Estimating composite functions by model selection

Yannick Baraud, Lucien Birgé

Estimating composite functions by model selection

Revised version

Yannick Baraud

Université de Nice Sophia-Antipolis
Laboratoire J-A Dieudonné

Lucien Birgé

Université Paris VI
Laboratoire de Probabilités et Modèles Aléatoires
U.M.R. C.N.R.S. 7599

July 1, 2012

Abstract

We consider the problem of estimating a function s on $[-1,1]^k$ for large values of k by looking for some best approximation of s by composite functions of the form $g \circ u$. Our solution is based on model selection and leads to a very general approach to solve this problem with respect to many different types of functions g, u and statistical frameworks. In particular, we handle the problems of approximating s by additive functions, single and multiple index models, neural networks, mixtures of Gaussian densities (when s is a density) among other examples. We also investigate the situation where $s = g \circ u$ for functions g and u belonging to possibly anisotropic smoothness classes. In this case, our approach leads to a completely adaptive estimator with respect to the regularity of s.

1 Introduction

In various statistical problems, we have at hand a random mapping $\boldsymbol{X}$ from a measurable space $(\Omega, \mathcal{A})$ to $(\mathbb{X}, \mathcal{X})$ with an unknown distribution P_s on $\mathbb{X}$ depending on some parameter $s \in \mathcal{S}$ which is a function from $[-1,1]^k$ to $\mathbb{R}$. For instance, s may be the density of an i.i.d. sample or the intensity of a Poisson process on $[-1,1]^k$

[0] *AMS 1991 subject classifications.* Primary 62G05
Key words and phrases. Curve estimation, model selection, composite functions.

or a regression function. The statistical problem amounts to estimating s by some estimator $\widehat{s} = \widehat{s}(\boldsymbol{X})$ the performance of which is measured by its quadratic risk,

$$R(s, \widehat{s}) = \mathbb{E}_s\left[d^2(s, \widehat{s})\right],$$

where d denotes a given distance on $\mathcal{S}$. To be more specific, we shall assume in this introduction that $\boldsymbol{X} = (X_1, \ldots, X_n)$ is a sample of density s^2 (with $s \geq 0$) with respect to some measure μ and d is the Hellinger distance. We recall that, given two probabilities P, Q dominated by μ with respective densities $f = dP/d\mu$ and $g = dQ/d\mu$, the Hellinger distance h between P and Q or, equivalently, between f and g (since it is independant of the choice of μ) is given by

$$h^2(P, Q) = h^2(f, g) = \frac{1}{2}\int \left(\sqrt{f} - \sqrt{g}\right)^2 d\mu. \tag{1.1}$$

It follows that $\sqrt{2}d(s, t)$ is merely the $\mathbb{L}_2$-distance between s and t.

A general method for constructing estimators $\widehat{s}$ is to choose a model S for s, i.e. do as if s belonged to S, and to build $\widehat{s}$ as an element of S. Sometimes the statistician really assumes that s belongs to S and that S is the true parameter set, sometimes he does not and rather considers S as an approximate model. This latter approach is somewhat more reasonable since it is in general impossible to be sure that s does belong to S. Given S and a suitable estimator $\widehat{s}$, as those built in Birgé (2006) for example, one can achieve a risk bound of the form

$$R(s, \widehat{s}) \leq C\left[\inf_{t \in S} d^2(s, t) + \tau \mathcal{D}(S)\right], \tag{1.2}$$

where C is a universal constant (independent of s and S), $\mathcal{D}(S)$ the dimension of the model S (with a proper definition of the dimension) and τ, which is equal to $1/n$ in the specific context of density estimation, characterizes the amount of information provided by the observation $\boldsymbol{X}$.

It is well-known that many classical estimation procedures suffer from the so-called "curse of dimensionality", which means that the risk bound (1.2) deteriorates when k increases and actually becomes very loose for even moderate values of k. This phenomenon is easy to explain and actually connected with the most classical way of choosing models for s. Typically, and although there is no way to check that such an assumption is true, one assumes that s belongs to some smoothness class (Hölder, Sobolev or Besov) of index α and such an assumption can be translated in terms of approximation properties with respect to the target function s of a suitable collection of linear spaces (generated by piecewise polynomials, splines, or wavelets for example). More precisely, there exists a collection $\mathbb{S}$ of models with the following property: for all $D \geq 1$, there exists a model $S \in \mathbb{S}$ with dimension D which approximates s with an error bounded by $cD^{-\alpha/k}$ for some c independent of D (but depending on s, α and k). With such a collection at hand, we deduce from (1.2) that whatever $D \geq 1$ one can choose a model $S = S(D) \in \mathbb{S}$ for which the estimator $\widehat{s} \in S$ achieves a risk bounded from above by $C\left[c^2 D^{-2\alpha/k} + \tau D\right]$. Besides, by using the elementary Lemma 1 below to be proved in Section 5.6, one can optimize the choice of D, and hence of the model S in $\mathbb{S}$, to build an estimator whose risk satisfies

$$R(s, \widehat{s}) \leq C \max\left\{3c^{2k/(2\alpha+k)}\tau^{2\alpha/(2\alpha+k)}; 2\tau\right\}. \tag{1.3}$$

Lemma 1 *For all positive numbers a, b and θ and $\mathbb{N}^\star$ the set of positive integers,*

$$\inf_{D\in\mathbb{N}^\star}\{aD^{-\theta}+bD\} \leq b+\min\left\{2a^{1/(\theta+1)}b^{\theta/(\theta+1)};a\right\} \leq \max\left\{3a^{1/(\theta+1)}b^{\theta/(\theta+1)};2b\right\}.$$

Since the risk bound (1.3) is achieved for D of order $\tau^{-k/(2\alpha+k)}$, as τ tends to 0, the deterioration of the rate $\tau^{2\alpha/(2\alpha+k)}$ when k increases comes from the fact that we use models of larger dimension to approximate s when k is large. Nevertheless, this phenomenon is only due to the previous approach based on smoothness assumptions for s. An alternative approach, assuming that s can be closely approximated by suitable parametric models the dimensions of which do not depend on k would not suffer from the same weaknesses. More generally, a structural assumption on s associated to a collection of models $\mathbb{S}'$, the approximation properties of which improve on those of $\mathbb{S}$, can only lead to a better risk bound and it is not clear at all that assuming that s belongs to a smoothness class is more realistic than directly assuming approximation bounds with respect to the models of $\mathbb{S}'$. Such structural assumptions that would amount to replacing the large models involved in the approximation of smooth functions by simpler ones have been used for many years, especially in the context of regression. Examples of such structural assumptions are provided by additive models, the single index model, the projection pursuit algorithm introduced by Friedman and Tuckey (1974), (an overview of the procedure is available in Huber (1985)) and artificial neural networks as in Barron (1993; 1994), among other examples. It actually appears that a large number of these alternative approaches (in particular those we just cited) can be viewed as examples of approximation by composite functions.

In any case, an unattractive feature of the previous approach based on an a priori choice of a model $S\in\mathbb{S}$ is that it requires to know suitable upper bounds on the distances between s and the models S in $\mathbb{S}$. Such a requirement is much too strong and an essential improvement can be brought by the modern theory of model selection. More precisely, given some prior probability π on $\mathbb{S}$, model selection allows to build an estimator $\widehat{s}$ with a risk bound

$$CR(s,\widehat{s}) \leq \inf_{S\in\mathbb{S}}\left\{\inf_{t\in S} d^2(s,t)+\tau\left[\mathcal{D}(S)+\log\left(1/\pi(S)\right)\right]\right\}, \tag{1.4}$$

for some universal constant $C>0$. If we neglect the influence of $\log\left(1/\pi(S)\right)$, which is connected to the complexity of the family $\mathbb{S}$ of models we use, the comparison between (1.2) and (1.4) indicates that the method selects a model in $\mathbb{S}$ leading approximately to the smallest risk bound.

With such a tool at hand that allows us to play with many models simultaneously and let the estimator choose a suitable one, we may freely introduce various models corresponding to various sorts of structural assumptions on s that avoid the "curse of dimensionality". We can, moreover, mix them with models which are based on pure smoothness assumptions that do suffer from this dimensional effect or even with simple parametric models. This means that we can so cumulate the advantages of the various models we introduce in the family $\mathbb{S}$.

The main purpose of this paper is to provide a method for building various sorts of models that may be used, in conjonction with other ones, to approximate functions on $[-1,1]^k$ for large values of k. The idea, which is not new, is to approximate the unknown s by a composite function $g\circ u$ where g and u have different approximation

properties. If, for instance, the true s can be closely approximated by a function $g \circ u$ where u goes from $[-1,1]^k$ to $[-1,1]$ and is very smooth and g, from $[-1,1]$ to $\mathbb{R}$, is rough, the overall smoothness of $g \circ u$ is that of g but the curse of dimensionality only applies to the smooth part u, resulting in a much better rate of estimation than what would be obtained by only considering $g \circ u$ as a rough function from $[-1,1]^k$ to $\mathbb{R}$. This is an example of the substantial improvement that might be brought by the use of models of composite functions.

Recent works in this direction can be found in Horowitz and Mammen (2007) or Juditsky, Lepski and Tsybakov (2009). Actually, our initial motivation for this research was a series of lectures given at CIRM in 2005 by Oleg Lepski about a former version of this last paper. There are, nevertheless, major differences between their approach and ours. They deal with estimation in the white noise model, kernel methods and the $\mathbb{L}_\infty$-loss. They also assume that the true unknown density s to be estimated can be written as $s = g \circ u$ where g and u have given smoothness properties and use these properties to build a kernel estimator which is better than those based on the overall smoothness of s. The use of the $\mathbb{L}_\infty$-loss indeed involves additional difficulties and the minimax rates of convergence happen to be substantially slower (not only by logarithmic terms) than the rates one gets for the $\mathbb{L}_2$-loss, as the authors mention on page 1369, comparing their results with those of Horowitz and Mammen (2007).

Our approach is radically different from the one of Juditsky, Lepski and Tsybakov and considerably more general as we shall see, but this level of generality has a price. While they provide a constructive estimator that can be computed in a reasonable amount of time, although based on supposedly known smoothness properties of g and u, we offer a general but abstract method that applies to many situations but does not provide practical estimators, only abstract ones. As a consequence, our results about the performance of these estimators are of a theoretical nature, to serve as benchmarks about what can be expected from good estimators in various situations.

We actually consider "curve estimation" with an unknown functional parameter s and measure the loss by $\mathbb{L}_2$-type distances. Our construction applies to various statistical frameworks (not only the Gaussian white noise but also all these for which a suitable model selection theorem is available). Besides, we do not assume that $s = g \circ u$ but rather approximate s by functions of the form $g \circ u$ and do not fix in advance the smoothness properties of g and u but rather let our estimator adapt to it. In order to give a simple account of our result, let us focus on pairs (u, g) with u mapping $[-1,1]^k$ into $[-1,1]$ and g $[-1,1]$ into $\mathbb{R}$. In this case, our main theorem says the following: consider two (at most) countable collections of models $\mathbb{T}$ and $\mathbb{F}$, endowed with the probabilities λ and γ respectively, in order to approximate such functions u and g respectively. There exists an estimator $\widehat{s}$ such that, whatever the choices of u and g with g at least L-Lipschitz for some $L > 0$,

$$\begin{aligned} C'(L)R(s,\widehat{s}) \quad \leq \quad & d^2(s, g \circ u) + \inf_{F \in \mathbb{F}} \left\{ \inf_{f \in F} d^2_\infty(g, f) + \tau \left[\mathcal{D}(F) + \log(1/\gamma(F)) \right] \right\} \\ & + \inf_{T \in \mathbb{T}} \left\{ \inf_{t \in T} d^2(u, t) + \tau \left[\mathcal{D}(T) \log \tau^{-1} + \log(1/\lambda(T)) \right] \right\}, \quad (1.5) \end{aligned}$$

where d_∞ denotes the distance based on the supremum norm. Compared to (1.4), this result says that, apart from the extra logarithmic terms and the constant C'

depending on L, if s were of the form $g \circ u$ the risk bound we get for estimating s is the maximum of those we would get for estimating g and u separately from a model selection procedure based on $(\mathbb{F}, \gamma)$ and $(\mathbb{T}, \lambda)$ respectively. A more general version of (1.5) allowing to handle less regular functions g and multivariate functions $u = (u_1, \ldots, u_l)$ with values in $[-1, 1]^l$ is available in Section 3. As a consequence, our approach leads to a completely adaptive method with many different possibilities to approximate s. It allows, in particular, to play with the smoothness properties of g and u or to mix purely parametric models with others based on smooth functions. Since methods and theorems about model selection are already available, our main task here will be to build suitable models for various forms of composite functions $g \circ u$ and check that they do satisfy the assumptions required for applying previous model selection results.

2 Our statistical framework

We observe a random element $\boldsymbol{X}$ from the probability space $(\Omega, \mathcal{A}, \mathbb{P}_s)$ to $(\mathbb{X}, \mathcal{X})$ with distribution P_s on $\mathbb{X}$ depending on an unknown parameter s. The set $\mathcal{S}$ of possible values of s is a subset of some space $\mathbb{L}_q(E, \mu)$ where μ is a given *probability* on the measurable space $(E, \mathcal{E})$. We shall mainly consider the case $q = 2$ even though one can also take $q = 1$ in the context of density estimation. We denote by d the distance on $\mathbb{L}_q(E, \mu)$ corresponding to the $\mathbb{L}_q(E, \mu)$-norm $\|\cdot\|_q$ (omitting the dependency of d with respect to q) and by $\mathbb{E}_s$ the expectation with respect to $\mathbb{P}_s$ so that the quadratic risk of an estimator $\widehat{s}$ is $\mathbb{E}_s\left[d^2\left(s, \widehat{s}\right)\right]$. The main objective of this paper, in order to estimate s by model selection, is to build special models S that consist of functions of the form $f \circ t$ where $t = (t_1 \ldots, t_l)$ is a mapping from E to $I \subset \mathbb{R}^l$, f is a continuous function on I and $I = \prod_{j=1}^{l} I_j$ is a product of compact intervals of $\mathbb{R}$. Without loss of generality, we may assume that $I = [-1, 1]^l$. Indeed, if $l = 1$, t takes its values in $I_1 = [\beta - \alpha, \beta + \alpha]$, $\alpha > 0$ and f is defined on I_1, we can replace the pair (f, t) by $(\bar{f}, \bar{t})$ where $\bar{t}(x) = \alpha^{-1}[t(x) - \beta]$ and $\bar{f}(y) = f(\alpha y + \beta)$ so that $\bar{t}$ takes its values in $[-1, 1]$ and $f \circ t = \bar{f} \circ \bar{t}$. The argument easily extends to the multidimensional case.

2.1 Notations and conventions

To perform our construction based on composite functions $f \circ t$, we introduce the following spaces of functions : $\mathcal{T} \subset \mathbb{L}_q(E, \mu)$ is the set of measurable mappings from E to $[-1, 1]$, $\mathcal{F}_{l,\infty}$ is the set of bounded functions on $[-1, 1]^l$ endowed with the distance d_∞ given by $d_\infty(f, g) = \sup_{x \in [-1,1]^l} |f(x) - g(x)|$ and $\mathcal{F}_{l,c}$ is the subset of $\mathcal{F}_{l,\infty}$ which consists of continuous functions on $[-1, 1]^l$. We denote by $\mathbb{N}^\star$ (respectively, $\mathbb{R}_+^\star$) the set of positive integers (respectively positive numbers) and set

$$\lfloor z \rfloor = \sup\{j \in \mathbb{Z} \,|\, j \leq z\} \quad \text{and} \quad \lceil z \rceil = \inf\{j \in \mathbb{N}^\star \,|\, j \geq z\}, \quad \text{for all } z \in \mathbb{R}.$$

The numbers $x \wedge y$ and $x \vee y$ stand for $\min\{x, y\}$ and $\max\{x, y\}$ respectively and $\log_+(x)$ stands for $(\log x) \vee 0$. The cardinality of a set A is denoted by $|A|$ and, by convention, "countable" means "finite or countable". We call *subprobability* on some countable set A any positive measure π on A with $\pi(A) \leq 1$ and, given π and $a \in A$, we set $\pi(a) = \pi(\{a\})$ and $\Delta_\pi(a) = -\log(\pi(a))$ with the convention $\Delta_\pi(a) = +\infty$

if $\pi(a) = 0$. The dimension of the linear space V is denoted by $\mathcal{D}(V)$. Given a compact subset K of $\mathbb{R}^k$ with $\overset{\circ}{K} \neq \emptyset$, we define the *Lebesgue probability* μ on K by $\mu(A) = \lambda(A)/\lambda(K)$ for $A \subset K$, where λ denotes the Lebesgue measure on $\mathbb{R}^k$.

For $x \in \mathbb{R}^m$, x_j denotes the j^{th} coordinate of x ($1 \leq j \leq m$) and, similarly, $x_{i,j}$ denotes the j^{th} coordinate of x_i if the vectors x_i are already indexed. We set $|x|^2 = \sum_{j=1}^m x_j^2$ for the squared Euclidean norm of $x \in \mathbb{R}^m$, without reference to the dimension m, and denote by $\mathcal{B}_m$ the corresponding closed unit ball in $\mathbb{R}^m$. Similarly, $|x|_\infty = \max\{|x_1|, \ldots, |x_m|\}$ for all $x \in \mathbb{R}^m$. For x in some metric space (M, d) and $r > 0$, $\mathcal{B}(x, r)$ denotes the closed ball of center x and radius r in M and for $A \subset M$, $d(x, A) = \inf_{y \in A} d(x, y)$. Finally, C stands for a universal constant while C' is a constant that depends on some parameters of the problem. We may make this dependence explicit by writing $C'(a, b)$ for instance. Both C and C' are generic notations for constants that may change from line to line.

2.2 A general model selection result

General model selection results apply to models which possess a *finite dimension* in a suitable sense. Throughout the paper, we assume that in the statistical framework we consider the following theorem holds.

Theorem 1 *Let $\mathbb{S}$ be a countable family of finite dimensional linear subspaces S of $\mathbb{L}_q(E, \mu)$ and let π be some subprobability measure on $\mathbb{S}$. There exists an estimator $\widehat{s} = \widehat{s}(\boldsymbol{X})$ with values in $\cup_{S \in \mathbb{S}} S$ satisfying, for all $s \in \mathcal{S}$,*

$$\mathbb{E}_s\left[d^2\left(s, \widehat{s}\right)\right] \leq C \inf_{S \in \mathbb{S}} \left\{d^2\left(s, S\right) + \tau\left[\left(\mathcal{D}(S) \vee 1\right) + \Delta_\pi(S)\right]\right\}, \tag{2.1}$$

where the positive constant C and parameter τ only depend on the specific statistical framework at hand.

Similar results often hold also for the loss function $d^r(s, \widehat{s})$ ($r \geq 1$) replacing $d^2(s, \widehat{s})$. In such a case, the results we prove below for the quadratic risk easily extend to the risk $\mathbb{E}_s\left[d^r(s, \widehat{s})\right]$. For simplicity, we shall only focus on the case $r = 2$.

2.3 Some illustrations

The previous theorem actually holds for various statistical frameworks. Let us provide a partial list.

Gaussian frameworks A prototype for Gaussian frameworks is provided by some Gaussian isonormal linear process as described in Section 2 of Birgé and Massart (2001). In such a case, $\boldsymbol{X}$ is a Gaussian linear process with a known variance τ, indexed by a subset $\mathcal{S}$ of some Hilbert space $\mathbb{L}_2(E, \mu)$. This means that $s \in \mathcal{S}$ determines the distribution P_s. Regression with Gaussian errors and Gaussian sequences can both be seen as particular cases of this framework. Then Theorem 1 is a consequence of Theorem 2 of Birgé and Massart (2001). In the regression setting, Baraud, Giraud and Huet (2009) considered the practical case of an unknown variance and proved that (2.1) holds under the assumption that $\mathcal{D}(S) \vee \Delta_\pi(S) \leq n/2$ for all $S \in \mathbb{S}$.

Density estimation Here $\boldsymbol{X} = (X_1, \ldots, X_n)$ is an n-sample with density s^2 with respect to μ and $\mathcal{S}$ is the set of nonnegative elements of norm 1 in $\mathbb{L}_2(E, \mu)$. Then $d(s,t) = \sqrt{2}h\left(s^2, t^2\right)$ where h denotes the Hellinger distance between densities defined by (1.1), $\tau = n^{-1}$ and Theorem 1 follows from Theorem 6 of Birgé (2006) or Corollary 8 of Baraud (2011). Alternatively, one can take for s the density itself, for $\mathcal{S}$ the set of nonnegative elements of norm 1 in $\mathbb{L}_1(E, \mu)$ and set $q = 1$. The result then follows from Theorem 8 of Birgé (2006). Under the additional assumption that $s \in \mathbb{L}_2(E, \mu) \cap \mathbb{L}_\infty(E, \mu)$, the case $q = 2$ follows from Theorem 6 of Birgé (2008) with $\tau = n^{-1} \|s\|_\infty (1 \vee \log \|s\|_\infty)$.

Regression with fixed design We observe $\boldsymbol{X} = \{(x_1, Y_1), \ldots, (x_n, Y_n)\}$ with $\mathbb{E}[Y_i] = s(x_i)$ where s is a function from $E = \{x_1, \ldots, x_n\}$ to $\mathbb{R}$ and the errors $\varepsilon_i = Y_i - s(x_i)$ are i.i.d. Here μ is the uniform distribution on E, hence $d^2(s,t) = n^{-1} \sum_{i=1}^n [s(x_i) - t(x_i)]^2$ and $\tau = 1/n$. When the errors ε_i are subgaussian, Theorem 1 follows from Theorem 3.1 in Baraud, Comte and Viennet (2001). For more heavy-tailed distributions (Laplace, Cauchy, etc.) we refer to Theorem 6 of Baraud (2011) when s takes its values in $[-1, 1]$.

Bounded regression with random design Let (X, Y) be a pair of random variables with values in $E \times [-1, 1]$ where X has distribution μ and $\mathbb{E}[Y|X = x] = s(x)$ is a function from E to $[-1, 1]$. Our aim here is to estimate s from the observation of n independent copies $\boldsymbol{X} = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ of (X, Y). Here the distance d corresponds to the $\mathbb{L}_2(E, \mu)$-distance and Theorem 1 follows from Corollary 8 in Birgé (2006) with $\tau = n^{-1}$.

Poisson processes In this case, $\boldsymbol{X}$ is a Poisson process on E with mean measure $s^2 \cdot \mu$, where s is a nonnegative element of $\mathbb{L}_2(E, \mu)$. Then $\tau = 1$ and Theorem 1 follows from Birgé (2007) or Corollary 8 of Baraud (2011).

3 The basic theorems

3.1 Models and their dimensions

If we assume that the unknown parameter s to be estimated is equal or close to some composite function of the form $g \circ u$ with $u \in \mathcal{T}^l$ and $g \in \mathcal{F}_{l,c}$ and if we wish to estimate $g \circ u$ by model selection we need to have at disposal a family $\mathbb{F}$ of models for approximating g and families $\mathbb{T}_j$, $1 \leq j \leq l$, to approximate the components u_j of u. Typical sets that are used for approximating elements of $\mathcal{F}_{l,c}$ or $\mathcal{T}^l$ are finite-dimensional linear spaces or subsets of them. Many examples of such spaces are described in books on Approximation Theory, like the one by DeVore and Lorentz (1993) and we need a theorem which applies to such classical approximation sets for which it will be convenient to choose the following definition of their dimension.

Definition 1 *Let H be a linear space and $S \subset H$. The dimension $\mathcal{D}(S) \in \mathbb{N} \cup \{\infty\}$ of S is 0 if $|S| = 1$ and is, otherwise, the dimension (in the usual sense) of the linear span of S.*

3.2 Some smoothness assumptions

In order to transfer the approximation properties of g by f and u by t into approximation of $g \circ u$ by $f \circ t$, we shall also require that g be somewhat smooth. The smoothness assumptions we need can be expressed in terms of moduli of continuity. We start with the definition of the modulus of continuity of a function g in $\mathcal{F}_{l,c}$.

Definition 2 *We say that* w *from* $[0,2]^l$ *to* $\mathbb{R}_+^l$ *is a modulus of continuity for a continuous function* g *on* $[-1,1]^l$ *if, for all* $z \in [0,2]^l$, $\mathrm{w}(z)$ *is of the form* $\mathrm{w}(z) = (\mathrm{w}_1(z_1), \ldots, \mathrm{w}_l(z_l))$ *where each function* w_j *with* $j = 1, \ldots, l$ *is continuous, nondecreasing and concave from* $[0,2]$ *to* $\mathbb{R}_+$, *satisfies* $\mathrm{w}_j(0) = 0$, *and*

$$|g(x) - g(y)| \leq \sum_{j=1}^{l} \mathrm{w}_j(|x_j - y_j|) \quad \textit{for all } x, y \in [-1,1]^l.$$

For $\boldsymbol{\alpha} \in (0,1]^l$ *and* $\mathbf{L} \in (0,+\infty)^l$, *we say that* g *is* $(\boldsymbol{\alpha}, \mathbf{L})$*-Hölderian if one can take* $\mathrm{w}_j(z) = L_j z^{\alpha_j}$ *for all* $z \in [0,2]$ *and* $j = 1, \ldots, l$. *It is said to be* $\mathbf{L}$*-Lipschitz if it is* $(\boldsymbol{\alpha}, \mathbf{L})$*-Hölderian with* $\boldsymbol{\alpha} = (1, \ldots, 1)$.

Note that our definition of a modulus of continuity implies that the w_j are subadditive, a property which we shall often use in the sequel and that, given g, one can always choose for w_j the least concave majorant of w_j where

$$w_j(z) = \sup_{x \in [-1,1]^l;\, x_j \leq 1-z} |g(x) - g(x_1, \ldots, x_{j-1}, x_j + z, x_{j+1}, \ldots, x_l)|.$$

Then $\mathrm{w}_j(z) \leq 2w_j(z)$ according to Lemma 6.1 p. 43 of DeVore and Lorentz (1993).

3.3 The main theorem

Our construction of estimators $\widehat{s}$ of $g \circ u$ will be based on some set $\mathfrak{S}$ of the following form:

$$\mathfrak{S} = \{l, \mathbb{F}, \gamma, \mathbb{T}_1, \ldots, \mathbb{T}_l, \lambda_1, \ldots, \lambda_l\}, \quad l \in \mathbb{N}^\star, \tag{3.1}$$

where $\mathbb{F}, \mathbb{T}_1, \ldots, \mathbb{T}_l$ denote families of models and γ, λ_j are measures on $\mathbb{F}$ and $\mathbb{T}_j$ respectively. In the sequel, we shall assume that $\mathfrak{S}$ satisfies the following requirements.

Assumption 1 *The set* $\mathfrak{S}$ *is such that*

i) the family $\mathbb{F}$ *is a countable set and consists of finite-dimensional linear subspaces* F *of* $\mathcal{F}_{l,\infty}$ *with respective dimensions* $\mathcal{D}(F) \geq 1$,

ii) for $j = 1, \ldots, l$, $\mathbb{T}_j$ *is a countable set of subsets of* $\mathbb{L}_q(E, \mu)$ *with finite dimensions,*

iii) the measure γ *is a subprobability on* $\mathbb{F}$,

iv) for $j = 1, \ldots, l$, λ_j *is a subprobability on* $\mathbb{T}_j$.

Given $\mathfrak{S}$, one can design an estimator $\widehat{s}$ with the following properties.

Theorem 2 *Assume that Theorem 1 holds and that* $\mathfrak{S}$ *satisfies Assumption 1. One can build an estimator* $\widehat{s} = \widehat{s}(\boldsymbol{X})$ *satisfying, for all* $u \in \mathcal{T}^l$ *and* $g \in \mathcal{F}_{l,c}$ *with modulus*

of continuity w_g,

$$C\mathbb{E}_s\left[d^2\left(s,\widehat{s}\right)\right] \leq \sum_{j=1}^{l} \inf_{T\in\mathbb{T}_j} \left\{ l\mathrm{w}_{g,j}^2\big(d(u_j,T)\big) + \tau\left[\Delta_{\lambda_j}(T) + i(g,j,T)\mathcal{D}(T)\right]\right\} + d^2(s,g\circ u) + \inf_{F\in\mathbb{F}} \left\{ d_\infty^2(g,F) + \tau\left[\Delta_\gamma(F)+\mathcal{D}(F)\right]\right\}, \quad (3.2)$$

where $i(g,j,T)=1$ *if* $\mathcal{D}(T)=0$ *and otherwise,*

$$i(g,j,T) = \inf\left\{ i\in\mathbb{N}^\star \,|\, l\mathrm{w}_{g,j}^2\left(e^{-i}\right) \leq \tau i\mathcal{D}(T)\right\} < +\infty. \quad (3.3)$$

Note that, since the risk bound (3.2) is valid for all $g\in\mathcal{F}_{l,c}$ and $u\in\mathcal{T}^l$, we can minimize the right-hand side of (3.2) with respect to g and u in order to optimize the bound. The proof of this theorem is postponed to Section 5.4.

Of special interest is the case where g is $\mathbf{L}$-Lipschitz. If one is mainly interested by the dependence of the risk bound with respect to τ as it tends to 0, one can check that $i(g,j,T)\leq \log\tau^{-1}$ for τ small enough (depending on l and $\mathbf{L}$) so that (3.2) becomes for such a small τ

$$C'\mathbb{E}_s\left[d^2\left(s,\widehat{s}\right)\right] \leq \sum_{j=1}^{l} \inf_{T\in\mathbb{T}_j} \left\{ d^2(u_j,T) + \tau\left(\Delta_{\lambda_j}(T) + \mathcal{D}(T)\log\tau^{-1}\right)\right\} + d^2(s,g\circ u) + \inf_{F\in\mathbb{F}} \left\{ d_\infty^2(g,F) + \tau\left[\mathcal{D}(F)+\Delta_\gamma(F)\right]\right\}.$$

If it were possible to apply Theorem 1 to the models F with the distance d_∞ and the models T with the distance d for each j separately, we would get risk bounds of this form, apart from the value of C' and the extra $\log\tau^{-1}$ factor. This means that, apart from this extra logarithmic factor, our procedure amounts to performing $l+1$ separate model selection procedures, one with the collection $\mathbb{F}$ for estimating g and the other ones with the collections $\mathbb{T}_j$ for the components u_j, finally getting the sum of the $l+1$ resulting risk bounds. The result is however slightly different when g is no longer Lipschitz. When g is $(\boldsymbol{\alpha},\mathbf{L})$-Hölderian then one can check that $i(g,j,T)\leq\mathcal{L}_{j,T}$ where $\mathcal{L}_{j,T}=1$ if $\mathcal{D}(T)=0$ and, if $\mathcal{D}(T)\geq 1$,

$$\mathcal{L}_{j,T} = \left[\alpha_j^{-1}\log\left(lL_j^2[\tau\mathcal{D}(T)]^{-1}\right)\right]\bigvee 1 \quad (3.4)$$

$$\leq C'(l,\alpha_j)\left[\log(\tau^{-1})\vee\log(L_j^2/\mathcal{D}(T))\vee 1\right]. \quad (3.5)$$

In this case, Theorem 2 leads to the following result.

Corollary 1 *Assume that the assumptions of Theorem 2 holds. For all* $(\boldsymbol{\alpha},\mathbf{L})$*-Hölderian function* g *with* $\boldsymbol{\alpha}\in(0,1]^l$ *and* $\mathbf{L}\in(\mathbb{R}_+^\star)^l$*, the estimator* $\widehat{s}$ *of Theorem 2 satisfies*

$$C\mathbb{E}_s\left[d^2\left(s,\widehat{s}\right)\right] \leq \sum_{j=1}^{l} \inf_{T\in\mathbb{T}_j} \left\{ lL_j^2 d^{2\alpha_j}(u_j,T) + \tau\left[\Delta_{\lambda_j}(T) + \mathcal{D}(T)\mathcal{L}_{j,T}\right]\right\} + d^2(s,g\circ u) + \inf_{F\in\mathbb{F}} \left\{ d_\infty^2(g,F) + \tau\left[\Delta_\gamma(F)+\mathcal{D}(F)\right]\right\}, \quad (3.6)$$

where $\mathcal{L}_{j,T}$ *is defined by (3.4) and bounded by (3.5).*

3.4 Mixing collections corresponding to different values of l

If it is known that s takes the special form $g \circ u$ for some unknown values of $g \in \mathcal{F}_{l,c}$ and $u \in \mathcal{T}^l$, or if s is very close to some function of this form, the previous approach is quite satisfactory. If we do not have such an information, we may apply the previous construction with several values of l simultaneously, approximating s by different combinations $g_l \circ u_l$ with u_l taking its values in $[-1,1]^l$, g_l a function on $[-1,1]^l$ and l varying among some subset I of $\mathbb{N}^\star$. To each value of l we associate, as before, $l+1$ collections of models and the corresponding subprobabilities, each l then leading to an estimator $\widehat{s}_l$ the risk of which is bounded by $\mathcal{R}(\widehat{s}_l, g_l, u_l)$ given by the right-hand side of (3.2). The model selection approach allows us to use all the previous collections of models for all values of l simultaneously in order to build a new estimator the risk of which is approximately as good as the risk of the best of the $\widehat{s}_l$. More generally, let us assume that we have at hand a countable family $\{\mathfrak{S}_\ell,\, \ell \in I\}$ of sets $\mathfrak{S}_\ell$ of the form (3.1) satisfying Assumption 1 for some $l = l(\ell) \geq 1$. To each such set, Theorem 2 associates an estimator $\widehat{s}_\ell$ with a risk bounded by

$$\mathbb{E}_s\big[d^2\left(s, \widehat{s}_\ell\right)\big] \leq \inf_{(g,u)} \mathcal{R}(\widehat{s}_\ell, g, u),$$

where $\mathcal{R}(\widehat{s}_\ell, g, u)$ denotes the right-hand side of (3.2) when $\mathfrak{S} = \mathfrak{S}_\ell$ and the infimum runs among all pairs (g,u) with $g \in \mathcal{F}_{l(\ell),c}$ and $u \in \mathcal{T}^{l(\ell)}$. We can then prove (in Section 5.5 below) the following result.

Theorem 3 *Assume that Theorem 1 holds and let I be a countable set and ν a subprobability on I. For each $\ell \in I$ we are given a set $\mathfrak{S}_\ell$ of the form (3.1) that satisfies Assumption 1 with $l = l(\ell)$ and a corresponding estimator $\widehat{s}_\ell$ provided by Theorem 2. One can then design a new estimator $\widehat{s} = \widehat{s}(\boldsymbol{X})$ satisfying*

$$C\mathbb{E}_s\big[d^2\left(s, \widehat{s}\right)\big] \leq \inf_{\ell \in I} \inf_{(g,u)} \left\{\mathcal{R}(\widehat{s}_\ell, g, u) + \tau \Delta_\nu(\ell)\right\},$$

where $\mathcal{R}(\widehat{s}_\ell, g, u)$ denotes the right-hand side of (3.2) when $\mathfrak{S} = \mathfrak{S}_\ell$ and the second infimum runs among all pairs (g,u) with $g \in \mathcal{F}_{l(\ell),c}$ and $u \in \mathcal{T}^{l(\ell)}$.

3.5 The main ideas underlying our construction

Let us assume here that $p = q = 2$ and $E = [-1,1]^k$ with $k > l \geq 1$. Our approach is based on the construction of a family of linear spaces with good approximation properties with respect to composite functions $g \circ u$. More precisely, if one considers a finite dimensional linear space $F \subset \mathcal{F}_{l,\infty}$ for approximating g and compact sets $T_j \subset \mathcal{T}$ for approximating the u_j, we shall show (see Proposition 4 in Section 5.1 below) that there exists some t in $\mathbf{T} = \prod_{j=1}^l T_j$ such that the linear space $S_t = \{f \circ t \,|\, f \in F\}$ approximates the composite function $g \circ u$ with an error bound

$$d(g \circ u, S_t) \leq d_\infty(g, F) + \sqrt{2} \sum_{j=1}^{l} \mathrm{w}_{g,j}\left(d(u_j, T_j)\right). \tag{3.7}$$

The case where the function g is Lipschitz, i.e. $\mathrm{w}_{g,j}(x) = Lx$ for all j, is of particular interest since, up to constants, the error bound we get is the sum of those for

approximating separately g by F (with respect to the $\mathbb{L}_\infty$-distance) and the u_j by T_j. In particular, if s were exactly of the form $s = g \circ u$ for some known functions u_j, we could use a linear space F of piecewise constant functions with dimension of order D to approximate g, and take $T_j = \{u_j\}$ for all j. In this case the linear space S_u whose dimension is also of order D would approximate $s = g \circ u$ with an error bounded by $D^{-1/l}$. Note that if the u_j were all $(\boldsymbol{\beta}, \mathbf{L})$-Hölderian with $\boldsymbol{\beta} \in (0,1]^k$, the overall regularity of the function $s = g \circ u$ could not be expected to be better than $\boldsymbol{\beta}$-Hölderian, since this regularity is already achieved by taking $g(y_1, \ldots, y_l) = y_1$. In comparison, an approach based on the overall smoothness of s, which would completely ignore the fact that $s = g \circ u$ and the knowledge of the u_j, would lead to an approximation bound of order $D^{-\overline{\beta}/k}$ with $\overline{\beta} = k \left(\sum_{j=1}^k \beta_j^{-1} \right)^{-1}$. The former bound, $D^{-1/l}$, based on the structural assumption that $s = g \circ u$ therefore improves on the latter since $\overline{\beta} \leq 1$ and $k > l$. Of course, one could argue that the former approach uses the knowledge of the u_j, which is quite a strong assumption. Actually, a more reasonable approach would be to assume that u is unknown but close to a parametric set $\overline{\mathbf{T}}$, in which case, it would be natural to replace the single model S_u used for approximating s, by the family of models $\mathbb{S}_{\overline{\mathbf{T}}}(F) = \{S_t \,|\, t \in \overline{\mathbf{T}}\}$ and, ideally, let the usual model selection techniques select some best linear space among it. Unfortunately, results such as Theorem 1 do not apply to this case, since the family $\mathbb{S}_{\overline{\mathbf{T}}}(F)$ has the same cardinality as $\overline{\mathbf{T}}$ and is therefore typically not countable. The main idea of our approach is to take advantage of the fact that the u_j take their values in $[-1,1]$ so that we can embed $\overline{\mathbf{T}}$ into a compact subset of $\mathcal{T}^l$. We may then introduce a suitably discretized version $\mathbf{T}$ of $\overline{\mathbf{T}}$ (more precisely, of its embedding) and replace the ideal collection $\mathbb{S}_{\overline{\mathbf{T}}}(F)$ by $\mathbb{S}_{\mathbf{T}}(F)$, for which similar approximation properties can be proved. The details of this discretization device will be given in the proofs of our main results. Finally, we shall let both $\overline{\mathbf{T}}$ and F vary into some collections of models and use all the models of the various resulting collections $\mathbb{S}_{\mathbf{T}}(F)$ together in order to estimate s at best.

4 Applications

The aim of this section is to provide various applications of Theorem 2 and its corollaries. We start with a brief overview of more or less classical collections of models commonly used for approximating smooth (and less smooth) functions on $[-1,1]^k$.

4.1 Classical models for approximating smooth functions

Along this section, d denotes the $\mathbb{L}_2$-distance in $\mathbb{L}_2([-1,1]^k, 2^{-k}dx)$, thus taking $q = 2$, $E = [-1,1]^k$ and μ the Lebesgue probability on E. Collections of models with the following property will be of special interest throughout this paper.

Assumption 2 *For each $D \in \mathbb{N}$ the number of elements with dimension D belonging to the collection $\mathbb{S}$ is bounded by $\exp[c(\mathbb{S})(D+1)]$ for some nonnegative constant $c(\mathbb{S})$ depending on $\mathbb{S}$ only.*

4.1.1 Approximating functions in Hölder spaces on $[-1,1]^k$

When $k = 1$, a typical smoothness condition for a function s on $[-1,1]$ is that it belongs to some Hölder space $\mathcal{H}^{\alpha}([-1,1])$ with $\alpha = r + \alpha'$, $r \in \mathbb{N}$ and $0 < \alpha' \leq 1$ which is the set of all functions f on $[-1,1]$ with a continuous derivative of order r satisfying, for some constant $L(f) > 0$,

$$\left| f^{(r)}(x) - f^{(r)}(y) \right| \leq L(f)|x-y|^{\alpha'} \quad \text{for all } x, y \in [-1,1].$$

This notion of smoothness extends to functions $f(x_1, \ldots, x_k)$ defined on $[-1,1]^k$, by saying that f belongs to $\mathcal{H}^{\boldsymbol{\alpha}}([-1,1]^k)$ with $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k) \in (0, +\infty)^k$ if, viewed as a function of x_i only, it belongs to $\mathcal{H}^{\alpha_i}([-1,1])$ for $1 \leq i \leq k$ with some constant $L(f)$ independent of both i and the variables x_j for $j \neq i$. The smoothness of a function s in $\mathcal{H}^{\boldsymbol{\alpha}}([-1,1]^k)$ is said to be *isotropic* if the α_i are all equal and *anisotropic* otherwise, in which case the quantity $\overline{\alpha}$ given by

$$\frac{1}{\overline{\alpha}} = \frac{1}{k} \sum_{i=1}^{k} \frac{1}{\alpha_i}$$

corresponds to the average smoothness of s. It follows from results in Approximation Theory that functions in the Hölder space $\mathcal{H}^{\boldsymbol{\alpha}}([-1,1]^k)$ can be well approximated by piecewise polynomials on k-dimensional hyperrectangles. More precisely, our next proposition follows from results in Dahmen, DeVore and Scherer (1980).

Proposition 1 *Let $(k, r) \in \mathbb{N}^{\star} \times \mathbb{N}$. There exists a collection of models $\mathbb{H}_{k,r} = \bigcup_{D \geq 1} \mathbb{H}_{k,r}(D)$ satisfying Assumption 2 such that for each positive integer D, the family $\mathbb{H}_{k,r}(D)$ consists of linear spaces S with dimensions $\mathcal{D}(S) \leq C_1'(k,r)D$ spanned by piecewise polynomials of degree at most r on k-dimensional hyperrectangles and for which*

$$\inf_{S \in \mathbb{H}_{k,r}(D)} d(s,S) \leq \inf_{S \in \mathbb{H}_{k,r}(D)} d_{\infty}(s,S) \leq C_2'(k,r) L(s) D^{-\overline{\alpha}/k},$$

for all $s \in \mathcal{H}^{\boldsymbol{\alpha}}([-1,1]^k)$ with $\sup_{1 \leq i \leq k} \alpha_i \leq r + 1$.

4.1.2 Approximating functions in anisotropic Besov spaces

Anisotropic Besov spaces generalize anisotropic Hölder spaces and are defined in a similar way by using directional moduli of smoothness, just as Hölder spaces are defined using directional derivatives. To be short, a function belongs to an anisotropic Besov space on $[-1,1]^k$ if, when all coordinates are fixed apart from one, it belongs to a Besov space on $[-1,1]$. A precise definition (restricted to $k = 2$ but which can be generalized easily) can be found in Hochmuth (2002). The general definition together with useful approximation properties by piecewise polynomials can be found in Akakpo (2012). For $0 < p \leq +\infty$, $k > 1$ and $\boldsymbol{\beta} \in (0, +\infty)^k$, let us denote by $\mathcal{B}^{\boldsymbol{\beta}}_{p,p}([-1,1]^k)$ the anisotropic Besov spaces. In particular, $\mathcal{B}^{\boldsymbol{\beta}}_{\infty,\infty}([-1,1]^k) = \mathcal{H}^{\boldsymbol{\beta}}([-1,1]^k)$. It follows from Akakpo (2012) that Proposition 1 can be generalized to Besov spaces in the following way.

Proposition 2 *Let $p > 0$, $k \in \mathbb{N}^\star$ and $r \in \mathbb{N}$. There exists a collection of models $\mathbb{B}_{k,r} = \bigcup_{D\geq 1} \mathbb{B}_{k,r}(D)$ satisfying Assumption 2 such that for each positive integer D, $\mathbb{B}_{k,r}(D)$ consists of linear spaces S with dimensions $\mathcal{D}(S) \leq C_1'(k,r)D$ spanned by piecewise polynomials of degree at most r on k-dimensional hyperrectangles and for which*

$$\inf_{S\in\mathbb{B}_{k,r}(D)} d(s,S) \leq C_2'(k,r,p)\,|s|_{\boldsymbol{\beta},p,p} D^{-\overline{\beta}/k}$$

for all $s \in \mathcal{B}^{\boldsymbol{\beta}}_{p,p}([-1,1]^k)$ with semi-norm $|s|_{\boldsymbol{\beta},p,p}$ and $\boldsymbol{\beta}$ satisfying

$$\sup_{1\leq i\leq k} \beta_i < r+1 \qquad \text{and} \qquad \overline{\beta} > k\left[\left(p^{-1}-2^{-1}\right)\vee 0\right]. \tag{4.1}$$

4.2 Estimation of smooth functions on $[-1,1]^k$

In this section, our aim is to establish risk bounds for our estimator $\widehat{s}$ when $s = g\circ u$ for some smooth functions g and u. We shall discuss the improvement, in terms of rates of convergence as τ tends to 0, when assuming such a structural hypothesis, as compared to a pure smoothness assumption on s. Throughout this section, we take $q = 2$, $E = [-1,1]^k$ and d as the $\mathbb{L}_2$-distance on $\mathbb{L}_2(E, 2^{-k}dx)$.

It follows from Section 4.1 that, for all $r \geq 0$, $\mathbb{H}_{k,r}$ satisfies Assumption 2 for some constant $c(\mathbb{H}_{k,r})$. Therefore the measure γ on $\mathbb{H}_{k,r}$ defined by

$$\Delta_\gamma(S) = (c(\mathbb{H}_{k,r})+1)(D+1) \quad \text{for all } S \in \mathbb{H}_{k,r}(D) \setminus \bigcup_{1\leq D'<D} \mathbb{H}_{k,r}(D') \tag{4.2}$$

is a subprobability since

$$\sum_{S\in\mathbb{H}_{k,r}} e^{-\Delta_\gamma(S)} \leq \sum_{D\geq 1} e^{-D}\left|\mathbb{H}_{k,r}(D)\right| e^{-c(\mathbb{H}_{k,r})(D+1)} \leq \sum_{D\geq 1} e^{-D} < 1.$$

We shall similarly consider the subprobability λ defined on $\mathbb{B}_{k,r}$ by

$$\Delta_\lambda(S) = (c(\mathbb{B}_{k,r})+1)(D+1) \quad \text{for all } S \in \mathbb{B}_{k,r}(D) \setminus \bigcup_{1\leq D'<D} \mathbb{B}_{k,r}(D'). \tag{4.3}$$

Finally, for $g \in \mathcal{H}^{\boldsymbol{\alpha}}([-1,1]^l) = \mathcal{B}^{\boldsymbol{\alpha}}_{\infty,\infty}([-1,1]^l)$ with $\boldsymbol{\alpha} \in (\mathbb{R}_+^\star)^l$, we set $\|g\|_{\boldsymbol{\alpha},\infty} = |g|_{\boldsymbol{\alpha},\infty,\infty} + \inf L'$ where the infimum runs among all numbers L' for which $\mathrm{w}_{g,j}(z) \leq L' z^{\alpha_j\wedge 1}$ for all $z \in [0,2]$ and $j = 1,\ldots,l$.

4.2.1 Convergence rates using composite functions

Let us consider here the set $\mathcal{S}_{k,l}(\boldsymbol{\alpha},\boldsymbol{\beta},\mathbf{p},L,\boldsymbol{R})$ gathering the composite functions $g\circ u$ with $g \in \mathcal{H}^{\boldsymbol{\alpha}}([-1,1]^l)$ satisfying $\|g\|_{\boldsymbol{\alpha},\infty} \leq L$ and $u_j \in \mathcal{B}^{\boldsymbol{\beta}_j}_{p_j,p_j}$ with semi-norms $|u_j|_{\boldsymbol{\beta}_j,p_j,p_j} \leq R_j$ for all $j = 1,\ldots,l$. The following result holds.

Theorem 4 *Assume that Theorem 1 holds with $q = 2$. There exists an estimator $\widehat{s}$ such that, for all $l \geq 1$, $\boldsymbol{\alpha},\boldsymbol{R} \in (\mathbb{R}_+^\star)^l$, $L > 0$, $\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_l \in (\mathbb{R}_+^\star)^k$ and $\mathbf{p} \in (0,+\infty]^l$*

with $\overline{\beta}_j > k\left[\left(p_j^{-1}-2^{-1}\right)\vee 0\right]$ *for* $1\le j\le l$,

$$\sup_{s\in\mathcal{S}_{k,l}(\boldsymbol{\alpha},\boldsymbol{\beta},\mathbf{p},L,\boldsymbol{R})} C'\mathbb{E}_s\left[d^2(s,\widehat{s})\right]$$
$$\le \sum_{j=1}^{l}\left(LR_j^{\alpha_j\wedge 1}\right)^{\frac{2k}{2\overline{\beta}_j(\alpha_j\wedge 1)+k}}\left[\tau\mathcal{L}\right]^{\frac{2\overline{\beta}_j(\alpha_j\wedge 1)}{2\overline{\beta}_j(\alpha_j\wedge 1)+k}} + L^{\frac{2l}{l+2\overline{\alpha}}}\tau^{\frac{2\overline{\alpha}}{l+2\overline{\alpha}}}+\tau\mathcal{L},$$

where $\mathcal{L}=\log(\tau^{-1})\vee\log(L^2)\vee 1$ *and* C' *depends on* $k,l,\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ *and* $\mathbf{p}$.

Let us recall that we need not assume that s is exactly of the form $g\circ u$ but rather, as we did before, that s can be approximated by a function $\overline{s}=g\circ u\in\mathcal{S}_{k,l}(\boldsymbol{\alpha},\boldsymbol{\beta},\mathbf{p},L,\boldsymbol{R})$. In such a case we simply get an additional bias term of the form $d^2(s,\overline{s})$ in our risk bounds.

Proof: Let us fix some value of $l\ge 1$ and take $s=g\circ u\in\mathcal{S}_{k,l}(\boldsymbol{\alpha},\boldsymbol{\beta},\mathbf{p},L,\boldsymbol{R})$ and define

$$r=r(\boldsymbol{\alpha},\boldsymbol{\beta})=1+\left\lfloor \max_{i=1,\ldots,l}\alpha_i\bigvee\max_{j=1,\ldots,l,\ell=1,\ldots,k}\beta_{j,\ell}\right\rfloor.$$

The regularity properties of g and the u_j together with Propositions 1 and 2 imply that for all $D\ge 1$, there exist $F\in\mathbb{H}_{l,r}(D)$ and sets $T_j\in\mathbb{B}_{k,r}(D)$ for $j=1,\ldots,l$ such that

$$\mathcal{D}(F)\le C_1'(l,\boldsymbol{\alpha},\boldsymbol{\beta})\,D;\qquad d_\infty(g,F)\le C_2'(l,\boldsymbol{\alpha},\boldsymbol{\beta})\,LD^{-\overline{\alpha}/l};$$

and, for $1\le j\le l$,

$$\mathcal{D}(T_j)\le C_3'\big(k,\boldsymbol{\alpha},\boldsymbol{\beta}_j,p_j\big)\,D;\qquad d(u_j,T_j)\le C_4'\big(k,\boldsymbol{\alpha},\boldsymbol{\beta}_j,p_j\big)\,R_jD^{-\overline{\beta_j}/k}.$$

Since the collections $\mathbb{H}_{l,r}$ and $\mathbb{B}_{k,r}$ satisfy Assumption 2 and $\mathrm{w}_{g,j}(z)\le Lz^{\alpha_j\wedge 1}$ for all j and $z\in[0,2]$, we may apply Corollary 1 with

$$\mathfrak{S}_{l,r}=(l,\mathbb{H}_{l,r},\gamma_r,\mathbb{B}_{k,r},\ldots,\mathbb{B}_{k,r},\lambda_r,\ldots,\lambda_r)$$

the subprobabilities $\gamma_{l,r}$ and $\lambda_{l,r}$ being given by (4.2) and (4.3) respectively. Besides, it follows from (3.5) that $\mathcal{L}_{j,T}\le C'(l,\boldsymbol{\alpha})\mathcal{L}$ for all j, so that (3.6) implies that the risk of the resulting estimator $\widehat{s}_{l,r}$ is bounded from above by

$$C'\mathcal{R}(\widehat{s}_{l,r},g,u)=\sum_{j=1}^{l}\inf_{D\ge1}\left[L^2R_j^{2(\alpha_j\wedge1)}D^{-2(\alpha_j\wedge1)\overline{\beta}_j/k}+D\tau\mathcal{L}\right]+\inf_{D\ge1}\left[L^2D^{-2\overline{\alpha}/l}+D\tau\right],$$

for some constant C' depending on $l,k,\boldsymbol{\alpha},\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_l$. We obtain the result by optimizing each term of the sum with respect to D by means of Lemma 1, and by using Theorem 3 with ν defined for $\ell=(l,r)\in\mathbb{N}^\star\times\mathbb{N}$ by $\nu(l,r)=e^{-(l+r+1)}$ for which $\Delta_\nu(l,r)\tau\le(l+r+1)\mathcal{R}(\widehat{s}_{l,r},g,u)$ for all l,r. □

4.2.2 Structural assumption versus smoothness assumption

In view of discussing the interest of the risk bounds provided by Theorem 4, let us focus here, for simplicity, on the case where $g\in\mathcal{H}^\alpha([-1,1])$ with $\alpha>0$ (hence $l=1$) and u is a function from $E=[-1,1]^k$ to $[-1,1]$ that belongs to $\mathcal{H}^{\boldsymbol{\beta}}([-1,1]^k)$ with $\boldsymbol{\beta}\in(\mathbb{R}_+^\star)^k$. The following proposition is to be proved in Section 5.7.

Proposition 3 *Let ϕ be the function defined on $(\mathbb{R}_+^\star)^2$ by*

$$\phi(x,y) = \begin{cases} xy & \text{if } x \vee y \leq 1; \\ x \wedge y & \text{otherwise.} \end{cases}$$

For all $k \geq 1$, $\alpha > 0$, $\boldsymbol{\beta} \in (\mathbb{R}_+^\star)^k$, $g \in \mathcal{H}^\alpha([-1,1])$ and $u \in \mathcal{H}^{\boldsymbol{\beta}}([-1,1]^k)$,

$$g \circ u \in \mathcal{H}^{\boldsymbol{\theta}}([-1,1]^k) \quad \text{with} \quad \theta_i = \phi(\beta_i, \alpha) \quad \text{for } 1 \leq i \leq k. \tag{4.4}$$

Moreover, $\boldsymbol{\theta}$ is the largest possible value for which (4.4) holds for all $g \in \mathcal{H}^\alpha([-1,1])$ and $u \in \mathcal{H}^{\boldsymbol{\beta}}([-1,1]^k)$ since, whatever $\boldsymbol{\theta}' \in (\mathbb{R}_+^\star)^k$ such that $\theta_i' > \theta_i$ for some $i \in \{1,\dots,k\}$, there exists some $g \in \mathcal{H}^\alpha([-1,1])$ and $u \in \mathcal{H}^{\boldsymbol{\beta}}([-1,1]^k)$ such that $g \circ u \notin \mathcal{H}^{\boldsymbol{\theta}'}([-1,1]^k)$.

Using the information that s belongs to $\mathcal{H}^{\boldsymbol{\theta}}([-1,1]^k)$ with $\boldsymbol{\theta}$ given by (4.4) and that we cannot assume that s belongs to some smoother class (although this may happen in special cases) since $\boldsymbol{\theta}$ is minimal, but ignoring the fact that $s = g \circ u$, we can estimate s at rate $\tau^{2\overline{\theta}/(2\overline{\theta}+k)}$ (as τ tends to 0) while, on the other hand, by using Theorem 4 and the structural information that $s = g \circ u$, we can achieve the rate

$$\tau^{2\alpha/(2\alpha+1)} + \left(\tau \left[\log \tau^{-1}\right]\right)^{2\overline{\beta}(\alpha\wedge 1)/(2\overline{\beta}(\alpha\wedge 1)+k)}.$$

Let us now compare these two rates. First note that it follows from (4.4) that $\theta_i \leq \alpha$ for all i, hence $\overline{\theta} \leq \alpha$ and, since $k > 1$, $2\alpha/(2\alpha+1) > 2\overline{\theta}/\left(2\overline{\theta}+k\right)$. Therefore the term $\tau^{2\alpha/(2\alpha+1)}$ always improves over $\tau^{2\overline{\theta}/\left(2\overline{\theta}+k\right)}$ when τ is small and, to compare the two rates, it is enough to compare $\overline{\theta}$ with $\overline{\beta}(\alpha \wedge 1)$. To do so, we use the following lemma (to be proved in Section 5.8).

Lemma 2 *For all $\alpha > 0$ and $\boldsymbol{\beta} \in (\mathbb{R}_+^\star)^k$, the smoothness index*

$$\boldsymbol{\theta} = \left(\phi(\alpha,\beta_1), \dots, \phi(\alpha,\beta_k)\right)$$

satisfies $\overline{\theta} \leq \overline{\beta}(\alpha \wedge 1)$ and equality holds if and only if $\sup_{1\leq i\leq k} \beta_i \leq \alpha \vee 1$.

When $\sup_{1\leq i\leq k} \beta_i \leq \alpha \vee 1$, our special strategy does not bring any improvement as compared to the standard one, it even slightly deteriorates the risk bound because of the extra $\log \tau^{-1}$ factor. On the opposite, if $\sup_{1\leq i\leq k} \beta_i > \alpha \vee 1$, our new strategy improves over the classical one and this improvement can be substantial if $\overline{\beta}$ is much larger than $\alpha \vee 1$. If, for instance, $\alpha = 1$ and $\overline{\beta} = k = \beta_j$ for all j, we get a bound of order $\left[\tau\left(\log \tau^{-1}\right)\right]^{2/3}$ which, apart from the extra $\log \tau^{-1}$ factor, corresponds to the minimax rate of estimation of a Lipschitz function on $[-1,1]$, instead of the risk bound $\tau^{2/(2+k)}$ that we would get if we estimated s as a Lipschitz function on $[-1,1]^k$. When our strategy does not improve over the classical one, i.e. when $\sup_{1\leq i\leq k} \beta_i \leq \alpha \vee 1$, the additional loss due to the extra logarithmic factor in our risk bound can be avoided by mixing the models used for the classical strategy with the models used for designing our estimator, following the recipe of Section 3.4.

4.3 Generalized additive models

In this section, we assume that $E = [-1,1]^k$, μ is the Lebesgue probability on E and $q = 2$. A special structure that has often been considered in regression corresponds to functions $s = g \circ u$ with

$$u(x_1, \ldots, x_k) = u_1(x_1) + \ldots + u_k(x_k); \quad s(x) = g\left(u_1(x_1) + \ldots + u_k(x_k)\right), \qquad (4.5)$$

where the u_j take their values in $[-1/k, 1/k]$ for all $j = 1, \ldots, k$. Such a model has been considered in Horowitz and Mammen (2007) and while their approach is non-adaptive, ours, based on Theorem 2 and a suitable choice of the collections of models, allows to derive a fully adaptive estimator with respect to the regularities of g and the u_j. More precisely, for $r \in \mathbb{N}$, let $\mathbb{T}_r$ be the collection of all models of the form $T = T_1 + \ldots + T_k$ where for $j = 1, \ldots, k$, T_j is the set of functions of the form $x \mapsto t_j(x_j)$ with $x \in E$ and t_j in $\mathbb{B}_{1,r}$. Using $\lambda_r = \lambda$ as defined by (4.3), we endow $\mathbb{T}_r$ with the subprobability $\lambda_r^{(k)}$ defined for $T \in \mathbb{T}_r$ by the infimum of the quantities $\prod_{i=1}^k \lambda_r(T_i)$ when $(T_1, \ldots, T_k)$ runs among all the k-uplets of $\mathbb{B}_{1,r}^k$ satisfying $T = T_1 + \ldots + T_k$. Finally, for $\alpha, L > 0$, $\boldsymbol{\beta}, \boldsymbol{R} \in (\mathbb{R}_+^\star)^k$ and $\mathbf{p} = (p_1, \ldots, p_k) \in (\mathbb{R}_+^\star)^k$, let $\mathcal{S}_k^{\mathrm{Add}}(\alpha, \boldsymbol{\beta}, \mathbf{p}, L, \boldsymbol{R})$ be the set of functions of the form (4.5) with $g \in \mathcal{H}^\alpha([-1,1])$ satisfying $\|g\|_{\alpha,\infty} \le L$ and $u_j \in \mathcal{B}_{p_j,p_j}^{\beta_j}([-1,1])$ with $|u_j|_{\beta_j,p_j,p_j} \le R_j k^{-1}$ for all $j = 1, \ldots, k$. Using the sets $\mathfrak{S}_r = (1, \mathbb{H}_{1,r}, \gamma_r, \mathbb{T}_r, \lambda_r^{(k)})$ with $r \in \mathbb{N}$ we can build an estimator with the following property.

Theorem 5 *Assume that Theorem 1 holds with $q = 2$. There exists an estimator $\widehat{s}$ which satisfies for all $\alpha, L > 0$, $\mathbf{p}, \boldsymbol{R} \in (\mathbb{R}_+^\star)^k$ and $\boldsymbol{\beta} \in (\mathbb{R}_+^\star)^k$ with $\beta_j > (1/p_j - 1/2)_+$ for all $j = 1, \ldots, k$,*

$$\sup_{s \in \mathcal{S}_k^{\mathrm{Add}}(\alpha, \boldsymbol{\beta}, \mathbf{p}, L, \boldsymbol{R})} C' \mathbb{E}_s\left[d^2(s, \widehat{s})\right]$$
$$\le \quad L^{\frac{2}{2\alpha+1}} \tau^{\frac{2\alpha}{2\alpha+1}} + \sum_{j=1}^k \left(L (R_j k^{-1/2})^{\alpha \wedge 1}\right)^{\frac{2}{2(\alpha\wedge 1)\beta_j + 1}} (\tau \mathcal{L})^{\frac{2(\alpha\wedge 1)\beta_j}{2(\alpha\wedge 1)\beta_j + 1}} + \tau \mathcal{L},$$

where $\mathcal{L} = \log\left(\tau^{-1}\right) \vee \log\left(L^2\right) \vee 1$ and C' is a constant that depends on $\alpha, \boldsymbol{\beta}, \mathbf{p}$ and k only.

If one is mainly interested in the rate of convergence as τ tends to 0, the bound we get is of order $\max\{\tau^{2\alpha/(2\alpha+1)}, [\tau \log(\tau^{-1})]^{2(\alpha\wedge 1)\beta/(2(\alpha\wedge 1)\beta+1)}\}$ where $\beta = \min\{\beta_1, \ldots, \beta_k\}$. In particular, if $\alpha \ge 1$, this rate is the same (up to a logarithmic factor) as that we would obtain for estimating a function on $[-1,1]$ with the smallest regularity among $\alpha, \beta_1, \ldots, \beta_k$.

Proof: Let us consider some $s = g \circ u \in \mathcal{S}_k^{\mathrm{Add}}(\alpha, \boldsymbol{\beta}, \mathbf{p}, L, R)$ and $r = 1 + \lceil \alpha \vee \beta_1 \vee \ldots \vee \beta_k \rceil$. For all $D, D_1, \ldots, D_k \ge 1$, there exist $F \in \mathbb{H}_{1,r}(D)$ and $T_j \in \mathbb{B}_{1,r}(D_j)$ for all $j = 1, \ldots, k$ such that

$$\mathcal{D}(F) \le C_1'(r) D; \qquad d_\infty(g, F) \le C_2'(r) L D^{-\alpha};$$

and, for $1 \le j \le k$,

$$\mathcal{D}(T_j) \le C_3'(k, r, \mathbf{p})\, D_j; \qquad d(u_j, T_j) \le C_4'(k, r, \mathbf{p})\, R_j k^{-1} D_j^{-\beta_j}.$$

If $T = T_1 + \ldots + T_k$, then $\mathcal{D}(T) \leq \sum_{j=1}^k \mathcal{D}(T_j)$, $\Delta_{\lambda_r^{(k)}}(T) \leq \sum_{j=1}^k \Delta_{\lambda_r}(T_j) \leq (c(\mathbb{B}_{1,r}) + 1)\sum_{j=1}^k (D_j + 1)$. Moreover, $d(u,T) \leq \sum_{j=1}^k d(u_j, T_k) \leq C_4' k^{-1} \sum_{j=1}^k R_j D_j^{-\beta_j}$, hence, $d^2(u,T) \leq (C_4')^2 k^{-1} \sum_{j=1}^k R_j^2 D_j^{-2\beta_j}$ and finally,

$$d^{2(\alpha\wedge 1)}(u,T) \leq (C_4')^{2(\alpha\wedge 1)} \sum_{j=1}^k (R_j k^{-1/2})^{2(\alpha\wedge 1)} D_j^{-2(\alpha\wedge 1)\beta_j}.$$

For all T, $\mathcal{L}_{1,T} \leq C'(\alpha)\mathcal{L}$ and since $\mathrm{w}_g(z) \leq Lz^\alpha$ for all $z \in [0,2]$, we may apply Corollary 1 with $l = 1$ and get that the risk of the resulting estimator $\widehat{s}_r$ satisfies

$$C'\mathcal{R}(\widehat{s}_r, g, u) = \sum_{j=1}^k \inf_{D \geq 1} \left[L^2 (R_j k^{-1/2})^{2(\alpha\wedge 1)} D^{-2(\alpha\wedge 1)\beta_j} + D\tau\mathcal{L} \right] + \inf_{D\geq 1} \left[L^2 D^{-2\alpha} + D\tau \right].$$

We conclude by arguing as in the proof of Theorem 4. □

4.4 Multiple index models and artificial neural networks

In this section, we assume that $E = [-1,1]^k$, $q = 2$ and d is the distance in $\mathbb{L}_2(E,\mu)$ where μ is the Lebesgue probability on E. We denote by $|\cdot|_1$ and $|\cdot|_\infty$ respectively the ℓ_1- and ℓ_∞-norms in $\mathbb{R}^k$ and $\mathcal{C}_k$ the unit ball for the ℓ_1-norm. As we noticed earlier, when s is an arbitrary function on E and k is large, there is no hope to get a nice estimator for s without some additional assumptions. A very simple one is that $s(x)$ can be written as $g(\langle\theta, x\rangle)$ for some $\theta \in \mathcal{C}_k$, which corresponds to the so-called *single index model*. More generally, we may pretend that s can be well approximated by some function $\overline{s}$ of the form

$$\overline{s}(x) = g\left(\langle\theta_1, x\rangle, \ldots, \langle\theta_l, x\rangle\right)$$

where $\theta_1, \ldots, \theta_l$ are l elements of $\mathcal{C}_k$ and g maps $[-1,1]^l$ to $\mathbb{R}$, l being possibly unknown and larger than k. When $\overline{s} = g \circ u$ is of this form, the coordinate functions $u_j(\cdot) = \langle\theta_j, \cdot\rangle$, for $1 \leq j \leq l$, belong to the set $T_0 \subset \mathcal{T}$ of functions on E of the form $x \mapsto \langle\theta, x\rangle$ with $\theta \in \mathcal{C}_k$, which is a subset of a k-dimensional linear subspace of $\mathbb{L}_2(E,\mu)$, hence $\mathcal{D}(T_0) \leq k$. A slight generalization of this situation leads to the following result.

Theorem 6 *Assume that Theorem 1 holds with $q = 2$. For $j \geq 1$, let T_j be a subset of $\mathcal{T}$ with finite dimension k_j and for $I \subset \mathbb{N}^*$ and $l \in I$, let $\mathbb{F}_l$ be a collection of models satisfying Assumptions 1-(i and iii) for some subprobability γ_l. There exists an estimator $\widehat{s}$ which satisfies*

$$C\mathbb{E}_s\left[d^2(s,\widehat{s})\right] \leq \inf_{l\in I} \inf_{g\in\mathcal{F}_{l,c},\, u\in\mathbf{T}_l} \left[d^2(s, g\circ u) + A(g, \mathbb{F}_l, \gamma_l) + \tau \sum_{j=1}^l k_j i(g,j,T_j) \right], \quad (4.6)$$

where $\mathbf{T}_l = T_1 \times \ldots \times T_l$, $i(g,j,T_j)$ is defined by (3.3) and

$$A(g, \mathbb{F}_l, \gamma_l) = \inf_{F\in\mathbb{F}_l} \left\{ d_\infty^2(g,F) + \tau\left[\mathcal{D}(F) + \Delta_{\gamma_l}(F)\right] \right\}.$$

In particular, for all $l \in I$ and $(\boldsymbol{\alpha}, \mathbf{L})$-Hölderian functions g with $\boldsymbol{\alpha} \in (0,1]^l$ and $\mathbf{L} \in (\mathbb{R}_+^\star)^l$

$$C\mathbb{E}_s\left[d^2\left(s,\widehat{s}\right)\right] \leq \inf_{u \in \mathbf{T}_l}\left[d^2(s, g\circ u) + A(g, \mathbb{F}_l, \gamma_l) + \tau\sum_{j=1}^{l} k_j\left[\frac{1}{\alpha_j}\log\left(lL_j^2(k_j\tau)^{-1}\right)\bigvee 1\right]\right]. \quad (4.7)$$

Let us comment on this result, fixing some value $l \in I$. The term $d(s, g\circ u)$ corresponds to the approximation of s by functions of the form $g(u_1(.), \ldots, u_l(.))$ with g in $\mathcal{F}_{l,c}$ and $u_1, \ldots, u_l$ in $T_1, \ldots, T_l$ respectively. As to the quantity $A(g, \mathbb{F}_l, \gamma_l)$, it corresponds to the estimation bound for estimating the function g alone if s were really of the previous form. Finally, the quantity $\tau\sum_{j=1}^{l} k_j i(g, j, T_j)$ corresponds to the sum of the statistical errors for estimating the u_j. If for all j, the dimensions of the T_j remain bounded by some integer $\overline{k}$ independent of τ, which amounts to making a parametric assumption on the u_j, and if g is smooth enough the quantity $\tau\sum_{j=1}^{l} k_j i(g, j, T_j)$ is then of order $\tau\log\tau^{-1}$ for small values of τ as seen in (4.7).

Proof of Theorem 6: For all j, we choose λ_j to be the Dirac mass at T_j so that $\Delta_{\lambda_j}(T_j) = 0 = d(u_j, T_j)$. The result follows by applying Theorem 2 (for a fixed value of $l \in I$) and then Theorem 3 with ν defined by $\nu(l) = e^{-l}$ for all $l \in I$. □

4.4.1 The multiple index model

As already mentioned, the multiple index model amounts to assuming that s is of the form

$$s(x) = g\left(\langle\theta_1, x\rangle, \ldots, \langle\theta_l, x\rangle\right) \quad \text{whatever } x \in E,$$

for some known $l \geq 1$ and $k_j = k$ for all j. For $L > 0$ and $\boldsymbol{\alpha} \in (\mathbb{R}_+^\star)^l$, let us denote by $\mathcal{S}_l^{\boldsymbol{\alpha}}(L)$ the set of functions s of this form with $g \in \mathcal{H}^{\boldsymbol{\alpha}}([-1,1]^l)$ satisfying $\|g\|_{\boldsymbol{\alpha},\infty} \leq L$. Applying Theorem 6 to this special case, we obtain the following result.

Corollary 2 *Assume that Theorem 1 holds with $q = 2$ and let $I \subset \mathbb{N}^\star$. There exists an estimator $\widehat{s}$ such that for all $l \in I$, $\boldsymbol{\alpha} \in (\mathbb{R}_+^\star)^l$ and $L > 0$,*

$$\sup_{s \in \mathcal{S}_l^{\boldsymbol{\alpha}}(L)} C'\mathbb{E}_s\left[d^2\left(s,\widehat{s}\right)\right] \leq L^{\frac{2}{2\overline{\alpha}+1}}\tau^{\frac{2\overline{\alpha}}{2\overline{\alpha}+1}} + k\tau\mathcal{L},$$

where $\mathcal{L} = \log(\tau^{-1}) \vee \log(L^2k^{-1}) \vee 1$ and C' is a constant depending on l and $\boldsymbol{\alpha}$ only.

The effect of the dimension k only appears in the remaining term. The latter is essentially proportional de $k\tau(\log(\tau^{-1}) \vee 1)$, at least for $k \geq L^2$. It is not difficult to see that there is no hope to get a faster rate than $k\tau$ over $\mathcal{S}_l^{\boldsymbol{\alpha}}(L)$. Indeed, by taking $l = L = 1$ for simplicity and g the identity function, we see that $\mathcal{S}_1^{\boldsymbol{\alpha}}(1)$ contains the unit ball of a k-dimensional linear space and this is enough to assert that, at least in the regression setting, the minimax rate is of order $k\tau$. As to the extra logarithmic factor $\log(\tau^{-1})$, we do not know whether it is necessary or not.

Proof: Fix $s = g \circ u \in \mathcal{S}_l^{\boldsymbol{\alpha}}(L)$ and apply Theorem 6 with $T_j = T_0$ for all $j \geq 1$, $I = \{l\}$, $\mathbb{F}_l = \mathbb{H}_{l,r}$ and γ_l defined by (4.2) with $k = l$ and $r = \lceil \alpha_1 \vee \ldots \vee \alpha_l \rceil$. Arguing as in the proof of Theorem 4, we obtain an estimator $\widehat{s}_{(l,r)}$ the risk of which satisfies

$$\mathcal{R}(\widehat{s}_{(l,r)}, g, u) = C' \left[\inf_{D \geq 1} \left(L^2 D^{-2\overline{\alpha}/l} + D\tau \right) + \tau k \mathcal{L} \right] \leq C'' \left[L^{\frac{2}{2\overline{\alpha}+1}} \tau^{\frac{2\overline{\alpha}}{2\overline{\alpha}+1}} + k\tau\mathcal{L} \right],$$

for constants C' and C'' depending on l and $\boldsymbol{\alpha}$ only. Finally, we conclude as in the proof of Theorem 4. □

4.4.2 Case of an additive function g

In the multiple index model, when the value of l is allowed to become large (typically not smaller than k) it is often assumed that g is additive, i.e. of the form

$$g(y_1, \ldots, y_l) = g_1(y_1) + \ldots + g_l(y_l) \quad \text{for all } y \in [-1,1]^l, \tag{4.8}$$

where the g_j are smooth functions from $[-1,1]$ to $\mathbb{R}$. Hereafter, we shall denote by $\mathcal{F}_{l,c}^{\mathrm{Add}}$ the set of such additive functions g. The functions $\overline{s} = g \circ u$ with $g \in \mathcal{F}_{l,c}^{\mathrm{Add}}$ and $u \in T_0^l$ hence take the form

$$\overline{s}(x) = \sum_{j=1}^{l} g_j\left(\langle \theta_j, x \rangle\right) \quad \text{for all } x \in E. \tag{4.9}$$

For each $j = 1, \ldots, l$, let $\mathbb{F}_j$ be a countable family of finite dimensional linear subspaces of $\mathcal{F}_{1,\infty}$ designed to approximate g_j and γ_j some subprobability measure on $\mathbb{F}_j$. Given $(F_1, \ldots, F_l) \in \prod_{j=1}^l \mathbb{F}_j$, we define the subspace F of $\mathcal{F}_{l,\infty}$ as

$$F = \left\{ f(y_1, \ldots, y_l) = f_1(y_1) + \ldots + f_l(y_l) \;\middle|\; f_j \in F_j \text{ for } 1 \leq j \leq l \right\} \tag{4.10}$$

and denote by $\mathbb{F}$ the set of all such F when $(F_1, \ldots, F_l)$ varies among $\prod_{j=1}^l \mathbb{F}_j$. Then, we define a subprobability measure γ on $\mathbb{F}$ by setting

$$\gamma(F) = \prod_{j=1}^{l} \gamma_j(F_j) \quad \text{or} \quad \Delta_\gamma(F) = \sum_{j=1}^{l} \Delta_{\gamma_j}(F_j),$$

when F is given by (4.10). For such an F, $d_\infty(g, F) \leq \sum_{j=1}^l d_\infty(g_j, F_j)$, hence $d_\infty^2(g, F) \leq l \sum_{j=1}^l d_\infty^2(g_j, F_j)$ and $\mathcal{D}(F) \leq \sum_{j=1}^l \mathcal{D}(F_j)$. We deduce from Theorem 6 the following result.

Corollary 3 *Assume that Theorem 1 holds with $q = 2$ and let $I \subset \mathbb{N}^\star$ and for $j \geq 1$, let $\mathbb{F}_j$ be a collection of finite dimensional linear subspaces of $\mathcal{F}_{1,\infty}$ satisfying Assumption 1-i) and-iii) for some subprobability γ_j. There exists an estimator $\widehat{s}$ such that*

$$C\mathbb{E}\left[d^2(s, \widehat{s})\right] \leq \inf_{l \in I} \inf_{g \in \mathcal{F}_{l,c}^{\mathrm{Add}}, u \in T_0^l} \left[d^2(s, g \circ u) + \sum_{j=1}^{l} \left(R_j(g, \mathbb{F}_j, \Delta_{\gamma_j}) + \tau k i(g, j, T_0) \right) \right],$$

where

$$R_j(g, \mathbb{F}_j, \Delta_{\gamma_j}) = \inf_{F_j \in \mathbb{F}_j} \left\{ d^2_\infty(g_j, F_j) + \tau \left[\mathcal{D}(F_j) + \Delta_{\gamma_j}(F_j) \right] \right\} \quad \textit{for } 1 \leq j \leq l.$$

Moreover, if s of the form (4.9) for some $l \in I$ and functions $g_j \in \mathcal{H}^{\alpha_j}([-1,1])$ satisfying $\|g_j\|_{\alpha_j,\infty} \leq L_j$ for $\alpha_j, L_j > 0$ and all $j = 1, \ldots, l$, one can choose the $\mathbb{F}_j$ and γ_j in such a way that

$$\mathbb{E}_s \left[d^2(s, \widehat{s}) \right] \leq C' \left[\sum_{j=1}^{l} L_j^{\frac{2}{2\alpha_j+1}} \tau^{\frac{2\alpha_j}{2\alpha_j+1}} \; + \; k\tau\mathcal{L} \right], \tag{4.11}$$

where $\mathcal{L} = \log\left(\tau^{-1}\right) \vee 1 \vee \left[\bigvee_{j=1}^{l} \log\left(L_j^2 k^{-1}\right) \right]$ and C' is a constant depending on l and $\alpha_1, \ldots, \alpha_l$ only.

For $j \geq 1$, $R_j = R_j(g, \mathbb{F}_j, \Delta_{\gamma_j})$ corresponds to the risk bound for the estimation of the function g_j alone when we use the family of models $\mathbb{F}_j$, i.e. what we would get if we knew θ_j and that $g_i = 0$ for all $i \neq j$. In short, $\sum_{j=1}^{l} R_j$ corresponds to the estimation rate of the additive function g. If each g_j belongs to some smoothness class, this rate is similar to that of a real-valued function defined on the line with smoothness given by the worst component of g, as seen in (4.11).

Proof of Corollary 3: The first part is a straightforward consequence of Theorem 6. For the second part, fix $s = g \circ u$ and $r = \lfloor \alpha_1 \vee \ldots \vee \alpha_l \rfloor$. Since the g_j are $(\alpha_j \wedge 1, L_j)$-Hölderian, $i(g, j, T_0) \leq C'\mathcal{L}$ for some C' depending on the α_j only. By using Proposition 1, Lemma 1 and the collection $\mathbb{F}_{j,r} = \mathbb{H}_{1,r}$ with $\gamma_{j,r}$ defined by (4.2), for all $j = 1, \ldots, l$, $R_j \leq C' \inf_{D \geq 1} \{ L_j^2 D^{-2\alpha_j} + D\tau \} \leq C''(L_j^{2/(2\alpha_j+1)} \tau^{2\alpha_j/(2\alpha_j+1)} + \tau)$ for some constants C', C'' depending on the α_j only. Putting these bounds together, we end up with an estimator $\widehat{s}_r$ the risk of which is bounded from above by the right-hand side of (4.11). We get the result for all values of r by using Theorem 3 and arguing as in the proof of Theorem 4. □

4.4.3 Artificial neural networks

In this section, we consider approximations of s on $E = [-1,1]^k$ by functions of the form

$$\bar{s}(x) = \sum_{j=1}^{l} R_j \psi\left(\langle a_j, x \rangle + b_j \right) \quad \text{with } |b_j| + |a_j|_1 \leq 2^r, \tag{4.12}$$

for given values of $(l, r) \in I = (\mathbb{N}^\star)^2$. Here, $R = (R_1, \ldots, R_l) \in \mathbb{R}^l$, $a_j \in \mathbb{R}^k$, $b_j \in \mathbb{R}$ for $j = 1, \ldots, l$ and ψ is a given uniformly continuous function on $\mathbb{R}$ with modulus of continuity w_ψ. We denote by $\mathcal{S}_{l,r}$ the set of all functions $\bar{s}$ of the form (4.12).

Let us now set $\psi_r(y) = \psi\left(2^r y\right)$ for $y \in \mathbb{R}$ and, for $x \in E$, $u_j(x) = 2^{-r}\left(\langle a_j, x \rangle + b_j\right)$, so that $u_j \in \mathcal{T}$ belongs to the $(k+1)$-dimensional spaces of functions of the form $x \mapsto \langle a, x \rangle + b$. We can then rewrite $\bar{s}$ in the form $g \circ u$ with $g(y_1, \ldots, y_l) = \sum_{j=1}^{l} R_j \psi_r(y_j)$. Since g belongs to the l-dimensional linear space F spanned by the functions $\psi_r(y_j)$, we may set $\mathbb{F} = \{F\}$, $\Delta_\gamma(F) = 0$ and apply Theorem 6. With $\mathrm{w}_{g,j}(y) = |R_j| \mathrm{w}_\psi\left(2^r y\right)$,

(4.6) becomes,

$$C\mathbb{E}_s\big[d^2\left(s,\widehat{s}_{l,r}\right)\big] \le d^2(s,\bar{s}) + \tau(k+1)\sum_{j=1}^{l}\inf\left\{i\in\mathbb{N}^\star \;\big|\; lR_j^2\mathrm{w}_\psi^2\left(2^re^{-i}\right)\le(k+1)\tau i\right\}.$$

If $\mathrm{w}_\psi(y)\le Ly^\alpha$ for some $L>0$, $0<\alpha\le 1$ and all $y\in\mathbb{R}_+$, then, according to (3.4),

$$\begin{aligned} C\mathbb{E}_s\big[d^2\left(s,\widehat{s}_{l,r}\right)\big] &\le d^2(s,\bar{s}) + k\tau\left(\sum_{j=1}^{l}\left[\alpha^{-1}\log\big(lR_j^2L^2 2^{2r\alpha}[k\tau]^{-1}\big)\right]\bigvee 1\right) \\ &\le d^2(s,\bar{s}) + lk\tau\left[r\log 4 + \alpha^{-1}\log_+\Big(l\,|R|_\infty^2\,L^2[k\tau]^{-1}\Big)\right]. \end{aligned} \tag{4.13}$$

These bounds being valid for all $(l,r)\in I$ and $\bar{s}\in\mathcal{S}_{l,r}$, we may apply Theorem 3 to the family of all estimators $\widehat{s}_{l,r}$, $(l,r)\in I$, with ν given by $\nu(l,r)=e^{-l-r}$. We then get the following result.

Theorem 7 *Assume that Theorem 1 holds with $q=2$ and that ψ is a continuous function with modulus of continuity $\mathrm{w}_\psi(y)$ bounded by Ly^α for some $L>0$, $0<\alpha\le 1$ and all $y\in\mathbb{R}_+$. Then one can build an estimator $\widehat{s}=\widehat{s}(\mathbf{X})$ such that*

$$\begin{aligned} &C\mathbb{E}_s\big[d^2\left(s,\widehat{s})\right)\big] \\ &\quad\le \inf_{(l,r)\in I}\inf_{\bar{s}\in\mathcal{S}_{l,r}}\left\{d^2(s,\bar{s}) + lk\tau r\left[1+(r\alpha)^{-1}\log_+\Big(l\,|R|_\infty^2\,L^2[k\tau]^{-1}\Big)\right]\right\}. \end{aligned} \tag{4.14}$$

Approximation by functions of the form (4.12). Various authors have provided conditions on the function s so that it can be approximated within η by functions $\bar{s}$ of the form (4.12) for a given function ψ. An extensive list of authors and results is provided in Section 4.2.2 of Barron, Birgé and Massart (1999) and some proofs are provided in Section 8.2 of that paper. The starting point of such approximations is the assumed existence of a Fourier representation of s of the form

$$s(x) = K_s\int_{\mathbb{R}^k}\cos\big(\langle a,x\rangle+\delta(a)\big)\,dF_s(a), \quad K_s\in\mathbb{R}, \quad |\delta(a)|\le\pi,$$

for some *probability* measure F_s on $\mathbb{R}^k$. To each given function ψ that can be used for the approximation of s is associated a positive number $\beta=\beta(\psi)>0$ and one has to assume that

$$c_{s,\beta} = \int |a|_1^\beta dF_s(a) < +\infty, \tag{4.15}$$

in order to control the approximation of s by functions of the form (4.12). A careful inspection of the proof of Proposition 6 in Barron, Birgé and Massart (1999) shows that, when (4.15) holds, one can derive the following approximation result for s. There exist constants $r_\psi\ge 1$, $\gamma_\psi>0$ and $C_\psi>0$ depending on ψ only, a number $R_{s,\beta}\ge 1$ depending on $c_{s,\beta}$ only and some $\bar{s}\in\mathcal{S}_{l,r}$ with $|R|_1\le R_{s,\beta}$ such that

$$d\left(s,\bar{s}\right)\le K_sC_\psi\left[2^{-r\gamma_\psi}+R_{s,\beta}l^{-1/2}\right] \quad \text{for } r\ge r_\psi. \tag{4.16}$$

Putting this bound into (4.14) and omitting the various indices for simplicity, we get a risk bound of the form

$$\mathcal{R}(l,r) = CK^2\left[2^{-2r\gamma} + R^2 l^{-1} + K^{-2} lk\tau r\left[1 + (r\alpha)^{-1}\log_+\left(lR^2L^2[k\tau]^{-1}\right)\right]\right],$$

to be optimized with respect to $l \geq 1$ and $r \geq r_\psi$. We shall actually perform the optimization with respect to the first three terms, omitting the logarithmic one.

Let us first note that, if $RK \leq \sqrt{r_\psi k\tau}$, one should set $r = r_\psi$ and $l = 1$, which leads to

$$\mathcal{R}(1, r_\psi) \leq Ck\tau r_\psi\left[1 + (r_\psi\alpha)^{-1}\log_+\left(R^2L^2[k\tau]^{-1}\right)\right].$$

Otherwise $\sqrt{r_\psi k\tau} < RK$ and we set

$$r = r^* = \inf\left\{r \geq r_\psi \;\middle|\; 2^{-2r\gamma} \leq (R/K)\sqrt{rk\tau}\right\} \qquad \text{and} \qquad l = l^* = \left\lceil \frac{RK}{\sqrt{r^*k\tau}} \right\rceil.$$

If $l^* > 1$, then $RK(r^*k\tau)^{-1/2} \leq l^* < 2RK(r^*k\tau)^{-1/2}$ hence

$$\mathcal{R}(l^*, r^*) \leq CRK\sqrt{r^*k\tau}\left[1 + \frac{1}{r^*\alpha}\log_+\left(\frac{2R^3L^2K}{(k\tau)^{3/2}\sqrt{r^*}}\right)\right]. \tag{4.17}$$

If $l^* = 1$, then $R^2 \leq K^{-2}r^*k\tau$ and $\sqrt{r_\psi k\tau} < RK \leq \sqrt{r^*k\tau}$, hence $r^* > r_\psi$ and $r^* - 1 \geq r^*/2$. Then, from the definition of r^*,

$$RK^{-1}\sqrt{(r^*/2)k\tau} \leq RK^{-1}\sqrt{(r^*-1)k\tau} < 2^{-2(r^*-1)\gamma} \leq 2^{-2\gamma},$$

hence $\sqrt{r^*k\tau} < (K/R)2^{-2\gamma+(1/2)} < \sqrt{2}K$ and (4.17) still holds. To conclude, we observe that either $-2\gamma r_\psi \log 2 \leq \log\left(RK^{-1}\sqrt{r_\psi k\tau}\right)$ and $r^* = r_\psi$ or the solution z_0 of the equation

$$2z\gamma\log 2 = \log\left(K/\left\lfloor R\sqrt{k\tau}\right\rfloor\right) - (1/2)\log z$$

satisfies $r_\psi < z_0 \leq r^*$. Since $\log z_0 \leq z_0/e$, it follows that

$$r^* \geq \log\left(K/\left\lfloor R\sqrt{k\tau}\right\rfloor\right) / \left(2\gamma\log 2 + e^{-1}\right)$$

and, by monotonicity, that

$$\frac{1}{r^*}\log_+\left(\frac{2R^3L^2K}{(k\tau)^{3/2}\sqrt{r^*}}\right) \leq \mathcal{L} = \left(2\gamma\log 2 + e^{-1}\right)\log_+\left(\frac{2R^3L^2K}{(k\tau)^{3/2}\sqrt{r_\psi}}\right)\left[\log\left(\frac{K}{R\sqrt{k\tau}}\right)\right]^{-1}$$

where $\mathcal{L}$ is a bounded function of $k\tau$. One can also check that

$$r^* \leq \overline{r} = \left\lceil \frac{\log\left(K/\left\lfloor R\sqrt{r_\psi k\tau}\right\rfloor\right)}{2\gamma\log 2} \right\rceil$$

and (4.17) finally leads, when $r^* > r_\psi$, to

$$\mathcal{R}(l^*, r^*) \leq CRK\left(k\tau\left\lceil \frac{\log\left(K/\left\lfloor R\sqrt{r_\psi k\tau}\right\rfloor\right)}{2\gamma\log 2}\right\rceil\right)^{1/2}\left[1 + \alpha^{-1}\mathcal{L}\right]. \tag{4.18}$$

In the asymptotic situation where τ converges to zero, (4.18) prevails and we get a risk bound of order $[-k\tau\log(k\tau)]^{1/2}$.

4.5 Estimation of a regression function and PCA

We consider here the regression framework

$$Y_i = s(X_i) + \varepsilon_i, \quad i = 1, \dots, n,$$

where the X_i are random variables with values in some known compact subset K of $\mathbb{R}^k$ (with $k > 1$ to avoid trivialities) the ε_i are i.i.d. centered random variables of common variance 1 for simplicity and s is an unknown function from $\mathbb{R}^k$ to $\mathbb{R}$. By a proper origin and scale change on the X_i, mapping K into the unit ball $\mathcal{B}_k$ of $\mathbb{R}^k$, one may assume that the X_i belong to $\mathcal{B}_k$, hence that $E = \mathcal{B}_k$, which we shall do from now on. We also assume that the X_i are either i.i.d. with common distribution μ on E (random design) or deterministic ($X_i = x_i$, fixed design), in which case $\mu = n^{-1} \sum_{i=1}^n \delta_{x_i}$, where δ_x denotes the Dirac measure at x. In both cases, we choose for d the distance in $\mathbb{L}_2(E, \mu)$. As already mentioned in Section 2.3, Theorem 1 with $\tau = n^{-1}$ applies to this framework, at least in the two cases when the design is fixed and the errors Gaussian (or subgaussian) or when the design is random and the Y_i are bounded, say with values in $[-1, 1]$.

4.5.1 Introducing PCA

Our aim is to estimate s from the observation of the pairs (X_i, Y_i) for $i = 1, \dots, n$, assuming that s belongs to some smoothness class. More precisely, given $A \subset \mathbb{R}^k$ and some concave modulus of continuity w on $\mathbb{R}_+$, we define $\mathcal{H}_{\mathrm{w}}(A)$ to be the class of functions h on A such that

$$|h(x) - h(y)| \leq \mathrm{w}(|x - y|) \quad \text{for all } x, y \in A.$$

Here we assume that s is defined on $\mathcal{B}_k$ and belongs to $\mathcal{H}_{\mathrm{w}}(\mathcal{B}_k)$, in which case it can be extended to an element of $\mathcal{H}_{\mathrm{w}}\left(\mathbb{R}^k\right)$, which we shall use when needed. Typically, if $\mathrm{w}(z) = Lz^\alpha$ with $\alpha \in (0, 1]$ and the X_i are i.i.d. with uniform distribution μ on E, the minimax risk bound over $\mathcal{H}_{\mathrm{w}}(\mathcal{B}_k)$ with respect to the $\mathbb{L}_2(E, \mu)$-loss is $C' L^{2k/(k+2\alpha)} n^{-2\alpha/(k+2\alpha)}$ (where C' depends on k and the distribution of the ε_i). It can be quite slow if k is large (see Stone (1982)), although no improvement is possible from the minimax point of view if the distribution of the X_i is uniform on $\mathcal{B}_k$. Nevertheless, if the data X_i were known to belong to an affine subspace V of $\mathbb{R}^k$ the dimension l of which is small as compared to k, so that $\mu(V) = 1$, estimating the function s with $\mathbb{L}_2(E, \mu)$-loss would amount to estimating $s \circ \Pi_V$ (where Π_V denotes the orthogonal projector onto V) and one would get the much better rate $n^{-2\alpha/(l+2\alpha)}$ with respect to n for the quadratic risk. Such a situation is seldom encountered in practice but we may assume that it is approximately satisfied for some well-chosen V. It therefore becomes natural to look for an affine space V with dimension $l < k$ such that s and $s \circ \Pi_V$ are close with respect to the $\mathbb{L}_2(E, \mu)$-distance. For $s \in \mathcal{H}_{\mathrm{w}}\left(\mathbb{R}^k\right)$, it follows from Lemma 3 below that,

$$\begin{aligned} \int_E |s(x) - s \circ \Pi_V(x)|^2 \, d\mu(x) &\leq \int_E \mathrm{w}^2 \left(|x - \Pi_V x|\right) d\mu(x) \\ &\leq 2\mathrm{w}^2 \left[\left(\int_E |x - \Pi_V x|^2 \, d\mu(x) \right)^{1/2} \right], \end{aligned}$$

and minimizing the right-hand side amounts to finding an affine space V with dimension l for which $\int_E |x - \Pi_V x|^2 \, d\mu(x)$ is minimum. This way of reducing the dimension is usually known as PCA (for Principal Components Analysis). When the X_i are deterministic and $\mu = n^{-1} \sum_{i=1}^n \delta_{X_i}$, the solution to this minimization problem is given by the affine space $V_l = a + W_l$ where the origin $a = \overline{X}_n = n^{-1} \sum_{i=1}^n X_i \in \mathcal{B}_k$ and W_l is the linear space generated by the eigenvectors associated to the l largest eigenvalues (counted with their multiplicity) of XX^* (where X is the $k \times n$ matrix with columns $X_i - \overline{X}_n$ and X^* is the transpose of X). In the general case, it suffices to set $a = \int_E x d\mu$ (so that $a \in E$) and replace XX^* by the matrix

$$\Gamma = \int_E (x - a)(x - a)^* \, d\mu(x).$$

If $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_k \geq 0$ are the eigenvalues of Γ in nonincreasing order, then

$$\inf_{\{V \mid \dim(V) = l\}} \int_E |x - \Pi_V x|^2 \, d\mu(x) = \sum_{j=l+1}^{k} \lambda_j \tag{4.19}$$

(with the convention $\sum_{\varnothing} = 0$) and therefore

$$\inf_{\{V \mid \dim(V) = l\}} \|s - s \circ \Pi_V\|_2^2 \leq \|s - s \circ \Pi_{V_l}\|_2^2 \leq 2\mathrm{w}^2 \left(\sqrt{\sum_{j=l+1}^{k} \lambda_j} \right). \tag{4.20}$$

4.5.2 PCA and composite functions

In order to put the problem at hand into our framework, we have to express $s \circ \Pi_{V_l}$ in the form $g \circ u$. To do so we consider an orthonormal basis $\overline{u}_1, \ldots, \overline{u}_k$ of eigenvectors of XX^* or Γ (according to the situation) corresponding to the ordered eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_k \geq 0$. For a given value of $l < k$ we denote by $a^\perp$ the component of a which is orthogonal to the linear span W_l of $\overline{u}_1, \ldots, \overline{u}_l$ and for $x \in \mathcal{B}_k$, we define $u_j(x) = \langle x, \overline{u}_j \rangle$ for $j = 1, \ldots, l$. This results in an element $u = (u_1, \ldots, u_l)$ of $\mathcal{T}^l$ and $a^\perp + \sum_{j=1}^l u_j(x) \overline{u}_j = \Pi_{V_l}(x)$ is the projection of x onto the affine space $V_l = a^\perp + W_l$. Setting

$$g(z) = s \left(a^\perp + \sum_{j=1}^{l} z_j \overline{u}_j \right) \quad \text{for } z \in [-1, 1]^l,$$

leads to a function $g \circ u$ with $u \in \mathcal{T}^l$ and $g \in \mathcal{F}_{l,c}$ which coincides with $s \circ \Pi_{V_l}$ on $\mathcal{B}_k$ as required. Consequently, the right-hand side of (4.20) provides a bound on the distance between s and $g \circ u$. Moreover, since $s \in \mathcal{H}_{\mathrm{w}}\left(\mathbb{R}^k\right)$,

$$\left| g(z) - g(z') \right| \leq \mathrm{w} \left(\left| \sum_{j=1}^{l} z_j \overline{u}_j - \sum_{j=1}^{l} z'_j \overline{u}_j \right| \right) = \mathrm{w} \left(\left| \sum_{j=1}^{l} (z_j - z'_j) \overline{u}_j \right| \right) = \mathrm{w}(|z - z'|), \tag{4.21}$$

so that we may set $\mathrm{w}_{g,j} = \mathrm{w}$ for all $j \in \{1, \ldots, l\}$.

In the following sections we shall use this preliminary result in order to establish risk bounds for estimators $\widehat{s}_l$ of s, distinguishing between the two situations where μ is known and μ is unknown.

4.5.3 Case of a known μ

For $D \in \mathbb{N}^\star$, we consider the partition $\mathcal{P}_{l,D}$ of $[-1,1]^l$ into D^l cubes with edge length $2/D$ and denote by $F_{l,D}$ the linear space of functions which are piecewise constant on each element of $\mathcal{P}_{l,D}$ so that $\mathcal{D}(F_{l,D}) = D^l$ for all $D \in \mathbb{N}^\star$. This leads to the family $\mathbb{F} = \{F_{l,D},\, D \in \mathbb{N}^\star\}$ and we set $\gamma(F_{l,D}) = e^{-D}$ for all $D \geq 1$. We define u_j as in the previous section and take for $\mathbb{T}_j$ the family reduced to the single model $T_j = \{u_j\}$ for $j = 1, \ldots, l$. Then $\mathcal{D}(T_j) = 0$ for all j and we take for λ_j the Dirac measure on $\mathbb{T}_j$. This leads to a set $\mathfrak{S}$ which satisfies Assumption 1 and we may therefore apply Theorem 2 which leads to an estimator $\widehat{s}_l$ with a risk bounded by

$$C\mathbb{E}_s\left[\|s - \widehat{s}_l\|_2^2\right] \leq d^2(s, g \circ u) + \inf_{D \geq 1}\left\{d_\infty^2(g, F_{l,D}) + \frac{D^l + D}{n}\right\}.$$

Since $s \circ \Pi_{V_l}$ and $g \circ u$ coincide on $\mathcal{B}_k$, it follows from (4.20) that

$$\|s - g \circ u\|_2^2 = \|s - s \circ \Pi_{V_l}\|_2^2 \leq 2\mathrm{w}^2\left(\sqrt{\sum_{j=l+1}^{k} \lambda_j}\right).$$

Moreover, for all cubes $I \in \mathcal{P}_{l,D}$ and $x \in I$, the Euclidean distance between x and the center of I is at most $\sqrt{l}D^{-1}$, hence by (4.21), $d_\infty(g, F_{l,D}) \leq \mathrm{w}\left(\sqrt{l}D^{-1}\right)$ for all $D \geq 1$. Putting these inequalities together we see that the risk of $\widehat{s}_l$ is bounded by

$$C\mathbb{E}_s\left[\|s - \widehat{s}_l\|_2^2\right] \leq \mathrm{w}^2\left(\sqrt{\sum_{j=l}^{k} \lambda_j}\right) + \inf_{D \geq 1}\left\{\mathrm{w}^2\left(\sqrt{l}D^{-1}\right) + \frac{D^l}{n}\right\}. \tag{4.22}$$

4.5.4 Case of an unknown μ

When μ corresponds to an unknown distribution of the X_i, the matrix Γ is unknown, its eigenvectors $\overline{u}_1, \ldots, \overline{u}_k$ and the vector a as well and therefore also the elements $u_1, \ldots, u_l$ of $\mathcal{T}$. In order to cope with this problem, we have to approximate the unknown u_j which requires to modify the definition of T_j given in the previous section, keeping all other things unchanged. For each $v \in \mathbb{R}^k$ with $|v| \leq 1$, we denote by t_v the linear map, element of $\mathcal{T}$, given by $t_v(x) = \langle x, v\rangle$. Denoting by $\mathcal{B}_k^\circ$ the unit sphere in $\mathbb{R}^k$ we then set, for all j, $T_j = T = \{t_v,\, v \in \mathcal{B}_k^\circ\}$ which is a subset of a k-dimensional linear subspace of $\mathbb{L}_2(\mu)$. It follows that Assumption 1 remains satisfied but now with $\mathcal{D}(T_j) = k$. Since $u_j \in T_j$ for all j, an application of Theorem 2 leads to

$$C\mathbb{E}_s\left[d^2\left(s, \widehat{s}\right)\right] \leq \frac{k}{n}\sum_{j=1}^{l} i(g, j, T) + d^2(s, g \circ u) + \inf_{D \geq 1}\left\{d_\infty^2(g, F_{l,D}) + \frac{D^l + D}{n}\right\},$$

where $i(g, j, T)$ is given by (3.3). Since, by (4.21), $\mathrm{w}_{g,j} = \mathrm{w}$ for all $j \in \{1, \ldots, l\}$,

$$i(g, j, T) = \underline{i} = \inf\left\{i \in \mathbb{N}^\star \,|\, l\mathrm{w}^2\left(e^{-i}\right) \leq \frac{ik}{n}\right\}.$$

Arguing as in the case of a known μ, we get

$$C\mathbb{E}_s\left[\|s - \widehat{s}_l\|_2^2\right] \leq \frac{kl\underline{i}}{n} + \mathrm{w}^2\left(\sqrt{\sum_{j=l+1}^{k} \lambda_j}\right) + \inf_{D \geq 1}\left\{\mathrm{w}^2\left(\sqrt{l}D^{-1}\right) + \frac{D^l}{n}\right\}.$$

Let $i_D = \left\lceil \log\left(D/\sqrt{l}\right)\right\rceil$. If $\underline{i} \le i_D$, then $kl\underline{i}/n \le klD/n$ since $i_D \le D$. Otherwise, $\underline{i} \ge i_D + 1 \ge 2$ and

$$l^2 \mathrm{w}^2\left(e^{-i_D}\right) \ge l^2 \mathrm{w}^2\left(e^{-\underline{i}+1}\right) > \frac{kl(\underline{i}-1)}{n} \ge \frac{kl\underline{i}}{2n},$$

which shows that $kl\underline{i}/n < 2l^2\mathrm{w}^2\left(\sqrt{l}D^{-1}\right) + klD/n$. Finally

$$\begin{aligned} C\mathbb{E}_s\left[\|s-\widehat{s}_l\|_2^2\right] &\le \mathrm{w}^2\left(\sqrt{\sum_{j=l+1}^{k}\lambda_j}\right) + \inf_{D\ge 1}\left\{l^2\mathrm{w}^2\left(\sqrt{l}D^{-1}\right) + \frac{D^l + klD}{n}\right\} \\ &\le \mathrm{w}^2\left(\sqrt{\sum_{j=l+1}^{k}\lambda_j}\right) + lk\inf_{D\ge 1}\left\{\mathrm{w}^2\left(\sqrt{l}D^{-1}\right) + \frac{2D^l}{n}\right\}, \end{aligned}$$

which is, up to constants, the same as (4.22). We do not know whether the multiplicative factor lk arising here and missing in (4.22) can be improved or not.

4.5.5 Varying l

The previous bounds are valid for all values of $l \in I = \{1,\dots,k\}$ but we do not know which value of l will lead to the best estimator. We may therefore apply Theorem 3 with $\nu(l) = l^{-2}/2$ for $l \in I$ which leads to the following risk bound for the new estimator $\widehat{s}$ in the case of a known μ:

$$C\mathbb{E}_s\left[\|s-\widehat{s}\|_2^2\right] \le \inf_{l\in\{1,\dots,k\}}\inf_{D\ge 1}\left[\mathrm{w}^2\left(\sqrt{\sum_{j=l+1}^{k}\lambda_j}\right) + \mathrm{w}^2\left(\sqrt{l}D^{-1}\right) + \frac{D^l + \log l}{n}\right].$$

Apart from multiplicative constants depending only on k, the same result holds when μ is unknown. If $\mathrm{w}(z) = Lz^\alpha$ for some $L > 0$ and $\alpha \in (0,1]$, we get, since $\sum_{j=l+1}^{k}\lambda_j \le (k-l)\lambda_{l+1}$ (with the convention $\lambda_{k+1} = 0$),

$$C\mathbb{E}_s\left[\|s-\widehat{s}\|_2^2\right] \le \inf_{l\in\{1,\dots,k\}}\inf_{D\ge 1}\left\{L^2[(k-l)\lambda_{l+1}]^\alpha + L^2 l^\alpha D^{-2\alpha} + \frac{D^l + \log l}{n}\right\}.$$

Assuming that $n \ge L^{-2}$ to avoid trivialities and choosing $D = \left\lfloor \left(nL^2l^\alpha\right)^{1/(l+2\alpha)}\right\rfloor$, we finally get

$$C\mathbb{E}_s\left[\|s-\widehat{s}\|_2^2\right] \le \inf_{l\in\{1,\dots,k\}}\left\{L^2[(k-l)\lambda_{l+1}]^\alpha + \frac{\log l}{n} + \frac{L^{2l/(l+2\alpha)}}{n^{2\alpha/(l+2\alpha)}}\right\}.$$

For $l = k$, we recover (up to constants) the minimax risk bound over $\mathcal{H}_{\mathrm{w}}(\mathcal{B}_k)$, namely $C'(k)L^{2k/(k+2\alpha)}n^{-2\alpha/(k+2\alpha)}$. Therefore our procedure can only improve the risk as compared to the minimax approach.

4.6 Introducing parametric models

In this section, we approximate s by functions of the form $\overline{s} = g \circ u$ where g belongs to $\mathcal{F}_{l,c}$ and the components u_j of u to parametric models $\mathbf{T}_j = \{u_j(\boldsymbol{\theta}, .),\ \boldsymbol{\theta} \in \Theta_j\} \subset \mathcal{T}$ indexed by subsets Θ_j of $\mathbb{R}^{k_j}$ with $k_j \geq 1$. Besides, we assume that the following holds.

Assumption 3 *For each $j = 1, \ldots, l$, $\Theta_j \subset \mathcal{B}_{k_j}(0, M_j)$ for some positive number M_j and the mapping $\boldsymbol{\theta} \mapsto u_j(\boldsymbol{\theta}, .)$ from Θ_j to $(\mathcal{T}, d)$ is (β_j, R_j)-Hölderian for $\beta_j \in (0, 1]$ and $R_j > 0$ which means that*

$$d(u_j(\boldsymbol{\theta}, .), u_j(\boldsymbol{\theta}', .)) \leq R_j \left|\boldsymbol{\theta} - \boldsymbol{\theta}'\right|^{\beta_j} \quad \text{for all} \quad \boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta_j. \tag{4.23}$$

Under such an assumption, the following result holds.

Theorem 8 *Assume that Theorem 1 holds and let $l \geq 1$, $\mathbf{T}_1, \ldots, \mathbf{T}_l$ be parametric sets satisfying Assumption 3, $\mathbb{F}$ be a collection of models satisfying Assumption 1-i) and γ be a subprobability on $\mathbb{F}$. There exists an estimator $\widehat{s}$ such that*

$$\begin{aligned} C\mathbb{E}_s\left[d^2(s, \widehat{s})\right] & \\ \leq\ & d^2(s, g \circ u) + \inf_{F \in \mathbb{F}} \left[d^2_\infty(g, F) + \tau(\Delta_\gamma(F) + \mathcal{D}(F))\right] \\ & + \tau \left[\sum_{j=1}^{l} k_j \log(1 + 2M_j R_j^{1/\beta_j})\right] + \sum_{j=1}^{l} \inf_{i \geq 1} \left[l \mathrm{w}^2_{g,j}\left(e^{-i}\right) + i\tau\left(1 + k_j \beta_j^{-1}\right)\right], \end{aligned}$$

for all $g \in \mathcal{F}_{l,c}$ and $u_j \in \mathbf{T}_j$, $j = 1, \ldots, l$.

In particular, for all $(\boldsymbol{\alpha}, \mathbf{L})$-Hölderian functions g with $\boldsymbol{\alpha} \in (0, 1]^l$ and $\mathbf{L} \in (\mathbb{R}_+^\star)^l$,

$$\begin{aligned} C\mathbb{E}_s\left[d^2(s, \widehat{s})\right] \ \leq\ & d^2(s, g \circ u) + \inf_{F \in \mathbb{F}} \left[d^2_\infty(g, F) + \tau(\Delta_\gamma(F) + \mathcal{D}(F))\right] \\ & + \tau \sum_{j=1}^{l} \left[k_j \log(1 + 2M_j R_j^{1/\beta_j}) + (\mathcal{L}_j \vee 1)\left(1 + k_j \beta_j^{-1}\right)\right], \end{aligned} \tag{4.24}$$

where

$$\mathcal{L}_j = \frac{1}{2\alpha_j} \log\left(\frac{l L_j^2}{(1 + k_j \beta_j^{-1})\tau}\right) \quad \text{for } j = 1, \ldots, l. \tag{4.25}$$

Although this theorem is stated for a given value of l, we may, arguing as before, let l vary and design a new estimator which achieves the same risk bounds (apart for the constant C) whatever the value of l.

As usual, the quantity $\inf_{F \in \mathbb{F}} \left[d^2_\infty(g, F) + \tau(\Delta_\gamma(F) + \mathcal{D}(F))\right]$ corresponds to the estimation rate for the function g alone by using the collection $\mathbb{F}$. In particular, if $g \in \mathcal{H}^{\boldsymbol{\alpha}}([-1, 1]^l)$ with $\boldsymbol{\alpha} \in (\mathbb{R}_+^\star)^l$, this bound is of order $\tau^{2\overline{\alpha}/(2\overline{\alpha}+l)}$ as τ tends to 0 for a classical choice of $\mathbb{F}$ (see Section 4.1). Since for all j, g is also $(\alpha_j \wedge 1)$-Hölderian as a function of x_j alone, the last term in the right-hand side of (4.24), which is of order $-\tau \log \tau$, becomes negligible as compared to $\tau^{2\overline{\alpha}/(2\overline{\alpha}+l)}$ and therefore, when s is really of the form $g \circ u$ with $g \in \mathcal{H}^{\boldsymbol{\alpha}}([-1, 1]^l)$ the rate we get for estimating s is the same as that for estimating g.

Proof of Theorem 8: For $\eta > 0$ and $j = 1, \ldots, l$, let $\Theta_j[\eta]$ be a maximal subset of Θ_j satisfying $|t - t'| > \eta$ for all t, t' in $\Theta_j[\eta]$. Since Θ_j is a subset of the Euclidean ball in $\mathbb{R}^{k_j}$ centered at 0 with radius M_j, it follows from classical entropy computations (see Lemma 4 in Birgé (2006)) that $\log|\Theta_j[\eta]| \leq k_j \log(1 + 2M_j\eta^{-1})$. For all $i \in \mathbb{N}^\star$, let $T_{j,i}$ be the image of $\Theta_{j,i} = \Theta_j[(R_j e^i)^{-1/\beta_j}]$ by the mapping $\boldsymbol{\theta} \mapsto u_j(\boldsymbol{\theta}, .)$. Clearly,

$$\log|T_{j,i}| \leq \log|\Theta_{j,i}| \leq k_j \log\left(1 + 2M_j R_j^{1/\beta_j} e^{i/\beta_j}\right) \leq k_j \left[\log(1 + 2M_j R_j^{1/\beta_j}) + i\beta_j^{-1}\right]$$

and because of the maximality of $\Theta_{j,i}$ and (4.23), for all $\boldsymbol{\theta} \in \Theta_j$ there exists $\overline{\boldsymbol{\theta}} \in \Theta_{j,i}$ such that $d(u_j(\boldsymbol{\theta}, .), u_j(\overline{\boldsymbol{\theta}}, .)) \leq R_j \left|\boldsymbol{\theta} - \overline{\boldsymbol{\theta}}\right|^{\beta_j} \leq e^{-i}$ so that $T_{j,i}$ is an e^{-i}-net for $\mathbf{T}_j$. For $j = 1, \ldots, l$, we set $\mathbb{T}_j = \bigcup_{i \geq 1} T_{j,i}$ so that the models in $\mathbb{T}_j$ are merely the elements of the sets $T_{j,i}$. For a model T that belongs to $T_{j,i} \setminus \bigcup_{1 \leq i' < i} T_j[e^{-i'}]$ (with the convention $\bigcup_\varnothing = \varnothing$) we set

$$\Delta_{\lambda_j}(T) = \log|T_{j,i}| + i \leq k_j \log\left(1 + 2M_j R_j^{1/\beta_j}\right) + i\left(1 + k_j\beta_j^{-1}\right)$$

which defines a measure λ_j on $\mathbb{T}_j$ satisfying

$$\sum_{T \in \mathbb{T}_j} \lambda_j(T) \leq \sum_{i \geq 1} \sum_{t \in T_{j,i}} \lambda_j(\{t\}) \leq \sum_{i \geq 1} e^{-i} < 1.$$

Since for all j and $T \in \mathbb{T}_j$, $\mathcal{D}(T) = 0$, we get the first risk bound by applying Theorem 2 to the corresponding set $\mathfrak{S}$. To prove (4.24), let us set $i(j) = \lfloor \mathcal{L}_j \rfloor \vee 1$ for $j = 1, \ldots, l$ with $\mathcal{L}_j$ given by (4.25), so that $1 \leq i(j) \leq \mathcal{L}_j \vee 1$ and notice that, if $z \geq \mathcal{L}_j \vee 1$, then $lL_j^2 e^{-2\alpha_j z} \leq z\tau\left(1 + k_j\beta_j^{-1}\right)$. If $\mathcal{L}_j \geq 1$, then $\mathcal{L}_j \leq i(j) + 1 \leq 2\mathcal{L}_j$, hence

$$lL_j^2 e^{-2\alpha_j(i(j)+1)} \leq (i(j) + 1)\tau\left(1 + k_j\beta_j^{-1}\right) \leq 2\mathcal{L}_j\tau\left(1 + k_j\beta_j^{-1}\right)$$

and

$$l\mathrm{w}_{g,j}^2(e^{-i(j)}) \leq lL_j^2 e^{-2\alpha_j i(j)} \leq 2e^{2\alpha_j}\mathcal{L}_j\tau\left(1 + k_j\beta_j^{-1}\right) \leq 2e^2\mathcal{L}_j\tau\left(1 + k_j\beta_j^{-1}\right).$$

Otherwise, $\mathcal{L}_j < 1$, $i(j) = 1 \geq \mathcal{L}_j \vee 1$ and

$$l\mathrm{w}_{g,j}^2(e^{-i(j)}) \leq lL_j^2 e^{-2\alpha_j} \leq \tau\left(1 + k_j\beta_j^{-1}\right),$$

so that in both cases $l\mathrm{w}_{g,j}^2(e^{-i(j)}) \leq 2e^2(\mathcal{L}_j \vee 1)\tau\left(1 + k_j\beta_j^{-1}\right)$, which leads to the conclusion. □

4.6.1 Estimating a density by a mixture of Gaussian densities

In this section, we consider the problem of estimating a bounded density s with respect to some probability μ (to be specified later) on $E = \mathbb{R}^k$, d denoting, as before, the $\mathbb{L}_2$-distance on $\mathbb{L}_2(E, \mu)$. We recall from Section 2.3 that Theorem 1 applies to this situation with $\tau = n^{-1}\|s\|_\infty(1 \vee \log\|s\|_\infty)$. A common way of modeling a density on

$E = \mathbb{R}^k$ is to assume that it is a mixture of Gaussian densities (or close enough to it). More precisely, we wish to approximate s by functions $\overline{s}$ of the form

$$\overline{s}(x) = \sum_{j=1}^{l} q_j p(m_j, \Sigma_j, x) \quad \text{for all} \quad x \in \mathbb{R}^k, \tag{4.26}$$

where $l \geq 1$, $\boldsymbol{q} = (q_1, \ldots, q_l) \in [0,1]^l$ satisfies $\sum_{j=1}^{l} q_j = 1$ and for $j = 1, \ldots, l$, $p(m_j, \Sigma_j, .) = d\mathcal{N}(m_j, \Sigma_j^2)/d\mu$ denotes the density (with respect to μ) of the Gaussian distribution $\mathcal{N}(m_j, \Sigma_j^2)$ centered at $m_j \in \mathbb{R}^k$ with covariance matrix Σ_j^2 for some symmetric positive definite matrix Σ_j. Throughout this section, we shall restrict to means m_j with Euclidean norms not larger than some positive number r and to matrices Σ_j with eigenvalues ρ satisfying $\underline{\rho} \leq \rho \leq \overline{\rho}$ for positive numbers $\underline{\rho} < \overline{\rho}$. In order to parametrize the corresponding densities, we introduce the set Θ gathering the elements θ of the form $\theta = (m, \Sigma)$ where Σ is a positive symmetric matrix with eigenvalues in $[\underline{\rho}, \overline{\rho}]$ and $m \in \mathcal{B}_k(0, r)$. We shall consider Θ as a subset of $\mathbb{R}^{k(k+1)}$ endowed with the Euclidean distance. In particular, the set M_k of square $k \times k$ matrices of dimension k is identified to $\mathbb{R}^{k^2}$ and endowed with the Euclidean distance and the corresponding norm N defined by

$$N^2(A) = \sum_{i=1}^{k} \sum_{j=1}^{k} A_{i,j}^2 \quad \text{if } A = \left(A_{i,j}\right)_{\substack{1 \leq i \leq k \\ 1 \leq j \leq k}}.$$

This norm derives from the inner product $[A, B] = \mathrm{tr}(AB^*)$ (where B^* denotes the transpose of B) on M_k and satisfies $N(AB) \leq N(A)N(B)$ (by Cauchy-Schwarz inequality) and $N(A) = N(UAU^{-1})$ for all orthogonal matrices U. In particular, if A is symmetric and positive with eigenvalues bounded from above by c, $N(A) \leq \sqrt{k}c$. We shall use these properties later on. For $b = r^2/(2\overline{\rho}^2) + k \log(\sqrt{2}\overline{\rho}/\underline{\rho})$ and μ the Gaussian distribution $\mathcal{N}(0, 2\overline{\rho}^2 I_k)$ on $\mathbb{R}^k$ (where I_k denotes the identity matrix) we define the parametric set $\mathbf{T}$ by

$$\mathbf{T} = \left\{ u(\theta, .) = e^{-b/2} \sqrt{p(\theta, .)},\ \theta \in \Theta \right\}.$$

For parameters $\theta_1 = (m_1, \Sigma_1), \ldots, \theta_l = (m_l, \Sigma_l)$ in Θ, the density $\overline{s}$ can be viewed as a composite function $g \circ u$ with

$$g(y_1, \ldots, y_l) = e^b q_1 y_1^2 + \ldots + e^b q_l y_l^2 \tag{4.27}$$

and $u = (u_1, \ldots, u_l)$ with $u_j(.) = u(\theta_j, .)$ for $j = 1, \ldots, l$. With our choices of b and μ, $u(\theta, .) \in \mathcal{T}$ for all $\theta = (m, \Sigma) \in \Theta$ as required, since for all $x \in E$

$$\begin{aligned} p(\theta, x) &= \frac{(2\overline{\rho}^2)^{k/2}}{\det \Sigma} \exp\left[\frac{|x|^2}{4\overline{\rho}^2} - \frac{\left|\Sigma^{-1}(x-m)\right|^2}{2} \right] \\ &\leq (2\overline{\rho}^2 \underline{\rho}^{-2})^{k/2} \exp\left[\frac{|x-m|^2}{2\overline{\rho}^2} + \frac{|m|^2}{2\overline{\rho}^2} - \frac{\left|\Sigma^{-1}(x-m)\right|^2}{2} \right] \\ &\leq (2\overline{\rho}^2 \underline{\rho}^{-2})^{k/2} e^{r^2/(2\overline{\rho}^2)} \leq e^b. \end{aligned}$$

An application of Theorem 8 leads to the following result.

Corollary 4 *Let s be a bounded density in $\mathbb{L}_2(E,\mu)$, $d(.,.)$ be the $\mathbb{L}_2$-distance, $\tau = n^{-1}\|s\|_\infty(1\vee\log\|s\|_\infty)$, $M=\sqrt{k}\overline{\rho}+r$, $b=r^2/(2\overline{\rho}^2)+k\log(\sqrt{2}\overline{\rho}/\underline{\rho})$, $R=\sqrt{k/2}e^{-b/2}\underline{\rho}^{-1}$ and*

$$\mathcal{L}(\tau)=\frac{1}{2}\log\left(\frac{4le^{2b}\tau^{-1}}{1+k(k+1)}\right).$$

There exists an estimator $\widehat{s}$ satisfying for some universal constant $C>0$

$$C\mathbb{E}_s\left[d^2(s,\widehat{s})\right]\leq\inf\left[d^2(s,g\circ u)\right]+lk(k+1)\tau\left[\log(1+2MR)+(\mathcal{L}(\tau)\vee 1)\right], \quad (4.28)$$

where the infimum runs among all functions $u=(u_1,\ldots,u_l)\in\mathbf{T}^l$ and g of the form (4.27).

The second term in the right-hand side of (4.28) does not depend on g nor u and is of order $-\tau\log\tau$ as τ tends to 0. As already mentioned, one can also consider many values of l simultaneously and find the best one by using Theorem 3. Up to a possibly different constant C, the risk of the resulting estimator then satisfies (4.28) for all $l\geq 1$ simultaneously. The problem of estimating the parameters involved in a mixture of Gaussian densities in $\mathbb{R}^k$ has also been considered by Maugis and Michel (2011). Their approach is based on model selection among a family of parametric models consisting of densities of the form (4.26). Nevertheless, they restrict to Gaussian densities with specific forms of covariance matrices only.

Proof of Corollary 4: First note that for all $\theta\in\Theta$, $|\theta|=|m|+N(\Sigma)\leq r+\sqrt{k}\overline{\rho}$. Hence, if we can prove that for all $\theta_0=(m_0,\Sigma_0),\theta_1=(m_1,\Sigma_1)$ in Θ

$$d\left(u(\theta_0,.),u(\theta_1,.)\right)\leq\frac{\sqrt{k/2}\ e^{-b/2}}{\underline{\rho}}\left|\theta_0-\theta_1\right|, \quad (4.29)$$

Assumption 3 will be satisfied with

$$M_j=M=r+\sqrt{k}\overline{\rho}\ \text{ and }\ R_j=\sqrt{k/2}e^{-b/2}\underline{\rho}^{-1}=R\quad\text{for } j=1,\ldots,l.$$

We shall therefore be able to apply Theorem 8 with $\mathbf{T}_j=\mathbf{T}$ for all j, $\tau=n^{-1}\|s\|_\infty(1\vee\log\|s\|_\infty)$, $\mathbb{F}=\{F\}$ where F is the linear span of dimension $\mathcal{D}(F)=l$ of functions g of the form (4.27) and γ the Dirac mass at F. Since the functions g of the form (4.27) are $\mathbf{L}$-Lipschitz with $L_j=2q_je^b\leq 2e^b$ for all j, we shall finally deduce (4.28) from (4.24). We therefore only have to prove (4.29). Let us first note that

$$d^2(u(\theta_0,.),u(\theta_1,.))=2e^{-b}h^2\left(\mathcal{N}(m_0,\Sigma_0^2),\mathcal{N}(m_1,\Sigma_1^2)\right), \quad (4.30)$$

where h denotes the Hellinger distance defined by (1.1). Some classical calculations show that

$$h^2\left(\mathcal{N}(m_0,\Sigma_0^2),\mathcal{N}(m_1,\Sigma_1^2)\right)=1-\frac{\exp\left[-\frac{1}{4}\langle m_1-m_0,(\Sigma_0^2+\Sigma_1^2)^{-1}(m_1-m_0)\rangle\right]}{\sqrt{\det\left(\frac{\Sigma_0^{-1}\Sigma_1+\Sigma_0\Sigma_1^{-1}}{2}\right)}},$$

and from the inequalities, $1-e^{-z}\leq z$ and $\log(\det A)\leq\mathrm{tr}(A-I_k)$ which hold for all $z\in\mathbb{R}$ and all matrices A such that $\det A>0$, by setting $\Sigma^2=\Sigma_0^2+\Sigma_1^2$ we deduce

that

$$
\begin{aligned}
&4h^2\left(\mathcal{N}(m_0,\Sigma_0^2),\mathcal{N}(m_1,\Sigma_1^2)\right)\\
&\quad\leq\ 2\log\left[\det\left(\frac{\Sigma_0^{-1}\Sigma_1+\Sigma_0\Sigma_1^{-1}}{2}\right)\right]+\langle m_1-m_0,\Sigma^{-2}(m_1-m_0)\rangle\\
&\quad\leq\ \mathrm{tr}\left(\Sigma_0^{-1}\Sigma_1+\Sigma_0\Sigma_1^{-1}-2I_k\right)+\langle m_1-m_0,\Sigma^{-2}(m_1-m_0)\rangle\\
&\quad=\ \mathrm{tr}\left((\Sigma_0-\Sigma_1)\Sigma_0^{-1}(\Sigma_0-\Sigma_1)\Sigma_1^{-1}\right)+\langle m_1-m_0,\Sigma^{-2}(m_1-m_0)\rangle\ \ =\ \ U_1+U_2,
\end{aligned}
$$

with

$$
U_1=\mathrm{tr}\left((\Sigma_0-\Sigma_1)\Sigma_0^{-1}(\Sigma_0-\Sigma_1)\Sigma_1^{-1}\right)\ \text{ and }\ U_2=\langle m_1-m_0,\Sigma^{-2}(m_1-m_0)\rangle.
$$

It remains to bound U_1 and U_2 from above. For U_1, taking $A=(\Sigma_0-\Sigma_1)\Sigma_0^{-1}$ and $B=\Sigma_1^{-1}(\Sigma_0-\Sigma_1)$ and using the fact that the eigenvalues of Σ_0^{-1} and Σ_1^{-1} are not larger than $\underline{\rho}^{-1}$, we get

$$
\begin{aligned}
U_1\ &=\ [A,B]\ \leq\ N(A)N(B)\ =\ N((\Sigma_0-\Sigma_1)\Sigma_0^{-1})N(\Sigma_1^{-1}(\Sigma_0-\Sigma_1))\\
&\leq\ N(\Sigma_0^{-1})N(\Sigma_1^{-1})N^2(\Sigma_0-\Sigma_1)\ \leq\ \frac{kN^2(\Sigma_0-\Sigma_1)}{\underline{\rho}^2}.
\end{aligned}
$$

Let us now turn to U_2. It follows from the same arguments that the symmetric matrix $\Sigma^2=\Sigma_0^2+\Sigma_1^2$ satisfies for all $x\in\mathbb{R}^k$,

$$
\langle\Sigma^2x,x\rangle=|\Sigma_0x|^2+|\Sigma_1x|^2\geq 2\underline{\rho}^2\,|x|^2\,,
$$

hence

$$
U_2=\langle m_1-m_0,\Sigma^{-2}(m_1-m_0)\rangle\leq\frac{|m_0-m_1|^2}{2\underline{\rho}^2}.
$$

Putting these bounds together, we obtain that

$$
4h^2\left(\mathcal{N}(m_0,\Sigma_0^2),\mathcal{N}(m_1,\Sigma_1^2)\right)\leq\frac{k}{\underline{\rho}^2}\left(N^2(\Sigma_1-\Sigma_0)+|m_0-m_1|^2\right)=\frac{k}{\underline{\rho}^2}\,|\boldsymbol{\theta}_0-\boldsymbol{\theta}_1|^2\,,
$$

which, together with (4.30), leads to (4.29). □

5 Proofs of the main results

Let us recall that, in this section, d denotes the distance associated to the $\|\ \|_q$ norm of $\mathbb{L}_q(E,\mu)$ and d_∞ the distance associated to the supnorm on $\mathcal{F}_{l,\infty}$.

5.1 Preliminary approximation results

The purpose of this section is to see how well $f\circ t$ approximates $g\circ u$ when we know how well f approximates g and $t=(t_1,\ldots,t_l)$ approximates u.

Proposition 4 *Let $p\geq 1$, $g\in\mathcal{F}_{l,c}$, $f\in\mathcal{F}_{l,\infty}$ and $t,u\in\mathcal{T}^l$. If w_g is a modulus of continuity for g, then*

$$
\|g\circ u-f\circ t\|_p\leq d_\infty(g,f)+2^{1/p}\sum_{j=1}^{l}\mathrm{w}_{g,j}\left(\|u_j-t_j\|_p\right)
$$

with the convention $2^{1/\infty}=1$.

Proof: It relies on the following lemma the proof of which is postponed to the end of the section.

Lemma 3 *Let $(E, \mathcal{E}, \mu)$ be some probability space and w some nondecreasing and nonnegative concave function on $\mathbb{R}_+$ such that $w(0) = 0$. For all $p \in [1, +\infty]$ and $h \in \mathbb{L}_p(\mu)$,*

$$\|w(|h|)\|_p \leq 2^{1/p} w(\|h\|_p),$$

with the convention $2^{1/\infty} = 1$.

We argue as follows. For all $y, y' \in [-1,1]^l$, $|g(y) - g(y')| \leq \sum_{j=1}^l \mathrm{w}_{g,j}(|y_j - y'_j|)$ and, since μ is a probability on E,

$$\begin{aligned}
\|g \circ u - f \circ t\|_p &\leq \|g \circ u - g \circ t\|_p + \|g \circ t - f \circ t\|_p \\
&\leq \left\| \sum_{j=1}^l \mathrm{w}_{g,j}(|u_j - t_j|) \right\|_p + \|g \circ t - f \circ t\|_p \\
&\leq \sum_{j=1}^l \|\mathrm{w}_{g,j}(|u_j - t_j|)\|_p + \sup_{y \in [-1,1]^l} |g(y) - f(y)| \\
&\leq 2^{1/p} \sum_{j=1}^l \mathrm{w}_{g,j}\left(\|u_j - t_j\|_p\right) + d_\infty(g, f),
\end{aligned}$$

which proves the proposition. □

Proof of Lemma 3: Since there is nothing to prove if $\|h\|_p = 0$, we shall assume that $\|h\|_p > 0$. The assumptions on w imply that, for all $0 < a < b$, $b^{-1}w(b) \leq a^{-1}w(a)$ and $w(a) \leq w(b)$. Consequently, for $p \in [1, +\infty[$,

$$\begin{aligned}
\int_E w^p(|h|)\, d\mu &= \int_E w^p(|h|) \mathbb{1}_{|h| \leq b}\, d\mu + \int_E w^p(|h|) \mathbb{1}_{|h| > b}\, d\mu \\
&\leq w^p(b) + \int_E \frac{w^p(|h|)}{|h|^p} |h|^p \mathbb{1}_{|h| > b}\, d\mu \leq w^p(b) + \frac{w^p(b)}{b^p} \int_E |h|^p\, d\mu,
\end{aligned}$$

and the result follows by choosing $b = \|h\|_p$. The case $p = \infty$ can be deduced by letting p go to $+\infty$. □

5.2 Basic theorem

We shall first prove a general theorem of independent interest that applies to finite models $\mathbf{T}$ for functions in $\mathcal{T}^l$ and is at the core of all our further developments.

Theorem 9 *Let I be a countable set and ν a subprobability on I. Assume that, for each $\ell \in I$, we are given two countable families $\mathbb{T}_\ell$ and $\mathbb{F}_\ell$ of subsets of $\mathcal{T}^l$ and $\mathcal{F}_{l,\infty}$ respectively such that each element $\mathbf{T}$ of $\mathbb{T}_\ell$ is finite and each $F \in \mathbb{F}_\ell$ is a linear subspace of dimension $\mathcal{D}(F) \geq 1$ of $\mathcal{F}_{l,\infty}$. Let λ_ℓ and γ_ℓ be subprobabilities on $\mathbb{T}_\ell$ and $\mathbb{F}_\ell$ respectively. One can design an estimator $\widehat{s} = \widehat{s}(\boldsymbol{X})$ satisfying, for all $\ell \in I$, all*

$u \in \mathcal{T}^l$ and $g \in \mathcal{F}_{l,c}$ with modulus of continuity w_g,

$$C\mathbb{E}_s\left[d^2\left(s,\widehat{s}\right)\right] \leq \inf_{\mathbf{T}\in\mathbb{T}_\ell}\left\{ l \inf_{t\in\mathbf{T}} \sum_{j=1}^{l} \mathrm{w}_{g,j}^2\left(\|u_j - t_j\|_p\right) + \tau\left[\Delta_{\lambda_\ell}(\mathbf{T}) + \log|\mathbf{T}| + \Delta_\nu(\ell)\right]\right\} + d^2(s, g\circ u) + \inf_{F\in\mathbb{F}_\ell}\left\{d_\infty^2(g,F) + \tau\left[\mathcal{D}(F) + \Delta_{\gamma_\ell}(F)\right]\right\}.$$

Proof: For each $t \in \bigcup_{\mathbf{T}\in\mathbb{T}_\ell} \mathbf{T}$ and $F \in \mathbb{F}_\ell$ we consider the set $F_t = \{f\circ t,\ f\in F\} \subset \mathbb{L}_q(E,\mu)$, which is a $\mathcal{D}(F)$-dimensional linear space. This leads to a new countable family of models $\mathbb{S}_\ell$ together with a subprobability π_ℓ on $\mathbb{S}_\ell$ given by

$$\mathbb{S}_\ell = \left\{ F_t,\, t\in \bigcup_{\mathbf{T}\in\mathbb{T}_\ell} \mathbf{T},\, F\in\mathbb{F}_\ell \right\};\quad \pi_\ell(F_t) = \gamma_\ell(F) \inf_{\mathbf{T}\in\mathbb{T}_\ell,\,\mathbf{T}\ni t} |\mathbf{T}|^{-1}\lambda_\ell(\mathbf{T}). \tag{5.1}$$

We then set

$$\mathbb{S} = \bigcup_{\ell\in I} \mathbb{S}_\ell \qquad \text{and} \qquad \pi(F_t) = \nu(\ell)\pi_\ell(F_t) \quad \text{for } F_t \in \mathbb{S}_\ell.$$

It follows that

$$\Delta_\pi(F_t) = \Delta_{\gamma_\ell}(F) + \inf_{\mathbf{T}\in\mathbb{T}_\ell,\,\mathbf{T}\ni t} \left[\Delta_{\lambda_\ell}(\mathbf{T}) + \log(|\mathbf{T}|)\right] + \Delta_\nu(\ell) \quad \text{for } F_t\in\mathbb{S}_\ell.$$

Applying Theorem 1 to $\mathbb{S}$ and π leads to an estimator $\widehat{s}$ satisfying, for each $\ell \in I$,

$$C\mathbb{E}_s\left[d^2\left(s,\widehat{s}\right)\right] \leq \inf_{F\in\mathbb{F}_\ell,\,\mathbf{T}\in\mathbb{T}_\ell,\,t\in\mathbf{T}} \left\{d^2(s,F_t) + \tau\left[\mathcal{D}(F) + \Delta_{\gamma_\ell}(F) + \Delta_{\lambda_\ell}(\mathbf{T}) + \log|\mathbf{T}| + \Delta_\nu(\ell)\right]\right\}.$$

We now use Proposition 4 which implies that, for each $f\circ t \in F_t$,

$$\begin{aligned} d^2(s, f\circ t) &\leq \left(\|s - g\circ u\|_q + \|g\circ u - f\circ t\|_q\right)^2 \\ &\leq \left(\|s - g\circ u\|_q + d_\infty(g,f) + 2^{1/q}\sum_{j=1}^{l} \mathrm{w}_{g,j}\left(\|u_j - t_j\|_q\right)\right)^2 \\ &\leq 3\left(\|s - g\circ u\|_q^2 + d_\infty^2(g,f) + 4l\sum_{j=1}^{l} \mathrm{w}_{g,j}^2\left(\|u_j - t_j\|_q\right)\right), \end{aligned}$$

for some universal constant C since $2^{1/q} \leq 2$. The conclusion follows from a minimization over all possible choices for f and t. □

5.3 Building new models

In order to use Theorem 9, which applies to finite sets $\mathbf{T}$, starting from the models T which satisfy Assumption 1, we need to derive new models from the original ones. Let us first observe that, since u_j takes its values in $[-1,1]$ and μ is a probability on E, $d(0,u_j) \leq 1$. It is consequently useless to try to approximate u_j by elements of $\mathbb{L}_q(E,\mu)$ that do not belong to $\mathcal{B}(0,2)$ since 0 always does better. We may therefore

replace $T \subset \mathbb{L}_q(E,\mu)$ by $\big(T \cap \mathcal{B}(0,2)\big) \cup \{0\}$, denoting again the resulting set, which remains a subset of some $\mathcal{D}(T)$-dimensional linear space, by T. Moreover, this modification can only decrease the value of $d(T,u_j)$. Since now $T \subset \mathcal{B}(0,2)$, we can use the discretization argument described by the following lemma.

Lemma 4 *Let $T \subset \mathcal{B}(0,2)$ be either a singleton (in which case $\mathcal{D}(T)=0$) or a subset of some $\mathcal{D}(T)$-dimensional linear subspace of $\mathbb{L}_q(E,\mu)$ with $\mathcal{D}(T) \geq 1$. For each $\eta \in (0,1]$, one can find a subset $T[\eta]$ of $\mathcal{T}$ with cardinality bounded by $(5/\eta)^{\mathcal{D}(T)}$ such that*

$$\inf_{t\in T[\eta]} d(t,v) \leq \inf_{t\in T} d(t,v) + [\eta \wedge \mathcal{D}(T)] \quad \textit{for all } v \in \mathcal{T}. \tag{5.2}$$

Proof: If $\mathcal{D}(T)=0$, then $T=\{t\}$, we set $T[\eta]=\{(-1\vee t)\wedge 1\}$ and the result is immediate since v takes its values in $[-1,1]$. Otherwise, let T' be a maximal subset of T such that $d(t,t') > \eta$ for each pair (t,t') of distinct points in T'. Then, for each $t \in T$ there exists $t' \in T'$ such that $d(t,t') \leq \eta$ and it follows from Lemma 4 in Birgé (2006) that $|T'| \leq (5/\eta)^{\mathcal{D}(T)}$. Now set $T[\eta]=\{(-1\vee t)\wedge 1,\, t\in T'\}$. Then (5.2) holds since $\mathcal{D}(T)\geq 1$. □

We are now in a position to build discrete models for approximating the elements of $\mathcal{T}^l$. Given $j \in \{1,\ldots,l\}$, T_j in $\mathbb{T}_j$ and some $i \in \mathbb{N}^\star$, the previous lemma provides a set $T_j\big[e^{-i}\big]$ satisfying $\big|T_j\big[e^{-i}\big]\big| \leq \exp\left[\mathcal{D}(T_j)\,(i+\log 5)\right]$. Moreover,

$$d\left(u_j, T_j\left[e^{-i}\right]\right) \leq d\left(u_j,T_j\right) + \left[e^{-i} \wedge \mathcal{D}(T_j)\right] \quad \text{for all } u \in \mathcal{T}^l \text{ and } i \in \mathbb{N}^\star. \tag{5.3}$$

We then define the family $\mathbb{T}$ of models by

$$\mathbb{T} = \left\{ \mathbf{T} = \prod_{j=1}^{l} T_j\big[e^{-i_j}\big] \text{ with } (i_j,T_j) \in \mathbb{N}^\star \times \mathbb{T}_j \text{ for } j=1,\ldots,l \right\}. \tag{5.4}$$

Then each $\mathbf{T} = T_1\left[e^{-i_1}\right] \times \ldots \times T_l\left[e^{-i_l}\right]$ in $\mathbb{T}$ has a finite cardinality bounded by

$$\log|\mathbf{T}| \leq \sum_{j=1}^{l} \mathcal{D}(T_j)\,(i_j+\log 5). \tag{5.5}$$

5.4 Proof of Theorem 2

Starting from the families $\mathbb{T}_j$, $1 \leq j \leq l$, we build the set $\mathbb{T}$ given by (5.4) as indicated in the previous section and we apply Theorem 9 to $\mathbb{F}$ and $\mathbb{T}$. This requires to define a suitable subprobability λ on $\mathbb{T}$, which can be done by setting, for each $\mathbf{T} = T_1\left[e^{-i_1}\right] \times \ldots \times T_l\left[e^{-i_l}\right]$ in $\mathbb{T}$,

$$\lambda(\mathbf{T}) = \prod_{j=1}^{l} \lambda_j(T_j) \exp\left[-i_j \mathcal{D}(T_j)\right] \quad \text{or} \quad \Delta_\lambda(\mathbf{T}) = \sum_{j=1}^{l} \left[\Delta_{\lambda_j}(T_j) + i_j \mathcal{D}(T_j)\right].$$

Applying Theorem 9 to $\mathbb{F}$ and $\mathbb{T}$ with I reduced to a single element and ν the Dirac measure and using (5.5) and (5.3) which implies that

$$\begin{aligned} \inf_{t_j \in T_j\left[e^{-i_j}\right]} \mathrm{w}_{g,j}\left(\|u_j - t_j\|_p\right) &\leq \mathrm{w}_{g,j}\left(\left[e^{-i_j} \wedge \mathcal{D}(T_j)\right] + d\left(u_j,T_j\right)\right) \\ &\leq \mathrm{w}_{g,j}\left(e^{-i_j} \wedge \mathcal{D}(T_j)\right) + \mathrm{w}_{g,j}\left(d(u_j,T_j)\right) \end{aligned}$$

by the subadditivity property of the modulus of continuity $\mathrm{w}_{g,j}$, we get the risk bound

$$\begin{aligned} C\mathbb{E}_s\left[d^2\left(s,\widehat{s}\right)\right] \;\leq\; & \inf \sum_{j=1}^{l} \Big\{ 2l \left[\mathrm{w}_{g,j}^2(d(u_j,T_j)) + \mathrm{w}_{g,j}^2\big(e^{-i_j} \wedge \mathcal{D}(T_j)\big)\right] \\ & \qquad + \tau \left[\Delta_{\lambda_j}(T_j) + (2i_j + \log 5)\mathcal{D}(T_j)\right] \Big\} \\ & + d^2(s, g\circ u) + \inf_{F\in\mathbb{F}} \left\{ d_\infty^2(g,F) + \tau\left[\mathcal{D}(F) + \Delta_\gamma(F)\right]\right\}, \end{aligned}$$

where the first infimum runs among all $T_j \in \mathbb{T}_j$ and all $i_j \in \mathbb{N}^\star$ for $j = 1,\ldots,l$. Setting $i_j = i(g,j,T_j)$ implies that $l\mathrm{w}_{g,j}^2\left(e^{-i_j} \wedge \mathcal{D}(T_j)\right) \leq \tau i_j \mathcal{D}(T_j)$, which proves (3.2). As to (3.6), it simply derives from the fact that, if $\mathcal{D}(T) \geq 1$, then

$$i(g,j,T) \leq \left\lceil (2\alpha_j)^{-1} \log\left(lL_j^2[\tau\mathcal{D}(T)]^{-1}\right)\right\rceil \leq \left[\alpha_j^{-1}\log\left(lL_j^2[\tau\mathcal{D}(T)]^{-1}\right)\right] \bigvee 1 = \mathcal{L}_{j,T}.$$

5.5 Proof of Theorem 3

It follows exactly the line of proof of Theorem 2 via Theorem 9 with an additional step in order to mix the different families of models corresponding to the various sets $\mathfrak{S}_\ell$. To each $\mathfrak{S}_\ell$ corresponds a family of models $\mathbb{S}_\ell$ and a subprobability π_ℓ on $\mathbb{S}_\ell$ given by (5.1). We again apply Theorem 9 with I and ν as given in Theorem 3.

5.6 Proof of Lemma 1

If $D = 1$, we get the bound $a + b$. When $a > b$, we can choose D such that $(a/b)^{1/(\theta+1)} \leq D < (a/b)^{1/(\theta+1)} + 1$, so that

$$aD^{-\theta} + bD < a(a/b)^{-\theta/(\theta+1)} + b\left[(a/b)^{1/(\theta+1)} + 1\right] = b + 2a^{1/(\theta+1)}b^{\theta/(\theta+1)}$$

and the bound $b + \left[2a^{1/(\theta+1)}b^{\theta/(\theta+1)} \wedge a\right]$ follows. If $b \geq a$, the bound $2b$ holds, otherwise $b < a^{1/(\theta+1)}b^{\theta/(\theta+1)}$ and the conclusion follows.

5.7 Proof of Proposition 3

It suffices to show that for all $i \in \{1,\ldots,k\}$ and $x \in [-1,1]^k$, the map $g\circ u_x(t) = g\circ u(x_1,\ldots,x_{i-1},t,x_{i+1},\ldots,x_k)$ from $[-1,1]$ into $\mathbb{R}$ belongs to $\mathcal{H}^{\boldsymbol{\theta}}([-1,1]^k)$. If at least α or β_i are not larger than 1, the result is clear. Otherwise both are larger than 1 and we can write $\beta_i = b_i + \beta_i'$ and $\alpha = a + \alpha'$ with $a, b \in \mathbb{N}^\star$ and $\beta_i', \alpha' \in (0,1]$. Both functions g and u_x are $b_i \wedge a$ times differentiable and the derivatives $g^{(\ell)}\circ u_x$ and $u_x^{(\ell)}$ for $\ell = 0,\ldots,b_i\wedge a$ are Hölderian with smoothness $\rho = (\beta_i - b_i\wedge\alpha)\wedge(\alpha - b_i\wedge a) \in (0,1]$. Since the derivative of order $b_i \wedge a$ of $g\circ u_x$ is a polynomial with respect to these functions, we derive (4.4) from the fact that the set $(\mathcal{H}^\rho([-1,1]^k),+,.)$ is an algebra on $\mathbb{R}$.

We shall prove the second part of the proposition for the case $k = 1$ only since the general case can be proved by similar arguments. For $\rho > 0$, let $h_\rho \in \mathcal{H}^\rho([-1,1]) \setminus \bigcup_{\rho'>\rho}\mathcal{H}^{\rho'}([-1,1])$. Given $\alpha,\beta > 0$, we distinguish between the cases below and the reader can check that for each of these $g \in \mathcal{H}^\alpha([-1,1])$, $u \in \mathcal{H}^\beta([-1,1])$, $g\circ u \in \mathcal{H}^\theta([-1,1])$ with $\theta = \phi(\alpha,\beta)$ but $g\circ u \notin \mathcal{H}^{\theta'}([-1,1])$ whatever $\theta' > \theta$. If $\alpha,\beta \leq 1$,

take $g(x) = |x|^\alpha$ and $u(y) = |y|^\beta$ for all $x, y \in [-1, 1]$, if $1 < \beta$ and $\alpha \leq \beta$, take $g = h_\alpha$ and $u(y) = y$ for all $y \in [-1, 1]$, finally, if $\alpha > 1$ and $\alpha > \beta$, take $g(x) = x$ for all $x \in [-1, 1]$ and $u = h_\beta$.

5.8 Proof of Lemma 2

For all $\alpha > 0$, the map defined for y in $(0, +\infty)$ by

$$\phi_\alpha(y) = \frac{1}{\phi(\alpha, 1/y)} = \begin{cases} y(\alpha \wedge 1)^{-1} & \text{if } y \geq (\alpha \vee 1)^{-1}; \\ \alpha^{-1} & \text{otherwise,} \end{cases}$$

is positive, piecewise linear and convex. Hence,

$$\frac{1}{\overline{\theta}} = \frac{1}{k} \sum_{i=1}^{k} \phi_\alpha \left(\frac{1}{\beta_i} \right) \geq \phi_\alpha \left(\frac{1}{\overline{\beta}} \right) = \frac{1}{\phi(\alpha, \overline{\beta})}$$

and equality holds if and only if $\beta_i \leq (\alpha \vee 1)$ for all i or if for all i, $\beta_i \geq (\alpha \vee 1)$. We conclude by using the fact that $\phi(\alpha, z) \leq z(\alpha \wedge 1)$ for all positive number z and that equality holds if and only if $z \leq \alpha \vee 1$.

References

Akakpo, N. (2012). Adaptation to anisotropy and inhomogeneity via dyadic piecewise polynomial selection. *Math. Methods Statist.*, 21(1):1-28.

Baraud, Y. (2011). Estimator selection with respect to Hellinger-type risks. *Probab. Theory Related Fields*, 151(1-2):353–401.

Baraud, Y., Comte, F., and Viennet, G. (2001). Model selection for (auto-)regression with dependent data. *ESAIM Probab. Statist.*, 5:33–49 (electronic).

Baraud, Y., Giraud, C., and Huet, S. (2009). Gaussian model selection with an unknown variance. *Ann. Statist.*, 37(2):630–672.

Barron, A., Birgé, L., and Massart, P. (1999). Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, 113(3):301–413.

Barron, A. R. (1993). Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Trans. Inform. Theory*, 39(3):930–945.

Barron, A. R. (1994). Approximation and estimation bounds for artificial neural networks. *Machine Learning*, 14:115–133.

Birgé, L. (2006). Model selection via testing: an alternative to (penalized) maximum likelihood estimators. *Ann. Inst. H. Poincaré Probab. Statist.*, 42(3):273–325.

Birgé, L. (2007). Model selection for Poisson processes. In *Asymptotics: Particles, Processes and Inverse Problems, Festschrift for Piet Groeneboom*, number 55, pages 32–64. E. Cator, G. Jongbloed, C. Kraaikamp, R. Lopuhaä and J. Wellner, eds. IMS Lecture Notes – Monograph Series.

Birgé, L. (2008). Model selection for density estimation with $\mathbb{L}_2$-loss. *ArXiv e-prints*.

Birgé, L. and Massart, P. (2001). Gaussian model selection. *J. Eur. Math. Soc. (JEMS)*, 3(3):203–268.

Dahmen, W., DeVore, R., and Scherer, K. (1980). Multidimensional spline approximation. *SIAM J. Numer. Anal.*, 17(3):380–402.

DeVore, R. and Lorentz, G. (1993). *Constructive Approximation*. Springer-Verlag.

Friedman, J. and Tuckey, J. (1974). A projection pursuit algorithm for exploratory data analysis. *IEEE Trans. Comput.*, C-23(881-889).

Hochmuth, R. (2002). Wavelet characterizations for anisotropic Besov spaces. *Appl. Comput. Harmon. Anal.*, 12(2):179–208.

Horowitz, J. L. and Mammen, E. (2007). Rate-optimal estimation for a general class of nonparametric regression models with unknown link functions. *Ann. Statist.*, 35(6):2589–2619.

Huber, P. J. (1985). Projection pursuit. *Ann. Statist.*, 13(2):435–525. With discussion.

Juditsky, A. B., Lepski, O. V., and Tsybakov, A. B. (2009). Nonparametric estimation of composite functions. *Ann. Statist.*, 37(3):1360–1404.

Maugis, C. and Michel, B. (2011). A non asymptotic penalized criterion for Gaussian mixture model selection *ESAIM Probab. Statist.*, 15:41–68.

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.*, 10(4):1040–1053.