepsilon^{4\beta/(2\beta+1)},\ m\varepsilon^2\log\Big(\frac{d}{m}\Big)\Big). \tag{9}$$

Note that when m is proportional to d, this lower bound departs from the upper bound in a logarithmic way. It should also be noted that the upper bounds in these papers are achieved by estimators that are not adaptive in the sense that they require the knowledge of the smoothness index β.

In this paper, we establish non-asymptotic upper and lower bounds on the minimax risk for the model with Sobolev smoothness class $\boldsymbol{\Sigma} = \boldsymbol{W}(\beta, L)$. We will prove that, up to a multiplicative constant, the minimax risk behaves itself as

$$\max\left\{mL^{s/(2\beta+s)}\varepsilon^{4\beta/(2\beta+s)},\ ms\varepsilon^2\log\left(\frac{d}{sm^{1/s}}\right)\right\}\wedge L \tag{10}$$

(we assume here $d/(sm^{1/s}) > 1$, otherwise a constant factor greater than 1 should be inserted under the logarithm, cf. the results below). In addition, we demonstrate that this rate can be reached in an adaptive way that is without the knowledge of β, s, and m. The rate (10) is non-asymptotic, which explains, in particular, the presence of minimum with constant L in (10). For $s = 1$, *i.e.*, for the additive regression model, our rate matches the lower bound of [21].

For $m = 1$, *i.e.*, when $f(\boldsymbol{x}) = f_V(\boldsymbol{x}_V)$ for some unknown $V \subseteq \{1, \dots, d\}$ (the single atom model), the minimax rate of convergence takes the form

$$\max\left\{L^{s/(2\beta+s)}\varepsilon^{4\beta/(2\beta+s)},\ s\varepsilon^2\log\left(\frac{d}{s}\right)\right\}\wedge L. \tag{11}$$

This rate accounts for two effects, namely, the accuracy of nonparametric estimation of f for fixed macroscopic structure parameter $\boldsymbol{\eta}$, cf. the first term $\sim \varepsilon^{4\beta/(2\beta+s)}$, and the complexity of the structure itself (irrespective to the nonparametric nature of microscopic components $f_V(\boldsymbol{x}_V)$). In particular, the second term $\sim s\varepsilon^2\log(d/s)$ in (11) coincides with the optimal rate of prediction in linear regression model under the standard sparsity assumption. This is what we obtain in the limiting case when β tends to infinity. It is important to note that the optimal rates depend only logarithmically on the ambient dimension d. Thus, even if d is large, the rate optimal estimators achieve nice performance under the sparsity scenario when s and m are small.

3. The estimator and upper bounds on the minimax risk

In this section, we suggest an estimator attaining the minimax rate. It is constructed in the following two steps.

Constructing weak estimators. At this step, we proceed as if the macroscopic structure parameter $\boldsymbol{\eta}$ was known and denote by $V_1, \dots, V_m$ the elements of the support of $\boldsymbol{\eta}$. The goal is to provide for each $\boldsymbol{\eta}$ a family of "simple" estimators of f—indexed by some parameter $\boldsymbol{t}$—containing a rate-minimax one. To this end, we first project $\boldsymbol{Y}$ onto the basis functions $\{\varphi_{\boldsymbol{j}} : |\boldsymbol{j}|_\infty \le \varepsilon^{-2}\}$ and denote

$$\boldsymbol{Y}_\varepsilon = (Y_{\boldsymbol{j}} \triangleq Y(\varphi_{\boldsymbol{j}}) : \boldsymbol{j} \in \mathbb{Z}^d,\ |\boldsymbol{j}|_\infty \le \varepsilon^{-2}). \tag{12}$$

Then, we consider a collection $\{\widehat{\boldsymbol{\theta}}_{\boldsymbol{t},\boldsymbol{\eta}} :\ \boldsymbol{t} \in \mathbb{Z}^m \cap [1, \varepsilon^{-2}]^m\}$ of projection estimators of the vector $\boldsymbol{\theta}_\varepsilon = (\theta_{\boldsymbol{j}}[f])_{\boldsymbol{j}\in\mathbb{Z}^d:|\boldsymbol{j}|_\infty\le\varepsilon^{-2}}$. The role of each component t_ℓ of $\boldsymbol{t}$ is to indicate the cut-off level of the coefficients $\theta_{\boldsymbol{j}}$ corresponding to the atom f_{V_ℓ}, that is the level of indices beyond of which the coefficients are estimated by 0.

To be more precise, for an integer-valued vector $\boldsymbol{t} = (t_{V_\ell}, \ell = 1, \dots, m) \in [0, \varepsilon^{-2}]^m$ we set $\widehat{\boldsymbol{\theta}}_{\boldsymbol{t},\boldsymbol{\eta}} = (\widehat{\theta}_{\boldsymbol{t},\boldsymbol{\eta},\boldsymbol{j}} : \boldsymbol{j} \in \mathbb{Z}^d, |\boldsymbol{j}|_\infty \le \varepsilon^{-2})$, where $\widehat{\theta}_{\boldsymbol{t},\boldsymbol{\eta},\mathbf{0}} = Y_{\mathbf{0}}$ and

$$\widehat{\theta}_{\boldsymbol{t},\boldsymbol{\eta},\boldsymbol{j}} = \begin{cases} Y_{\boldsymbol{j}}, & \exists \ell \text{ s.t. } \operatorname{supp}(\boldsymbol{j}) \subseteq V_\ell \,,\ |\boldsymbol{j}|_\infty \in [1, t_{V_\ell}], \\ 0, & \text{otherwise} \end{cases}$$

if $\boldsymbol{j} \neq \mathbf{0}$. Based on these estimators of the coefficients of f, we recover the function f using the estimator

$$\widehat{f}_{\boldsymbol{t},\boldsymbol{\eta}}(\boldsymbol{x}) = \sum\nolimits_{\boldsymbol{j} \in \mathbb{Z}^d : |\boldsymbol{j}|_\infty \le \varepsilon^{-2}} \widehat{\theta}_{\boldsymbol{t},\boldsymbol{\eta},\boldsymbol{j}} \varphi_{\boldsymbol{j}}(\boldsymbol{x}).$$

Smoothness- and structure-adaptive estimation: The goal in this step is to combine the weak estimators $\{\widehat{f}_{\boldsymbol{t},\boldsymbol{\eta}}\}_{\boldsymbol{t},\boldsymbol{\eta}}$ in order to get a structure and smoothness adaptive estimator of f with a risk which is as small as possible. To this end, we use a version of exponentially weighted aggregate [18, 10, 11] in the spirit of sparsity pattern aggregation as described in [23, 24]. More precisely, for every pair of integers (s, m) such that $s \in \{1, \dots, d\}$ and $m \in \{1, \dots, M_{d,s}\}$, we define prior probabilities for $(\boldsymbol{t}, \boldsymbol{\eta}) \in [\![0, \varepsilon^{-2}]\!]^m \times (\mathcal{B}^d_{s,m} \setminus \mathcal{B}^d_{s-1,m})$ by

$$\pi_{\boldsymbol{t},\boldsymbol{\eta}} = \frac{2^{-sm}}{H_d (1 + [\varepsilon^{-2}])^m |\mathcal{B}^d_{s,m} \setminus \mathcal{B}^d_{s-1,m}|}, \qquad H_d = \sum_{s=0}^{d} \sum_{m=1}^{M_{d,s}} 2^{-sm} \le e. \tag{13}$$

For $s = 0$ and the unique $\boldsymbol{\eta}_0 \in \mathcal{B}^d_{0,1}$ we consider only one weak estimator $\widehat{\boldsymbol{\theta}}_{\boldsymbol{t},\boldsymbol{\eta}_0}$ with all entries zero except for the entry $\widehat{\theta}_{\boldsymbol{t},\boldsymbol{\eta}_0,\mathbf{0}}$, which is equal to $Y_{\mathbf{0}}$. We set $\pi_{\boldsymbol{t},\boldsymbol{\eta}_0} = 1/H_d$. It is easy to see that $\boldsymbol{\pi} = \Big(\pi_{\boldsymbol{t},\boldsymbol{\eta}}; (\boldsymbol{t}, \boldsymbol{\eta}) \in \bigcup_{s,m} \{[\![0, \varepsilon^{-2}]\!]^m \times \mathcal{B}^d_{s,m}\}\Big)$ defines a probability distribution. For any pair $(\boldsymbol{t}, \boldsymbol{\eta})$ we introduce the penalty function

$$\operatorname{pen}(\boldsymbol{t}, \boldsymbol{\eta}) = 2\varepsilon^2 \prod_{V \in \operatorname{supp}(\boldsymbol{\eta})} (2t_V + 1)^{|V|}$$

and define the vector of coefficients $\widehat{\boldsymbol{\theta}}_\varepsilon = (\widehat{\theta}_{\varepsilon,\boldsymbol{j}} : \boldsymbol{j} \in \mathbb{Z}^d, |\boldsymbol{j}|_\infty \le \varepsilon^{-2})$ by

$$\widehat{\boldsymbol{\theta}}_\varepsilon = \sum_{s=1}^{d} \sum_{m=1}^{M_{d,s}} \sum_{(\boldsymbol{t},\boldsymbol{\eta})} \widehat{\boldsymbol{\theta}}_{\boldsymbol{t},\boldsymbol{\eta}} \frac{\exp\big\{ -\frac{1}{4\varepsilon^2} \big(|\boldsymbol{Y}_\varepsilon - \widehat{\boldsymbol{\theta}}_{\boldsymbol{t},\boldsymbol{\eta}}|_2^2 + \operatorname{pen}(\boldsymbol{t}, \boldsymbol{\eta}) \big) \big\} \pi_{\boldsymbol{t},\boldsymbol{\eta}}}{\sum_{\bar{s}=1}^{d} \sum_{\bar{m}=1}^{M_{d,\bar{s}}} \sum_{(\bar{\boldsymbol{t}},\bar{\boldsymbol{\eta}})} \exp\big\{ -\frac{1}{4\varepsilon^2} \big(|\boldsymbol{Y}_\varepsilon - \widehat{\boldsymbol{\theta}}_{\bar{\boldsymbol{t}},\bar{\boldsymbol{\eta}}}|_2^2 + \operatorname{pen}(\bar{\boldsymbol{t}}, \bar{\boldsymbol{\eta}}) \big) \big\} \pi_{\bar{\boldsymbol{t}},\bar{\boldsymbol{\eta}}}}, \tag{14}$$

where the summations $\sum_{(\boldsymbol{t},\boldsymbol{\eta})}$ and $\sum_{(\bar{\boldsymbol{t}},\bar{\boldsymbol{\eta}})}$ correspond to $(\boldsymbol{t}, \boldsymbol{\eta}) \in [\![0, \varepsilon^{-2}]\!]^m \times (\mathcal{B}^d_{s,m} \setminus \mathcal{B}^d_{s-1,m})$ and $(\bar{\boldsymbol{t}}, \bar{\boldsymbol{\eta}}) \in [\![0, \varepsilon^{-2}]\!]^{\bar{m}} \times (\mathcal{B}^d_{\bar{s},\bar{m}} \setminus \mathcal{B}^d_{\bar{s}-1,\bar{m}})$, respectively. The final estimator of f is

$$\widehat{f}_\varepsilon(\boldsymbol{x}) = \sum\nolimits_{\boldsymbol{j} \in \mathbb{Z}^d : |\boldsymbol{j}|_\infty \le \varepsilon^{-2}} \widehat{\theta}_{\varepsilon,\boldsymbol{j}} \varphi_{\boldsymbol{j}}(\boldsymbol{x}), \qquad \forall \boldsymbol{x} \in [0,1]^d.$$

Note that each $\widehat{\boldsymbol{\theta}}_{\boldsymbol{t},\boldsymbol{\eta}}$ is a projection estimator of the vector $\boldsymbol{\theta} = (\theta_{\boldsymbol{j}}[f])_{\boldsymbol{j} \in \mathbb{Z}^d}$. Hence, $\widehat{f}_\varepsilon$ is a convex combination of projection estimators. We also note that, to construct $\widehat{f}_\varepsilon$, we only need to know ε and d. Therefore, the estimator is adaptive to all other parameters of the model, such as s, m, the parameters that define the class $\boldsymbol{\Sigma}$ and the choice of a particular subset $\tilde{\mathcal{B}}$ of $\mathcal{B}^d_{s,m}$.

The following theorem gives an upper bound on the risk of the estimator $\widehat{f}_\varepsilon$ when $\boldsymbol{\Sigma} = \boldsymbol{W}(\beta, L)$.

Theorem 1. *Let $\beta > 0$ and $L > 0$ be such that $\log(\varepsilon^{-2}) \ge (2\beta)^{-1} \log(L)$, $L > \varepsilon^2 \log(e\varepsilon^{-2})^{\frac{2\beta+s}{s}}$. Let $\tilde{\mathcal{B}}$ be any subset of $\mathcal{B}^d_{s,m}$. Assume that condition (2) holds. Then, for some constant $C(\beta) > 0$ depending only on β we have*

$$\sup_{f \in \mathcal{F}_{s,m}(\boldsymbol{W}(\beta,L))} \mathbf{E}_f[\|\widehat{f}_\varepsilon - f\|_2^2] \le (6L) \wedge \Big(m \Big\{ C(\beta) L^{\frac{s}{2\beta+s}} \varepsilon^{\frac{4\beta}{2\beta+s}} + 4s\varepsilon^2 \log\Big(\frac{2e^3 d}{s m^{1/s}} \Big) \Big\} \Big). \tag{15}$$

Proof. Since the functions $\varphi_{\boldsymbol{j}}$ are orthonormal, $\boldsymbol{Y}_\varepsilon$ is composed of independent Gaussian random variables with common variance equal to ε^2. Thus, the array $\boldsymbol{Y}_\varepsilon$ defined by (12) obeys the Gaussian sequence model studied in [18]. Therefore, using Parseval's theorem and [18, Cor. 6] we obtain that the estimator $\widehat{f}_\varepsilon$ satisfies, for all f,

$$\mathbf{E}_f[\|\widehat{f}_\varepsilon - f\|_2^2] \le \min_{\boldsymbol{t},\boldsymbol{\eta}} \Big(\mathbf{E}_f[\|\widehat{f}_{\boldsymbol{t},\boldsymbol{\eta}} - f\|_2^2] + 4\varepsilon^2 \log(\pi_{\boldsymbol{t},\boldsymbol{\eta}}^{-1})\Big), \tag{16}$$

where the minimum is taken over all $(\boldsymbol{t},\boldsymbol{\eta}) \in \bigcup_{s,m}\{[\![0,\varepsilon^{-2}]\!]^m \times \mathcal{B}_{s,m}^d\}$. Denote by $\boldsymbol{\eta}_0$ the unique element of $\mathcal{B}_{0,1}^d$ for which $\mathrm{supp}(\boldsymbol{\eta}) = \{\varnothing\}$. The corresponding estimator $\widehat{f}_{\boldsymbol{t},\boldsymbol{\eta}_0}$ coincides with the constant function equal to $Y_{\boldsymbol{0}}$ and its risk is bounded by $\varepsilon^2 + L$ for all $f \in \mathcal{F}_{s,m}(\boldsymbol{W}(\beta,L))$. Therefore,

$$\begin{aligned}\sup_{f\in\mathcal{F}_{s,m}(\boldsymbol{W}(\beta,L))} \mathbf{E}_f[\|\widehat{f}_\varepsilon - f\|_2^2] &\le \sup_{f\in\mathcal{F}_{s,m}(\boldsymbol{W}(\beta,L))} \mathbf{E}_f[\|\widehat{f}_{\boldsymbol{t},\boldsymbol{\eta}_0} - f\|_2^2] + 4\varepsilon^2\log(\pi_{\boldsymbol{t},\boldsymbol{\eta}_0}^{-1}) \\ &\le \varepsilon^2 + L + 4\varepsilon^2 \le 6L. \end{aligned}\tag{17}$$

Take now any $f \in \mathcal{F}_{s,m}(\boldsymbol{W}(\beta,L))$, and let $\boldsymbol{\eta}^* \in \tilde{\mathcal{B}} \subseteq \mathcal{B}_{s,m}^d$ be such that $f \in \mathcal{F}_{\boldsymbol{\eta}^*}(\boldsymbol{W}(\beta,L))$. Then it follows from (16) that

$$\begin{aligned}\mathbf{E}_f[\|\widehat{f}_\varepsilon - f\|_2^2] &\le \min_{\boldsymbol{t}\in[\![0,\varepsilon^{-2}]\!]^m} \Big(\mathbf{E}_f[\|\widehat{f}_{\boldsymbol{t},\boldsymbol{\eta}^*} - f\|_2^2] + 4\varepsilon^2\log(\pi_{\boldsymbol{t},\boldsymbol{\eta}^*}^{-1})\Big) \\ &\le \min_{\boldsymbol{t}\in[\![0,\varepsilon^{-2}]\!]^m} \mathbf{E}_f[\|\widehat{f}_{\boldsymbol{t},\boldsymbol{\eta}^*} - f\|_2^2] + 4\varepsilon^2\big(m\log(2\varepsilon^{-2}) + ms\log(2) + \log(e|\mathcal{B}_{s,m}^d|)\big). \end{aligned}\tag{18}$$

Note that for all $d, s \in \mathbb{N}$ such that $s \le d$ we have

$$M_{d,s} = \sum_{\ell=0}^{s}\binom{d}{\ell} \le \left(\frac{ed}{s}\right)^s \quad\text{and}\quad |\mathcal{B}_{s,m}^d| \le \binom{M_{d,s}}{m} \le \left(\frac{eM_{d,s}}{m}\right)^m \le \left(\frac{e^2 d}{sm^{1/s}}\right)^{ms}. \tag{19}$$

Also, we have the following bound on the risk of estimator $\widehat{f}_{\boldsymbol{t},\boldsymbol{\eta}}$ for each $\boldsymbol{\eta} \in \tilde{\mathcal{B}}$ and for an appropriate choice of the bandwidth parameter $\boldsymbol{t} \in [\![0,\varepsilon^{-2}]\!]^m$.

Lemma 1. *Let $\beta > 0$, $L \ge \varepsilon^2$ be such that $\log(\varepsilon^{-2}) \ge (2\beta)^{-1}\log(L)$. Let $\boldsymbol{t} \in [\![0,\varepsilon^{-2}]\!]^m$ be a vector with integer coordinates $t_{V_\ell} = [(L/(3^{|V_\ell|}\varepsilon^2))^{1/(2\beta+|V_\ell|)} \wedge \varepsilon^{-2}]$, $\ell = 1,\dots,m$. Assume that condition (2) holds. Then*

$$\sup_{f\in\mathcal{F}_{\boldsymbol{\eta}}(\boldsymbol{W}(\beta,L))} \mathbf{E}_f[\|\widehat{f}_{\boldsymbol{t},\boldsymbol{\eta}} - f\|_2^2] \le 2C_* 3^{2\beta\wedge s}\, m\, L^{s/(2\beta+s)}\varepsilon^{4\beta/(2\beta+s)}, \qquad \forall\, \boldsymbol{\eta} \in \tilde{\mathcal{B}} \subset \mathcal{B}_{s,m}^d. \tag{20}$$

Proof of this lemma is given in the appendix.

Combining (18) with (19) and (20) yields the following upper bound on the risk of $\widehat{f}_\varepsilon$:

$$\sup_{f\in\mathcal{F}_{s,m}(\boldsymbol{W}(\beta,L))} \mathbf{E}_f[\|\widehat{f}_\varepsilon - f\|_2^2] \le m\left\{C_\beta L^{\frac{s}{2\beta+s}}\varepsilon^{\frac{4\beta}{2\beta+s}} + 4\varepsilon^2\log(2\varepsilon^{-2}) + 4s\varepsilon^2\log\left(\frac{2e^3 d}{sm^{1/s}}\right)\right\}$$

where $C_\beta > 0$ is a constant depending only on β. The assumptions of the theorem guarantee that $\varepsilon^2\log(2\varepsilon^{-2}) \le L^{\frac{s}{2\beta+s}}\varepsilon^{\frac{4\beta}{2\beta+s}}$, so that the desired result follows from (17) and the last display.

The behavior of the estimator $\widehat{f}_\varepsilon$ in the case $\boldsymbol{\Sigma} = \boldsymbol{T}_{\mathcal{A}}$ is described in the next theorem.

Theorem 2. *Assume that $k = \max\{|\ell| : \ell \in \mathcal{A}\} < \varepsilon^{-2}$. Then*

$$\sup_{f\in\mathcal{F}_{s,m}(\boldsymbol{T}_{\mathcal{A}})} \mathbf{E}_f[\|\widehat{f}_\varepsilon - f\|_2^2] \le m\varepsilon^2\left\{(2k+1)^s + 4\log(2\varepsilon^{-2}) + 4s\log\left(\frac{2e^3 d}{sm^{1/s}}\right)\right\}. \tag{21}$$

Proof of Theorem 2 follows the same lines as that of Theorem 1. We take $f \in \mathcal{F}_{s,m}(\boldsymbol{T}_{\mathcal{A}})$, and let $\boldsymbol{\eta}^* \in \tilde{\mathcal{B}} \subseteq \mathcal{B}^d_{s,m}$ be such that $f \in \mathcal{F}_{\boldsymbol{\eta}^*}(\boldsymbol{T}_{\mathcal{A}})$. Let $\boldsymbol{t}^* \in \mathbb{R}^m$ be the vector with all coordinates equal to k. Then the same argument as in (18) yields

$$\mathbf{E}_f[\|\widehat{f}_\varepsilon - f\|_2^2] \le \mathbf{E}_f[\|\widehat{f}_{\boldsymbol{t}^*,\boldsymbol{\eta}^*} - f\|_2^2] + 4\varepsilon^2\big(m\log(2\varepsilon^{-2}) + ms\log(2) + \log(e|\mathcal{B}^d_{s,m}|)\big). \tag{22}$$

We can write $\mathrm{supp}(\boldsymbol{\eta}^*) = \{V_1, \dots, V_m\}$ where $|V_\ell| \le s$. Since the model is parametric, there is no bias term in the expression for the risk on the right hand side of (22) and we have (cf. (29)):

$$\mathbf{E}_f[\|\widehat{f}_{\boldsymbol{t}^*,\boldsymbol{\eta}^*} - f\|_2^2] \le \sum_{\ell=1}^m \sum_{\boldsymbol{j}:\mathrm{supp}(\boldsymbol{j})\subseteq V_\ell} \varepsilon^2 \mathbf{1}_{\{|\boldsymbol{j}|_\infty \le k\}} \le m\varepsilon^2(2k+1)^s.$$

Together with (22), this implies (21).

The bound of Theorem 2 is particularly interesting when k and s are small. For the examples of multilinear and polynomial systems [20, 15] we have $k = 1$. We also note that the result is much better than what can be obtained by using the Lasso. Indeed, consider the simplest case of single atom tensor product model ($m = 1$). Since we do not know s, we need to run the Lasso in the dimension $p = k^d$ and we can only guarantee the rate $\varepsilon^2 \log p = d\varepsilon^2 \log k$, which is linear in the dimension d. If d is very large and $s \ll d$, this is much slower than the rate of Theorem 2.

4. Lower bound

In this section, we prove a minimax lower bound on the risk of any estimator over the class $\mathcal{F}_{s,m}(\boldsymbol{W}(\beta, L))$. We will assume that $\{\varphi_{\boldsymbol{j}}\}$ is the tensor-product trigonometric basis and $\tilde{\mathcal{B}} = \tilde{\mathcal{B}}^d_{s,m}$ where $\tilde{\mathcal{B}}^d_{s,m}$ denotes the set of all $\boldsymbol{\eta} \in \mathcal{B}^d_{s,m}$ such that the sets $V \in \mathrm{supp}(\boldsymbol{\eta})$ are disjoint. Then condition (2) holds with equality and $C_* = 1$. We will split the proof into two steps. First, we establish a lower bound on the minimax risk in the case of known structure $\boldsymbol{\eta}$, *i.e.*, when f belongs to the class $\mathcal{F}_{\boldsymbol{\eta}}(\boldsymbol{W}(\beta, L))$ for some known parameters $\boldsymbol{\eta} \in \tilde{\mathcal{B}}$ and $\beta, L > 0$. We will show that the minimax risk tends to zero with the rate not faster than $m\varepsilon^{4\beta/(2\beta+s)}$. In a second step, we will prove that if $\boldsymbol{\eta}$ is unknown, then the minimax rate is bounded from below by $ms\varepsilon^2(1 + \log(d/(sm^{1/s})))$ if the function f belongs to $\mathcal{F}_{\boldsymbol{\eta}}(\Theta)$ for a set Θ spanned by the tensor products involving only the functions φ_1 and φ_{-1} of various arguments.

4.1. Lower bound for known structure $\boldsymbol{\eta}$

Proposition 1. *Let $\{\varphi_{\boldsymbol{j}}\}$ be the tensor-product trigonometric basis and let s, m, d be positive integers satisfying $d \ge sm$. Assume that $L \ge \varepsilon^2$. Then there exists an absolute constant $C > 0$ such that*

$$\inf_{\widehat{f}} \sup_{f \in \mathcal{F}_{\boldsymbol{\eta}}(\boldsymbol{W}(\beta,L))} \mathbf{E}_f[\|\widehat{f} - f\|_2^2] \ge CmL^{s/(2\beta+s)}\varepsilon^{4\beta/(2\beta+s)}, \qquad \forall\, \boldsymbol{\eta} \in \tilde{\mathcal{B}}^d_{s,m}.$$

Proof. Without loss of generality assume that $m = 1$. We will also assume that $L = 1$ (this is without loss of generality as well, since we can replace ε by $\varepsilon/\sqrt{L}$ and by our assumption this quantity is less than 1). After a renumbering if needed, we can assume that $\boldsymbol{\eta}$ is such that $\boldsymbol{\eta}_V = 1$ for $V = \{1, \dots, s\}$ and $\boldsymbol{\eta}_V = 0$ for $V \ne \{1, \dots, s\}$.

Let t be an integer not smaller than 4. Then, the set I of all multi-indices $\boldsymbol{k} \in \mathbb{Z}^s$ satisfying $|\boldsymbol{k}|_\infty \le t$ is of cardinality $|I| \ge 9$. For any $\boldsymbol{\omega} = (\omega_k, k \in I) \in \{0,1\}^I$, we set $f_{\boldsymbol{\omega}}(\boldsymbol{x}) = \gamma \sum_{\boldsymbol{k}\in I} \omega_{\boldsymbol{k}} \varphi_{\boldsymbol{k}}(x_1, \dots, x_s)$, where $\varphi_{\boldsymbol{k}}(x_1, \dots, x_s) = \prod_{j=1}^s \varphi_{k_j}(x_j)$, $\boldsymbol{k} = (k_1, \dots, k_s)$, is

an element of the tensor-product trigonometric basis and $\gamma > 0$ is a parameter to be chosen later. In view of the orthonormality of the basis functions $\varphi_{\boldsymbol{k}}$, we have

$$\|f_{\boldsymbol{\omega}}\|_2^2 = \gamma^2 |\boldsymbol{\omega}|_1, \quad \forall\, \boldsymbol{\omega} \in \{0,1\}^I. \tag{23}$$

Therefore, we have $\sum_{\boldsymbol{k}} |\boldsymbol{k}|_\infty^{2\beta} \theta_{\boldsymbol{k}}[f_{\boldsymbol{\omega}}]^2 \le t^{2\beta} \|f_{\boldsymbol{\omega}}\|_2^2 \le t^{2\beta}\gamma^2 (2t+1)^s \le \gamma^2 (2t+1)^{2\beta+s}$. Thus, the condition $\gamma^2(2t+1)^{2\beta+s} \le 1$ ensures that all the functions $f_{\boldsymbol{\omega}}$ belong to $W(\beta, 1)$.

Furthermore, for two vectors $\boldsymbol{\omega}, \boldsymbol{\omega}' \in \{0,1\}^I$ we have $\|f_{\boldsymbol{\omega}} - f_{\boldsymbol{\omega}'}\|_2^2 = \gamma^2 |\boldsymbol{\omega} - \boldsymbol{\omega}'|_1$. Note that the entries of the vectors $\boldsymbol{\omega}, \boldsymbol{\omega}'$ are either 0 or 1, therefore the ℓ_1 distance between these vectors coincides with the Hamming distance. According to the Varshamov-Gilbert lemma [28, Lemma 2.9], there exists a set $\Omega \subset \{0,1\}^I$ of cardinality at least $2^{|I|/8}$ such that it contains the zero element and the pairwise distances $|\boldsymbol{\omega} - \boldsymbol{\omega}'|_1$ are at least $|I|/8$ for any pair $\boldsymbol{\omega}, \boldsymbol{\omega}' \in \Omega$.

We can now apply Theorem 2.7 from [28] that asserts that if, for some $\tau > 0$, we have $\min_{\boldsymbol{\omega},\boldsymbol{\omega}' \in \Omega} \|f_{\boldsymbol{\omega}} - f_{\boldsymbol{\omega}'}\|_2 \ge 2\tau > 0$, and

$$\frac{1}{|\Omega|} \sum\nolimits_{\boldsymbol{\omega} \in \Omega} \mathcal{K}(\mathbf{P}_{f_{\boldsymbol{\omega}}}, \mathbf{P}_0) \le \frac{\log |\Omega|}{16}, \tag{24}$$

where $\mathcal{K}(\cdot,\cdot)$ denotes the Kullback-Leibler divergence, then $\inf_{\widehat{f}} \max_{\boldsymbol{\omega} \in \Omega} \mathbf{E}_{f_{\boldsymbol{\omega}}}[\|\widehat{f} - f_{\boldsymbol{\omega}}\|_2^2] \ge c'\tau^2$ for some absolute constant $c' > 0$. In our case, we set $\tau = \gamma\sqrt{|I|/32}$. Combining (23) and the fact that the Kullback-Leibler divergence between the Gaussian measures $\mathbf{P}_f$ and $\mathbf{P}_g$ is given by $\frac{1}{2}\varepsilon^{-2}\|f - g\|_2^2$, we obtain $\frac{1}{|\Omega|}\sum_{\boldsymbol{\omega}\in\Omega} \mathcal{K}(\mathbf{P}_{f_{\boldsymbol{\omega}}}, \mathbf{P}_0) \le \frac{1}{2}\varepsilon^{-2}\gamma^2 |I|$. If $\gamma^2 \le (\log 2)\varepsilon^2/64$, then (24) is satisfied and $\tau^2 = \gamma^2(2t+1)^s/32$ is a lower bound on the rate of convergence of the minimax risk.

To finish the proof, it suffices to choose $t \in \mathbb{N}$ and $\gamma > 0$ satisfying the following three conditions: $t \ge 4$, $\gamma^2 \le (2t+1)^{-2\beta-s}$ and $\gamma^2 \le \varepsilon^2 \log(2)/64$. For the choice $\gamma^{-2} = (2t+1)^{2\beta+s} + \varepsilon^{-2} 64/\log(2)$ and$t = [4\varepsilon^{-2/(2\beta+s)}]$ all these conditions are satisfied and $\tau^2 \ge c_1 \varepsilon^{4\beta/(2\beta+s)}$ for some absolute positive constant c_1.

4.2. Lower bound for unknown structure $\boldsymbol{\eta}$

Proposition 2. *Let the assumptions of Proposition 1 be satisfied. Then there exists an absolute constant $C' > 0$ such that*

$$\inf_{\widehat{f}} \sup_{f \in \mathcal{F}_{s,m}(\boldsymbol{W}(\beta,L))} \mathbf{E}_f[\|\widehat{f} - f\|_2^2] \ge C' \min\left\{ L,\, ms\varepsilon^2 \log\left(\frac{8\,d}{sm^{1/s}}\right)\right\}.$$

Proof. We use again Theorem 2.7 in [28] but with a choice of the finite subset of $\mathcal{F}_{s,m}(\boldsymbol{W}(\beta, L))$ different from that of Proposition 1. First, we introduce some additional notation. For every triplet $(m, s, d) \in \mathbb{N}_*^3$ satisfying $ms \le d$, let $\mathcal{P}_{s,m}^d$ be the set of collections $\pi = \{V_1, \dots, V_m\}$ such that each $V_\ell \subseteq \{1, \dots, d\}$ has exactly s elements and V_ℓ's are pairwise disjoint. We consider $\mathcal{P}_{s,m}^d$ as a metric space with the distance $\rho(\pi, \pi') = \frac{1}{m}\sum_{\ell=1}^m \mathbf{1}(V_\ell \notin \{V_1', \dots, V_m'\}) = \frac{|\pi \Delta \pi'|}{2m}$, where $\pi' = \{V_1', \dots, V_m'\} \in \mathcal{P}_{s,m}^d$. It is easy to see that $\rho(\cdot,\cdot)$ is a distance bounded by 1.

For any $\vartheta \in (0,1)$, let $\mathcal{N}_{s,m}^d(\vartheta)$ denote the logarithm of the packing number, *i.e.*, the logarithm of the largest integer K such that there are K elements $\pi^{(1)}, \dots, \pi^{(K)}$ of $\mathcal{P}_{s,m}^d$ satisfying $\rho(\pi^{(k)}, \pi^{(k')}) \ge \vartheta$. To each $\pi^{(k)}$ we associate a family of functions $\mathcal{U} = \{f_{k,\boldsymbol{\omega}} : \boldsymbol{\omega} \in \{-1,1\}^{ms},\, k = 1, \dots, K\}$ defined by

$$f_{k,\boldsymbol{\omega}}(\boldsymbol{x}) = \frac{\tau}{\sqrt{m}} \sum_{V \in \pi^{(k)}} \varphi_{\boldsymbol{\omega},V}(\boldsymbol{x}_V),$$

where $\tau = (1/4)\min\big(\varepsilon\sqrt{ms\log 2 + \log K}, \sqrt{L}\big)$ and $\varphi_{\boldsymbol{\omega},V}(\boldsymbol{x}_V) = \prod_{j\in V}\varphi_{\omega_j}(x_j)$. Using that $\{\varphi_{\boldsymbol{j}}\}$ is the tensor-product trigonometric basis it is easy to see that each $f_{k,\boldsymbol{\omega}}$ belongs to $\mathcal{F}_{s,m}(\boldsymbol{W}(\beta, L))$. Next, $|\mathcal{U}| = 2^{ms}K$ and, for any $f_{k,\boldsymbol{\omega}} \in \mathcal{U}$, the Kullback-Leibler divergence between $\mathbf{P}_{f_{k,\boldsymbol{\omega}}}$ and $\mathbf{P}_0$ is equal to $\mathcal{K}(\mathbf{P}_{f_{k,\boldsymbol{\omega}}}, \mathbf{P}_0) = \frac{1}{2}\varepsilon^{-2}\|f_{k,\boldsymbol{\omega}}\|_2^2 = \frac{\varepsilon^{-2}\tau^2}{2} \le \frac{\log|\mathcal{U}|}{16}$. Furthermore, the functions $f_{k,\boldsymbol{\omega}}$ are not too close to each other. Indeed, since $\{\varphi_{\boldsymbol{j}}\}$ is the tensor-product trigonometric basis we get that, for all $f_{k,\boldsymbol{\omega}}, f_{k',\boldsymbol{\omega}'} \in \mathcal{U}$,

$$\begin{aligned}\|f_{k,\boldsymbol{\omega}} - f_{k',\boldsymbol{\omega}'}\|_2^2 &= \tau^2 m^{-1}\Big(2m - \sum_{V\in\pi^{(k)}}\sum_{V'\in\pi^{(k')}}\int_{[0,1]^d}\varphi_{\boldsymbol{\omega},V}(\boldsymbol{x}_V)\varphi_{\boldsymbol{\omega}',V'}(\boldsymbol{x}_{V'})\,d\boldsymbol{x}\Big)\\ &= \tau^2\Big(2 - \frac{1}{m}\sum_{V\in\pi^{(k)}}\sum_{V'\in\pi^{(k')}}\mathbf{1}(V = V')\Big) = 2\tau^2\rho(\pi^{(k)}, \pi^{(k')}) \ge 2\vartheta\tau^2.\end{aligned}$$

These remarks and Theorem 2.7 in [28] imply that

$$\inf_{\widehat{f}}\sup_{f\in\mathcal{U}}\mathbf{E}_f[\|\widehat{f} - f\|_2^2] \ge c_3\vartheta\tau^2 = \frac{c_3\vartheta}{16}\min\Big\{L, \varepsilon^2(ms\log 2 + \log K)\Big\} \tag{25}$$

for some absolute constant $c_3 > 0$. Assume first that $d < 4sm^{1/s}$. Then $ms\log 2 \ge \frac{ms}{5}\log\big(\frac{8d}{sm^{1/s}}\big)$ and the result of the proposition is straightforward. If $d \ge 4sm^{1/s}$ we fix $\vartheta = 1/8$ and use the following lemma (cf. the Appendix for a proof) to bound $\log K = \mathcal{N}_{s,m}^d(\vartheta)$ from below.

Lemma 2. *For any $\vartheta \in (0, 1/8]$ we have $\mathcal{N}_{s,m}^d(\vartheta) \ge -m\log\big(\frac{8e^{7/8}s^{1/2}}{7}\big) + \frac{ms}{3}\log\big(\frac{d}{sm^{1/s}}\big)$.*

This yields

$$ms\log 2 + \mathcal{N}_{s,m}^d(\vartheta) \ge \frac{ms}{3}\log\Big(\frac{8d}{sm^{1/s}}\Big) - m\log\big((8/7)e^{7/8}s^{1/2}\big). \tag{26}$$

It is easy to check that $m\log\big((8/7)e^{7/8}s^{1/2}\big) \le 1.01ms$, while for $d \ge 4sm^{1/s}$ we have $\frac{1}{3}\log\Big(\frac{8d}{sm^{1/s}}\Big) \ge 1.15$. Combining these inequalities with (25) and (26) we get the result.

5. Discussion and outlook

We presented a new framework, called the compound functional model, for performing various statistical tasks such as prediction, estimation and testing in the context of high dimension. We studied the problem of estimation in this model from a minimax point of view when the data are generated by a Gaussian process. We established upper and lower bounds on the minimax risk that match up to a multiplicative constant. These bounds are nonasymptotic and are attained adaptively with respect to the macroscopic and microscopic sparsity parameters m and s, as well as to the complexity of the atoms of the model. In particular, we improve in several aspects upon the existing results for the sparse additive model, which is a special case of the compound functional model (only for this case the rates were previously explicitly treated in the literature):

- The exact expression for the optimal rate that we obtain reveals that the existing methods for the sparse additive model based on penalized least squares techniques have logarithmically suboptimal rates.
- On the difference from most of the previous work, we do not require restricted isometry type assumptions on the subspaces of the additive model; we need only a much weaker one-sided condition (2). Possible extensions to general compound model based on the existing literature would again suffer from the rate suboptimality and require such type of extra conditions.

– When specialized to the sparse additive model, our results are adaptive with respect to the smoothness of the atoms, while all the previous work about the rates considered the smoothness (or the reproducing kernel) as given in advance.

For the general compound model, the main difficulty is in the proof of the lower bounds of the order $ms\varepsilon^2\log(d/(sm^{1/s}))$ that are not covered by the standard tools such as the Varshamov-Gilbert lemma or k-selection lemma. Therefore, we developed here new tools for the lower bounds that can be of independent interest.

An important issue that remained out of scope of the present work but is undeniably worth studying is the possibility of achieving the minimax rates by computationally tractable procedures. Clearly, the complexity of exact computation of the procedure described in Section 3 scales as $\varepsilon^{-2m}2^{M_{d,s}}$, which is prohibitively large for typical values of d, s and m. It is possible, however, to approximate our estimator by using a Markov Chain Monte-Carlo (MCMC) algorithm similar to that of [23, 24]. The idea is to begin with an initial state $(\boldsymbol{t}_0, \boldsymbol{\eta}_0)$ and to randomly generate a new candidate $(\boldsymbol{u}, \boldsymbol{\zeta})$ according to the distribution $q(\cdot|\boldsymbol{t}_0, \boldsymbol{\eta}_0)$, where $q(\cdot|\cdot)$ is a given Markov kernel. Then, a Bernoulli random variable ξ with probability of the output 1 equal to $\alpha = 1 \wedge \frac{\widehat{\pi}(\boldsymbol{u},\boldsymbol{\zeta})}{\widehat{\pi}(\boldsymbol{t},\boldsymbol{\eta})}\frac{q(\boldsymbol{t},\boldsymbol{\eta}|\boldsymbol{u},\boldsymbol{\zeta})}{q(\boldsymbol{u},\boldsymbol{\zeta}|\boldsymbol{t},\boldsymbol{\eta})}$ is drawn and a new state $(\boldsymbol{t}_1, \boldsymbol{\eta}_1) = \xi\cdot(\boldsymbol{u},\boldsymbol{\zeta}) + (1-\xi)\cdot(\boldsymbol{t}_0,\boldsymbol{\eta}_0)$ is defined. This procedure is repeated K times producing thus a realization $\{(\boldsymbol{t}_k, \boldsymbol{\eta}_k); k = 0, \ldots, K\}$ of a reversible Markov chain. Then, the average value $\frac{1}{K}\sum_{k=1}^K \widehat{\boldsymbol{\theta}}_{\boldsymbol{t}_k,\boldsymbol{\eta}_k}$ provides an approximation to the estimator $\widehat{f}_\varepsilon$ defined in Section 3.

If s and m are small and $q(\cdot|\boldsymbol{t},\boldsymbol{\eta}')$ is such that all the mass of this distribution is concentrated on the nearest neighbors of the $\boldsymbol{\eta}'$ in the hypercube of $2^{M_{d,s}}$ all possible $\boldsymbol{\eta}$'s, then the computations can be performed in a polynomial time. For example, if $s = 2$, *i.e.*, if we allow only pairwise interactions, each step of the algorithm requires $\sim \varepsilon^{-2m}d^2$ computations, where the factor ε^{-2m} can be reduced to a power of $\log(\varepsilon^{-2})$ by a suitable modification of the estimator. How fast such MCMC algorithms converge to our estimator and what is the most appealing choice for the Markov kernel $q(\cdot|\cdot)$ are challenging open questions for future research.

Appendix

A. Proof of Lemma 1

Let $\boldsymbol{\eta} \in \tilde{\mathcal{B}}$ be such that $f \in \mathcal{F}_{\boldsymbol{\eta}}(\boldsymbol{W}(\beta, L))$ and $\mathrm{supp}(\boldsymbol{\eta}) = \{V_1, \ldots, V_m\}$ where $|V_\ell| \le s$. Then there exist a constant $\bar{f}$ and m functions $f_1, \ldots, f_m$ such that $f_\ell \in W_{V_\ell}(\beta, L)$, $\ell = 1, \ldots, m$, and $f = \bar{f} + f_1 + \ldots + f_m$. Set $\theta_{\boldsymbol{j},\ell} = \theta_{\boldsymbol{j}}[f_\ell]$, $(\boldsymbol{j}, \ell) \in \mathbb{Z}^d \times \{1, \ldots, m\}$. Using the notation $t_\ell = t_{V_\ell}$ and

$$\mathcal{J} = \{\boldsymbol{j} \in \mathbb{Z}^d : \exists\, \ell \in \{1, \ldots, m\} \text{ such that } \mathrm{supp}(\boldsymbol{j}) \subseteq V_\ell \text{ and } |\boldsymbol{j}|_\infty \le t_\ell\}$$

we get

$$\begin{aligned}
\widehat{f}_{\boldsymbol{t},\boldsymbol{\eta}} - f &= (Y_{\mathbf{0}} - \bar{f})\varphi_{\mathbf{0}} + \sum_{\boldsymbol{j}\in\mathcal{J}\setminus\mathbf{0}} \widehat{\theta}_{\boldsymbol{t},\boldsymbol{\eta},\boldsymbol{j}}\varphi_{\boldsymbol{j}} - \sum_{\ell=1}^m \sum_{\boldsymbol{j}\in\mathbb{Z}^d\setminus\mathbf{0}} \theta_{\boldsymbol{j},\ell}\varphi_{\boldsymbol{j}}\mathbf{1}_{\{\mathrm{supp}(\boldsymbol{j})\subseteq V_\ell; |\boldsymbol{j}|_\infty \le t_\ell\}} \\
&\quad - \sum_{\ell=1}^m \sum_{\boldsymbol{j}\in\mathbb{Z}^d} \theta_{\boldsymbol{j},\ell}\varphi_{\boldsymbol{j}}\mathbf{1}_{\{\mathrm{supp}(\boldsymbol{j})\subseteq V_\ell; |\boldsymbol{j}|_\infty > t_\ell\}} \\
&= \sum_{\boldsymbol{j}\in\mathcal{J}} \varepsilon\xi_{\boldsymbol{j}}\varphi_{\boldsymbol{j}} - \sum_{\ell=1}^m \sum_{\boldsymbol{j}\in\mathbb{Z}^d} \theta_{\boldsymbol{j},\ell}\varphi_{\boldsymbol{j}}\mathbf{1}_{\{\mathrm{supp}(\boldsymbol{j})\subseteq V_\ell; |\boldsymbol{j}|_\infty > t_\ell\}},
\end{aligned}$$

where $(\xi_{\boldsymbol{j}})_{\boldsymbol{j}\in\mathbb{Z}^d}$ are i.i.d. Gaussian random variables with zero mean and variance one. In view of the bias-variance decomposition and (2), we bound the risk of $\widehat{f}_{\boldsymbol{t},\boldsymbol{\eta}}$ as follows:

$$\begin{aligned}\mathbf{E}_f[\|\widehat{f}_{\boldsymbol{t},\boldsymbol{\eta}}-f\|_2^2] &\le \sum_{\boldsymbol{j}\in\mathcal{J}}\varepsilon^2 + C_*\sum_{\ell=1}^m\sum_{\boldsymbol{j}\in\mathbb{Z}^d}\theta_{\boldsymbol{j},\ell}^2\mathbf{1}_{\{\mathrm{supp}(\boldsymbol{j})\subseteq V_\ell;|\boldsymbol{j}|_\infty>t_\ell\}}\\ &\le \sum_{\ell=1}^m\sum_{\boldsymbol{j}\in\mathbb{Z}^d}\varepsilon^2\mathbf{1}_{\{\mathrm{supp}(\boldsymbol{j})\subseteq V_\ell;|\boldsymbol{j}|_\infty\le t_\ell\}} + C_*\sum_{\ell=1}^m\sum_{\boldsymbol{j}\in\mathbb{Z}^d}\theta_{\boldsymbol{j},\ell}^2\mathbf{1}_{\{\mathrm{supp}(\boldsymbol{j})\subseteq V_\ell;|\boldsymbol{j}|_\infty>t_\ell\}}\\ &\le m\max_{\ell=1,\dots,m}\sum_{\boldsymbol{j}:\mathrm{supp}(\boldsymbol{j})\subseteq V_\ell}\Big(\varepsilon^2\mathbf{1}_{\{|\boldsymbol{j}|_\infty\le t_\ell\}} + C_*\theta_{\boldsymbol{j},\ell}^2\mathbf{1}_{\{|\boldsymbol{j}|_\infty>t_\ell\}}\Big). \end{aligned}\tag{27}$$

In the right-hand side of (27), the first summand is the variance term, while the second summand is the (squared) bias term of the risk. We bound these two terms separately. For the bias contribution to the risk, we find:

$$\begin{aligned}\sum_{\boldsymbol{j}:\mathrm{supp}(\boldsymbol{j})\subseteq V_\ell}\theta_{\boldsymbol{j},\ell}^2\mathbf{1}_{\{|\boldsymbol{j}|_\infty>t_\ell\}} &\le (t_\ell+1)^{-2\beta}\sum_{\boldsymbol{j}:\mathrm{supp}(\boldsymbol{j})\subseteq V_\ell}|\boldsymbol{j}|_\infty^{2\beta}\theta_{\boldsymbol{j},\ell}^2\\ &\le L(t_\ell+1)^{-2\beta}\\ &\le L\Big(\varepsilon^{4\beta}\vee(L/(3^{|V_\ell|}\varepsilon^2))^{-2\beta/(2\beta+|V_\ell|)}\Big)\\ &\le 3^{2\beta\wedge s}\big(L\varepsilon^{4\beta}\vee L^{s/(2\beta+s)}\varepsilon^{4\beta/(2\beta+s)}\big).\end{aligned}\tag{28}$$

If $t_\ell\ge 1$, then the variance contribution to the risk is bounded as follows:

$$\sum_{\boldsymbol{j}:\mathrm{supp}(\boldsymbol{j})\subseteq V_\ell}\varepsilon^2\mathbf{1}_{\{|\boldsymbol{j}|_\infty\le t_\ell\}} = \varepsilon^2(2t_\ell+1)^{|V_\ell|}\le\varepsilon^2(3t_\ell)^{|V_\ell|}\le 3^{2\beta\wedge s}L^{|V_\ell|/(2\beta+|V_\ell|)}\varepsilon^{4\beta/(2\beta+|V_\ell|)},\tag{29}$$

where we have used that $t_\ell\le(L/3^{|V_\ell|}\varepsilon^2)^{1/(2\beta+|V_\ell|)}$ and $|V_\ell|\le s$. Finally, note that condition $\log(\varepsilon^{-2})\ge(2\beta)^{-1}\log(L)$ implies that $L\varepsilon^{4\beta}<L^{s/(2\beta+s)}\varepsilon^{4\beta/(2\beta+s)}$ in (28). Thus, inequality (27) together with (28) and (29) yields the lemma in the case $t_\ell\ge 1$. If $t_\ell<1$, *i.e.*, $t_\ell=0$, the same arguments imply that the bias is bounded by L and the variance is bounded by ε^2. Since $L\ge\varepsilon^2$, the sum at the right-hand side of (27) is bounded by $(1+C_*)L$. One can check that t_ℓ equals 0 only if $L<3^s\varepsilon^{-2}$, and in this case $L=L^{s/(2\beta+s)}L^{2\beta/(2\beta+s)}\le L^{s/(2\beta+s)}\varepsilon^{4\beta/(2\beta+s)}3^{2\beta s/(2\beta+s)}\le 3^{2\beta\wedge s}L^{s/(2\beta+s)}\varepsilon^{-4\beta/(2\beta+s)}$. This completes the proof.

B. Proof of Lemma 2

Prior to presenting a proof of Lemma 2, we need an additional result.

Lemma 3. *For a triplet $(m,s,d)\in\mathbb{N}_*^3$ satisfying $ms\le d$, let $\mathcal{P}_{s,m}^d$ be the set of all collections $\pi=\{A_1,\dots,A_m\}$ with $A_i\subseteq\{1,\dots,d\}$ such that $|A_i|=s$ for all i and $A_i\cap A_k=\varnothing$ for $i\ne k$. Then*

$$s^{-(m-1)/2}\left(\frac{d}{sm^{1/s}}\right)^{ms}\le|\mathcal{P}_{s,m}^d|\le\left(\frac{e^2d}{sm^{1/s}}\right)^{ms}.$$

Proof. Using standard combinatorial arguments we find

$$\mathcal{P}_{s,m}^d=\binom{d}{ms}\frac{(ms)!}{(s!)^m m!}\ge\left(\frac{d}{ms}\right)^{ms}\frac{(ms)!}{(s!)^m m!}.$$

If either $s = 1$ or $m = 1$ then $(ms)! = (s!)^m m!$ and the lower bound stated in the lemma is obviously true. Assume now that $m \geq 2$ and $s \geq 2$. Recall that according to the Stirling formula, for every $n \in \mathbb{N}$, $\sqrt{2\pi n}(n/e)^n \leq n! \leq \sqrt{2\pi n}(n/e)^n e^{1/12n}$. Therefore,

$$\begin{aligned}\frac{(ms)!}{m!(s!)^m} &\geq \frac{\sqrt{2\pi ms}(ms/e)^{ms}}{\sqrt{2\pi m}(m/e)^m e^{\frac{1}{12m}+\frac{m}{12s}}(\sqrt{2\pi s})^m (s/e)^{ms}} \\ &= \frac{m^{ms}}{m^m}\Big[e^{1-\frac{1}{12m^2}-\frac{1}{12s}}/\sqrt{2\pi}\Big]^m s^{-(m-1)/2}\,.\end{aligned}$$

Since the expression in square brackets in the last display is greater than 1 we obtain the desired lower bound on $|\mathcal{P}^d_{s,m}|$. The upper bound follows from (19) and the fact that $|\mathcal{P}^d_{s,m}| \leq |\mathcal{B}^d_{s,m}|$.

Proof of Lemma 2. Consider first the case $m = 1$. The set $\mathcal{P}^d_{s,1}$ is the collection of all subsets of $\{1, \ldots, d\}$ having exactly s elements. The distance ρ is then 0 if the sets coincide and 1 otherwise. Thus, we need to bound from below the logarithm of $|\mathcal{P}^d_{s,1}| = \binom{d}{s}$. It is enough to use the inequality $\log\binom{d}{s} \geq s\log(d/s)$.

Assume now that $m \geq 2$. Since $\pi^{(1)}, \ldots, \pi^{(K)}$ is a *maximal* ϑ-separated set of $\mathcal{P}^d_{s,m}$ we have that $\mathcal{P}^d_{s,m}$ is covered by the union of ρ-balls $B(\pi^{(k)}, \vartheta)$ of radius ϑ centered at $\pi^{(k)}$'s. Therefore,

$$|\mathcal{P}^d_{s,m}| \leq \sum_{k=1}^K |B(\pi^{(k)}, \vartheta)|.$$

It is clear that the cardinality of the ball $|B(\pi^{(k)}, \vartheta)|$ does not depend on $\pi^{(k)}$. This yields

$$K \geq \frac{|\mathcal{P}^d_{s,m}|}{|B(\pi^0, \vartheta)|}$$

where $\pi^0 = \{A^0_1, \ldots, A^0_m\}$ such that $A^0_i = \{(i-1)s + 1, \ldots, is\}$. We have already established a lower bound on $|\mathcal{P}^d_{s,m}|$ in Lemma 3. We now find an upper bound on the cardinality of the ball $B(\pi^0, \vartheta)$. Let m_ϑ be the smallest integer greater than or equal to $(1-\vartheta)m$. Consider some $\pi = \{A_1, \ldots, A_m\} \in \mathcal{P}^d_{s,m}$. Note that $\pi \in B(\pi^0, \vartheta)$ if and only if

$$\sum_{i=1}^m \mathbf{1}(A_i \in \{A^0_1, \ldots, A^0_m\}) \geq m_\vartheta.$$

This means that there are m_ϑ indexes $i_1, \ldots, i_{m_\vartheta}$ such that the m_ϑ sets $A^0_{i_j}$ are in π and the remaining $m - m_\vartheta$ elements of π are chosen as an arbitrary collection of $m - m_\vartheta$ disjoint subsets of $\{1, \ldots, d\} \setminus \bigcup_{j=1}^{m_\vartheta} A^0_{i_j}$, each of which is of cardinality s. There are $\binom{m}{m_\vartheta}$ ways of choosing $\{i_1, \ldots, i_{m_\vartheta}\}$ and once this choice is fixed, there are $|\mathcal{P}^{d-sm_\vartheta}_{s,m-m_\vartheta}|$ ways of choosing the remaining parts. Thus, $|B(\pi^0, \vartheta)| \leq \binom{m}{m_\vartheta}|\mathcal{P}^{d-sm_\vartheta}_{s,m-m_\vartheta}|$. Using this inequality and Lemma 3 we obtain

$$\begin{aligned}K &\geq \frac{|\mathcal{P}^d_{s,m}|}{\binom{m}{m_\vartheta}|\mathcal{P}^{d-sm_\vartheta}_{s,m-m_\vartheta}|} \geq \frac{s^{-(m-1)/2}\Big(\frac{d}{sm^{1/s}}\Big)^{ms}}{\Big(\frac{em}{m_\vartheta}\Big)^{m_\vartheta}\Big(\frac{e^2(d-sm_\vartheta)}{sm_\vartheta^{1/s}}\Big)^{s(m-m_\vartheta)}} \\ &\geq s^{-(m-1)/2}e^{2s(m_\theta - m) - m_\theta}\Big(\frac{d}{sm^{1/s}}\Big)^{sm_\vartheta}\Big(\frac{m_\vartheta}{m}\Big)^m\Big(1 + \frac{sm_\theta}{d - sm_\theta}\Big)^{s(m-m_\theta)}.\end{aligned}$$

Since $\vartheta \leq 1/8$ we have $m_\vartheta \geq m\big(1-\vartheta\big) \geq 7m/8$ and after some algebra we deduce from the previous display that

$$\log(K) \geq -\frac{ms}{4} - m\log\left(\frac{8e^{7/8}s^{1/2}}{7}\right) + \frac{7ms}{8}\log\left(\frac{d}{sm^{1/s}}\right). \tag{30}$$

Assume first that $s \geq 3$. Since also $m \geq 2$ we have $2^{1-1/3} \leq m^{1-1/s} = \frac{ms}{sm^{1/s}} \leq \frac{d}{sm^{1/s}}$. Hence

$$\log(K) \geq -\frac{ms}{4\log(2^{2/3})} \log\left(\frac{d}{sm^{1/s}}\right) - m\log\left(\frac{8e^{7/8}s^{1/2}}{7}\right) + \frac{7ms}{8}\log\left(\frac{d}{sm^{1/s}}\right)$$

and the result of the lemma follows from the inequality $7/8 - 1/\log(2^{8/3}) \geq 1/3$. It remains to consider the case $s \in \{1, 2\}$. If the right-hand side of the inequality of the lemma is negative, then the result is trivial. If the right-hand side is positive, we have $\log(8e^{7/8}/7) \leq \frac{2}{3}\log\left(\frac{d}{sm^{1/s}}\right)$ for $s \in \{1, 2\}$. Therefore, from (30) we obtain

$$\log(K) \geq -\frac{ms}{6\log(8e^{7/8}/7)} \log\left(\frac{d}{sm^{1/s}}\right) - m\log\left(\frac{8e^{7/8}s^{1/2}}{7}\right) + \frac{7ms}{8}\log\left(\frac{d}{sm^{1/s}}\right)$$

and the result of the lemma follows from the inequality $7/8 - (6\log(8e^{7/8}/7))^{-1} \geq 1/2$.

Acknowledgments

The authors acknowledge the support of the French Agence Nationale de la Recherche (ANR) under the grant PARCIMONIE.

References

1. Francis Bach. High-dimensional non-linear variable selection through hierarchical kernel learning. Technical report, arXiv:0909.0844, 2009.
2. Karine Bertin and Guillaume Lecué. Selection of variables and dimension reduction in high-dimensional non-parametric regression. *Electron. J. Stat.*, 2:1224–1241, 2008.
3. Peter J. Bickel, Ya'acov Ritov, and Alexandre B. Tsybakov. Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.*, 37(4):1705–1732, 2009.
4. Peter J. Bickel, Ya'acov Ritov, and Alexandre B. Tsybakov. Hierarchical selection of variables in sparse high-dimensional regression. In *Borrowing strength: theory powering applications—a Festschrift for Lawrence D. Brown*, volume 6 of *Inst. Math. Stat. Collect.*, pages 56–69. Inst. Math. Statist., Beachwood, OH, 2010.
5. Lawrence D. Brown and Mark G. Low. Asymptotic equivalence of nonparametric regression and white noise. *Ann. Statist.*, 24(6):2384–2398, 1996.
6. Laëtitia Comminges and Arnak S. Dalalyan. Tight conditions for consistent variable selection in high dimensional nonparametric regression. *Journal of Machine Learning Research - Proceedings Track*, 19:187–206, 2011.
7. Laëtitia Comminges and Arnak S. Dalalyan. Tight conditions for consistency of variable selection in the context of high dimensionality. *Ann. Statist.*, (in press), 2012.
8. Dong Dai, Philippe Rigollet, and Tong Zhang. Deviation optimal learning using greedy Q-aggregation. *Ann. Stat.*, 40(3):1878–1905, 2012.
9. Arnak Dalalyan and Markus Reiß. Asymptotic statistical equivalence for scalar ergodic diffusions. *Probab. Theory Related Fields*, 134(2):248–282, 2006.
10. Arnak S. Dalalyan and Alexandre B. Tsybakov. Aggregation by exponential weighting, sharp PAC-Bayesian bounds and sparsity. *Machine Learning*, 72(1-2):39–61, 2008.
11. Arnak S. Dalalyan and Alexandre B. Tsybakov. Sparse regression learning by aggregation and Langevin Monte-Carlo. *J. Comput. System Sci.*, 78(5):1423–1443, 2012.
12. Ghislaine Gayraud and Yuri Ingster. Detection of sparse variable functions. *Electron. J. Statist.*, 6:1409–1448, 2012.
13. Georgi K. Golubev, Michael Nussbaum, and Harrison H. Zhou. Asymptotic equivalence of spectral density estimation and Gaussian white noise. *Ann. Statist.*, 38(1):181–214, 2010.
14. Yu. Ingster and O. Lepski. Multichannel nonparametric signal detection. *Math. Methods Statist.*, 12(3):247–275, 2003.

15. Vassilis Kekatos and Georgios B. Giannakis. Sparse Volterra and polynomial regression models: Recoverability and estimation. *IEEE Transactions on Signal Processing*, 59(12):5907–5920, 2011.
16. Vladimir Koltchinskii and Ming Yuan. Sparsity in multiple kernel learning. *Ann. Statist.*, 38 (6):3660–3695, 2010.
17. Guillaume Lecué and Shahar Mendelson. On the optimality of the aggregate with exponential weights for low temperatures. *Bernoulli*, (to appear), 2012.
18. Gilbert Leung and Andrew R. Barron. Information theory and mixing least-squares regressions. *IEEE Trans. Inform. Theory*, 52(8):3396–3410, 2006.
19. Lukas Meier, Sara van de Geer, and Peter Bühlmann. High-dimensional additive modeling. *Ann. Statist.*, 37(6B):3779–3821, 2009.
20. Bobak Nazer and Robert Nowak. Sparse interactions: Identifying high-dimensional multilinear systems via compressed sensing. In *Proc. of the Allerton Conf.*, Monticello, IL, 2010.
21. Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *J. Mach. Learn. Res.*, 13:389–427, 2012.
22. Markus Reiß. Asymptotic equivalence for nonparametric regression with multivariate and random design. *Ann. Statist.*, 36(4):1957–1982, 2008.
23. Philippe Rigollet and Alexandre B. Tsybakov. Exponential Screening and optimal rates of sparse estimation. *Ann. Statist.*, 39(2):731–771, 2011.
24. Philippe Rigollet and Alexandre B. Tsybakov. Sparse estimation by exponential weighting. *Statist. Sci.*, 27(4):558–575, 2012.
25. Lorenzo Rosasco, Silvia Villa, Sofia Mosci, Matteo Santoro, and Alessandro Verri. Nonparametric sparsity and regularization. Technical report, arXiv:1208.2572v1, 2009.
26. Charles J. Stone. Additive regression and other nonparametric models. *Ann. Statist.*, 13(2): 689–705, 1985.
27. Taiji Suzuki. PAC-Bayesian bound for gaussian process regression and multiple kernel additive model. In *COLT*, arXiv:1102.3616v1 [math.ST], 2012.
28. Alexandre B. Tsybakov. *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York, 2009.
29. Sara van de Geer and Peter Bühlmann. *Statistics for High-Dimensional Data*. Springer Texts in Statistics. Springer-Verlag, New York, second edition, 2011.