Electronic Journal of Statistics
Vol. 1 (2007) 538–573
ISSN: 1935-7524
DOI: 10.1214/07-EJS077

Optimal rates and adaptation in the single-index model using aggregation

Stéphane Gaïffas

Université Paris 6
Laboratoire de Statistique Théorique et Appliquée
175 rue du Chevaleret
75013 Paris
e-mail: gaiffas@ccr.jussieu.fr

Guillaume Lecué

Université Paris 6
Laboratoire de Probabilités et Modèles Aléatoires
175 rue du Chevaleret
75013 Paris
e-mail: lecue@ccr.jussieu.fr

Abstract: We want to recover the regression function in the single-index model. Using an aggregation algorithm with local polynomial estimators, we answer in particular to the second part of Question 2 from Stone (1982) on the optimal convergence rate. The procedure constructed here has strong adaptation properties: it adapts both to the smoothness of the link function and to the unknown index. Moreover, the procedure locally adapts to the distribution of the design. We propose new upper bounds for the local polynomial estimator (which are results of independent interest) that allows a fairly general design. The behavior of this algorithm is studied through numerical simulations. In particular, we show empirically that it improves strongly over empirical risk minimization.

AMS 2000 subject classifications: Primary 62G08; secondary 62H12.
Keywords and phrases: Nonparametric regression, single-index, aggregation, adaptation, minimax, random design, oracle inequality.

Received July 2007.

1. Introduction

The single-index model is standard in statistical literature. It is widely used in several fields, since it provides a simple trade-off between purely nonparametric and purely parametric approaches. Moreover, it is well-known that it allows to deal with the so-called "curse of dimensionality" phenomenon. Within the minimax theory, this phenomenon is explained by the fact that the minimax rate linked to this model (which is multivariate, in the sense that the number of explanatory variables is larger than 1) is the same as in the univariate model. Indeed, if n is the sample size, the minimax rate over an isotropic s-Hölder ball is $n^{-2s/(2s+d)}$ for mean integrated square error (MISE) in the d-dimensional regression model without the single-index constraint, while in the single-index model, this rate is conjectured to be $n^{-2s/(2s+1)}$ by Stone (1982). Hence, even for small values of d (larger than 2), the dimension has a strong impact on the quality of estimation when no prior assumption on the structure of the multivariate regression function is made. In this sense, the single-index model provides a simple way to reduce the dimension of the problem.

Let $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$ be a random variable satisfying

$$Y = g(X) + \sigma(X)\varepsilon, \tag{1.1}$$

where ε is independent of X with law $N(0, 1)$ and where $\sigma(\cdot)$ is such that $\sigma_0 < \sigma(X) \leq \sigma_1$ a.s. for some $\sigma_0 > 0$ and a known $\sigma_1 > 0$. We denote by P the probability distribution of (X, Y) and by P_X the margin law in X or *design* law. In the single-index model, the regression function as a particular structure. Indeed, we assume that g can be written has

$$g(x) = f(\vartheta^\top x) \tag{1.2}$$

for all $x \in \mathbb{R}^d$, where $f : \mathbb{R} \to \mathbb{R}$ is the *link* function and where the direction $\vartheta \in \mathbb{R}^d$, or *index*. In order to make the representation (1.2) unique (identifiability), we assume the following (see for instance the survey paper by Geenens and Delecroix (2005), or Chapter 2 in Horowitz (1998)):

- f is not constant over the support of $\vartheta^\top X$;
- X admits at least one continuously distributed coordinate (w.r.t. the Lebesgue measure);
- the support of X is not contained in any linear subspace of $\mathbb{R}^d$;
- $\vartheta \in S_+^{d-1}$, where S_+^{d-1} is the half-unit sphere defined by

$$S_+^{d-1} = \big\{ v \in \mathbb{R}^d \mid \|v\|_2 = 1 \text{ and } v_d \geq 0 \big\}, \tag{1.3}$$

where $\| \cdot \|_2$ is the Euclidean norm over $\mathbb{R}^d$.

We assume that the available data

$$D_n := [(X_i, Y_i); 1 \leq i \leq n] \tag{1.4}$$

is a sample of n i.i.d. copies of (X, Y) satisfying (1.1) and (1.2). In this model, we can focus on the estimation of the index ϑ based on D_n when the link function f is unknown, or we can focus on the estimation of the regression g when both f and ϑ are unknown. In this paper, we consider the latter problem. It is assumed below that f belongs to some family of Hölder balls, that is, we do not suppose its smoothness to be known.

Statistical literature on this model is wide. Among many other references, see Horowitz (1998) for applications in econometrics, an application in medical science can be found in Xia and Härdle (2006), see also Delecroix et al. (2003), Delecroix et al. (2006) and the survey paper by Geenens and Delecroix (2005). For the estimation of the index, see for instance Hristache et al. (2001); for testing the parametric versus the nonparametric single-index assumption, see Stute and Zhu (2005). See also a chapter in Györfi et al. (2002) which is devoted to dimension reduction techniques in the bounded regression model. While the literature on single-index modelling is vast, several problems remain open. For instance, the second part of Question 2 from Stone (1982) concerning the minimax rate over Hölder balls in model (1.1),(1.2) is still open. The first part, concerning additive modelling is handled in Yang (2000a) and Yang and Barron (1999).

This paper provides new minimax results about the single-index model, which provides an answer, in particual, to the latter question. Indeed, we prove that in model (1.1),(1.2), we can achieve the rate $n^{-2s/(2s+1)}$ for a link function in a whole family of Hölder balls with smoothness s, see Theorem 1. The optimality of this rate is proved in Theorem 2. To prove the upper bound, we use an estimator which adapts both to the index parameter and to the smoothness of the link function. This result is stated under fairly general assumptions on the design, which include any "non-pathological" law for P_X. Moreover, this estimator has a nice "design-adaptation" property, since it does not depend within its construction on P_X.

2. Construction of the procedure

The procedure developed here for recovering the regression does not use a plugin estimator by direct estimation of the index. Instead, it *adapts* to it, by aggregating several univariate estimators based on projected samples

$$D_m(v) := [(v^\top X_i, Y_i), 1 \leq i \leq m], \tag{2.1}$$

where $m < n$, for several v in a regular lattice of S_+^{d-1}. This "adaptation to the direction" uses a split of the sample. We split the whole sample D_n into a *training sample*

$$D_m := [(X_i, Y_i); 1 \leq i \leq m]$$

and a *learning sample*

$$D_{(m)} := [(X_i, Y_i); m+1 \leq i \leq n].$$

The choice of the split size can be quite general (see Section 3 for details). In the numerical study (conducted in Section 4 below), we consider simply $m = 3n/4$ (the learning sample size is a quarter of the whole sample), which provides good results, but other splits can be considered as well.

Using the training sample, we compute a family $\{\bar{g}^{(\lambda)}; \lambda \in \Lambda\}$ of linear (or *weak*) estimators of the regression g. Each of these estimators depend on a parameter $\lambda = (v, s)$ which make them work based on the data "as if" the true underlying index were v and "as if" the smoothness of the link function were s (in the Hölder sense, see Section 3).

Then, using the learning sample, we compute a weight $w(\bar{g}) \in [0, 1]$ for each $\bar{g} \in \{\bar{g}^{(\lambda)}; \lambda \in \Lambda\}$, satisfying $\sum_{\lambda \in \Lambda} w(\bar{g}^{(\lambda)}) = 1$. These weights give a level of significance to each weak estimator. Finally, the adaptive, or *aggregated* estimator, is simply the convex combination of the weak estimators:

$$\hat{g} := \sum_{\lambda \in \Lambda} w(\bar{g}^{(\lambda)}) \bar{g}^{(\lambda)}.$$

The family of weak estimators consists of univariate local polynomial estimators (LPE), with a data-driven bandwidth that fits locally to the amount of data. In the next section the parameter $\lambda = (v, s)$ is fixed and known: we contruct a univariate LPE based on the sample $D_m(v) = [(Z_i, Y_i); 1 \leq i \leq m] = [(v^\top X_i, Y_i); 1 \leq i \leq m]$.

2.1. *Weak estimators*: *univariate LPE*

The LPE is standard in statistical literature, see for instance Fan and Gijbels (1996, 1995), among many others. We construct an estimator $\bar{f}$ of f based on i.i.d. copies $[(Z_i, Y_i); 1 \leq i \leq m]$ of a couple $(Z, Y) \in \mathbb{R} \times \mathbb{R}$ such that

$$Y = f(Z) + \sigma(Z)\epsilon, \tag{2.2}$$

where ϵ is standard Gaussian noise independent of Z, $\sigma : \mathbb{R} \to [\sigma_0, \sigma_1] \subset (0, +\infty)$ and $f \in H(s, L)$ where $H(s, L)$ is the set of s-Hölderian functions such that

$$|f^{(\lfloor s \rfloor)}(z_1) - f^{(\lfloor s \rfloor)}(z_2)| \leq L|z_1 - z_2|^{s - \lfloor s \rfloor}$$

for any $z_1, z_2 \in \mathbb{R}$, where $L > 0$ and $\lfloor s \rfloor$ stands for the largest integer smaller than s. This Hölder assumption is standard in nonparametric literature.

Let $r \in \mathbb{N}$ and $h > 0$ be fixed. If z is fixed, we consider the polynomial $\bar{P}_{(z,h)} \in \mathrm{Pol}_r$ (the set of real polynomials with degree at most r) which minimizes in P:

$$\sum_{i=1}^{m} \big(Y_i - P(Z_i - z)\big)^2 \mathbf{1}_{Z_i \in I(z,h)}, \tag{2.3}$$

where $I(z, h) := [z - h, z + h]$ and we define the LPE at z by

$$\bar{f}(z, h) := \bar{P}_{(z,h)}(0).$$

The polynomial $\bar{P}_{(z,h)}$ is well-defined and unique when the symmetrical matrix $\bar{\mathbf{Z}}_m(z, h)$, with entries

$$(\bar{\mathbf{Z}}_m(z, h))_{a,b} := \frac{1}{m\bar{P}_Z[I(z, h)]} \sum_{i=1}^{m} \Big(\frac{Z_i - z}{h}\Big)^{a+b} \mathbf{1}_{Z_i \in I(z,h)} \tag{2.4}$$

for $(a, b) \in \{0, \ldots, R\}^2$, is definite positive, where $\bar{P}_Z$ is the empirical distribution of $(Z_i)_{1 \leq i \leq m}$, given by

$$\bar{P}_Z[A] := \frac{1}{m} \sum_{i=1}^{m} \mathbf{1}_{Z_i \in A} \tag{2.5}$$

for any $A \subset \mathbb{R}$. When $\bar{\mathbf{Z}}_m(z,h)$ is degenerate, we simply take $\bar{f}(z,h) := 0$. The tuning parameter $h > 0$, which is called *bandwidth*, localizes the least square problem around the point z in (2.3). Of course, the choice of h is of first importance in this estimation method (as with any linear method). An important remark is then about the design law. Indeed, the law of $Z = v^\top X$ varies with v strongly: even if P_X is very simple (for instance uniform over some subset of $\mathbb{R}^d$ with positive Lebesgue measure), $P_{v^\top X}$ can be "far" from the uniform law, namely with a density that can vanish at the boundaries of its support, or inside the support, see the examples in Figure 1. This remark motivates the following choice for the bandwidth.

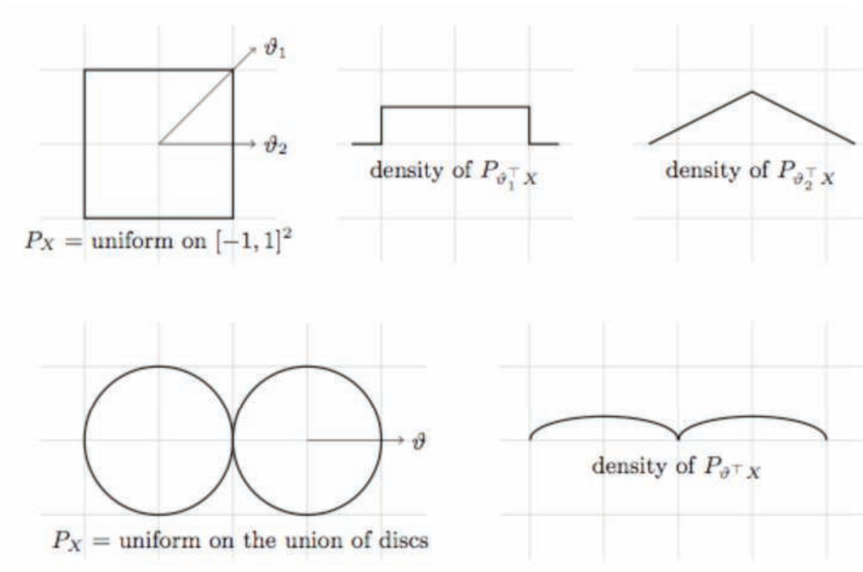

FIG 1. *Simple design examples*

If $f \in H(s,L)$ for known s and L, a "natural" bandwidth, which makes the balance between the bias and the variance of the LPE is given by

$$H_m(z) := \operatorname*{argmin}_{h \in (0,1)} \Big\{ L h^s \geq \frac{\sigma_1}{(m \bar{P}_Z[I(z,h)])^{1/2}} \Big\}. \tag{2.6}$$

This bandwidth choice stabilizes the LPE, since it fits point-by-point to the local amount of data. We consider then

$$\bar{f}(z) := \bar{f}(z, H_m(z)), \tag{2.7}$$

for any $z \in \mathbb{R}$, which is in view of Theorem 3 (see Section 3) a minimax estimator over $H(s,L)$ in model (2.2).

Remark 1. The reason why we consider local polynomials instead of some other method (like smoothing splines, for instance) is theoretical. It is linked with the fact that we need minimax weak estimators under the general design Assumption (D), so that the aggregated estimator is also minimax.

2.2. Adaptation by aggregation

If $\lambda := (v, s)$ is fixed, we consider the LPE $\bar{f}^{(\lambda)}$ given by (2.7), and we take

$$\bar{g}^{(\lambda)}(x) := \tau_Q(\bar{f}^{(\lambda)}(\vartheta^\top x)), \tag{2.8}$$

for any $x \in \mathbb{R}^d$ as an estimator of g, where $\tau_Q(f) := \max(-Q, \min(Q, f))$ is the truncation operator by $Q > 0$. The reason why we need to truncate the weak estimators is related to the theoretical results concerning the aggregation procedure described below, see Theorem 4 in Section 3. In order to adapt to the index ϑ and to the smoothness s of the link function, we aggregate the weak estimators from the family $\{\bar{g}^{(\lambda)}; \lambda \in \Lambda\}$ with the following algorithm: we take the convex combination

$$\hat{g} := \sum_{\lambda \in \Lambda} w(\bar{g}^{(\lambda)})\bar{g}^{(\lambda)} \tag{2.9}$$

where for a function $\bar{g} \in \{\bar{g}^{(\lambda)}; \lambda \in \Lambda\}$, the weight is given by

$$w(\bar{g}) := \frac{\exp\big(-TR_{(m)}(\bar{g})\big)}{\sum_{\lambda \in \Lambda} \exp\big(-TR_{(m)}(\bar{g}^{(\lambda)})\big)}, \tag{2.10}$$

with a *temperature* parameter $T > 0$ and

$$R_{(m)}(\bar{g}) := \sum_{i=m+1}^{n} (Y_i - \bar{g}(X_i))^2, \tag{2.11}$$

which is the empirical sum of squares of $\bar{g}$ over the training sample (up to a division by the sample size). This aggregation algorithm (with Gibbs weights) can be found in Leung and Barron (2006) in the regression framework, for projection-type weak estimators. Cumulative versions of this algorithm can be found in Catoni (2001), Juditsky et al. (2005a), Juditsky et al. (2005b), Yang (2000b) and Yang (2004).

We can understand the aggregation algorithm in the following way: first, we compute the least squares of each weak estimators. This is the most natural way of assessing the level of significance of some estimator among the other ones. Then, we put a Gibbs law over the set of weak estimators. The mass of each estimator relies on its least squares (over the learning sample). Finally, the aggregate is simply the mean expected estimator according to this law.

If T is small, the weights (2.10) are close to the uniform law over the set of weak estimators, and of course, the resulting aggregate is inaccurate. If T is large, only one weight will equal 1, and the others equal to 0: in this situation, the aggregate is equal to the estimator obtained by empirical risk minimization (ERM). This behavior can be also explained by equation (5.10) in the proof of

Theorem 4. Indeed, the exponential weights (2.10) realize an optimal tradeoff between the ERM procedure and the uniform weights procedure. Hence, T is somehow a regularization parameter of this tradeoff.

The ERM already gives good results, but if T is chosen carefully, we expect to obtain an estimator which outperforms the ERM. It has been proved theoretically in Lecué (2007) that an aggregation procedure outperforms the ERM in the regression framework. This fact is confirmed by the numerical study conducted in Section 4, where the choice of T is done using a simple leave-one-out cross-validation algorithm over the whole sample for aggregates obtained with several T. Namely, we consider the temperature

$$\hat{T} := \operatorname*{argmin}_{T \in \mathcal{T}} \sum_{j=1}^{n} \sum_{i \neq j} \big(Y_i - \hat{g}_{-i}^{(T)}(X_i) \big)^2, \tag{2.12}$$

where $\hat{g}_{-i}^{(T)}$ is the aggregated estimator (2.9) with temperature T, based on the sample $D_n^{-i} = [(X_j, Y_j); j \neq i]$, and where $\mathcal{T}$ is some set of temperatures (in Section 4, we take $\mathcal{T} = \{0.1, 0.2, \ldots, 4.9, 5\}$).

The set of parameters Λ is given by $\Lambda := \bar{S} \times G$, where G is the grid with step $(\log n)^{-1}$ given by

$$G := \big\{ s_{\min}, s_{\min} + (\log n)^{-1}, s_{\min} + 2(\log n)^{-1}, \ldots, s_{\max} \big\}. \tag{2.13}$$

The tuning parameters $s_{\min}$ and $s_{\max}$ correspond to the minimum and maximum "allowed" smoothness for the link function: for this grid choice, the aggregated estimator converges with the optimal rate for a link function in $H(s, L)$ for any $s \in [s_{\min}, s_{\max}]$ in view of Theorem 1.

The set $\bar{S} = \bar{S}_{\Delta}^{d-1}$ is the regular lattice of the half unit-sphere S_{+}^{d-1} with discretization step Δ which is constructed as follows. Let us introduce $\Phi(\delta) := \cup_{\ell \geq 0} \{\ell \delta\} \cap [0, \pi]$ and consider the function $p : [0, \pi]^{d-1} \to S^{d-1}$ defined by $p(\phi_1, \ldots, \phi_{d-1}) = (x_1, \ldots, x_d)$, where

$$\begin{cases} x_1 = \cos(\phi_1)\cos(\phi_2) \times \cdots \times \cos(\phi_{d-1}) \\ x_2 = \sin(\phi_1)\cos(\phi_2) \times \cdots \times \cos(\phi_{d-1}) \\ \qquad \vdots \\ x_\ell = \sin(\phi_{\ell-1})\cos(\phi_\ell) \times \cdots \times \cos(\phi_{d-1}) \\ \qquad \vdots \\ x_{d-1} = \sin(\phi_{d-2})\cos(\phi_{d-1}) \\ x_d = \sin(\phi_{d-1}). \end{cases}$$

Then, the regular lattice $\bar{S}_{\Delta}^{d-1}$ is constructed using Algorithm 1. In Figure 2 we show $\bar{S}_{\Delta}^{d-1}$ for $\Delta = 0.1$ and $d = 2, 3$. The step is taken as

$$\Delta = (n \log n)^{-1/(2 s_{\min})}, \tag{2.14}$$

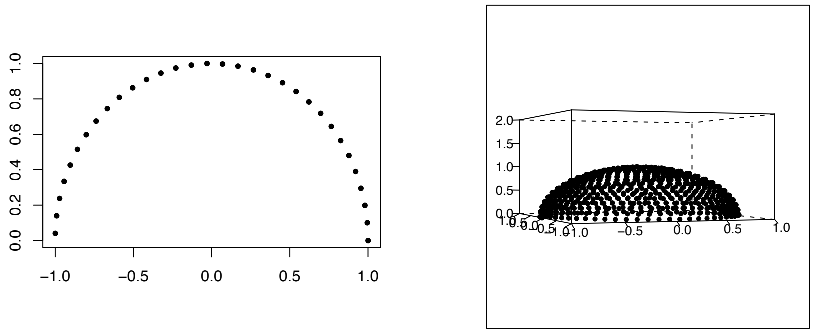

FIG 2. *Lattices* $\bar{S}_{\Delta}^{d-1}$ *for* $\Delta = 0.1$ *and* $d = 2, 3$

which relies on the minimal allowed smoothness of the link function. For instance, if we want the estimator to be adaptive for link functions at least Lipschitz, we take $\Delta = (n \log n)^{-1/2}$.

Input: d (dimension parameter) and Δ (discretization step)
Output: $\bar{S}_{\Delta}^{d-1}$ (regular discretization of S^{d-1})
$\bar{S}_{\Delta}^{d-1} = \emptyset$
$\Phi_{d-1} = \Phi(\arccos(1 - \Delta^2/2))$
foreach $\phi_{d-1} \in \Phi_{d-1}$ **do**
  $\Phi_{d-2} = \Phi(\Delta / \arccos(\phi_{d-1}))$
  foreach $\phi_{d-2} \in \Phi_{d-2}$ **do**
    $\Phi_{d-3} = \Phi(\Delta / \arccos(\phi_{d-2}))$
    $\vdots$
    foreach $\phi_2 \in \Phi_2$ **do**
      $\Phi_1 = \Phi(\Delta / \arccos(\phi_2))$
      foreach $\phi_1 \in \Phi_1$ **do**
        add the point of coordinates $h(\phi_1, \ldots, \phi_{d-1})$ in $\bar{S}_{\Delta}^{d-1}$
      end
    end
  end
end

Algorithm 1: Construction of the regular lattice $\bar{S}_{\Delta}^{d-1}$.

2.3. Reduction of the complexity of the algorithm

The adaptive procedure described previously requires the computation of the LPE for each parameter $\lambda \in \tilde{\Lambda} := \Lambda \times \mathcal{L}$ (actually, we do also a grid $\mathcal{L}$ over the radius parameter L in the simulations). Hence, there are $|\bar{S}_{\Delta}^{d-1}| \times |G| \times |\mathcal{L}|$ LPE to compute. Namely, this is $(\pi/\Delta)^{d-1} \times |G| \times |\mathcal{L}|$, which equals, if $|G| =$

$|\mathcal{L}| = 4$ and $\Delta = (n \log n)^{-1/2}$ (see Section 4) to 1079 when $d = 2$ and to 72722 when $d = 3$, which is much too large. Hence, the complexity of this procedure must be reduced: we propose a recursive algorithm which improves strongly the complexity of the estimator. Actually, the coefficients $w(\bar{g}^{(\lambda)})$ are very close to zero (see Figures 7 and 8 in Section 4) when $\lambda = (v, s)$ is such that v is "far" from the true index ϑ. Hence, these coefficients should not be computed at all, since the corresponding weak estimators do not contribute to the aggregated estimator (2.9). Thus, instead of using a lattice of the whole half unit-sphere for detecting the index, we only build a part of it, which corresponds to the coefficients which are the most significative. This is done with an iterative algorithm, see Alogorithm 2, which makes a preselection of weak estimators to aggregate ($B^d(v, \delta)$ stands for the ball in $(\mathbb{R}^d, \|\cdot\|_2)$ centered at v with radius δ and $R_{(m)}(\bar{g})$ is given by (2.11)).

Input: (X_i, Y_i) (Data), G (smoothness grid)
Output: $\hat{S}$ (a section of $\bar{S}_\Delta^{d-1}$)
Put $\Delta = (n \log n)^{-1/2}$ and $\Delta_0 = (2dn)^{-1/(2(d-1))}$
Compute the lattice $\hat{S} = \bar{S}_{\Delta_0}^{d-1}$ and put $\hat{\Lambda} := \hat{S} \times G$
while $\Delta_0 > \Delta$ **do**
    find the point $\hat{v}$ such that $(\hat{v}, \hat{s}) = \hat{\lambda} = \mathrm{argmin}_{\lambda \in \hat{\Lambda}} R_{(m)}(\bar{g}^{(\lambda)})$
    put $\Delta_0 = \Delta_0/2$
    put $\hat{S} = \bar{S}_{\Delta_0}^{d-1} \cap B^d(\hat{v}, 2\Delta_0)$ and $\hat{\Lambda} := \hat{S} \times G$;
end

Algorithm 2: Preselection of the coefficients

When the algorithm exits, $\hat{S}$ is a section of the lattice $\bar{S}_\Delta^{d-1}$ centered at $\hat{v}$ with radius $2^{d-1}\Delta$, which contains (with a high probability) the points $v \in \bar{S}_\Delta^{d-1}$ corresponding to the largest coefficients $w(\bar{g}^{(\lambda)})$ where $\lambda = (v, s, L) \in \bar{S}_\Delta^{d-1} \times G \times \mathcal{L}$. The aggegate is then computed for a set of parameters $\hat{\Lambda} = \hat{S} \times G \times \mathcal{L}$ using (2.9) with weights (2.10). The parameter Δ_0 is chosen so that the surface of $B^d(v, \Delta_0)$ is $C_d(2dn)^{-1/2}$: n is not a power of d. Moreover, the number of iterations is $O(\log n)$, thus the complexity is much smaller than the full aggregation algorithm. This procedure gives nice empirical results, see Section 4. We show the iterative construction of $\hat{S}$ in Figure 3.

3. Main results

The error of estimation is measured with the $L^2(P_X)$-norm, defined by

$$\|g\|_{L^2(P_X)} := \Big(\int_{\mathbb{R}^d} g(x)^2 P_X(dx) \Big)^{1/2},$$

where we recall that P_X is the design law. We consider the set $H^Q(s, L) := H(s, L) \cap \{f : \mathbb{R} \to \mathbb{R} \mid \|f\|_\infty := \sup_x |f(x)| \le Q\}$. Since we want the adaptive procedure to work whatever $\vartheta \in S_+^{d-1}$ is, we need to work with as general assumptions on the law of $\vartheta^\top X$ as possible. The following assumption generalizes

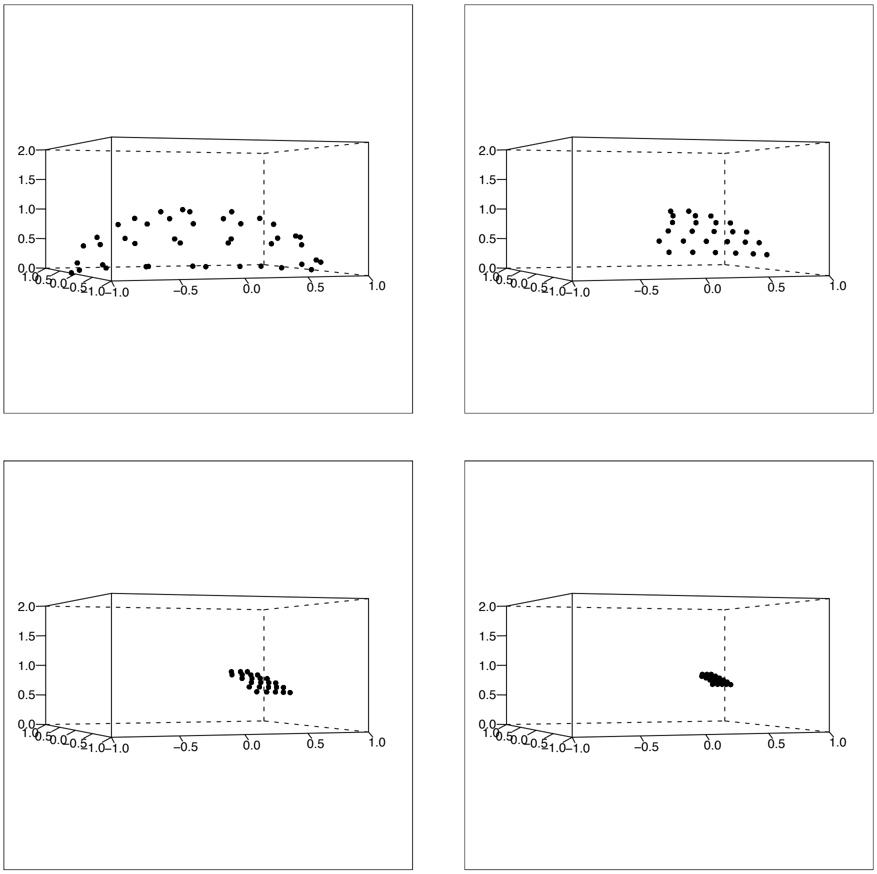

Fig 3. *Iterative construction of the set $\hat{S}$ of preselectioned weak estimators indexes. Weak estimators are aggregated only for $v \in \hat{S}$ (bottom right), which is concentrated around the true index.*

the usual assumptions on random designs (when P_X has a density with respect to the Lebesgue measure) that can be met in literature. Namely, we do not assume that the density of $P_{v^\top X}$ is bounded away from zero. Indeed, even with a very simple P_X, this assumption holds for specific v only (see Figure 1). We say that a real random variable Z satisfies Assumption (D) if:

Assumption (D). *There is a density μ of P_Z with respect to the Lebesgue measure which is continuous. Moreover, we assume that*

- *μ is compactly supported*;
- *There is a finite number of z in the support of μ such that $\mu(z) = 0$*;
- *For any such z, there is an interval $I_z = [z - a_z, z + b_z]$ such that μ is decreasing over $[z - a_z, z]$ and increasing over $[z, z + b_z]$*;

- *There is $\beta \geq 0$ and $\gamma > 0$ such that*

$$P_Z[I] \geq \gamma |I|^{\beta+1}$$

for any I, where $|I|$ stands for the length of I.

This assumption includes any design with continuous density with respect to the Lebesgue measure that can vanish at several points, but not faster than some power function.

3.1. Upper and lower bounds

The next Theorem provides an upper bound for the adaptive estimator constructed in Section 2. This upper bound holds for quite general tuning parameters. The temperature $T > 0$ can be arbitrary (but not in practice of course). The training sample size is given by

$$m = [n(1 - \ell_n)], \tag{3.1}$$

where $[x]$ is the integral part of x, and where ℓ_n is a positive sequence such that for all n, $(\log n)^{-\alpha} \leq \ell_n < 1$ with $\alpha > 0$. Note that in methods involving data splitting, the optimal choice of the split size is open. The degree r of the LPE and the grid choice G must be such that $s_{\max} \leq r + 1$.

The upper bound below shows that the estimator converges with the optimal rate for a link function in a whole family of Hölder classes, and for any index. In what follows, E^n stands for the expectation with respect to the joint law P^n of the whole sample D_n.

Theorem 1. *Let $\hat{g}$ be the aggregated estimator given by* (2.9) *with the weights* (2.10). *If for all $v \in S_+^{d-1}$, $v^\top X$ satisfies Assumption (D), we have*

$$\sup_{\vartheta \in S_+^{d-1}} \sup_{f \in H^Q(s,L)} E^n \|\hat{g} - g\|_{L^2(P_X)}^2 \leq C n^{-2s/(2s+1)},$$

for any $s \in [s_{\min}, s_{\max}]$ when n is large enough, where we recall that $g(\cdot) = f(\vartheta^\top \cdot)$. The constant $C > 0$ depends on $\sigma_1, L, s_{\min}, s_{\max}$ and P_X only.

Note that $\hat{g}$ does not depend within its construction on the index ϑ, nor the smoothness s of the link function f, nor the design law P_X. The assumption that $v^\top X$ satisfies Assumption (D) for any $v \in S_+^{d-1}$ holds, for instance, for the multivariate designs from Figure 1. More generally, this property holds for any uniform law over a support that does not have very "spiky" boundary. Note that this assumption is more general than the one considered in Audibert and Tsybakov (2007).

In Theorem 2 below, we prove in our setting (when Assumption (D) holds on the design) that $n^{-2s/(2s+1)}$ is a lower bound for a link function in $H(s, L)$ in the single-index model.

Theorem 2. *Let $s, L, Q > 0$ and $\vartheta \in S_+^{d-1}$ be such that $\vartheta^\top X$ satisfies Assumption (D). We have*

$$\inf_{\tilde g} \sup_{f \in H^Q(s,L)} E^n \|\tilde g - g\|_{L^2(P_X)}^2 \geq C' n^{-2s/(2s+1)},$$

where the infimum is taken among all estimators based on data from (1.1),(1.2), *and where $C' > 0$ is a constant depending on σ_1, s, L and $P_{\vartheta^\top X}$ only.*

Theorem 1 and Theorem 2 together entail that $n^{-2s/(2s+1)}$ is the minimax rate for the estimation of g in model (1.1) under the constraint (1.2) when the link function belongs to an s-Hölder class. It answers in particular to Question 2 from Stone (1982).

3.2. A new result for the LPE

In this section, we give upper bounds for the LPE in the univariate regression model (2.2). Despite the fact that the literature about LPE is wide, the Theorem below is new. It provides a minimax optimal upper bound for the $L^2(P_Z)$-integrated risk of the LPE over Hölder balls under Assumption (D), which is a general assumption for random designs having a density with respect to the Lebesgue measure.

In this section, the smoothness s is supposed known and fixed, and we assume that the degree r of the local polynomials satisfies $r + 1 \geq s$. First, we give an upper bound for the pointwise risk conditionally on the design. Then, we derive from it an upper bound for the $L^2(P_Z)$-integrated risk, using standard tools from empirical process theory (see Appendix). Here, E^m stands for the expectation with respect to the joint law P^m of the observations $[(Z_i, Y_i); 1 \leq i \leq m]$ from model (2.2). Let us define the matrix

$$\bar{\mathbf{Z}}_m(z) := \bar{\mathbf{Z}}_m(z, H_m(z))$$

where $\bar{\mathbf{Z}}_m(z, h)$ is given by (2.4) and $H_m(z)$ is given by (2.6). Let us denote by $\lambda(M)$ the smallest eigenvalue of a matrix M and introduce $Z_1^m := (Z_1, \ldots, Z_m)$.

Theorem 3. *For any $z \in \operatorname{Supp} P_Z$, let $\bar f(z)$ be given by* (2.7). *We have on the event $\{\lambda(\bar{\mathbf{Z}}_m(z)) > 0\}$:*

$$\sup_{f \in H(s,L)} E^m\big[(\bar f(z) - f(z))^2 | Z_1^m\big] \leq 2\lambda(\bar{\mathbf{Z}}_m(z))^{-2} L^2 H_m(z)^{2s}. \tag{3.2}$$

Moreover, if Z satisfies Assumption (D), we have

$$\sup_{f \in H^Q(s,L)} E^m\big[\|\tau_Q(\bar f) - f\|_{L^2(P_Z)}^2\big] \leq C_2 m^{-2s/(2s+1)} \tag{3.3}$$

for m large enough, where we recall that τ_Q is the truncation operator by $Q > 0$ and where $C_2 > 0$ is a constant depending on s, Q, and P_Z only.

Remark 2. While inequality (3.2) in Theorem 3 is stated over $\{\lambda(\bar{\mathbf{Z}}_m(z)) > 0\}$, which entails the existence and the unicity of a solution to the linear system (2.3) (this inequality is stated conditionally on the design), we only need Assumption (D) for inequality (3.3) to hold.

3.3. Oracle inequality

In this section, we provide an oracle inequality for the aggregation algorithm (2.9) with weights (2.10). This result, which is of independent interest, is stated for a general finite set $\{\bar{g}^{(\lambda)}; \lambda \in \Lambda\}$ of deterministic functions such that $\|\bar{g}^{(\lambda)}\|_\infty \leq Q$ for all $\lambda \in \Lambda$. These functions are for instance weak estimators computed with the training sample (or *frozen* sample), which is independent of the learning sample. Let $D := [(X_i, Y_i); 1 \leq i \leq |D|]$ (where $|D|$ stands for the cardinality of D) be an i.i.d. sample of (X, Y) from the multivariate regression model (1.1), where no particular structure like (1.2) is assumed.

The aim of aggregation schemes is to mimic (up to an additive residual) the oracle in $\{\bar{g}^{(\lambda)}; \lambda \in \Lambda\}$. This aggregation framework has been considered, among others, by Birgé (2005), Catoni (2001), Juditsky and Nemirovski (2000), Leung and Barron (2006), Nemirovski (2000), Tsybakov (2003b) and Yang (2000b).

Theorem 4. *The aggregation procedure $\hat{g}$ based on the learning sample D defined by* (2.9) *and* (2.10) *satisfies*

$$E^D \|\hat{g} - g\|^2_{L^2(P_X)} \leq (1+a) \min_{\lambda \in \Lambda} \|\bar{g}^{(\lambda)} - g\|^2_{L^2(P_X)} + \frac{C \log |\Lambda| (\log |D|)^{1/2}}{|D|}$$

for any $a > 0$, where $|\Lambda|$ denotes the cardinality of Λ, where E^D stands for the expectation with respect to the joint law of D, and where $C := 3[8Q^2(1+a)^2/a + 4(6Q^2 + 2\sigma_1 2\sqrt{2})(1+a)/3] + 2 + 1/T$.

This theorem is a model-selection type oracle inequality for the aggregation procedure given by (2.9) and (2.10). Sharper oracle inequalities for more general models can be found in Juditsky et al. (2005a), where the algorithm used therein requires an extra cumulative sum.

Remark 3. Inspection of the proof of Theorem 4 shows that the ERM (which is the estimator minimizing the empirical risk $R_{(m)}(g) := \sum_{i=m+1}^n (Y_i - g(X_i))^2$ over all g in $\{\bar{g}^{(\lambda)}; \lambda \in \Lambda\}$) satisfies the same oracle inequality. Nevertheless, it has been proved in Lecué (2007) that the ERM is theoretically suboptimal in this framework, when we want to mimic the oracle without the extra factor $1 + a$ in front of the biais term $\min_{\lambda \in \Lambda} \|\bar{g}^{(\lambda)} - g\|^2_{L^2(P_X)}$. The simulation study of Section 4 (especially Figures 4, 5, 6) confirms this suboptimality.

4. Numerical illustrations

We implemented the procedure described in Section 2 using the `R` software (see `http://www.r-project.org/`). In order to increase computation speed, we implemented the computation of local polynomials and the bandwidth selection (2.6) in `C` language. The simulated samples satisfy (1.1),(1.2), where the noise is centered Gaussian with homoscedastic variance

$$\sigma = \Big[\sum_{1 \leq i \leq n} f(\vartheta^\top X_i)^2 / (n \times \texttt{rsnr}) \Big]^{1/2},$$

where $\texttt{rsnr} = 5$. This choice of σ makes the root-signal-to-noise ratio, which is a commonly used assessment of the complexity of estimation, equals to 5. We consider the following link functions (see the dashed lines in Figures 9 and 10):

- $\texttt{oscsine}(x) = 4(x+1)\sin(4\pi x^2)$,
- $\texttt{hardsine}(x) = 2\sin(1+x)\sin(2\pi x^2+1)$.

The simulations are done with a uniform design on $[-1,1]^d$, with dimensions $d \in \{2,3,4\}$ and we consider several indexes ϑ that make $P_{\vartheta^\top X}$ not uniform.

In all the computations below, the parameters for the procedure are $\Lambda = \hat{S} \times G \times \mathcal{L}$ where $\hat{S}$ is computed using the algorithm described in Section 2.3 and where $G = \{1,2,3,4\}$ and $\mathcal{L} = \{0.1, 0.5, 1, 1.5\}$. The degree of the local polynomials is $r = 5$. The learning sample has size $[n/4]$, and is chosen randomly in the whole sample. We do not use a jackknife procedure (that is, the average of estimators obtained with several learning subsamples), since the results are stable enough (at least when $n \geq 100$) when we consider only one learning sample.

In Tables 1, 2, 3 and Figures 4, 5, 6, we show the mean MISE for 100 replications and its standard deviation for several Gibbs temperatures, sev-

Table 1
MISE against the Gibbs temperature ($f = \texttt{hardsine}$, $d = 2$, $\vartheta = (1/\sqrt{2}, 1/\sqrt{2})$.)

Temperature	0.1	0.5	0.7	1.0	1.5	2.0	ERM	`aggCVT`
n = 100	0.026	0.017	0.015	**0.014**	**0.014**	0.015	0.034	0.015
	(.009)	(.006)	(.006)	(.005)	(.005)	(.006)	(.018)	(.005)
n = 200	0.015	0.009	**0.008**	**0.008**	0.009	0.011	0.027	0.009
	(.004)	(.002)	(.003)	(.003)	(.005)	(.007)	(.014)	(.004)
n = 400	0.006	0.005	**0.004**	0.005	0.006	0.007	0.016	0.005
	(.001)	(.001)	(.001)	(.001)	(.002)	(.002)	(.003)	(.002)

Table 2
MISE against the Gibbs temperature ($f = \texttt{hardsine}$, $d = 3$, $\vartheta = (2/\sqrt{14}, 1/\sqrt{14}, 3/\sqrt{14})$).

Temperature	0.1	0.5	0.7	1.0	1.5	2.0	ERM	`aggCVT`
n = 100	0.029	0.021	0.019	0.018	**0.017**	0.018	0.037	0.020
	(.011)	(.008)	(.008)	(.007)	(.008)	(.009)	(.022)	(.008)
n = 200	0.016	0.010	0.010	**0.009**	**0.009**	0.010	0.026	0.010
	(.005)	(.003)	(.003)	(.002)	(.002)	(.003)	(0.008)	(.003)
n = 400	0.007	0.006	**0.005**	**0.005**	0.006	0.007	0.017	0.006
	(.002)	(.001)	(.001)	(.001)	(.001)	(.002)	(.003)	(.001)

Table 3
MISE against the Gibbs temperature ($f = \texttt{hardsine}$, $d = 4$, $\vartheta = (1/\sqrt{21}, -2/\sqrt{21}, 0, 4/\sqrt{21})$)

Temperature	0.1	0.5	0.7	1.0	1.5	2.0	ERM	`aggCVT`
n = 100	0.038	0.027	0.021	0.019	**0.017**	**0.017**	0.038	0.020
	(.016)	(.010)	(.009)	(.008)	(.007)	(.007)	(.025)	(.010)
n = 200	0.019	0.013	**0.012**	**0.012**	0.013	0.014	0.031	0.013
	(.014)	(.009)	(.010)	(.011)	(.012)	(.012)	(.016)	(.010)
n = 400	0.009	0.006	**0.005**	**0.005**	0.006	0.007	0.017	0.006
	(.002)	(.001)	(.001)	(.001)	(.001)	(.002)	(.004)	(.001)

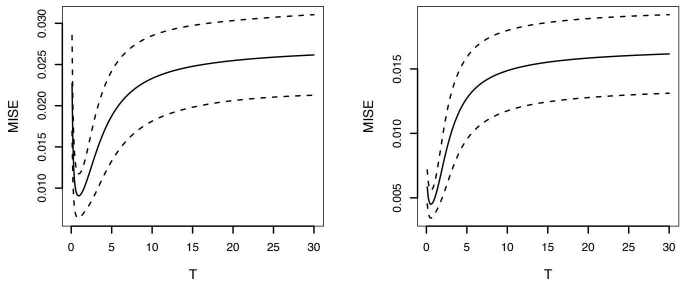

FIG 4. *MISE against the Gibbs temperature for* $f =$ `hardsine`, $\vartheta = (1/\sqrt{2}, 1/\sqrt{2})$, $n = 200, 400$ (*solid line = mean of the MISE for 100 replications, dashed line = mean MISE* $\pm$ *standard deviation.*)

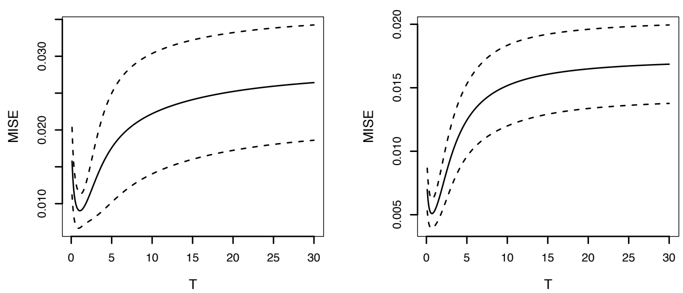

FIG 5. *MISE against the Gibbs temperature for* $f =$ `hardsine`, $\vartheta = (2/\sqrt{14}, 1/\sqrt{14}, 3/\sqrt{14})$, $n = 200, 400$ (*solid line = mean of the MISE for 100 replications, dashed line = mean MISE* $\pm$ *standard deviation.*)

eral sample sizes and indexes. These results empirically prove that the aggregated estimator outperforms the ERM (which is computed as the aggregated estimator with a large temperature $T = 30$) since in each case, the aggregated estimator with cross-validated temperature (`aggCVT`, given by (2.12), with $\mathcal{T} = \{0.1, 0.2, \ldots, 4.9, 5\}$), has a MISE much smaller than the MISE of the ERM. Moreover, `aggCVT` is more stable than the ERM in view of the standard deviations (in brackets). Note also that as expected, the dimension parameter has no impact on the accuracy of estimation: the MISEs are barely the same when $d = 2, 3, 4$.

The aim of Figures 7 and 8 is to give an illustration of the aggregation phenomenon. In these figures, we show the points

$$\big\{(1 + w(\bar{g}^{(\lambda)}))\vartheta \text{ for } \lambda = (\vartheta, s, L) \in \Lambda = \bar{S}_{\Delta}^{d-1} \times \{3\} \times \{1\}\big\} \tag{4.1}$$

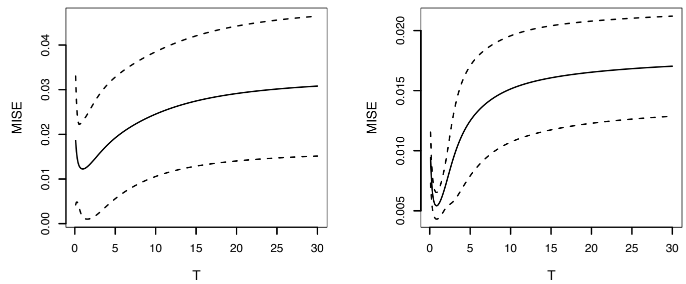

Fig 6. *MISE against the Gibbs temperature for* $f =$ `hardsine`*,* $\vartheta = (2/\sqrt{14}, 1/\sqrt{14}, 3/\sqrt{14})$*,* $n = 200, 400$.

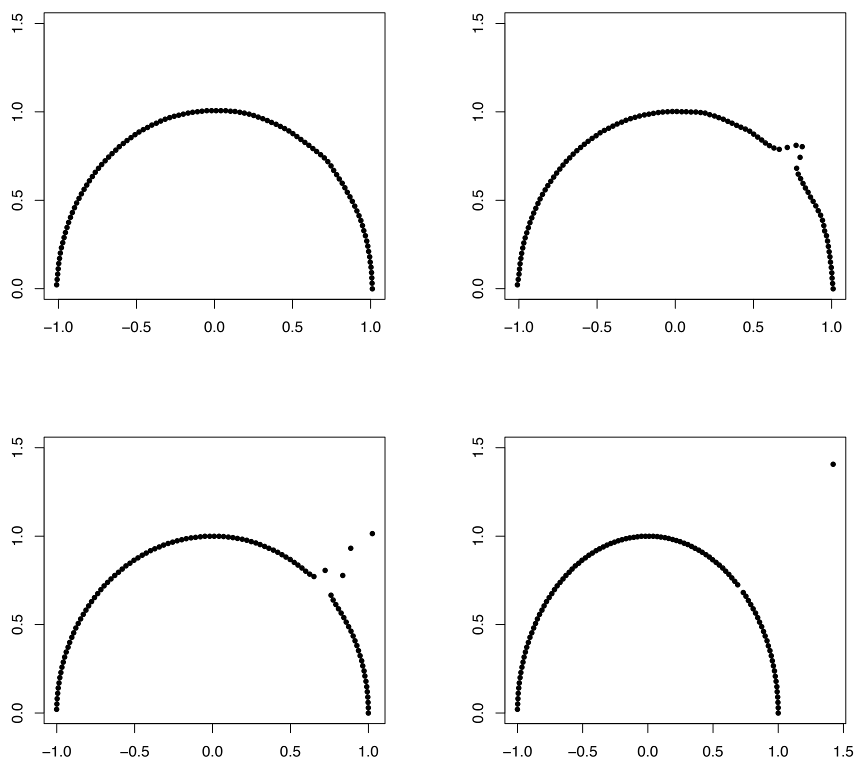

Fig 7. *Weights associated to each point (see* (4.1)*) of the lattice* $\bar{S}^1_\Delta$ *for* $\Delta = 0.03$*,* $\vartheta = (1/\sqrt{2}, 1/\sqrt{2})$ *and* $T = 0.05, 0.2, 0.5, 10$ *(from top to bottom and left to right.)*

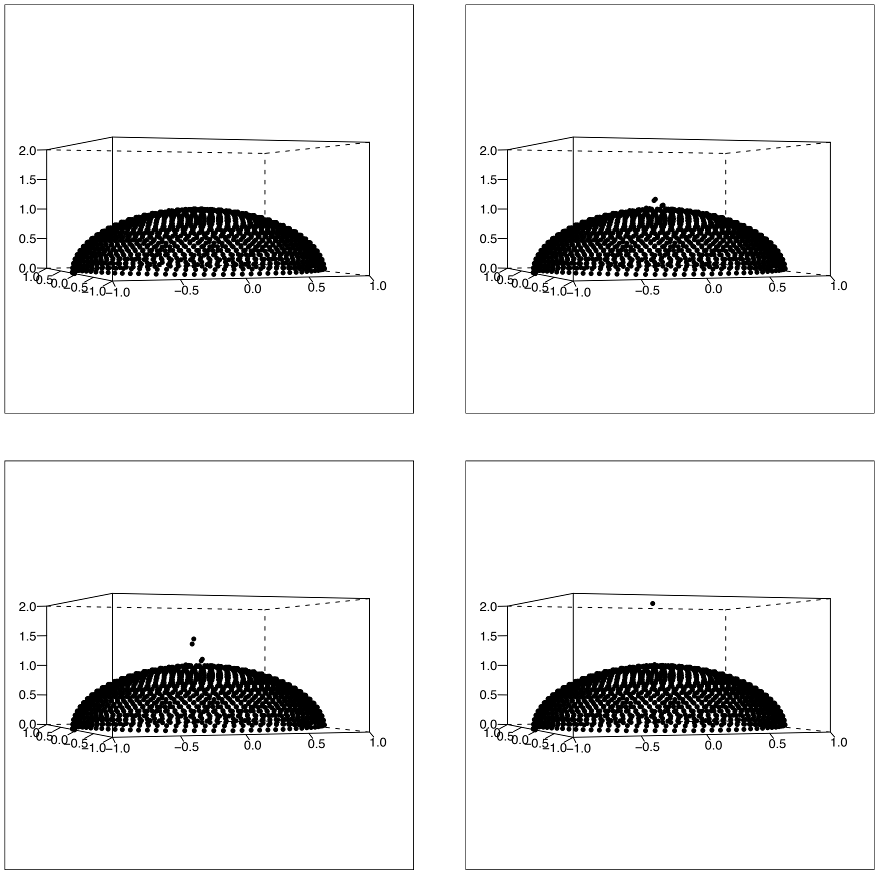

FIG 8. *Weights associated to each points (see* (4.1)*) of the lattice* $\bar{S}^2_\Delta$ *for* $\Delta = 0.07$*,* $\vartheta = (0, 0, 1)$*, and* $T = 0.05, 0.3, 0.5, 10$ *(from top to bottom and left to right).*

obtained for a single run (that is, we take $s = 3$ and $L = 1$ in the bandwidth choice (2.6) and we do not use the reduction of complexity algorithm). These figures motivates the use of the complexity reduction algorithm, since only the weights corresponding to a point of $\bar{S}^{d-1}_\Delta$ which is close to the true index are significant (at least numerically). Moreover, these weights provide information about the true index: the direction $v \in \bar{S}^{d-1}_\Delta$ corresponding to the largest coefficient $w(\bar{g}^{(\lambda)})$ for $\lambda = (v, s, L)$ is an accurate estimator of the index, see Figures 7 and 8. Finally, we show typical realisations for several index functions, indexes and sample sizes in Figures 9, 10, 11, 12.

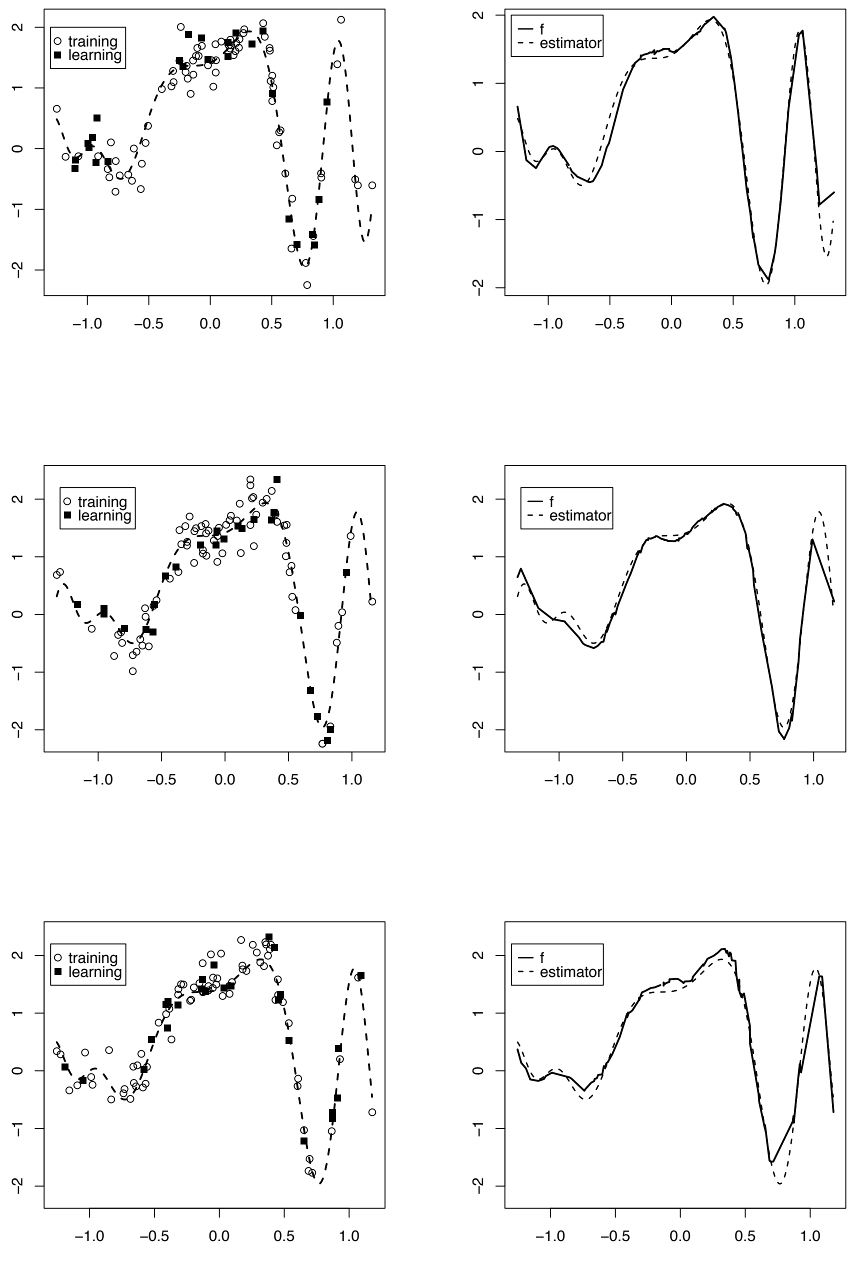

Fig 9. *Simulated datasets and aggregated estimators with cross-validated temperature for* $f=$ hardsine*,* $n=100$*, and indexes* $\vartheta=(1/\sqrt{2},1/\sqrt{2})$*,* $\vartheta=(2/\sqrt{14},1/\sqrt{14},3/\sqrt{14})$*,* $\vartheta=(1/\sqrt{21},-2/\sqrt{21},0,4/\sqrt{21})$ *from top to bottom.*

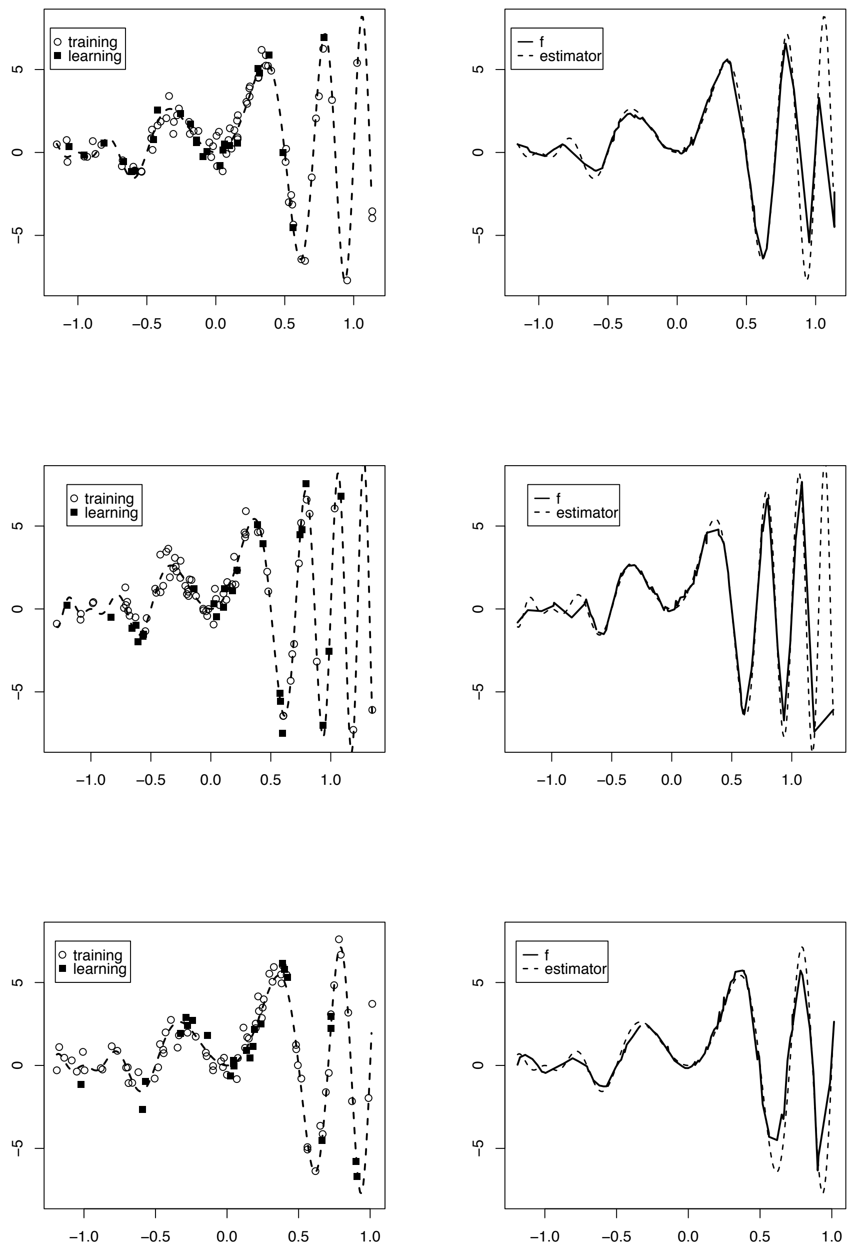

Fig 10. *Simulated datasets and aggregated estimators with cross-validated temperature for* f = `oscsine`, $n = 100$, *and indexes* $\vartheta = (1/\sqrt{2}, 1/\sqrt{2})$, $\vartheta = (2/\sqrt{14}, 1/\sqrt{14}, 3/\sqrt{14})$, $\vartheta = (1/\sqrt{21}, -2/\sqrt{21}, 0, 4/\sqrt{21})$ *from top to bottom.*

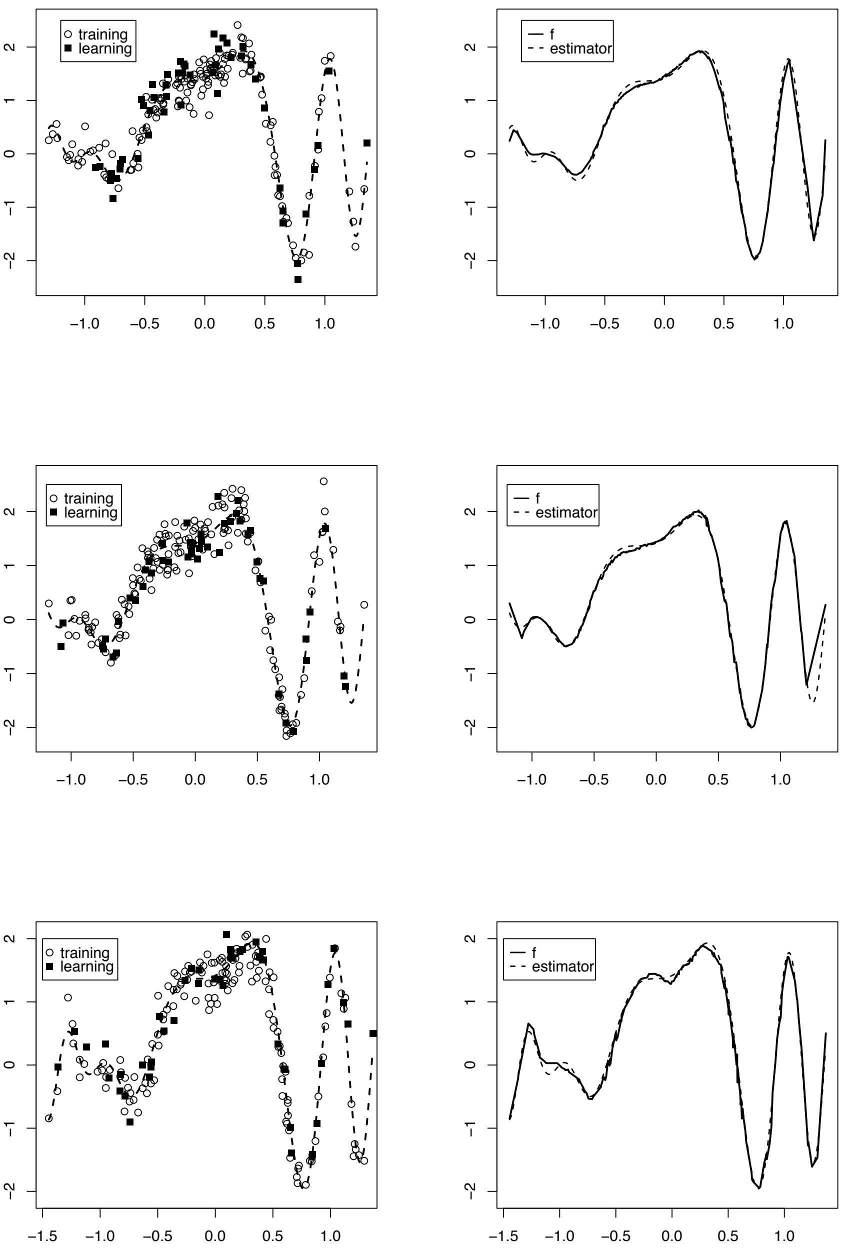

FIG 11. *Simulated datasets and aggregated estimators with cross-validated temperature for* $f = \texttt{hardsine}$, $n = 200$, *and indexes* $\vartheta = (1/\sqrt{2}, 1/\sqrt{2})$, $\vartheta = (2/\sqrt{14}, 1/\sqrt{14}, 3/\sqrt{14})$, $\vartheta = (1/\sqrt{21}, -2/\sqrt{21}, 0, 4/\sqrt{21})$ *from top to bottom.*

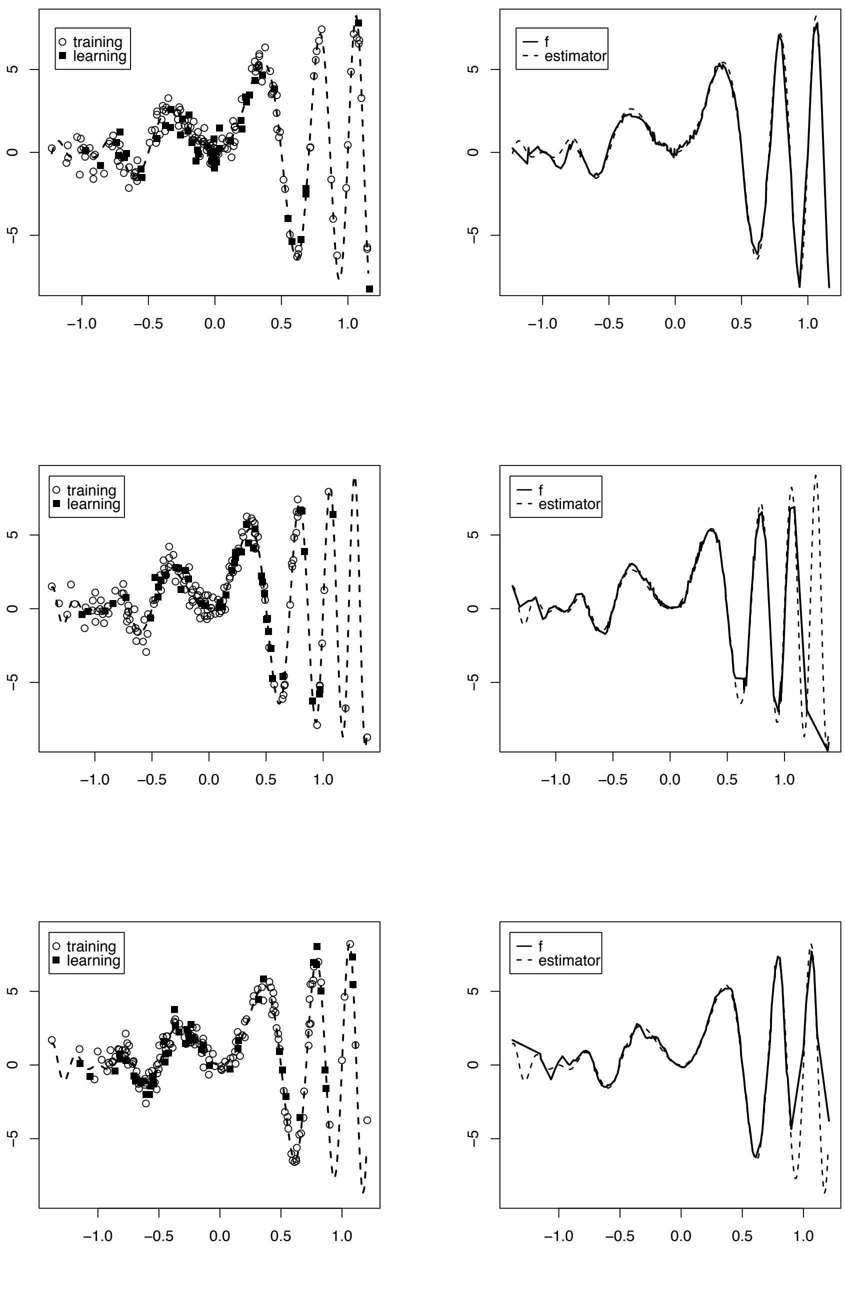

FIG 12. *Simulated datasets and aggregated estimators with cross-validated temperature for* $f = \texttt{oscsine}$, $n = 200$, *and indexes* $\vartheta = (1/\sqrt{2}, 1/\sqrt{2})$, $\vartheta = (2/\sqrt{14}, 1/\sqrt{14}, 3/\sqrt{14})$, $\vartheta = (1/\sqrt{21}, -2/\sqrt{21}, 0, 4/\sqrt{21})$ *from top to bottom.*

5. Proofs

Proof of Theorem 1

The functions $\bar{g}^{(\lambda)}$ are given by (2.8). They are computed based on the training (or "frozen") sample D_m, which is independent of the learning sample $D_{(m)}$. If $E^{(m)}$ denotes the integration with respect to the joint law of $D_{(m)}$, we obtain using Theorem 4:

$$E^{(m)}\|\hat{g} - g\|_{L^2(P_X)}^2 \leq (1+a)\min_{\lambda\in\Lambda}\|\bar{g}^{(\lambda)} - g\|_{L^2(P_X)}^2 + \frac{C\log|\Lambda|(\log|D_{(m)}|)^{1/2}}{|D_{(m)}|}$$
$$\leq (1+a)\|\bar{g}^{(\bar{\lambda})} - g\|_{L^2(P_X)}^2 + o(n^{-2s/(2s+1)}),$$

since $\log|\Lambda|(\log|D_{(m)}|)^{1/2}/|D_{(m)}| \leq d(\log n)^{3/2+\gamma}/(2s_{\min}n)$ (see (3.1) and (2.14)), and where $\bar{\lambda} = (\bar{\vartheta}, \bar{s}) \in \Lambda$ is such that $\|\bar{\vartheta} - \vartheta\|_2 \leq \Delta$ and $\lfloor \bar{s} \rfloor = \lfloor s \rfloor$ with $s \in [\bar{s}, \bar{s} + (\log n)^{-1}]$. By integration with respect to P^m, we obtain

$$E^n\|\hat{g} - g\|_{L^2(P_X)}^2 \leq (1+a)E^m\|\bar{g}^{(\bar{\lambda})} - g\|_{L^2(P_X)}^2 + o(n^{-2s/(2s+1)}). \quad (5.1)$$

The choice of $\bar{\lambda}$ entails $H^Q(s,L) \subset H^Q(\bar{s},L)$ and

$$n^{-2\bar{s}/(2\bar{s}+1)} \leq e^{1/2}n^{-2s/(2s+1)}.$$

Thus, together with (3.1) and (5.1), the Theorem follows if we prove that

$$\sup_{f\in H^Q(\bar{s},L)} E^m\|\bar{g}^{(\bar{\lambda})} - g\|_{L^2(P_X)}^2 \leq Cm^{-2\bar{s}/(2\bar{s}+1)}. \quad (5.2)$$

for n large enough, where $C > 0$. We cannot use directly Theorem 3 to prove this, since the weak estimator $\bar{g}^{(\bar{\lambda})}$ works based on data $D_m(\bar{\vartheta})$ (see (2.1)) while the true index is ϑ. In order to clarify the proof, we write $\bar{g}^{(\bar{\vartheta})}$ instead of $\bar{g}^{(\bar{\lambda})}$ since in (5.2), the estimator uses the "correct" smoothness parameter $\bar{s}$. We have

$$\|\bar{g}^{(\bar{\vartheta})} - g\|_{L^2(P_X)}^2 \leq 2\big(\|\bar{g}^{(\bar{\vartheta})}(\cdot) - f(\bar{\vartheta}^\top\cdot)\|_{L^2(P_X)}^2 + \|f(\bar{\vartheta}^\top\cdot) - f(\vartheta^\top\cdot)\|_{L^2(P_X)}^2\big)$$

and using together (2.14) and $f \in H^Q(s,L)$ for $s \geq s_{\min}$, we obtain

$$\|f(\bar{\vartheta}^\top\cdot) - f(\vartheta^\top\cdot)\|_{L^2(P_X)}^2 \leq L^2\int\|x\|_2^{2s_{\min}}P_X(dx)\Delta^{2s_{\min}} \leq C(n\log n)^{-1}.$$

Let us denote by $Q_\vartheta(\cdot|X_1^m)$ the joint law of $(X_i,Y_i)_{1\leq i\leq m}$ from model (1.1) (when the index is ϑ) conditional on the $(X_i)_{1\leq i\leq m}$, which is given by

$$Q_\vartheta(dy_1^m|x_1^m) := \prod_{i=1}^m \frac{1}{(\sigma(x_i)(2\pi)^{1/2})}\exp\Big(-\frac{(y_i - f(\vartheta^\top x_i))^2}{2\sigma(x_i)^2}\Big)dy_i.$$

Under $Q_{\bar\vartheta}(\cdot|X_1^m)$, we have

$$
\begin{aligned}
L_X(\vartheta,\bar\vartheta) &:= \frac{dQ_{\vartheta}(\cdot|X_1^m)}{dQ_{\bar\vartheta}(\cdot|X_1^m)} \\
&\overset{\text{(law)}}{=} \exp\Big(-\sum_{i=1}^m \frac{\epsilon_i(f(\bar\vartheta^\top X_i) - f(\vartheta^\top X_i))}{\sigma(X_i)} - \frac{1}{2}\sum_{i=1}^m \frac{(f(\bar\vartheta^\top X_i) - f(\vartheta^\top X_i))^2}{\sigma(X_i)^2}\Big).
\end{aligned}
$$

Hence, if P_X^m denotes the joint law of $(X_1, \dots, X_m)$,

$$
\begin{aligned}
&E^m \|\bar g^{(\bar\vartheta)}(\cdot) - f(\bar\vartheta^\top \cdot)\|_{L^2(P_X)}^2 \\
&\qquad = \int\int \|\bar g^{(\bar\vartheta)}(\cdot) - f(\bar\vartheta^\top\cdot)\|_{L^2(P_X)}^2 L_X(\vartheta,\bar\vartheta) dQ_{\bar\vartheta}(\cdot|X_1^m) dP_X^m \\
&\qquad \leq C\int\int \|\bar f^{(\bar\vartheta)}(\bar\vartheta^\top\cdot) - f(\bar\vartheta^\top\cdot)\|_{L^2(P_X)}^2 dQ_{\bar\vartheta}(\cdot|X_1^m) dP_X^m \qquad (5.3)\\
&\qquad + 4Q^2 \int\int L_X(\vartheta,\bar\vartheta)\mathbf{1}_{\{L_X(\vartheta,\bar\vartheta)\geq C\}} dQ_{\bar\vartheta}(\cdot|X_1^m) dP_X^m,
\end{aligned}
$$

where we decomposed the integrand over $\{L_X(\vartheta,\bar\vartheta) \geq C\}$ and $\{L_X(\vartheta,\bar\vartheta)\leq C\}$ for some constant $C\geq 3$, and where we used the fact that $\|\bar g^{(\bar\vartheta)}\|_\infty, \|f\|_\infty \leq Q$. Under $Q_{\bar\vartheta}(\cdot|X_1^m)$, the (X_i,Y_i) have the same law as (X,Y) from model (1.1) where the index is $\bar\vartheta$. Moreover, we assumed that $P_{\bar\vartheta^\top X}$ satisfies Assumption (D). Hence, Theorem 3 entails that, uniformly for $f\in H^Q(\bar s, L)$,

$$
\int\int \|\bar f^{(\bar\vartheta)}(\bar\vartheta^\top\cdot) - f(\bar\vartheta^\top\cdot)\|_{L^2(P_X)}^2 dQ_{\bar\vartheta}(\cdot|X_1^m) dP_X^m \leq C' m^{-2\bar s/(2\bar s+1)}.
$$

Moreover, the second term in the right hand side of (5.3) is smaller than

$$
4Q^2\int\Big(\int L_X(\vartheta,\bar\vartheta)^2 dQ_{\bar\vartheta}(\cdot|X_1^m)\Big)^{1/2} Q_{\bar\vartheta}\big[L_X(\vartheta,\bar\vartheta)\geq C|X_1^m\big]^{1/2} dP_X^m.
$$

Since $f\in H^Q(s,L)$ for $s\geq s_{\min}$, since P_X is compactly supported and since $\sigma(X) > \sigma_0$ a.s., we obtain using (2.14):

$$
\int L_X(\vartheta,\bar\vartheta)^2 dQ_{\bar\vartheta}(\cdot|X_1^m) \leq \exp\Big(\frac{1}{2}\sum_{i=1}^m \frac{(f(\bar\vartheta^\top X_i) - f(\vartheta^\top X_i))^2}{\sigma(X_i)^2}\Big) \leq 1
$$

P_X^m-a.s. when m is large enough. Moreover, with the same arguments we have

$$
Q_{\bar\vartheta}\big[L_X(\vartheta,\bar\vartheta)\geq C|X_1^m\big] \leq m^{-(\log C)^2/2} \leq m^{-4\bar s/(2\bar s+1)}
$$

for C large enough, where we use the standard Gaussian deviation $P[N(0,b^2)\geq a]\leq \exp(-a^2/(2b^2))$. This concludes the proof of Theorem 1. □

Proof of Theorem 2

We want to bound the minimax risk

$$\inf_{\tilde{g}} \sup_{f \in H^Q(s,L)} E^n \int \big(\tilde{g}(x) - f(\vartheta^\top x)\big)^2 P_X(dx) \tag{5.4}$$

from below, where the infimum is taken among all estimators $\mathbb{R}^d \to \mathbb{R}$ based on data from model (1.1),(1.2). We recall that $\vartheta^\top X$ satisfies Assumption (D). We consider $\vartheta^{(2)}, \dots, \vartheta^{(d)}$ in $\mathbb{R}^d$ such that $(\vartheta, \vartheta^{(2)}, \dots, \vartheta^{(d)})$ is an orthogonal basis of $\mathbb{R}^d$. We denote by $\mathbf{O}$ the matrix with columns $\vartheta, \vartheta^{(2)}, \dots, \vartheta^{(d)}$. We define $Y := \mathbf{O}X = (Y^{(1)}, \dots, Y^{(d)})$ and $Y_2^d := (Y^{(2)}, \dots, Y^{(d)})$. By the change of variable $y = \mathbf{O}x$, we obtain

$$\begin{aligned}
&\int_{\mathbb{R}^d} \big(\tilde{g}(x) - f(\vartheta^\top x)\big)^2 P_X(dx) \\
&\qquad = \int_{\mathbb{R}^d} \big(\tilde{g}(\mathbf{O}^{-1}y) - f(y^{(1)})\big)^2 P_Y(dy) \\
&\qquad = \int_{\mathbb{R}} \int_{\mathbb{R}^{d-1}} \big(\tilde{g}(\mathbf{O}^{-1}y) - f(y^{(1)})\big)^2 P_{Y_2^d|Y^{(1)}}(dy_2^d|y^{(1)}) P_{Y^{(1)}}(dy^{(1)}) \\
&\qquad \geq \int_{\mathbb{R}} \big(\tilde{f}(y^{(1)}) - f(y^{(1)})\big)^2 P_{\vartheta^\top X}(dy^{(1)}),
\end{aligned}$$

where $\tilde{f}(y^{(1)}) := \int \tilde{g}(\mathbf{O}^{-1}y) P_{Y_2^d|Y^{(1)}}(dy_2^d|y^{(1)})$. Hence, if $Z := \vartheta^\top X$, (5.4) is larger than

$$\inf_{\tilde{f}} \sup_{f \in H^Q(s,L)} E^n \int \big(\tilde{f}(z) - f(z)\big)^2 P_Z(dz), \tag{5.5}$$

where the infimum is taken among all estimators $\mathbb{R} \to \mathbb{R}$ based on data from model (1.1) with $d = 1$ (univariate regression). In order to bound (5.5) from below, we use the following Theorem, from Tsybakov (2003a), which is a standard tool for the proof of such a lower bound. We say that ∂ is a *semi-distance* on some set Θ if it is symmetric, if it satisfies the triangle inequality and if $\partial(\theta, \theta) = 0$ for any $\theta \in \Theta$. We consider $K(P|Q) := \int \log(\frac{dP}{dQ}) dP$ the Kullback-Leibler divergence between probability measures P and Q.

Theorem 5. *Let (Θ, ∂) be a set endowed with a semi-distance ∂. We suppose that $\{P_\theta; \theta \in \Theta\}$ is a family of probability measures on a measurable space $(\mathcal{X}, \mathcal{A})$ and that $(v_n)_{n \in \mathbb{N}}$ is a sequence of positive numbers. If there exist $\{\theta_0, \dots, \theta_M\} \subset \Theta$, with $M \geq 2$, such that*

- $\partial(\theta_j, \theta_k) \geq 2v_n \quad \forall 0 \leq j < k \leq M$
- $P_{\theta_j} \ll P_{\theta_0} \quad \forall 1 \leq j \leq M$,
- $\frac{1}{M} \sum_{j=1}^M K(P_{\theta_j}^n | P_{\theta_0}^n) \leq \alpha \log M$ *for some* $\alpha \in (0, 1/8)$,

then

$$\inf_{\tilde{\theta}_n} \sup_{\theta \in \Theta} E_\theta^n [(v_n^{-1} \partial(\tilde{\theta}_n, \theta))^2] \geq \frac{\sqrt{M}}{1 + \sqrt{M}} \left(1 - 2\alpha - 2\sqrt{\frac{\alpha}{\log M}}\right),$$

where the infimum is taken among all estimators based on a sample of size n.

Let us define $m := \lfloor c_0 n^{1/(2s+1)} \rfloor$, the largest integer smaller than $c_0 n^{1/(2s+1)}$, where $c_0 > 0$. Let $\varphi : \mathbb{R} \to [0, +\infty)$ be a function in $H^Q(s, 1/2; \mathbb{R})$ with support in $[-1/2, 1/2]$. We take $h_n := m^{-1}$ and $z_k := (k - 1/2)/m$ for $k \in \{1, \ldots, m\}$. For $\omega \in \Omega := \{0, 1\}^m$, we consider the functions

$$f(\cdot; \omega) := \sum_{k=1}^{m} \omega_k \varphi_k(\cdot) \quad \text{where} \quad \varphi_k(\cdot) := L h_n^s \varphi\Big(\frac{\cdot - z_k}{h_n}\Big).$$

We have

$$\begin{aligned} \|f(\cdot; \omega) - f(\cdot; \omega')\|_{L^2(P_Z)} &= \Big(\sum_{k=1}^{m} (\omega_k - \omega_{k'})^2 \int \varphi_k(z)^2 P_Z(dz) \Big)^{1/2} \\ &\geq \mu_0^{1/2} \rho(\omega, \omega') L^2 h_n^{2s+1} \int_{S_\mu} \varphi(u)^2 du, \end{aligned}$$

where $S_\mu := \operatorname{Supp} P_Z - \cup_z [a_z, b_z]$ (the union is over the z such that $\mu(z) = 0$, see Assumption (D)), where $\mu_0 := \min_{z \in S_\mu} \mu(z) > 0$ and where

$$\rho(\omega, \omega') := \sum_{k=1}^{m} \mathbf{1}_{\omega_k \neq \omega'_k}$$

is the Hamming distance on Ω. Using a result of Varshamov-Gilbert (see Tsybakov (2003a)) we can find a subset $\{\omega^{(0)}, \ldots, \omega^{(M)}\}$ of Ω such that $\omega^{(0)} = (0, \ldots, 0)$, $\rho(\omega^{(j)}, \omega^{(k)}) \geq m/8$ for any $0 \leq j < k \leq M$ and $M \geq 2^{m/8}$. Hence, we have

$$\|f(\cdot; \omega^{(j)}) - f(\cdot; \omega^{(k)})\|_{L^2(P_Z)} \geq D n^{-s/(2s+1)},$$

where $D = \mu_0^{1/2} \int_{S_\mu} \varphi(u)^2 du / (8 c_0^{2s}) \geq 2$ for c_0 small enough. Moreover,

$$\begin{aligned} \frac{1}{M} \sum_{k=1}^{M} K(P^n_{f(\cdot, \omega^{(0)})} | P^n_{f(\cdot, \omega^{(k)})}) &\leq \frac{n}{2M\sigma_0^2} \sum_{k=1}^{M} \|f(\cdot; \omega^{(0)}) - f(\cdot; \omega^{(k)})\|^2_{L^2(P_Z)} \\ &\leq \frac{n}{2\sigma_0^2} L^2 h_n^{2s+1} \|\varphi\|_2^2 m \leq \alpha \log M, \end{aligned}$$

where $\alpha := (L^2 \|\varphi\|_2^2)/(\sigma^2 c_0^{2s+1} \log 2) \in (0, 1/8)$ for c_0 small enough. The conclusion follows from Theorem 5. □

Proof of Theorem 3

We recall that $r = \lfloor s \rfloor$ is the largest integer smaller than s, and that $\lambda(M)$ stands for the smallest eigenvalue of a matrix M.

Proof of (3.2)

First, we prove a bias-variance decomposition of the LPE at a fixed point $z \in \operatorname{Supp} P_Z$. This kind of result is commonplace, see for instance Fan and Gijbels (1995, 1996). We introduce the following weighted pseudo-inner product, for fixed $z \in \mathbb{R}$ and $h > 0$:

$$\langle f, g \rangle_h := \frac{1}{m \bar{P}_Z[I(z,h)]} \sum_{i=1}^{m} f(Z_i) g(Z_i) \mathbf{1}_{Z_i \in I(z,h)},$$

where we recall that $I(z,h) = [z-h, z+h]$, and that $\bar{P}_Z$ is given by (2.5). We consider the associated pseudo-norm $\|g\|_h^2 := \langle g, g \rangle_h$. We introduce the power functions $\varphi_a(\cdot) := ((\cdot - z)/h)^a$ for $a \in \{0, \ldots, r\}$, which satisfy $\|\varphi_a\|_h \le 1$.

Note that the entries of the matrix $\bar{\mathbf{Z}}_m = \bar{\mathbf{Z}}_m(z,h)$ (see (2.4)) satisfy $(\bar{\mathbf{Z}}_m(z,h))_{a,b} := \langle \varphi_a, \varphi_b \rangle_h$ for $(a,b) \in \{0, \ldots, r\}^2$. Hence, (2.3) is equivalent to find $\bar{P} \in \mathrm{Pol}_r$ such that

$$\langle \bar{P}, \varphi_a \rangle_h = \langle Y, \varphi_a \rangle_h \tag{5.6}$$

for any $a \in \{0, \ldots, r\}$, where $\langle Y, \varphi \rangle_h := (m\bar{P}_Z[I(z,h)])^{-1} \sum_{i=1}^m Y_i \varphi(Z_i) \mathbf{1}_{Z_i \in I(z,h)}$. In other words, $\bar{P}$ is the projection of Y onto Pol_r with respect to the inner product $\langle \cdot, \cdot \rangle_h$. For $e_1 := (1, 0, \ldots, 0) \in \mathbb{R}^{r+1}$, we have

$$\bar{f}(z) - f(z) = e_1^\top \bar{\mathbf{Z}}_m^{-1} \bar{\mathbf{Z}}_m (\bar{\theta} - \theta)$$

whenever $\lambda(\bar{\mathbf{Z}}_m) > 0$, where $\bar{\theta}$ is the coefficient vector of $\bar{P}$ and θ is the coefficient vector of the Taylor polynomial P of f at z with degree r. In view of (5.6):

$$(\bar{\mathbf{Z}}_m(\bar{\theta} - \theta))_a = \langle \bar{P} - P, \varphi_a \rangle_h = \langle Y - P, \varphi_a \rangle_h,$$

thus $\bar{\mathbf{Z}}_m(\bar{\theta} - \theta)) = B + V$ where $(B)_a := \langle f - P, \varphi_a \rangle_h$ and $(V)_a := \langle \sigma(\cdot)\xi, \varphi_a \rangle_h$. The bias term satisfies $|e_1^\top \bar{\mathbf{Z}}_m^{-1} B| \le (r+1)^{1/2} \|\bar{\mathbf{Z}}_m^{-1}\| \|B\|_\infty$ where for any $a \in \{0, \ldots, r\}$

$$|(B)_a| \le \|f - P\|_h \le L h^s / r!.$$

Let $\bar{\mathbf{Z}}_m^\sigma$ be the matrix with entries $(\bar{\mathbf{Z}}_m^\sigma)_{a,b} := \langle \sigma(\cdot)\varphi_a, \sigma(\cdot)\varphi_b \rangle_h$. Since V is, conditionally on $Z_1^m = (Z_1, \ldots, Z_m)$, centered Gaussian with covariance matrix $(m\bar{P}_Z[I(z,h)])^{-1} \bar{\mathbf{Z}}_m^\sigma$, we have that $e_1^\top \bar{\mathbf{Z}}_m^{-1} V$ is centered Gaussian with variance smaller than

$$(m\bar{P}_Z[I(z,h)])^{-1} e_1^\top \bar{\mathbf{Z}}_m^{-1} \bar{\mathbf{Z}}_m^\sigma \bar{\mathbf{Z}}_m^{-1} e_1 \le \sigma_1^2 (m\bar{P}_Z[I(z,h)])^{-1} \lambda(\bar{\mathbf{Z}}_m)^{-1}$$

where we used $\sigma(\cdot) \le \sigma_1$. Hence, if $C_r := (r+1)^{1/2}/r!$, we obtain

$$E^m[(\bar{f}(z) - f(z))^2 | Z_1^m] \le \lambda(\bar{\mathbf{Z}}_m(z,h))^{-2} \big(C_r L h^s + \sigma_1 (m\bar{P}_Z[I(z,h)])^{-1/2} \big)^2$$

for any z, and the bandwidth choice (2.6) entails (3.2).

Proof of (3.3)

Let us consider the sequence of positive curves $h_m(\cdot)$ defined as the point-by-point solution to

$$Lh_m(z)^s = \frac{\sigma_1}{(mP_Z[I(z, h_m(z))])^{1/2}} \tag{5.7}$$

for all $z \in \operatorname{Supp} P_Z$, where we recall $I(z,h) = [z-h, z+h]$, and let us define

$$r_m(z) := Lh_m(z)^s.$$

The sequence $h_m(\cdot)$ is the deterministic equivalent to the bandwidth $H_m(\cdot)$ given by (2.6). Indeed, with a large probability, $H_m(\cdot)$ and $h_m(\cdot)$ are close to each other in view of Lemma 1 below. Under Assumption (D) we have $P_Z[I] \geq \gamma |I|^{\beta+1}$, which entails together with (5.7) that

$$h_m(z) \leq D_1 m^{-1/(1+2s+\beta)} \tag{5.8}$$

uniformly for $z \in \operatorname{Supp} P_Z$, where $D_1 = (\sigma_1/L)^{2/(1+2s+\beta)}(\gamma 2^{\beta+1})^{-1/(1+2s+\beta)}$. Moreover, since P_Z has a continuous density μ with respect to the Lebesgue measure, we have

$$h_m(z) \geq D_2 m^{-1/(1+2s)} \tag{5.9}$$

uniformly for $z \in \operatorname{Supp} P_Z$, where $D_2 = (\sigma_1/L)^{2/(1+2s)}(2\mu_\infty)^{-1/(2s+1)}$. We recall that P_Z^m stands for the joint law of $(Z_1, \ldots, Z_m)$.

Lemma 1. *If Z satisfies Assumption (D), we have for any $\epsilon \in (0, 1/2)$*

$$P_Z^m \Big[\sup_{z \in \operatorname{Supp}(P_Z)} \Big| \frac{H_m(z)}{h_m(z)} - 1 \Big| > \epsilon \Big] \leq \exp(-D\epsilon^2 m^\alpha)$$

for m large enough, where $\alpha := 2s/(1+2s+\beta)$ and D is a constant depending on σ_1 and L.

The next lemma provides an uniform control on the smallest eigenvalue of $\bar{\mathbf{Z}}_m(z) := \bar{\mathbf{Z}}_m(z, H_m(z))$ under Assumption (D).

Lemma 2. *If Z satisfies Assumption (D), there exists $\lambda_0 > 0$ depending on β and s only such that*

$$P_Z^m \big[\inf_{z \in \operatorname{Supp} P_Z} \lambda(\bar{\mathbf{Z}}_m(z)) \leq \lambda_0 \big] \leq \exp(-Dm^\alpha),$$

for m large enough, where $\alpha = 2s/(1+2s+\beta)$, and D is a constant depending on $\gamma, \beta, s, L, \sigma_1$.

The proofs Lemmas 1 and 2 are given in Section 6. We consider the event

$$\Omega_m(\epsilon) := \big\{ \inf_{z \in \operatorname{Supp} P_Z} \lambda(\bar{\mathbf{Z}}_m(z)) > \lambda_0 \big\} \cap \big\{ \sup_{z \in \operatorname{Supp} P_Z} |H_m(z)/h_m(z) - 1| \leq \epsilon \big\},$$

where $\epsilon \in (0, 1/2)$. We have for any $f \in H^Q(s, L)$

$$E^m[\|\tau_Q(\bar{f}) - f\|^2_{L^2(P_Z)} \mathbf{1}_{\Omega_m(\epsilon)}] \leq \lambda_0^{-2}(1+\epsilon)^{2s} \frac{\sigma_1^2}{m} \int \frac{P_Z(dz)}{\int_{z-h_m(z)}^{z+h_m(z)} P_Z(dt)},$$

where we used together the definition of $\Omega_m(\epsilon)$, (3.2) and (5.7). Let us denote $I := \operatorname{Supp} P_Z$ and let I_{z*} be the intervals from Assumption (D). Using together the fact that $\min_{z \in I - \cup_{z*} I_{z*}} \mu(z) > 0$ and (5.9), we obtain

$$\frac{\sigma_1^2}{m} \int_{I - \cup_{z*} I_{z*}} \frac{P_Z(dz)}{\int_{z-h_m(z)}^{z+h_m(z)} P_Z(dt)} \leq C m^{-2s/(2s+1)}.$$

Using the monoticity constraints from Assumption (D), we obtain

$$\begin{aligned}
&\frac{\sigma_1^2}{m} \int_{I_{z*}} \frac{P(dz)}{\int_{z-h_m(z)}^{z+h_m(z)} P_Z(dt)} \\
&\quad \leq \frac{\sigma_1^2}{m} \Big(\int_{z*-a_{z*}}^{z*} \frac{\mu(z)dz}{\int_{z-h_m(z)}^{z} \mu(t)dt} + \int_{z*}^{z*+b_{z*}} \frac{\mu(z)dz}{\int_{z}^{z+h_m(z)} \mu(t)dt} \Big) \\
&\quad \leq \frac{\sigma_1^2}{m} \int_{I_{z*}} h_m(z)^{-1} dz \leq C m^{-2s/(2s+1)},
\end{aligned}$$

hence $E^m[\|\tau_Q(\bar{f}) - f\|^2_{L^2(P_Z)} \mathbf{1}_{\Omega_m(\epsilon)}] \leq C m^{-2s/(2s+1)}$ uniformly for $f \in H^Q(s, L)$. Using together Lemmas 1 and 2, we obtain $E^m[\|\tau_Q(\bar{f}) - f\|^2_{L^2(P_Z)} \mathbf{1}_{\Omega_m(\epsilon)^\complement}] = o(n^{-2s/(2s+1)})$, and (3.3) follows. □

Proof of Theorem 4

In model (1.1), when the noise ϵ is centered and such that $E(\epsilon^2) = 1$, the risk of a function $\bar{g} : \mathbb{R}^d \to \mathbb{R}$ is given by

$$A(\bar{g}) := E[(Y - \bar{g}(X))^2] = E[\sigma(X)^2] + \|\bar{g} - g\|^2_{L^2(P_X)},$$

where g is the regression function. Therefore, the excess risk satisfies

$$A(\bar{g}) - A = \|\bar{g} - g\|^2_{L^2(P_X)},$$

where $A := A(g) = E[\sigma(X)^2]$. Let us introduce $n := |D|$ the size of the learning sample, and $M := |\Lambda|$ the size of the dictionary of functions $\{\bar{g}^{(\lambda)}; \lambda \in \Lambda\}$. The least squares of $\bar{g}$ over the learning sample is given by

$$A_n(\bar{g}) := \frac{1}{n} \sum_{i=1}^n (Y_i - \bar{g}(X_i))^2.$$

We begin with a linearization of these risks. We consider the convex set

$$\mathcal{C} := \Big\{ (\theta_\lambda)_{\lambda \in \Lambda} \text{ such that } \theta_\lambda \geq 0 \text{ and } \sum_{\lambda \in \Lambda} \theta_\lambda = 1 \Big\},$$

and define the linearized risks on $\mathcal{C}$ as

$$\tilde{A}(\theta) := \sum_{\lambda \in \Lambda} \theta_\lambda A(\bar{g}^{(\lambda)}), \quad \tilde{A}_n(\theta) := \sum_{\lambda \in \Lambda} \theta_\lambda A_n(\bar{g}^{(\lambda)}),$$

which are linear versions of the risk A and its empirical version A_n. The exponential weights $w = (w_\lambda)_{\lambda \in \Lambda} := (w(\bar{g}^{(\lambda)}))_{\lambda \in \Lambda}$ are actually the unique solution of the minimization problem

$$\min \Big(\tilde{A}_n(\theta) + \frac{1}{Tn} \sum_{\lambda \in \Lambda} \theta_\lambda \log \theta_\lambda \;\big|\; (\theta_\lambda) \in \mathcal{C} \Big), \tag{5.10}$$

where $T > 0$ is the temperature parameter in the weights (2.10), and where we use the convention $0 \log 0 = 0$. Let $\hat{\lambda} \in \Lambda$ be such that $A_n(\bar{g}^{(\hat{\lambda})}) = \min_{\lambda \in \Lambda} A_n(\bar{g}^{(\lambda)})$. Since $\sum_{\lambda \in \Lambda} w_\lambda \log \big(\frac{w_\lambda}{1/M} \big) = K(w|u) \geq 0$ where $K(w|u)$ denotes the Kullback-Leibler divergence between the weights w and the uniform weights $u := (1/M)_{\lambda \in \Lambda}$, we have together with (5.10):

$$\begin{aligned}
\tilde{A}_n(w) &\leq \tilde{A}_n(w) + \frac{1}{Tn} K(w|u) \\
&= \tilde{A}_n(w) + \frac{1}{Tn} \sum_{\lambda \in \Lambda} w_\lambda \log w_\lambda + \frac{\log M}{Tn} \\
&\leq \tilde{A}_n(e_{\hat{\lambda}}) + \frac{\log M}{Tn},
\end{aligned}$$

where $e_\lambda \in \mathcal{C}$ is the vector with 1 for the λ-th coordinate and 0 elsewhere. Let $a > 0$ and $A_n := A_n(g)$. For any $\lambda \in \Lambda$, we have

$$\begin{aligned}
\tilde{A}(w) - A &= (1+a)(\tilde{A}_n(w) - A_n) + \tilde{A}(w) - A - (1+a)(\tilde{A}_n(w) - A_n) \\
&\leq (1+a)(\tilde{A}_n(e_\lambda) - A_n) + (1+a)\frac{\log M}{Tn} \\
&\quad + \tilde{A}(w) - A - (1+a)(\tilde{A}_n(w) - A_n).
\end{aligned}$$

Let us denote by E_K the expectation with respect to P_K, the joint law of the learning sample for a noise ϵ which is bounded almost surely by $K > 0$. We have

$$\begin{aligned}
E_K\big[\tilde{A}(w) - A\big] \leq (1+a) \min_{\lambda \in \Lambda} (\tilde{A}_n(e_\lambda) - A_n) + (1+a)\frac{\log M}{Tn} \\
+ E_K\big[\tilde{A}(w) - A - (1+a)(\tilde{A}_n(w) - A_n)\big].
\end{aligned}$$

Using the linearity of $\tilde{A}$ on $\mathcal{C}$, we obtain

$$\tilde{A}(w) - A - (1+a)(\tilde{A}_n(w) - A_n) \leq \max_{g \in \mathcal{G}_\Lambda} \big(A(g) - A - (1+a)(A_n(g) - A_n) \big),$$

where $\mathcal{G}_\Lambda := \{\bar{g}^{(\lambda)} \, ; \, \lambda \in \Lambda\}$. Then, using Bernstein inequality, we obtain for all $\delta > 0$

$$\begin{aligned}
&P_K\big[\tilde{A}(w) - A - (1+a)(\tilde{A}_n(w) - A_n) \geq \delta\big] \\
&\quad \leq \sum_{g \in \mathcal{G}_\Lambda} P_K\Big[A(g) - A - (A_n(g) - A_n) \geq \frac{\delta + a(A(g) - A)}{1+a}\Big] \\
&\quad \leq \sum_{g \in \mathcal{G}_\Lambda} \exp\Big(- \frac{n(\delta + a(A(g)-A))^2(1+a)^{-1}}{8Q^2(1+a)(A(g)-A) + 2(6Q^2 + 2\sigma K)(\delta + a(A(g)-A))/3}\Big).
\end{aligned}$$

Moreover, we have for any $\delta > 0$ and $g \in \mathcal{G}_\Lambda$,

$$\frac{(\delta + a(A(g)-A))^2(1+a)^{-1}}{8Q^2(A(g)-A) + 2(6Q^2(1+a) + 2\sigma K)(\delta + a(A(g)-A))/3} \geq C(a,K)\delta,$$

where $C(a,K) := \big(8Q^2(1+a)^2/a + 4(6Q^2 + 2\sigma K)(1+a)/3\big)^{-1}$, thus

$$E_K\big[\tilde{A}(w) - A - (1+a)(\tilde{A}_n(w) - A_n)\big] \leq 2u + M\frac{\exp(-nC(a,K)u)}{nC(a,K)}.$$

If we denote by γ_A the unique solution of $\gamma = A\exp(-\gamma)$, where $A > 0$, we have $(\log A)/2 \leq \gamma_A \leq \log A$. Thus, if we take $u = \gamma_M/(nC(a,K))$, we obtain

$$E_K\big[\tilde{A}(w) - A - (1+a)(\tilde{A}_n(w) - A_n)\big] \leq \frac{3\log M}{C(a,K)n}.$$

By convexity of the risk, we have

$$\tilde{A}(w) - A \geq A(\hat{g}) - A,$$

thus

$$E_K\big[\|\hat{g} - g\|^2_{L^2(P_X)}\big] \leq (1+a)\min_{\lambda \in \Lambda} \|\bar{g}^{(\lambda)} - g\|^2_{L(P_X)} + C_1\frac{\log M}{n},$$

where $C_1 := (1+a)(T^{-1} + 3C(a,K)^{-1})$. It remains to prove the result when the noise is Gaussian. Let us denote $\epsilon^n_\infty := \max_{1 \leq i \leq n}|\epsilon_i|$. For any $K > 0$, we have

$$\begin{aligned}
E\big[\|\hat{g} - g\|^2_{L^2(P_X)}\big] &= E\big[\|\hat{g} - g\|^2_{L^2(P_X)}\mathbf{1}_{\epsilon^n_\infty \leq K}\big] + E\big[\|\hat{g} - g\|^2_{L^2(P_X)}\mathbf{1}_{\epsilon^n_\infty > K}\big] \\
&\leq E_K\big[\|\hat{g} - g\|^2_{L^2(P_X)}\big] + 2Q^2 P[\epsilon^n_\infty > K].
\end{aligned}$$

For $K = K_n := 2(2\log n)^{1/2}$, we obtain using standard results about the maximum of Gaussian vectors that $P[\epsilon^n_\infty > K_n] \leq P[\epsilon^n_\infty - E[\epsilon^n_\infty] > (2\log n)^{1/2}] \leq 1/n$, which concludes the proof of the Theorem. □

6. Proof of the lemmas

Proof of Lemma 1

Using together (2.6) and (5.7), if $I_m^\epsilon(z) := [z - (1+\epsilon)h_m(z), z + (1+\epsilon)h_m(z)]$ and $I_m(z) := I_m^0(z)$, we obtain for any $\epsilon \in (0, 1/2)$:

$$\begin{aligned}\{H_m(z) \le (1+\epsilon)h_m(z)\} &= \big\{(1+\epsilon)^{2s}\bar{P}_Z[I_m^\epsilon(z)] \ge P_Z[I_m(z)]\big\} \\ &\supset \big\{(1+\epsilon)^{2s}\bar{P}_Z[I_m(z)] \ge P_Z[I_m(z)]\big\},\end{aligned}$$

where we used the fact that $\epsilon \mapsto P_Z[I_m^\epsilon(z)]$ is nondecreasing. Similarly, we have on the other side

$$\{H_m(z) > (1-\epsilon)h_m(z)\} \supset \big\{(1-\epsilon)^{2s}\bar{P}_Z[I_m(z)] \le P_Z[I_m(z)]\big\}.$$

Thus, if we consider the set of intervals

$$\mathcal{I}_m := \bigcup_{z \in \text{Supp } P_Z} \big\{I_m(z)\big\},$$

we obtain

$$\Big\{\sup_{z\in \text{Supp } P_Z}\Big|\frac{H_m(z)}{h_m(z)} - 1\Big| \ge \epsilon\Big\} \subset \Big\{\sup_{I\in\mathcal{I}_m}\Big|\frac{\bar{P}_Z[I]}{P_Z[I]} - 1\Big| \ge \epsilon/2\Big\}.$$

Using together (5.7) and (5.8), we obtain

$$P_Z[I_m(z)] = \sigma_1^2/(mL^2h_m(z)^{2s}) \ge Dm^{-(\beta+1)/(1+2s+\beta)} =: \alpha_m. \tag{6.1}$$

Hence, if $\epsilon' := \epsilon(1+\epsilon/2)/(\epsilon+2)$, we have

$$\begin{aligned}\Big\{\sup_{I\in\mathcal{I}_m}\Big|\frac{\bar{P}_Z[I]}{P_Z[I]} - 1\Big| \ge \epsilon/2\Big\} &\subset \Big\{\sup_{I\in\mathcal{I}_m}\frac{\bar{P}_Z[I] - P_Z[I]}{\sqrt{\bar{P}_Z[I]}} \ge \epsilon'\alpha_m^{1/2}\Big\} \\ &\quad \cup \Big\{\sup_{I\in\mathcal{I}_m}\frac{P_Z[I] - \bar{P}_Z[I]}{\sqrt{P_Z[I]}} \ge \epsilon\alpha_m^{1/2}/2\Big\}.\end{aligned}$$

Then, Theorem 6 (see Appendix) and the fact that the shatter coefficient satisfies $\mathcal{S}(\mathcal{I}_m, m) \le m(m+1)/2$ entails the Lemma. □

Proof of Lemma 2

Let us denote $\bar{\mathbf{Z}}_m(z) := \bar{\mathbf{Z}}_m(z, H_m(z))$ where $\bar{\mathbf{Z}}_m(z,h)$ is given by (2.4) and where $H_m(z)$ is given by (2.6). Let us define the matrix $\tilde{\mathbf{Z}}_m(z) := \tilde{\mathbf{Z}}_m(z, h_m(z))$ where

$$(\tilde{\mathbf{Z}}_m(z,h))_{a,b} := \frac{1}{mP_Z[I(z,h)]}\sum_{i=1}^m \Big(\frac{Z_i - z}{h}\Big)^{a+b}\mathbf{1}_{Z_i \in I(z,h)}.$$

Step 1. Let us define for $\epsilon \in (0,1)$ the event

$$\Omega_1(\epsilon) := \Big\{ \sup_{z \in \mathrm{Supp}\, P_Z} \Big| \frac{H_m(z)}{h_m(z)} - 1 \Big| \le \epsilon \Big\} \cap \Big\{ \sup_{z \in \mathrm{Supp}\, P_Z} \Big| \frac{\bar P_Z[I(z, H_m(z))]}{P_Z[I(z, h_m(z))]} - 1 \Big| \le \epsilon \Big\}.$$

For a matrix A, we denote $\|A\|_\infty := \max_{a,b} |(A)_{a,b}|$. We can prove that on $\Omega_1(\epsilon)$, we have

$$\|\bar{\mathbf{Z}}_m(z) - \tilde{\mathbf{Z}}_m(z)\|_\infty \le \epsilon.$$

Moreover, using Lemma 1, we have $P_Z^m[\Omega_1(\epsilon)^{\complement}] \le C \exp(-D\epsilon^2 m^\alpha)$. Hence, on $\Omega_1(\epsilon)$, we have for any $v \in \mathbb{R}^d$, $\|v\|_2 = 1$

$$v^\top \bar{\mathbf{Z}}_m(z) v \ge v^\top \tilde{\mathbf{Z}}_m(z) v - \epsilon$$

uniformly for $z \in \mathrm{Supp}\, P_Z$.
Step 2. We define the deterministic matrix $\mathbf{Z}(z) := \mathbf{Z}(z, h_m(z))$ where

$$(\mathbf{Z}(z,h))_{a,b} := \frac{1}{P_Z[I(z,h)]} \int_{I(z,h)} \Big(\frac{t-z}{h} \Big)^{a+b} P_Z(dt),$$

and

$$\lambda_0 := \mathrm{liminf}_m \inf_{z\, \mathrm{Supp}\, P_Z} \lambda\big(\mathbf{Z}(z, h_m(z))\big).$$

We prove that $\lambda_0 > 0$. Two cases can occur: either $\mu(z) = 0$ or $\mu(z) > 0$. We show that in both cases, the liminf is positive. If $\mu(z) > 0$, the entries $(\mathbf{Z}(z, h_m(z)))_{a,b}$ have limit $(1+(-1)^{a+b})/(2(a+b+1))$, which defines a positive definite matrix. If $\mu(z) = 0$, we know that the density $\mu(\cdot)$ of P_Z behaves as the power function $|\cdot - z|^{\beta(z)}$ around z for $\beta(z) \in (0, \beta)$. In this case, $(\mathbf{Z}(z, h_m(z)))_{a,b}$ has limit $(1+(-1)^{a+b})(\beta(z)+1)/[2(1+a+b+\beta(z))]$, which defines also a definite positive matrix.
Step 3. We prove that

$$P_Z^m[\sup_{z \in \mathrm{Supp}\, P_Z} \|\tilde{\mathbf{Z}}_m(z) - \mathbf{Z}(z)\|_\infty > \epsilon] \le \exp(-D\epsilon^2 m^\alpha).$$

We consider the sets of nonnegative functions (we recall that $I(z,h) = [z - h, z+h]$)

$$F^{(\mathrm{even})} := \bigcup_{\substack{z \in \mathrm{Supp}\, P_Z \\ a \text{ even and } 0 \le a \le 2r}} \Big\{ \Big(\frac{\cdot - z}{h_m(z)} \Big)^a \mathbf{1}_{I(z, h_m(z))}(\cdot) \Big\},$$

$$F_+^{(\mathrm{odd})} := \bigcup_{\substack{z \in \mathrm{Supp}\, P_Z \\ a \text{ odd and } 0 \le a \le 2r}} \Big\{ \Big(\frac{\cdot - z}{h_m(z)} \Big)^a \mathbf{1}_{[z, z+h_m(z)]}(\cdot) \Big\},$$

$$F_-^{(\mathrm{odd})} := \bigcup_{\substack{z \in \mathrm{Supp}\, P_Z \\ a \text{ odd and } 0 \le a \le 2r}} \Big\{ \Big(\frac{z - \cdot}{h_m(z)} \Big)^a \mathbf{1}_{[z-h_m(z), z]}(\cdot) \Big\}.$$

Writing $I(z, h_m(z)) = [z - h_m(z), z) \cup [z, z + h_m(z)]$ when $a+b$ is odd, and since

$$P_Z[I(z, h_m(z))] \geq Ef(Z_1)$$

for any $f \in F := F^{(\text{even})} \cup F_+^{(\text{odd})} \cup F_-^{(\text{odd})}$, we obtain

$$\|\tilde{\mathbf{Z}}_m(z) - \mathbf{Z}(z)\|_\infty \leq \sup_{f \in F} \frac{|\frac{1}{m}\sum_{i=1}^m f(Z_i) - Ef(Z_1)|}{Ef(Z_1)}.$$

Hence, since $x \mapsto x/(x+\alpha)$ is increasing for any $\alpha > 0$, and since $\alpha := Ef(Z_1) \geq Dm^{-(\beta+1)/(1+2s+\beta)} =: \alpha_m$ (see (6.1)), we obtain

$$\Big\{ \sup_{z \in \operatorname{Supp} P_Z} \|\tilde{\mathbf{Z}}_m(z) - \mathbf{Z}(z)\|_\infty > \epsilon \Big\} \subset \Big\{ \sup_{f \in F} \frac{|\frac{1}{m}\sum_{i=1}^m f(Z_i) - Ef(Z_1)|}{\alpha_m + \frac{1}{m}\sum_{i=1}^m f(Z_i) + Ef(Z_1)} > \epsilon/2 \Big\}.$$

Then, using Theorem 7 (note that any $f \in F$ is non-negative), we obtain

$$P_Z^m[\sup_{z \in \operatorname{Supp} P_Z} \|\tilde{\mathbf{Z}}_m(z) - \mathbf{Z}(z)\|_\infty > \epsilon] \leq 4E[\mathcal{N}_1(\alpha_m \epsilon/8, F, Z_1^m)] \exp\big(-D\epsilon^2 m^{2s/(1+2s+\beta)}\big).$$

Together with the inequality

$$E[\mathcal{N}_1(\alpha_m \epsilon/8, F, Z_1^m)] \leq D(\alpha_m \epsilon)^{-1} m^{1/(2s+1)+(\beta-1)/(2s+\beta)}, \tag{6.2}$$

(see the proof below), this entails the Lemma. □

Proof of (6.2)

It suffices to prove the inequality for $F^{(\text{even})}$ and a fixed $a \in \{0, \ldots, 2r\}$, since the proof is the same for $F_+^{(\text{odd})}$ and $F_-^{(\text{odd})}$. We denote $f_z(\cdot) := ((\cdot - z)/h_m(z))^a \mathbf{1}_{I(z,h_m(z))}(\cdot)$. We prove the following statement:

$$\mathcal{N}(\epsilon, F, \|\cdot\|_\infty) \leq D\epsilon^{-1} m^{1/(2s+1)+(\beta-1)/(2s+\beta)},$$

which is stronger than (6.2), where $\|\cdot\|_\infty$ is the uniform norm over the support of P_Z. Let $z, z_1, z_2 \in \operatorname{Supp} P_Z$. We have

$$|f_{z_1}(z) - f_{z_2}(z)| \leq \max(a, 1) \Big| \frac{z - z_1}{h_1} - \frac{z - z_2}{h_2} \Big| \mathbf{1}_{I_1 \cup I_2},$$

where $h_j := h_m(z_j)$ and $I_j := [z_j - h_j, z_j + h_j]$ for $j = 1, 2$. Hence,

$$|f_{z_1}(z) - f_{z_2}(z)| \leq \frac{|h_1 - h_2| + |z_1 - z_2|}{\min(h_1, h_2)}.$$

Using (5.7) together with a differentiation of $z \mapsto h_m(z)^{2s} P_Z[I(z, h_m(z))]$, we obtain that

$$
\begin{aligned}
&|h_m(z_1) - h_m(z_2)| \\
&\leq \sup_{z_1 \leq z \leq z_2} \Big| \frac{h_m(z)^{2s+1}(\mu(z - h_m(z)) - \mu(z + h_m(z)))}{(2s\sigma_1^2)/(mL) + h_m(z)^{2s+1}(\mu(z - h_m(z)) + \mu(z + h_m(z)))} \Big| |z_1 - z_2|,
\end{aligned}
$$

for any $z_1 < z_2$ in $\operatorname{Supp}\mu$. This entails together with Assumption (D), (5.8) and (5.9):

$$
|h_m(z_1) - h_m(z_2)| \leq \frac{\mu_\infty}{2s(\gamma L)^{(2s+1)/(2s+\beta+1)}} \Big(\frac{m}{\sigma_1^2}\Big)^{\frac{\beta}{2s+\beta+1}} |z_1 - z_2|,
$$

for any $z_1 < z_2$ in $\operatorname{Supp}\mu$. Hence,

$$
|f_{z_1}(z) - f_{z_2}(z)| \leq D m^{\frac{1}{2s+1} + \frac{\beta-1}{2s+\beta}} |z_1 - z_2|,
$$

which concludes the proof of (6.2). □

Appendix A: Some tools from empirical process theory

Let $\mathcal{A}$ be a set of Borelean subsets of $\mathbb{R}$. If $x_1^n := (x_1, \ldots, x_n) \in \mathbb{R}^n$, we define

$$
N(\mathcal{A}, x_1^n) := \big|\big\{\{x_1, \ldots, x_n\} \cap A | A \in \mathcal{A}\big\}\big|
$$

and we define the *shatter* coefficient

$$
S(\mathcal{A}, n) := \max_{x_1^n \in \mathbb{R}^n} N(\mathcal{A}, (x_1, \ldots, x_n)). \tag{A.1}
$$

For instance, if $\mathcal{A}$ is the set of all the intervals $[a, b]$ with $-\infty \leq a < b \leq +\infty$, we have $S(\mathcal{A}, n) = n(n+1)/2$.

Let $X_1, \ldots, X_n$ be i.i.d. random variables with values in $\mathbb{R}$, and let us define $\mu[A] := P(X_1 \in A)$ and $\bar{\mu}_n[A] := n^{-1} \sum_{i=1}^n \mathbf{1}_{X_i \in A}$. The following inequalities for relative deviations are due to Vapnik and Chervonenkis (1974), see for instance in Vapnik (1998).

Theorem 6 (Vapnik and Chervonenkis (1974)). *We have*

$$
P\Big[\sup_{A \in \mathcal{A}} \frac{\mu(A) - \bar{\mu}_n(A)}{\sqrt{\mu(A)}} > \epsilon\Big] \leq 4\mathcal{S}(\mathcal{A}, 2n) \exp(-n\epsilon^2/4)
$$

and

$$
P\Big[\sup_{A \in \mathcal{A}} \frac{\bar{\mu}_n(A) - \mu(A)}{\sqrt{\bar{\mu}_n(A)}} > \epsilon\Big] \leq 4\mathcal{S}(\mathcal{A}, 2n) \exp(-n\epsilon^2/4)
$$

where $\mathcal{S}_{\mathcal{A}}(2n)$ is the shatter coefficient of $\mathcal{A}$ defined by (A.1).

Let $(\mathcal{X}, \tau)$ be a measured space and $\mathcal{F}$ be a class of functions $f : \mathcal{X} \to [-K, K]$. Let us fix $p \geq 1$ and $z_1^n \in \mathcal{X}^n$. Define the semi-distance $d_p(f, g)$ between f and g by

$$d_p(f,g) := \Big(\frac{1}{n}\sum_{i=1}^n |f(z_i) - g(z_i)|^p\Big)^{1/p}$$

and denote by $B^p(f, \epsilon)$ the d_p-ball with center f and radius ϵ. The $\epsilon-$covering number of $\mathcal{F}$ w.r.t d_p is defined as

$$\mathcal{N}_p(\epsilon, \mathcal{F}, z_1^n) := \min\big(N \mid \exists f_1, \ldots, f_N \text{ s.t. } \mathcal{F} \subseteq \cup_{j=1}^M B^p(f_j, \epsilon)\big).$$

Theorem 7 (Haussler (1992)). *If $\mathcal{F}$ consists of functions $f : \mathcal{X} \to [0, K]$, we have*

$$P\Big[\sup_{f\in\mathcal{F}} \frac{\big|E[f(X_1)] - \frac{1}{n}\sum_{i=1}^n f(X_i)\big|}{\alpha + E[f(X_1)] + \frac{1}{n}\sum_{i=1}^n f(X_i)} \geq \epsilon\Big] \leq 4E[\mathcal{N}_p(\alpha\epsilon/8, \mathcal{F}, X_1^n)] \exp\Big(-\frac{n\alpha\epsilon^2}{16K^2}\Big).$$

References

Audibert, J.-Y. and Tsybakov, A. (2007). Fast learning rates for plug-in estimators under the margin condition. *The Annals of Statistics*, **35**. MR2336861

Birgé, L. (2005). Model selection via testing: an alternative to (penalized) maximum likelihood estimators. Available at `http://www.proba.jussieu.fr/mathdoc/textes/PMA-862.pdf`. MR2219712

Catoni, O. (2001). *Statistical Learning Theory and Stochastic Optimization*. Ecole d'été de Probabilités de Saint-Flour 2001, Lecture Notes in Mathematics, Springer, N.Y. MR2163920

Delecroix, M., Härdle, W. and Hristache, M. (2003). Efficient estimation in conditional single-index regression. *J. Multivariate Anal.*, **86** 213–226. MR1997761

Delecroix, M., Hristache, M. and Patilea, V. (2006). On semiparametric *M*-estimation in single-index regression. *J. Statist. Plann. Inference*, **136** 730–769. MR2181975

Fan, J. and Gijbels, I. (1995). Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. *Journal of the Royal Statistical Society. Series B. Methodological*, **57** 371–394. MR1323345

Fan, J. and Gijbels, I. (1996). *Local polynomial modelling and its applications*. Monographs on Statistics and Applied Probability, Chapman & Hall, London. MR1383587

Geenens, G. and Delecroix, M. (2005). A survey about single-index models theory. URL `http://www.stat.ucl.ac.be/ISpub/dp/2005/dp0508.pdf`.

Györfi, L., Kohler, M., Krzyżak, A. and Walk, H. (2002). *A distribution-free theory of nonparametric regression*. Springer Series in Statistics, Springer-Verlag, New York.

Horowitz, J. L. (1998). *Semiparametric methods in econometrics*, vol. 131 of *Lecture Notes in Statistics*. Springer-Verlag, New York. MR1624936

Hristache, M., Juditsky, A. and Spokoiny, V. (2001). Direct estimation of the index coefficient in a single-index model. *Ann. Statist.*, **29** 595–623. MR1865333

Juditsky, A. and Nemirovski, A. (2000). Functional aggregation for nonparametric estimation. *Ann. Statist.*, **28(3)** 681–712. MR1792783

Juditsky, A., Rigollet, P. and Tsybakov, A. (2005a). Learning by mirror averaging. URL `http://arxiv.org/abs/math/0511468`.

Juditsky, A. B., Nazin, A. V., Tsybakov, A. B. and Vayatis, N. (2005b). Recursive aggregation of estimators by the mirror descent method with averaging. *Problemy Peredachi Informatsii*, **41** 78–96. MR2198228

Lecué, G. (2007). Suboptimality of penalized empirical risk minimization in classification. 20th Annual Conference On Learning Theory, COLT07, Springer, 142–156.

Leung, G. and Barron, A. R. (2006). Information theory and mixing least-squares regressions. *IEEE Trans. Inform. Theory*, **52** 3396–3410. MR2242356

Nemirovski, A. (2000). *Topics in Non-parametric Statistics*, vol. 1738 of *Ecole d'été de Probabilités de Saint-Flour 1998, Lecture Notes in Mathematics*. Springer, N.Y. MR1775640

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, **10** 1040–1053. MR0673642

Stute, W. and Zhu, L.-X. (2005). Nonparametric checks for single-index models. *Ann. Statist.*, **33** 1048–1083. MR2195628

Tsybakov, A. (2003a). *Introduction l'estimation non-paramerique*. Springer.

Tsybakov, A. B. (2003b). Optimal rates of aggregation. *Computational Learning Theory and Kernel Machines. B.Schölkopf and M.Warmuth, eds. Lecture Notes in Artificial Intelligence*, **2777** 303–313. Springer, Heidelberg.

Vapnik, V. N. (1998). *Statistical learning theory*. Adaptive and Learning Systems for Signal Processing, Communications, and Control, John Wiley & Sons Inc., New York. , A Wiley-Interscience Publication. MR1641250

Xia, Y. and Härdle, W. (2006). Semi-parametric estimation of partially linear single-index models 1162–1184. MR2276153

Yang, Y. (2000a). Combining different procedures for adaptive regression. *J. Multivariate Anal.*, **74** 135–161. MR1790617

Yang, Y. (2000b). Mixing strategies for density estimation. *Ann. Statist.*, **28** 75–87. MR1762904

Yang, Y. (2004). Aggregating regression procedures to improve performance. *Bernoulli*, **10** 25–47. MR2044592

Yang, Y. and Barron, A. (1999). Information-theoretic determination of minimax rates of convergence. *Ann. Statist.*, **27** 1564–1599. MR1742500