Sparse Single-Index Model

Pierre Alquier[1]

LPMA[2]
Université Paris Diderot – Paris VII
Boîte 188, 175 rue du Chevaleret
75013 Paris, France

CREST-LS
3 avenue Pierre Larousse
92240 Malakoff, France

alquier@math.jussieu.fr

Gérard Biau

LSTA & LPMA[3]
Université Pierre et Marie Curie – Paris VI
Boîte 158, Tour 15-25, 2ème étage
4 place Jussieu, 75252 Paris Cedex 05, France

DMA[4]
Ecole Normale Supérieure
45 rue d'Ulm
75230 Paris Cedex 05, France

gerard.biau@upmc.fr

Abstract

The single-index model is known to offer a flexible way to model a variety of high-dimensional real-world phenomena. However, despite its relative simplicity, this dimension reduction scheme is faced with severe complications as soon as the underlying dimension becomes larger than the number of observations ("p larger than n" paradigm). To circumvent this difficulty, we consider the single-index model estimation problem from a sparsity perspective using a PAC-Bayesian

[1]Corresponding author.

[2]Research partially supported by the French "Agence Nationale pour la Recherche" under grant ANR-09-BLAN-0128 "PARCIMONIE".

[3]Research partially supported by the French "Agence Nationale pour la Recherche" under grant ANR-09-BLAN-0051-02 "CLARA".

[4]Research carried out within the INRIA project "CLASSIC" hosted by Ecole Normale Supérieure and CNRS.

approach. On the theoretical side, we offer a sharp oracle inequality, which is more powerful than the best known oracle inequality for other common procedures of single-index recovery. The proposed method is implemented by means of the reversible jump Markov chain Monte Carlo technique and its performance is compared with that of standard procedures.

Index Terms — Single-index model, sparsity, regression estimation, PAC-Bayesian, oracle inequality, reversible jump Markov chain Monte Carlo method.

2010 Mathematics Subject Classification: 62G08, 62G05, 62G20.

1 Introduction

Let $\mathcal{D}_n = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$ be a collection of independent observations, distributed as a generic independent pair $(\mathbf{X}, Y)$ taking values in $\mathbb{R}^p \times \mathbb{R}$ and satisfying $\mathbb{E}|Y| < \infty$. Throughout, we let $\mathbf{P}$ be the distribution of the pair $(\mathbf{X}, Y)$ and note that the sample $\mathcal{D}_n$ is distributed according to $\mathbf{P}^{\otimes n}$. In the regression function estimation problem, the mission is to use the data $\mathcal{D}_n$ in order to construct an estimate $r_n : \mathbb{R}^p \to \mathbb{R}$ of the regression function $r(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$. In the classical parametric linear model, one assumes

$$Y = \theta^{\star T}\mathbf{X} + \mathbf{W},$$

where $\theta^\star = (\theta_1^\star, \ldots, \theta_p^\star)^T \in \mathbb{R}^p$ and $\mathbb{E}[\mathbf{W}|\mathbf{X}] = 0$. Here

$$r(\mathbf{x}) = \theta^{\star T}\mathbf{x} = \sum_{j=1}^{p} \theta_j^\star x_j$$

is a linear function of the components of $\mathbf{x} = (x_1, \ldots, x_p)^T$. More generally, we might define

$$Y = f^\star(\theta^{\star T}\mathbf{X}) + \mathbf{W}, \tag{1.1}$$

where $f^\star$ is an unknown univariate measurable function. This is the celebrated single-index model, which is recognized as a particularly useful variation of the linear formulation and can easily be interpreted: the model changes only in the direction $\theta^\star$, and the way it changes in this direction is described by the function $f^\star$. This semiparametric paradigm has applications to a variety of fields, such as discrete choice analysis in econometrics and dose response models in biometrics, where high-dimensional regression models are often employed. There are too many references to be included here, but the monographs of McCullagh and Nelder [30] and Horowitz [24] together with the references [22, 25, 18, 15, 26] will provide the reader with good introductions to the general subject area.

One of the main advantages of the single-index model is its supposed ability to deal with the so-called curse of dimensionality (Bellman [5]), which means that estimating the regression function is especially difficult whenever the dimension p of $\mathbf{X}$ becomes large. As a matter of fact, the optimal mean squared convergence rate $n^{-2k/(2k+p)}$ for the estimation of a k-times differentiable regression function converges to zero dramatically slowly if the dimension p of $\mathbf{X}$ is large compared to k. This leads to an unsatisfactory accuracy of estimation for moderate sample sizes, and the only possibility to circumvent this problem is to impose additional assumptions on the regression function. Thus, in particular, if $r(\mathbf{x}) = f^\star(\theta^{\star T}\mathbf{x})$ holds for every $\mathbf{x} \in \mathbb{R}^p$, then the underlying structural dimension of the model is 1 (instead of p) and the estimation of r can hopefully be performed easier. In this regard, Gaïffas and Lecué show in [19] that the optimal rate of convergence over the single-index model class is $n^{-2k/(2k+1)}$ (instead of $n^{-2k/(2k+p)}$), thereby answering a conjecture of Stone [36].

Nevertheless, practical estimation of the link function $f^\star$ and the index $\theta^\star$ still requires a degree of statistical smoothing. Perhaps the most common approach to reach this goal is to use a nonparametric smoother (for instance, a kernel or a local polynomial method) to construct an approximation $\hat{f}_n$ of $f^\star$, then substitute $\hat{f}_n$ into an empirical version $R_n(\theta)$ of the mean squared error $R(\theta) = \mathbb{E}[Y - f(\theta^T\mathbf{X})]^2$, and finally choose $\hat{\theta}_n$ to minimize $R_n(\theta)$ (see e.g. Härdle, Hall and Ichimura [22] and Delecroix, Hristache and Patilea [18] where the procedure is discussed in detail). The rationale behind this type of two-stage approach is that it produces a root-n consistent estimate of θ, thereby devolving the difficulty to the simpler problem of computing a good estimate for the one-dimensional function $f^\star$. However, the relative simplicity of this strategy is faced with severe difficulties when the dimension p becomes larger than the number of observations n. Indeed, finding a good estimate for $\theta^\star$ necessitates to solve a minimization problem which turns out to be inefficient as soon as p is larger than n. In fact, this drawback considerably reduces the ability of the single-index model to behave as an effective dimension reduction technology.

On the other hand, there is empirical evidence that many signals in high-dimensional spaces admit a sparse representation. As an example, wavelet coefficients of images often exhibit exponential decay, and a relatively small subset of all wavelet coefficients allow for a good approximation of the original image. Such signals have few nonzero coefficients and can therefore be described as sparse in the signal domain (see for instance [7]). Similarly, recent advances in high-throughput technologies such as microarrays and aCGH arrays indicate that, despite the huge dimensionality of problems, only a small number of genes may play a role in determining the outcome and be required to create good predictors ([39] for instance). Sparse estimation is playing an increasingly important role in the statistics and machine learning communities, and several methods have recently been developed in both fields, which rely upon the notion of sparsity (e.g. penalty methods like the Lasso and Dantzig selector, see [37, 9, 8, 6, 13] and the references therein).

In the present document, we consider the single-index model (1.1) from a sparsity perspective, i.e., we assume that $\theta^\star$ has only a few coordinates different from 0. In the dimension reduction scenario we have in mind, the ambient dimension p can be very large, much larger than the sample size n, but we believe that the representation is sparse, i.e., that very few coordinates of $\theta^\star$ are nonzero. This assumption is helpful at least for two reasons: if p is large and the number of nonzero coordinates is small enough, then the model is easier to interpret and its efficient estimation becomes possible. Our modus operandi will rely on the so-called PAC-Bayesian approach, originally developed in the classification context by Shawe-Taylor and Williamson [35], McAllester [29] and Catoni [10, 11]. This strategy was further investigated for regression by Audibert [4] and Alquier [1] and, more recently, worked out in the sparsity framework by Dalalyan and Tsybakov [16, 17] and Alquier and Lounici [2]. The main message of [16, 17, 2] is that aggregation with a properly chosen prior is able to deal nicely with the sparsity issue. Contrary to procedures such as the Lasso, the Dantzig selector and other penalized least square methods, which are provably consistent under rather restrictive assumptions on the Gram matrix associated to the predictors, PAC-Bayesian aggregation requires only minimal assumptions on the model. Besides, it is computationally feasible even for a large p and exhibits good statistical performance.

The paper is organized as follows. In Section 2, we first set out some notation and introduce the single-index estimation procedure. Then we state our main result (Theorem 2.1), which offers a sparsity oracle inequality more powerful than the best known oracle inequality for other common procedures of single-index recovery. Section 3 is devoted to the practical implementation of the estimate via a reversible jump Markov chain Monte Carlo (MCMC) algorithm, and to numerical experiments on both simulated and real-life data sets. In order to preserve clarity, proofs have been postponed to Section 4 and the description of the MCMC method in its full length is given in the Appendix Section 5.

Note finally that our techniques extend to the case of multiple-index models, of the form

$$Y = f^\star(\theta_1^{\star T}\mathbf{X}, \ldots, \theta_m^{\star T}\mathbf{X}) + \mathbf{W},$$

where the underlying structural dimension m is supposed to be larger than 1 but substantially smaller than p. However, to keep things simple, we let $m = 1$ and leave the reader the opportunity to adapt the results to the more general situation $m \geq 1$.

2 Sparse single-index estimation

We start this section with some notation and basic requirements.

2.1 Notation

Throughout the document, we assume that the recorded data $\mathcal{D}_n$ is generated according to the single-index model (1.1). More precisely, for each $i = 1, \ldots, n$,

$$Y_i = f^\star(\theta^{\star T}\mathbf{X}_i) + \mathbf{W}_i,$$

where $f^\star$ is a univariate measurable function, $\theta^\star$ is a p-variate vector, and $\mathbf{W}_1, \ldots, \mathbf{W}_n$ are independent real random variables satisfying $\mathbb{E}[\mathbf{W}_i|\mathbf{X}_i] = 0$. Within this framework, $\mathbb{E}[\mathbf{W}_i] = 0$. However, the distribution of $\mathbf{W}_i$ (in particular, the variance) may depend on $\mathbf{X}_i$. We shall not precisely specify this dependence, and will rather require the following condition on the common distribution of the $\mathbf{W}_i$.

Assumption N. There exist two positive constants σ and L such that, for all integers $k \geq 2$,

$$\mathbb{E}\left[|\mathbf{W}|^k \,|\, \mathbf{X}\right] \leq \frac{k!}{2}\sigma^2 L^{k-2}.$$

Observe that Assumption **N** holds if $\mathbf{W} = \sigma(\mathbf{X})\varepsilon$, where ε is a standard Gaussian random variable independent of $\mathbf{X}$ and $\sigma(\mathbf{X})$ is almost surely bounded.

Throughout the document, we will require that the random variable $\mathbf{X}$ is almost surely bounded by a constant which, without loss of generality, can be taken equal to 1. Moreover, it will also be assumed that the link function $f^\star$ is bounded by some known positive constant C. Thus, denoting by $\|\mathbf{X}\|_\infty$ the supremum norm of $\mathbf{X}$ and by $\|f^\star\|_\infty$ the functional supremum norm of $f^\star$, we set

Assumption B. The condition $\|\mathbf{X}\|_\infty \leq 1$ holds almost surely and there exists a positive constant C larger than 1 such that $\|f^\star\|_\infty \leq C$.

Let $\|\theta\|_1$ denote the ℓ_1-norm of the vector $\theta = (\theta_1, \ldots, \theta_p)^T$, i.e., $\|\theta\|_1 = \sum_{j=1}^p |\theta_j|$. Without loss of generality, it will be assumed throughout the document that the index $\theta^\star$ belongs to $\mathcal{S}^p_{1,+}$, where $\mathcal{S}^p_{1,+}$ is the set of all $\theta \in \mathbb{R}^p$ with $\|\theta\|_1 = 1$ and $\theta_{j(\theta)} > 0$, where $j(\theta)$ is the smallest $j \in \{1, \ldots, p\}$ such that $\theta_j \neq 0$.

In order to approximate the link function $f^\star$, we shall use the vector space $\mathcal{F}$ spanned by a given countable dictionary of measurable functions $\{\varphi_j\}_{j=1}^\infty$. Put differently, the approximation space $\mathcal{F}$ is the set of (finite) linear combinations of functions of the dictionary. Each φ_j of the collection is assumed to be defined on $[-1, 1]$ and take values in $[-1, 1]$. To avoid getting into too much technicalities, we will also assume that each φ_j is differentiable and such that, for some positive constant ℓ, $\|\varphi_j'\|_\infty \leq \ell j$. This assumption is typically satisfied by the (non-normalized) trigonometric system

$$\varphi_1(t) = 1,\ \varphi_{2j}(t) = \cos(\pi j t),\ \varphi_{2j+1}(t) = \sin(\pi j t), \quad j = 1, 2, \ldots$$

Finally, for any $f : \mathbb{R}^p \to \mathbb{R}$ and $\theta \in \mathcal{S}^p_{1,+}$, we let

$$R(\theta, f) = \mathbb{E}\left[\left(Y - f(\theta^T\mathbf{X})\right)^2\right]$$

and denote by

$$R_n(\theta, f) = \frac{1}{n}\sum_{i=1}^n \left(Y_i - f(\theta^T\mathbf{X}_i)\right)^2$$

the empirical counterpart of $R(\theta, f)$ based on the sample $\mathcal{D}_n$.

2.2 Estimation procedure

We are now in a position to describe our estimation procedure. The method which is presented here is a version of the Gibbs estimate introduced by Catoni in [10, 11]. The approach strongly relies on the choice of a probability measure π on $\mathcal{S}^p_{1,+} \times \mathcal{F}$, called the prior, which in our framework should enforce the sparsity properties of the target regression function. With this objective in mind, we first let

$$\mathrm{d}\pi(\theta, f) = \mathrm{d}\mu(\theta)\mathrm{d}\nu(f),$$

i.e., we assume implicitly that the distribution over the indexes is independent of the distribution over the link functions. With respect to the parameter θ, we put

$$\mathrm{d}\mu(\theta) = \frac{\displaystyle\sum_{i=1}^{p} 2^{-i} \sum_{I\subset\{1,\ldots,p\},|I|=i} \binom{p}{i}^{-1} \mathrm{d}\mu_I(\theta)}{1-(\frac{1}{2})^p}, \tag{2.1}$$

where $|I|$ denotes the cardinality of I and $\mathrm{d}\mu_I(\theta)$ is the uniform probability measure on the set

$$\mathcal{S}_{1,+}^p(I) = \{\theta = (\theta_1,\ldots,\theta_p) \in \mathcal{S}_{1,+}^p : \theta_j = 0 \text{ if and only if } j \notin I\}.$$

We see that $\mathcal{S}_{1,+}^p(I)$ may be interpreted as the set of active coordinates in the single-index regression of Y on $\mathbf{X}$, and note that the prior on $\mathcal{S}_{1,+}^p$ is a convex combination of uniform probability measures on the subsets $\mathcal{S}_{1,+}^p(I)$. The weights of this combination depend only on the size of the active coordinate subset I. As such, the value $|I|$ characterizes the sparsity of the model: the smaller $|I|$, the sparser $r(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = x]$. Thus, the larger the number of active coordinates, the fewer the magnitude of the prior, in accordance with the sparsity idea.

The choice of the prior $\mathrm{d}\nu(f)$ on $\mathcal{F}$ is more involved. To begin with, we define, for any positive integer $M \le n$ and all $r > 0$,

$$\mathcal{B}_M(r) = \left\{(\beta_1,\ldots,\beta_M) \in \mathbb{R}^M : \sum_{j=1}^{M} j|\beta_j| \le r \text{ and } \beta_M \neq 0\right\}.$$

Next, we let, for each positive integer $M \le n$, $\mathcal{F}_M(C) \subset \mathcal{F}$ be the image of $\mathcal{B}_M(C+1)$ by the map

$$\begin{array}{rlcl} \Phi_M & : \mathbb{R}^M & \to & \mathcal{F} \\ & (\beta_1,\ldots,\beta_M) & \mapsto & \sum_{j=1}^M \beta_j\varphi_j. \end{array}$$

Finally, we define $\nu_M(\mathrm{d}f)$ on the set $\mathcal{F}_M(C)$ as the image of the uniform measure on $\mathcal{B}_M(C+1)$ induced by the map Φ_M, and take

$$\mathrm{d}\nu(f) = \frac{\displaystyle\sum_{M=1}^{n} 10^{-M}\mathrm{d}\nu_M(f)}{1-(\frac{1}{10})^n}. \tag{2.2}$$

Some comments are in order here. At first, we note once and for all that the prior π is defined on $\mathcal{S}_{1,+}^p \times \mathcal{F}_n(C)$ endowed with its canonical Borel σ-field. The integer M should be interpreted as a measure of the "dimension" of the function f—the larger M, the more complex the function—and the prior $\mathrm{d}\nu(f)$ adapts again to the sparsity idea by penalizing large-dimensional functions f. The coefficients 2^{-i} and 10^{-M} which appear in (2.1) and (2.2) show that the influence of the cardinality of I (that is, the sparsity) decreases geometrically with $|I|$. Note however that the values 2 and 10, which have been chosen because of their good practical results, are somehow arbitrary. They could be, in all generality, replaced by more general coefficients α and β at the price of a more technical analysis. Finally, we observe that, for each $f = \sum_{j=1}^{M} \beta_j \varphi_j \in \mathcal{F}_M(C)$,

$$\|f\|_\infty \leq \sum_{j=1}^{M} |\beta_j| \leq C + 1.$$

Now, let λ be a positive real number, called the temperature parameter hereafter. The estimates $\hat{\theta}_\lambda$ and $\hat{f}_\lambda$ of θ and f, respectively, are simply obtained by randomly drawing

$$(\hat{\theta}_\lambda, \hat{f}_\lambda) \sim \hat{\rho}_\lambda,$$

where $\hat{\rho}_\lambda$ is the so-called Gibbs posterior distribution defined by

$$\frac{\mathrm{d}\hat{\rho}_\lambda}{\mathrm{d}\pi}(\theta, f) = \frac{\exp\left[-\lambda R_n(\theta, f)\right]}{\displaystyle\int \exp\left[-\lambda R_n(\theta, f)\right] \mathrm{d}\pi(\theta, f)}.$$

[The notation $\mathrm{d}\hat{\rho}_\lambda / \mathrm{d}\pi$ means the density of $\hat{\rho}_\lambda$ with respect to π.] The estimate $(\hat{\theta}_\lambda, \hat{f}_\lambda)$ has a simple interpretation. Firstly, the level of significance of each pair (θ, f) is assessed via its least square error performance on the data $\mathcal{D}_n$. Secondly, a Gibbs distribution with respect to the prior π enforcing those pairs (θ, f) with the most empirical significance is assigned on the space $\mathcal{S}_{1,+}^p \times \mathcal{F}_n(C)$. Finally, the resulting estimate is just a random realization (conditional to the data) of this Gibbs posterior distribution.

2.3 Sparsity oracle inequality

For any $I \subset \{1, \ldots, p\}$ and any positive integer $M \leq n$, we set

$$\left(\theta_{I,M}^\star, f_{I,M}^\star\right) \in \arg \min_{(\theta,f) \in \mathcal{S}_{1,+}^p(I) \times \mathcal{F}_M(C)} R(\theta, f).$$

The main result of the paper is the following theorem. Here and everywhere, the wording "with probability $1 - \varepsilon$" means the probability evaluated with respect to the distribution $\mathbf{P}^{\otimes n}$ of the data $\mathcal{D}_n$ *and* the conditional probability measure $\hat{\rho}_\lambda$.

Theorem 2.1 *Assume that Assumption* **N** *and Assumption* **B** *hold. Set* $w = 8(2C+1)\max[L, 2C+1]$ *and take*

$$\lambda = \frac{n}{w + 2\left[(2C+1)^2 + 4\sigma^2\right]}. \tag{2.3}$$

Then, for all $\varepsilon \in]0,1[$*, with probability at least* $1-2\varepsilon$ *we have*

$$R(\hat{\theta}_\lambda, \hat{f}_\lambda) - R(\theta^\star, f^\star) \le \mathcal{C} \inf_{\substack{I \subset \{1,\ldots,p\} \\ 1 \le M \le n}} \left\{ R(\theta^\star_{I,M}, f^\star_{I,M}) - R(\theta^\star, f^\star) \right.$$
$$\left. + \frac{M \log(Cn) + |I| \log\left(\frac{pn}{|I|}\right) + \log\left(\frac{1}{\varepsilon}\right)}{n} \right\},$$

where $\mathcal{C}$ *is a positive constant function of* L*,* C*,* σ *and* ℓ *only.*

Theorem 2.1 can be simply interpreted. Indeed, we see that if there is a "small" I and a "small" M such that $R(\theta^\star_{I,M}, f^\star_{I,M})$ is close to $R(\theta^\star, f^\star)$, then $R(\hat{\theta}_\lambda, \hat{f}_\lambda)$ is also close to $R(\theta^\star, f^\star)$ up to terms of the order $1/n$. However, if no such I or M exists, then one of the terms $M\log(Cn)/n$ and $|I|\log(pn/|I|)/n$ starts to dominate, thereby deteriorating the general quality of the bound. A good approximation with a "small" I is typically possible when $\theta^\star$ is sparse or, at least, when it can be approximated by a sparse parameter. On the other hand, a good approximation with a "small" M is possible if $f^\star$ has a sufficient degree of regularity.

To illustrate the latter remark, assume for instance that $\{\varphi_j\}_{j=1}^\infty$ is the (non-normalized) trigonometric system and suppose that the target $f^\star$ belongs to the Sobolev ellipsoid

$$\mathcal{W}\left(k, \frac{6(C+1)^2}{\pi^2}\right)$$
$$= \left\{ f \in L_2([-1,1]) : f = \sum_{j=1}^{\infty} \beta_j \varphi_j \text{ and } \sum_{j=1}^{\infty} j^{2k} \beta_j^2 \le \frac{6(C+1)^2}{\pi^2} \right\}$$

for some unknown regularity parameter $k \ge 2$ (see, e.g., Tsybakov [38]). Observe that, in this context, the approximation sets $\mathcal{F}_M(C)$ take the form

$$\mathcal{F}_M(C)$$
$$= \left\{ f \in L_2([-1,1]) : f = \sum_{j=1}^{M} \beta_j \varphi_j, \sum_{j=1}^{M} j|\beta_j| \le C+1 \text{ and } \beta_M \ne 0 \right\}.$$

We insist on the fact that the regularity parameter k is assumed to be unknown, and this casts our results in the so-called adaptive setting. The following additional assumption will be needed:

Assumption D. The random variable $\theta^{\star T}\mathbf{X}$ has a probability density on $[-1, 1]$, bounded by a positive constant B.

Last, we let $I^\star$ be the set I such that $\theta^\star \in \mathcal{S}^p_{1,+}(I)$ and set $\|\theta^\star\|_0 = |I^\star|$.

Corollary 2.1 *Assume that Assumption* $\mathbf{N}$*, Assumption* $\mathbf{B}$ *and Assumption* $\mathbf{D}$ *hold. Suppose also that* $f^\star$ *belongs to the Sobolev ellipsoid* $\mathcal{W}(k, 6(C+1)^2/\pi^2)$*, where* $k \geq 2$ *is an unknown regularity parameter. Set* $w = 8(2C+1)\max[L, 2C+1]$ *and take* λ *as in (2.3). Then, for all* $\varepsilon \in]0,1[$*, with probability at least* $1-2\varepsilon$ *we have*

$$R(\hat{\theta}_\lambda, \hat{f}_\lambda) - R(\theta^\star, f^\star) \leq \mathcal{C}\left\{\left(\frac{\log(Cn)}{n}\right)^{\frac{2k}{2k+1}} + \frac{\|\theta^\star\|_0 \log\left(\frac{pn}{\|\theta^\star\|_0}\right)}{n} + \frac{\log\left(\frac{1}{\varepsilon}\right)}{n}\right\}, \tag{2.4}$$

where $\mathcal{C}$ *is a positive constant function of* L*,* C*,* σ*,* ℓ *and* B *only.*

We see that, asymptotically, the leading term on the right-hand side of inequality (2.4) is $(\log(n)/n)^{\frac{2k}{2k+1}}$. This is the minimax rate of convergence over a Sobolev class, up to a $\log(n)$. However, when n is "small" and $\theta^\star$ is not sparse (i.e., $\|\theta^\star\|_0$ is not "small"), the term $\|\theta^\star\|_0 \log(pn/\|\theta^\star\|_0)/n$ starts to emerge and cannot be neglected. Put differently, in large dimension, the estimation of $\theta^\star$ itself is a problem—this phenomenon is not taken into account by asymptotic studies.

Finally, it should be stressed that the choice of λ in Theorem 4.3 and Corollary 2.1 is not the best possible and may eventually be improved, at the price of a more technical analysis however.

3 Implementation and numerical results

A series of experiments were conducted, both on simulated and real-life data sets, in order to assess the practical capabilities of the proposed method and compare its performance with that of standard procedures. Prior to analysis, we first need to discuss its concrete implementation, which has been carried out via a Markov Chain Monte Carlo (MCMC) method.

3.1 Implementation via reversible jump MCMC

The use of MCMC methods has become a popular way to compute Bayesian estimates. For an introduction to the domain, one should refer to the comprehensive monograph of Marin and Robert [27] and the references therein. Importantly, in this computational framework, an adaptation of the well-known Hastings-Metropolis algorithm to the case where the posterior distribution gives mass to several models of different dimensions was proposed by Green [20] under the name Reversible Jump MCMC (RJMCMC) method. In the PAC-Bayesian setting, MCMC procedures were first considered by Catoni [10], whereas Dalalyan and Tsybakov [16, 17] and Alquier and Lounici [2] explore their practical implementation in the sparse context using Langevin Monte Carlo and RJMCMC, respectively. Regarding the single-index model, MCMC algorithms were used to compute Bayesian estimates by Antoniadis, Grégoire and McKeague [3] and, more recently, by Wang [40], who develop a fully Bayesian method to analyse the single-index model. Our implementation technique is close in spirit to the one of Wang [40].

As a starting point for the approximate computation of our estimate, we used the RJMCMC method of Green [20], which is in fact an adaptation of the Hastings-Metropolis algorithm to the case where the objective posterior probability distribution (here, $\hat{\rho}_\lambda$) assigns mass to several different models. The idea is to start from an initial given pair $(\theta^{(0)}, f^{(0)}) \in \mathcal{S}_{1,+}^p \times \mathcal{F}_n(C)$ and then, at each step, to iteratively compute $(\theta^{(t+1)}, f^{(t+1)})$ from $(\theta^{(t)}, f^{(t)})$ via the following chain of rules:

- Sample a random pair $(\tau^{(t)}, h^{(t)})$ according to some proposal conditional density $k_t(\,.\,|(\theta^{(t)}, f^{(t)}))$ with respect to the prior π;
- Take
$$(\theta^{(t+1)}, f^{(t+1)}) = \begin{cases} (\tau^{(t)}, h^{(t)}) & \text{with probability } \alpha_t \\ (\theta^{(t)}, f^{(t)}) & \text{with probability } 1-\alpha_t, \end{cases}$$
  where
$$\alpha_t = \min\left(1, \frac{\frac{\mathrm{d}\hat{\rho}_\lambda}{\mathrm{d}\pi}(\tau^{(t)}, h^{(t)}) \times k_t\left((\theta^{(t)}, f^{(t)})|(\tau^{(t)}, h^{(t)})\right)}{\frac{\mathrm{d}\hat{\rho}_\lambda}{\mathrm{d}\pi}(\theta^{(t)}, f^{(t)}) \times k_t\left((\tau^{(t)}, h^{(t)})|(\theta^{(t)}, f^{(t)})\right)}\right).$$

This protocol ensures that the sequence $\{(\theta^{(t)}, f^{(t)})\}_{t=0}^{\infty}$ is a Markov chain with invariant probability distribution $\hat{\rho}_\lambda$ (see e.g. Marin and Robert [27]). A usual choice is to take $k_t \equiv k$, so that the Markov chain is homogeneous. However, in our context, it turns out to be more convenient to let $k_t = k_1$

if t is odd and $k_t = k_2$ if t is even. Roughly, k_1 is akin to modify the index $\theta^{(t)}$ while k_2 will essentially act on the link function $f^{(t)}$. While the ideas underlying the proposal densities k_1 and k_2 are quite simple, a precise description in its full length turns out to be more technical. Thus, in order to preserve the readability of the paper, the explicit construction of k_1 and k_2 has been postponed to the Appendix Section 5.

3.2 Simulation study

In this subsection, we illustrate the finite-sample performance of the presented estimation method on three synthetic data sets and compare its predictive capabilities with those of three standard statistical procedures. In all our experiments, we took as dictionary the (non-normalized) trigonometric system $\{\varphi_j\}_{j=1}^{\infty}$ and denote accordingly the resulting regression function estimate defined in Section 2 by $\hat{F}_{\text{Fourier}}$. In accordance with the order of magnitude indicated by the theoretical results, we set $\lambda = 4n$. This choice can undoubtedly be improved a bit but, as the numerical results show, it seems sufficient for our procedure to be fairly competitive.

The tested competing methods are the Lasso (Tibshirani [37]), the standard regression kernel estimate (Nadaraya [31, 32] and Watson [41], see also Tsybakov [38]), and the estimation strategy discussed in Härdle, Hall and Ichimura [22]. While the procedure of Härdle, Hall and Ichimura is specifically tailored for single-index models, the Lasso is designed to deal with the estimation of sparse linear models. On the other hand, the nonparametric kernel method is one of the best options when no obvious assumption (such as the single-index one) can be made on the shape of the targeted regression function.

We briefly recall that, for a linear model of form $Y = \theta^{\star T}\mathbf{X} + \mathbf{W}$, the Lasso estimate takes the form $\hat{F}_{\text{Lasso}}(\mathbf{x}) = \hat{\theta}_{\text{Lasso}}^T\mathbf{x}$, where

$$\hat{\theta}_{\text{Lasso}} \in \arg\min_{\theta\in\mathbb{R}^p} \left\{ \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \theta^T\mathbf{X}_i\right)^2 + \ell\sum_{j=1}^{p}|b_j| \right\}$$

and $\ell > 0$ is a regularization parameter. Theoretical results (see e.g. Bunea, Wegkamp and Tsybakov [8]) indicate that ℓ should be of the order $\ell^\star = \sigma\sqrt{\log(p)/n}$. Throughout, σ is assumed to be known, and we let $\ell = \ell^\star/3$, since this choice is known to give good practical results. The Nadaraya-Watson kernel estimate will be denoted by $\hat{F}_{\text{NW}}$. It is defined as

$$\hat{F}_{\text{NW}}(\mathbf{x}) = \frac{\sum_{i=1}^{n} Y_i K_h(\mathbf{x} - \mathbf{X}_i)}{\sum_{i=1}^{n} K_h(\mathbf{x} - \mathbf{X}_i)}$$

for some nonnegative kernel K on $\mathbb{R}^p$ and $K_h(\mathbf{z}) = K(\mathbf{z}/h)/h$. In the experiments, we let K be the Gaussian kernel $K(\mathbf{z}) = \exp(-\mathbf{z}^T\mathbf{z})$ and chose the smoothing parameter h via a classical leave-one-out procedure on the grid $\mathcal{G} = \{0.75^k, k = 0, \ldots, \lfloor \log(n) \rfloor\}$, see, e.g., Györfi, Kohler, Krzyżak and Walk [21] (notation $\lfloor . \rfloor$ stands for the floor function). Finally, the estimation procedure advocated in Härdle, Hall and Ichimura [22] takes the form

$$\hat{F}_{\mathrm{HHI}}(\mathbf{x}) = \frac{\sum_{i=1}^n Y_i G_{\hat{h}} \left(\hat{\theta}_{\mathrm{HHI}}^T (\mathbf{x} - \mathbf{X}_i) \right)}{\sum_{i=1}^n G_{\hat{h}} \left(\hat{\theta}_{\mathrm{HHI}}^T (\mathbf{x} - \mathbf{X}_i) \right)}$$

for some kernel G on $\mathbb{R}$, with $G_h(\mathbf{z}) = G(\mathbf{z}/h)/h$ and

$$\left(\hat{h}, \hat{\theta}_{\mathrm{HHI}} \right) \in \arg \min_{h>0, \theta \in \mathbb{R}^p} \sum_{i=1}^n \left[Y_i - \frac{\sum_{j \neq i} Y_j G_h \left(\theta^T (\mathbf{X}_j - \mathbf{X}_i) \right)}{\sum_{j \neq i} G_h \left(\theta^T (\mathbf{X}_j - \mathbf{X}_i) \right)} \right]^2 .$$

All calculations were performed with the Gaussian kernel. We used the grid $\mathcal{G}$ for the optimization with respect to h, whereas the best search for θ was implemented via a pathwise coordinate optimization.

The various methods were tested for the general regression model

$$Y_i = F(\mathbf{X}_i) + \mathbf{W}_i, \quad i = 1, \ldots, n,$$

for three different choices of F (single-index or not) and two values of n, namely $n = 20$ and $n = 100$. In each of these models, the observations $\mathbf{X}_i$ take values in $\mathbb{R}^p$, with $p = 10$, and have independent components uniformly distributed on $[-1, 1]$. The noise variables $\mathbf{W}_1, \ldots, \mathbf{W}_n$ are independently distributed according to a Gaussian $\mathcal{N}(0, \sigma^2)$, with $\sigma = 0.2$. It is worth pointing out that, for $n = 20$, p and n are of the same order, which means that the setting is nonasymptotic. It is essentially in this case that the use of sparse estimates, which reduce the variance, is expected to improve the performance over generalist methods. On the other hand, the situation $n = 100$ and $p = 10$ is less difficult and mimics the asymptotic setting.

The three examined functions $F(\mathbf{x})$, for $\mathbf{x} = (x_1, \ldots, x_{10})$, were the following ones:

[**Model 1**] A linear model $F_{\mathrm{Linear}}(\mathbf{x}) = 2\theta^{\star T}\mathbf{x}$.

[**Model 2**] A single-index function $F_{\mathrm{SI}}(\mathbf{x}) = 2(\theta^{\star T}\mathbf{x})^2 + \theta^{\star T}\mathbf{x}$.

[**Model 3**] A purely nonparametric model $F_{\mathrm{NP}}(\mathbf{x}) = 2|x_2|\sqrt{|x_1|} - x_3^3$,

$n = 20$		$\hat{F}_{\text{Fourier}}$	$\hat{F}_{\text{HHI}}$	$\hat{F}_{\text{Lasso}}$	$\hat{F}_{\text{NW}}$
F_{Linear}	median	0.091	0.297	**0.051**	0.354
	mean	0.104	0.320	**0.056**	0.400
	s.d.	0.083	0.129	0.028	178
F_{SI}	median	**0.083**	0.268	0.388	0.287
	mean	**0.264**	0.298	0.508	0.296
	s.d.	0.408	0.158	0.361	0.158
F_{NP}	median	0.503	0.406	1.480	**0.378**
	mean	0.576	0.420	1.389	**0.376**
	s.d.	0.318	0.180	0.471	0.126
$n = 100$		$\hat{F}_{\text{Fourier}}$	$\hat{F}_{\text{HHI}}$	$\hat{F}_{\text{Lasso}}$	$\hat{F}_{\text{NW}}$
F_{Linear}	median	0.051	0.056	**0.043**	0.129
	mean	0.059	0.058	**0.045**	0.247
	s.d.	0.011	0.011	0.006	0.039
F_{SI}	median	**0.047**	0.054	0.352	0.227
	mean	**0.049**	0.055	0.348	0.230
	s.d.	0.009	0.012	0.075	0.058
F_{NP}	median	0.316	0.356	0.809	**0.314**
	mean	**0.318**	0.360	0.822	0.3284
	s.d.	0.062	0.062	0.122	0.058

Table 1: Numerical results for the simulated data, with $n = 20$ and $n = 100$. The characters in bold indicate the best performance.

where, in the first and second model, $\theta^\star = (0.5, 0.5, 0, \dots, 0)^T$. Thus, in [**Model 1**] and [**Model 2**], even if the ambient dimension is $p = 10$, the intrinsic dimension of the model is in fact equal to 2.

For each experiment, a learning set of size n was generated to compute the estimates and their performance, in terms of mean squared prevision error, was evaluated on a separate test set of the same size. The results are shown in Table 1. As each experiment was repeated 20 times, this table also reports the median, the mean and the standard deviation (s.d.) of the prevision error of each procedure.

Some comments are in order. First, we note without surprise that

1. The Lasso performs well in the linear setting [**Model 1**].
2. The single-index methods F_{Fourier} and F_{HHI} are the best ones when the targeted regression function really involves a single-index model [**Model 2**].

3. The kernel method gives good results in the purely nonparametric setting [**Model 3**].

Interestingly, we note that $\hat{F}_{\text{Fourier}}$ provides slightly better results than the single-index-tailored estimate $\hat{F}_{\text{HHI}}$, especially in the nonasymptotic setting ($n = 20$, $p = 10$). This observation can be easily explained by the fact that $\hat{F}_{\text{HHI}}$ does not integrate any sparsity information regarding the parameter $\theta^{\star}$, whereas $\hat{F}_{\text{Fourier}}$ tries to focus on the dimension of the active coordinates, which is equal to 2 in this simulation.

As an important result, we observe in Table 1 that $\hat{F}_{\text{Fourier}}$ is the most robust of all the tested procedures. In fact, this method provides results in the linear and purely nonparametric cases which are close to the best ones. Consequently, in the next subsection, we decided to give up $\hat{F}_{\text{HHI}}$ and focus on the comparison of $\hat{F}_{\text{Fourier}}$ with $\hat{F}_{\text{Lasso}}$ and $\hat{F}_{\text{NW}}$ on real-life data sets.

3.3 Real data

The real-life data sets used in this second series of experiments are from two different sources. The first one, called **AIR-QUALITY** data ($n = 111$, $p = 3$), has been first used by Chambers, Cleveland, Kleiner and Tukey [12] and has been later considered as a benchmark in the study and comparison of single-index models (see, for example, Antoniadis, Grégoire and McKeague [3] and Wang [40], among others). This data set originated from an environmental study relating $n = 111$ ozone concentration measures to $p = 3$ meteorological variables, namely wind speed, temperature and radiation. The data is available as a package in the software **R** [34], which we employed in all the numerical experiments. The programs are available upon request from the authors.

The second category of data arises from the UC Irvine Machine Learning Repository `http://archive.ics.uci.edu/ml`, where the following packages have been downloaded from:

- **AUTO-MPG** (Quinlan [33], $n = 392$, $p = 7$).
- **CONCRETE** (Yeh [42], $n = 1030$, $p = 8$).
- **HOUSING** (Harrison and Rubinfeld [23], $n = 508$, $p = 13$).
- **SLUMP-1**, **SLUMP-2** and **SLUMP-3**, which correspond to the concrete slump test data introduced by Yeh [43] ($n = 51$, $p = 7$). Since there are 3 different output variables Y in the original data set, we created a single experiment for each of these variables (1 refers to the

output “slump”, 2 to the output “flow” and 3 to the output ’28-day Compressive Strength’).

- **WINE-RED** and **WINE-WHITE** (Cortez, Cerdeira, Almeida, Matos and Reis [14], $n = 1599$, $n = 4898$, $p = 11$).

We refer to the above-mentioned references for a precise description of the meaning of the variables involved in these data sets. For homogeneity reasons, all data were normalized to force the input variables to lie in $[-1, 1]$—in accordance with the setting of our method—and to ensure that all output variables have standard deviation 0.5. In two data-sets (**AIR-QUALITY** and **AUTO-MPG**) there were some missing values and the corresponding observations were simply removed.

For each method and each of the nine data sets, we split the observations in a learning and a test set of equal sizes, computed the estimate on the learning set, evaluated the prediction error on the test set, and repeated this protocol 20 times. The results are summarized in Table 2.

We see that all the tested methods provide reasonable results on most data sets. The Lasso procedure is very competitive, especially in the nonasymptotic framework. In particular, it is interesting to note that, according to these numerical results, there seems to be no particular advantage in using a single-index model in place of a linear one for the **AIR-QUALITY** data, which is a standard benchmark in the single-index literature.

The estimation procedure $\hat{F}_{\text{Fourier}}$ offers the best performance in terms of prediction error in 3 out of 9 experiments (**AUTO-MPG**, **CONCRETE** and **HOUSING**) and provides the worst result in only 1 experiment (**WHINE-WHITE**). Besides, when it is not the best, the method $\hat{F}_{\text{Fourier}}$ is very close to the best one, as for example in **SLUMP-3** and **WINE-RED**. As an illustrative example, the plot of the resulting fit of our procedure to the data set **AUTO-MPG** is shown in Figure 1.

Thus, as a general conclusion to this experimental section, we may say that our PAC-Bayesian oriented procedure has an excellent predictive ability, even in nonasymptotic/high-dimensional situations. It is fast, robust, and exhibits performance at the level of the gold standard Lasso.

Data set		$\hat{F}_{\text{Fourier}}$	$\hat{F}_{\text{Lasso}}$	$\hat{F}_{\text{NW}}$
AIR QUALITY	median	0.146	**0.136**	0.160
$n = 111$	mean	0.150	**0.137**	0.156
$p = 3$	s.d.	0.043	0.033	0.037
AUTO-MPG	median	**0.044**	0.071	0.066
$n = 392$	mean	**0.047**	0.072	0.069
$p = 7$	s.d.	0.010	0.009	0.007
CONCRETE	median	**0.088**	0.109	0.098
$n = 1030$	mean	**0.090**	0.108	0.097
$p = 8$	s.d.	0.008	0.005	0.57
HOUSING	median	**0.076**	0.091	0.088
$n = 508$	mean	**0.074**	0.090	0.094
$p = 13$	s.d.	0.012	0.011	0.022
SLUMP-1	median	0.217	**0.214**	0.225
$n = 51$	mean	0.233	**0.219**	0.228
$p = 7$	s.d.	0.069	0.027	0.030
SLUMP-2	median	0.182	**0.167**	0.223
$n = 51$	mean	0.192	**0.168**	0.215
$p = 7$	s.d.	0.047	0.027	0.026
SLUMP-3	median	0.061	**0.052**	0.131
$n = 51$	mean	0.066	**0.053**	0.136
$p = 7$	s.d.	0.024	0.011	0.026
WINE-RED	median	0.174	**0.171**	0.184
$n = 1599$	mean	0.172	**0.171**	0.183
$p = 11$	s.d.	0.010	0.007	0.009
WINE-WHITE	median	0.191	0.185	**0.184**
$n = 4898$	mean	0.202	0.186	**0.185**
$p = 11$	s.d.	0.045	0.004	0.004

Table 2: Numerical results for the real-life data sets. The characters in bold indicate the best performance.

4 Proofs

4.1 Preliminary results

Throughout this section, we let π be the prior probability measure on $\mathbb{R}^p \times \mathcal{F}_n(C)$ equipped with its canonical Borel σ-field, which is defined in Section 2. Recall that $\mathcal{F}_n(C) \subset \mathcal{F}$ and that, for each $f \in \mathcal{F}_n(C)$, we have $\|f\|_\infty \leq C+1$.

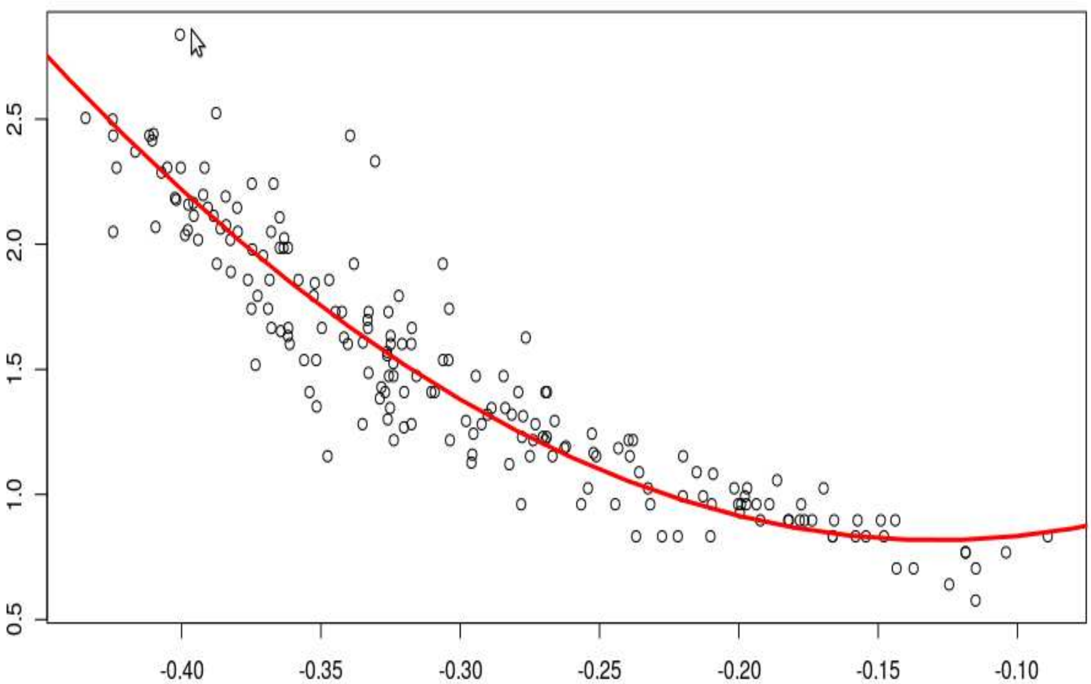

Figure 1: **AUTO-MPG** example: Estimated link function by the method $\hat{F}_{\mathsf{Fourier}}$.

We start with three technical lemmas. Lemma 4.1 is a version of Bernstein's inequality, whose proof can be found in Massart [28, Chapter 2, inequality (2.21)]. Lemma 4.2 is due to Catoni [11, page 4]. For a random variable Z, the notation $(Z)_+$ means the positive part of Z.

Lemma 4.1 *Let $T_1, \ldots, T_n$ be independent real-valued random variables. Assume that there exist two positive constants v and w such that, for all integers $k \geq 2$,*

$$\sum_{i=1}^{n} \mathbb{E}\left[(T_i)_+^k\right] \leq \frac{k!}{2} v w^{k-2}.$$

Then, for any $\zeta \in]0, 1/w[$,

$$\mathbb{E}\left[\exp\left(\zeta \sum_{i=1}^{n} [T_i - \mathbb{E}T_i]\right)\right] \leq \exp\left(\frac{v\zeta^2}{2(1 - w\zeta)}\right).$$

Given a measurable space $(E, \mathcal{E})$ and two probability measures μ_1 and μ_2 on $(E, \mathcal{E})$, we denote by $\mathcal{K}(\mu_1, \mu_2)$ the Kullback-Leibler divergence of μ_1 with respect to μ_2, defined as

$$\mathcal{K}(\mu_1, \mu_2) = \begin{cases} \displaystyle\int \log\left(\frac{\mathrm{d}\mu_1}{\mathrm{d}\mu_2}\right) \mathrm{d}\mu_1 & \text{if } \ \mu_1 \ll \mu_2, \\ \infty & \text{otherwise.} \end{cases}$$

Lemma 4.2 *Let $(E, \mathcal{E})$ be a measurable space. For any probability measure μ on $(E, \mathcal{E})$ and any measurable function $h : E \to \mathbb{R}$ such that $\int(\exp \circ h)\mathrm{d}\mu < \infty$, we have*

$$\log \int (\exp \circ h)\mathrm{d}\mu = \sup_m \left(\int h \mathrm{d}m - \mathcal{K}(m, \mu) \right), \tag{4.1}$$

where the supremum is taken over all probability measures on $(E, \mathcal{E})$ and, by convention, $\infty - \infty = -\infty$. Moreover, as soon as h is upper-bounded on the support of μ, the supremum with respect to m on the right-hand side of (4.1) is reached for the Gibbs distribution g given by

$$\frac{\mathrm{d}g}{\mathrm{d}\mu}(e) = \frac{\exp\left[h(e)\right]}{\displaystyle\int (\exp \circ h)\mathrm{d}\mu}, \quad e \in E.$$

Lemma 4.3 *Assume that Assumption* **N** *holds. Set $w = 8(2C+1)\max[L, 2C+1]$ and take*

$$\lambda \in \left]0, \frac{n}{w + [(2C+1)^2 + 4\sigma^2]}\right[.$$

Then, for all $\varepsilon \in]0,1[$ and any data-dependent probability measure $\hat{\rho}$ absolutely continuous with respect to π we have, with probability at least $1 - \varepsilon$,

$$R(\hat{\theta}, \hat{f}) - R(\theta^\star, f^\star) \leq$$
$$\frac{1}{1 - \frac{\lambda[(2C+1)^2+4\sigma^2]}{n - w\lambda}} \left(R_n(\hat{\theta}, \hat{f}) - R_n(\theta^\star, f^\star) + \frac{\log\left(\frac{\mathrm{d}\hat{\rho}}{\mathrm{d}\pi}(\hat{\theta}, \hat{f})\right) + \log\left(\frac{1}{\varepsilon}\right)}{\lambda} \right),$$

where the pair $(\hat{\theta}, \hat{f})$ is distributed according to $\hat{\rho}$.

Proof of Lemma 4.3. Fix $\theta \in \mathcal{S}^p_{1,+}$ and $f \in \mathcal{F}_n(C)$. The proof starts with an application of Lemma 4.1 to the random variables

$$T_i = -\left(Y_i - f(\theta^T \mathbf{X}_i)\right)^2 + \left(Y_i - f^\star(\theta^{\star T} \mathbf{X}_i)\right)^2, \quad i = 1, \ldots, n.$$

Note that these random variables are independent, identically distributed, and that

$$
\begin{aligned}
&\sum_{i=1}^{n} \mathbb{E} T_i^2 \\
&\quad = \sum_{i=1}^{n} \mathbb{E} \left\{ \left[2Y_i - f(\theta^T \mathbf{X}_i) - f^\star(\theta^{\star T} \mathbf{X}_i) \right]^2 \left[f(\theta^T \mathbf{X}_i) - f^\star(\theta^{\star T} \mathbf{X}_i) \right]^2 \right\} \\
&\quad = \sum_{i=1}^{n} \mathbb{E} \left\{ \left[2\mathbf{W}_i + f^\star(\theta^{\star T} \mathbf{X}_i) - f(\theta^T \mathbf{X}_i) \right]^2 \left[f(\theta^T \mathbf{X}_i) - f^\star(\theta^{\star T} \mathbf{X}_i) \right]^2 \right\} \\
&\quad \le 2 \sum_{i=1}^{n} \mathbb{E} \left\{ \left[4\mathbf{W}_i^2 + (2C+1)^2 \right] \left[f(\theta^T \mathbf{X}_i) - f^\star(\theta^{\star T} \mathbf{X}_i) \right]^2 \right\} \\
&\quad \le 2 \left[(2C+1)^2 + 4\sigma^2 \right] \sum_{i=1}^{n} \mathbb{E} \left\{ \left[f(\theta^T \mathbf{X}_i) - f^\star(\theta^{\star T} \mathbf{X}_i) \right]^2 \right\} \\
&\qquad \text{(by Assumption } \mathbf{N}\text{)} \\
&\quad = v,
\end{aligned}
$$

where we set

$$
v = 2n[(2C+1)^2 + 4\sigma^2] \left[R(\theta, f) - R(\theta^\star, f^\star) \right].
$$

More generally, for all integers $k \ge 3$,

$$
\begin{aligned}
&\sum_{i=1}^{n} \mathbb{E} \left[(T_i)_+^k \right] \\
&\quad \le \sum_{i=1}^{n} \mathbb{E} \left\{ \left| 2Y_i - f(\theta^T \mathbf{X}_i) - f^\star(\theta^{\star T} \mathbf{X}_i) \right|^k \left| f(\theta^T \mathbf{X}_i) - f^\star(\theta^{\star T} \mathbf{X}_i) \right|^k \right\} \\
&\quad = \sum_{i=1}^{n} \mathbb{E} \left\{ \left| 2\mathbf{W}_i + f^\star(\theta^{\star T} \mathbf{X}_i) - f(\theta^T \mathbf{X}_i) \right|^k \left| f(\theta^T \mathbf{X}_i) - f^\star(\theta^{\star T} \mathbf{X}_i) \right|^k \right\} \\
&\quad \le 2^{k-1} \sum_{i=1}^{n} \mathbb{E} \left\{ \left[2^k |\mathbf{W}_i|^k + (2C+1)^k \right] (2C+1)^{k-2} \left| f(\theta^T \mathbf{X}_i) - f^\star(\theta^{\star T} \mathbf{X}_i) \right|^2 \right\}.
\end{aligned}
$$

Therefore, by Assumption **N**,

$$\begin{aligned}
&\sum_{i=1}^{n} \mathbb{E}\left[(T_i)_+^k\right] \\
&\quad \le \sum_{i=1}^{n} \left[2^{2k-2} k! \sigma^2 L^{k-2} + 2^{k-1}(2C+1)^k\right] (2C+1)^{k-2} \left[R(\theta, f) - R(\theta^\star, f^\star)\right] \\
&\quad = v \times \frac{\left[2^{2k-2} k! \sigma^2 L^{k-2} + 2^{k-1}(2C+1)^k\right] (2C+1)^{k-2}}{2[(2C+1)^2 + 4\sigma^2]} \\
&\quad \le v \times \frac{8^{k-2} k! \max\left[L^{k-2}, (2C+1)^{k-2}\right] (2C+1)^{k-2}}{2} \\
&\quad = \frac{k!}{2} v w^{k-2},
\end{aligned}$$

with $w = 8(2C+1)\max[L, 2C+1]$.

Thus, for any temperature parameter $\lambda \in]0, n/w[$, setting $\zeta = \lambda/n$, we may write by Lemma 4.1

$$\begin{aligned}
&\mathbb{E}\Big\{ \exp\left[\lambda \left(R(\theta, f) - R(\theta^\star, f^\star) - R_n(\theta, f) + R_n(\theta^\star, f^\star)\right)\right] \Big\} \\
&\qquad \le \exp\left(\frac{v\lambda^2}{2n^2(1 - \frac{w\lambda}{n})} \right).
\end{aligned}$$

Therefore, using the definition of v, we obtain

$$\begin{aligned}
\mathbb{E}\Bigg\{ \exp\Bigg[&\left(\lambda - \frac{\lambda^2\left[(2C+1)^2 + 4\sigma^2\right]}{n(1 - \frac{w\lambda}{n})} \right) \left(R(\theta, f) - R(\theta^\star, f^\star)\right) \\
&\qquad + \lambda\left(-R_n(\theta, f) + R_n(\theta^\star, f^\star)\right) - \log\left(\frac{1}{\varepsilon}\right) \Bigg] \Bigg\} \le \varepsilon.
\end{aligned}$$

Next, we use a standard PAC-Bayesian approach (Catoni [10, 11], Audibert [4] and Alquier [1]). Let us remind the reader that π is a prior probability measure on the set $\mathcal{S}_{1,+}^p \times \mathcal{F}_n(C)$. We have

$$\begin{aligned}
\int \mathbb{E}\Bigg\{ \exp\Bigg[&\left(\lambda - \frac{\lambda^2\left[(2C+1)^2 + 4\sigma^2\right]}{n(1 - \frac{w\lambda}{n})} \right) \left(R(\theta, f) - R(\theta^\star, f^\star)\right) \\
&\qquad + \lambda\left(-R_n(\theta, f) + R_n(\theta^\star, f^\star)\right) - \log\left(\frac{1}{\varepsilon}\right) \Bigg] \Bigg\} \mathrm{d}\pi(\theta, f) \le \varepsilon
\end{aligned}$$

and consequently, using Fubini's Theorem,

$$\mathbb{E}\left\{\int \exp\left[\left(\lambda - \frac{\lambda^2\left[(2C+1)^2+4\sigma^2\right]}{n(1-\frac{w\lambda}{n})}\right)(R(\theta,f)-R(\theta^\star,f^\star))\right.\right.$$
$$\left.\left.+\lambda\left(-R_n(\theta,f)+R_n(\theta^\star,f^\star)\right)-\log\left(\frac{1}{\varepsilon}\right)\right]\mathrm{d}\pi(\theta,f)\right\}\le\varepsilon.$$

Therefore, for any data-dependent posterior probability measure $\hat{\rho}$ absolutely continuous with respect to π, adopting the convention $\infty\times 0=0$,

$$\mathbb{E}\left\{\int \exp\left[\left(\lambda - \frac{\lambda^2\left[(2C+1)^2+4\sigma^2\right]}{n(1-\frac{w\lambda}{n})}\right)(R(\theta,f)-R(\theta^\star,f^\star))\right.\right.$$
$$+\lambda\left(-R_n(\theta,f)+R_n(\theta^\star,f^\star)\right)$$
$$\left.\left.-\log\left(\frac{\mathrm{d}\hat{\rho}}{\mathrm{d}\pi}(\theta,f)\right)-\log\left(\frac{1}{\varepsilon}\right)\right]\mathrm{d}\hat{\rho}(\theta,f)\right\}$$
$$\le\varepsilon.$$

Recalling that $\mathbf{P}^{\otimes n}$ stands for the distribution of the sample $\mathcal{D}_n$, the latter inequality can be more conveniently written as

$$\mathbb{E}_{\mathcal{D}_n\sim\mathbf{P}^{\otimes n}}\mathbb{E}_{(\hat{\theta},\hat{f})\sim\hat{\rho}}\left\{\exp\left[\left(\lambda - \frac{\lambda^2\left[(2C+1)^2+4\sigma^2\right]}{n(1-\frac{w\lambda}{n})}\right)\left(R(\hat{\theta},\hat{f})-R(\theta^\star,f^\star)\right)\right.\right.$$
$$\left.\left.+\lambda\left(-R_n(\hat{\theta},\hat{f})+R_n(\theta^\star,f^\star)\right)-\log\left(\frac{\mathrm{d}\hat{\rho}}{\mathrm{d}\pi}(\hat{\theta},\hat{f})\right)-\log\left(\frac{1}{\varepsilon}\right)\right]\right\}$$
$$\le\varepsilon.$$

Thus, using the elementary inequality $\exp(\lambda x)\ge\mathbf{1}_{\mathbb{R}_+}(x)$ we obtain, with probability at most ε,

$$\left(1-\frac{\lambda\left[(2C+1)^2+4\sigma^2\right]}{n(1-\frac{w\lambda}{n})}\right)\left(R(\hat{\theta},\hat{f})-R(\theta^\star,f^\star)\right)$$
$$\ge R_n(\hat{\theta},\hat{f})-R_n(\theta^\star,f^\star)+\frac{\log\left(\frac{\mathrm{d}\hat{\rho}}{\mathrm{d}\pi}(\hat{\theta},\hat{f})\right)+\log\left(\frac{1}{\varepsilon}\right)}{\lambda},$$

where the probability is evaluated with respect to the distribution $\mathbf{P}^{\otimes n}$ of the data $\mathcal{D}_n$ *and* the conditional probability measure $\hat{\rho}$. Put differently, letting

$$\lambda\in\left]0,\frac{n}{w+\left[(2C+1)^2+4\sigma^2\right]}\right[,$$

we have, with probability at least $1-\varepsilon$,

$$R(\hat{\theta},\hat{f}) - R(\theta^\star, f^\star) \leq \frac{1}{1-\frac{\lambda[(2C+1)^2+4\sigma^2]}{n-w\lambda}}\left(R_n(\hat{\theta},\hat{f}) - R_n(\theta^\star, f^\star) + \frac{\log\left(\frac{\mathrm{d}\hat{\rho}}{\mathrm{d}\pi}(\hat{\theta},\hat{f})\right)+\log\left(\frac{1}{\varepsilon}\right)}{\lambda}\right).$$

This concludes the proof of Lemma 4.3. ■

4.2 Proof of Theorem 2.1

The proof relies on an application of Lemma 4.3 with $\hat{\rho} = \hat{\rho}_\lambda$ (the Gibbs distribution) as posterior distribution. More precisely, we know that, with probability larger than $1-\varepsilon$,

$$R(\hat{\theta}_\lambda,\hat{f}_\lambda) - R(\theta^\star, f^\star) \leq \frac{1}{1-\frac{\lambda[(2C+1)^2+4\sigma^2]}{n-w\lambda}}\left(R_n(\hat{\theta}_\lambda,\hat{f}_\lambda) - R_n(\theta^\star, f^\star) + \frac{\log\left(\frac{\mathrm{d}\hat{\rho}_\lambda}{\mathrm{d}\pi}(\hat{\theta}_\lambda,\hat{f}_\lambda)\right)+\log\left(\frac{1}{\varepsilon}\right)}{\lambda}\right),$$

where the probability is evaluated with respect to the distribution $\mathbf{P}^{\otimes n}$ of the data $\mathcal{D}_n$ *and* the conditional probability measure $\hat{\rho}_\lambda$. Observe that

$$\begin{aligned}\log\left(\frac{\mathrm{d}\hat{\rho}_\lambda}{\mathrm{d}\pi}(\hat{\theta}_\lambda,\hat{f}_\lambda)\right) &= \log\left(\frac{\exp\left[-\lambda R_n(\hat{\theta}_\lambda,\hat{f}_\lambda)\right]}{\displaystyle\int \exp\left[-\lambda R_n(\theta,f)\right]\mathrm{d}\pi(\theta,f)}\right)\\ &= -\lambda R_n(\hat{\theta}_\lambda,\hat{f}_\lambda) - \log\int \exp\left[-\lambda R_n(\theta,f)\right]\mathrm{d}\pi(\theta,f).\end{aligned}$$

Consequently, with probability at least $1-\varepsilon$,

$$\begin{aligned}&R(\hat{\theta}_\lambda,\hat{f}_\lambda) - R(\theta^\star, f^\star) \leq\\ &\quad\frac{1}{\lambda\left(1-\frac{\lambda[(2C+1)^2+4\sigma^2]}{n-w\lambda}\right)}\Bigg(-\log\int \exp\left[-\lambda R_n(\theta,f)\right]\mathrm{d}\pi(\theta,f)\\ &\qquad\qquad - \lambda R_n(\theta^\star, f^\star) + \log\left(\frac{1}{\varepsilon}\right)\Bigg).\end{aligned}$$

Using Lemma 4.2 we deduce that, with probability at least $1-\varepsilon$,

$$\begin{aligned}
&R(\hat{\theta}_\lambda, \hat{f}_\lambda) - R(\theta^\star, f^\star) \\
&\quad \leq \frac{1}{1-\frac{\lambda[(2C+1)^2+4\sigma^2]}{n-w\lambda}} \inf_{\hat{\rho}} \Bigg\{ \int R_n(\theta, f)\mathrm{d}\hat{\rho}(\theta, f) - R_n(\theta^\star, f^\star) \\
&\qquad\qquad + \frac{\mathcal{K}(\hat{\rho}, \pi) + \log\left(\frac{1}{\varepsilon}\right)}{\lambda} \Bigg\},
\end{aligned}$$

where the infimum is taken over all probability measures on $\mathcal{S}^p_{1,+} \times \mathcal{F}_n(C)$. In particular, letting $\mathcal{M}(I, M)$ be the set of all probability measures on $\mathcal{S}^p_{1,+}(I) \times \mathcal{F}_M(C)$, we have, with probability at least $1-\varepsilon$,

$$\begin{aligned}
&R(\hat{\theta}_\lambda, \hat{f}_\lambda) - R(\theta^\star, f^\star) \\
&\quad \leq \frac{1}{1-\frac{\lambda[(2C+1)^2+4\sigma^2]}{n-w\lambda}} \inf_{\substack{I\subset\{1,\ldots,p\} \\ 1\leq M\leq n}} \inf_{\hat{\rho}\in\mathcal{M}(I,M)} \Bigg\{ \int R_n(\theta, f)\mathrm{d}\hat{\rho}(\theta, f) - R_n(\theta^\star, f^\star) \\
&\qquad\qquad + \frac{\mathcal{K}(\hat{\rho}, \pi) + \log\left(\frac{1}{\varepsilon}\right)}{\lambda} \Bigg\}.
\end{aligned}$$

Next, observe that, for $\hat{\rho} \in \mathcal{M}(I, M)$,

$$\begin{aligned}
\mathcal{K}(\hat{\rho}, \pi) = \mathcal{K}(\hat{\rho}, \mu\otimes\nu) &= \mathcal{K}(\hat{\rho}, \mu_I\otimes\nu_M) + \log\left[\frac{\left(1-\left(\frac{1}{2}\right)^p\right)\left(1-\left(\frac{1}{10}\right)^n\right)\binom{p}{|I|}}{2^{-|I|}10^{-M}}\right] \\
&\leq \mathcal{K}(\hat{\rho}, \mu_I\otimes\nu_M) + \log\left[\frac{\binom{p}{|I|}}{2^{-|I|}10^{-M}}\right].
\end{aligned}$$

Therefore, with probability at least $1-\varepsilon$,

$$\begin{aligned}
&R(\hat{\theta}_\lambda, \hat{f}_\lambda) - R(\theta^\star, f^\star) \\
&\quad \leq \frac{1}{1-\frac{\lambda[(2C+1)^2+4\sigma^2]}{n-w\lambda}} \inf_{\substack{I\subset\{1,\ldots,p\} \\ 1\leq M\leq n}} \inf_{\hat{\rho}\in\mathcal{M}(I,M)} \Bigg\{ \int R_n(\theta, f)\mathrm{d}\hat{\rho}(\theta, f) - R_n(\theta^\star, f^\star) \\
&\qquad + \frac{\mathcal{K}(\hat{\rho}, \mu_I\otimes\nu_M) + \log\left[\frac{\binom{p}{|I|}}{2^{-|I|}10^{-M}}\right] + \log\left(\frac{1}{\varepsilon}\right)}{\lambda} \Bigg\}. \qquad (4.2)
\end{aligned}$$

Applying Lemma 4.1 again with $T_i = (Y_i - f(\theta^T\mathbf{X}_i))^2 - (Y_i - f^\star(\theta^{\star T}\mathbf{X}_i))^2$ and following the same method, we obtain, for any data-dependent distribution

$\hat{\rho} \in \mathcal{M}(I, M)$, with probability at least $1 - \varepsilon$,

$$\begin{aligned}
&\int R_n(\theta, f) \mathrm{d}\hat{\rho}(\theta, f) - R_n(\theta^\star, f^\star) \\
&\quad \le \left(1 + \frac{\lambda\left[(2C+1)^2 + 4\sigma^2\right]}{n - w\lambda}\right) \left(\int R(\theta, f) \mathrm{d}\hat{\rho}(\theta, f) - R(\theta^\star, f^\star)\right) \\
&\qquad + \frac{\mathcal{K}(\hat{\rho}, \mu_I \otimes \nu_M) + \log\left[\frac{\binom{p}{|I|}}{2^{-|I|}10^{-M}}\right] + \log\left(\frac{1}{\varepsilon}\right)}{\lambda}.
\end{aligned} \tag{4.3}$$

Combining inequalities (4.2) and (4.3), we may write, with probability at least $1 - 2\varepsilon$,

$$\begin{aligned}
&R(\hat{\theta}_\lambda, \hat{f}_\lambda) - R(\theta^\star, f^\star) \\
&\quad \le \frac{1}{1 - \frac{\lambda[(2C+1)^2 + 4\sigma^2]}{n - w\lambda}} \inf_{\substack{I \subset \{1, \ldots, p\} \\ 1 \le M \le n}} \inf_{\hat{\rho} \in \mathcal{M}(I,M)} \Bigg\{ \\
&\qquad \left(1 + \frac{\lambda\left[(2C+1)^2 + 4\sigma^2\right]}{n - w\lambda}\right) \left(\int R(\theta, f) \mathrm{d}\hat{\rho}(\theta, f) - R(\theta^\star, f^\star)\right) \\
&\qquad + 2\frac{\mathcal{K}(\hat{\rho}, \mu_I \otimes \nu_M) + \log\left[\frac{\binom{p}{|I|}}{2^{-|I|}10^{-M}}\right] + \log\left(\frac{1}{\varepsilon}\right)}{\lambda} \Bigg\}.
\end{aligned} \tag{4.4}$$

For any subset I of $\{1, \ldots, p\}$, any positive integer $M \le n$ and any $\delta, \gamma \in]0, 1/n]$, let the probability measure $\rho_{I,M,\delta,\gamma}$ be defined as

$$\mathrm{d}\rho_{I,M,\delta,\gamma}(\theta, f) = \mathrm{d}\rho^1_{I,M,\delta}(\theta) \mathrm{d}\rho^2_{I,M,\gamma}(f),$$

with

$$\frac{\mathrm{d}\rho^1_{I,M,\delta}}{\mathrm{d}\mu_I}(\theta) \propto \mathbf{1}_{[\|\theta - \theta^\star_{I,M}\|_1 \le \delta]}$$

and

$$\frac{\mathrm{d}\rho^2_{I,M,\gamma}}{\mathrm{d}\nu_M}(f) \propto \mathbf{1}_{[\|f - f^\star_{I,M}\|_M \le \gamma]}$$

where, for $f = \sum_{j=1}^M \beta_j \varphi_j \in \mathcal{F}_M(C)$, we put

$$\|f\|_M = \sum_{j=1}^M j|\beta_j|.$$

With this notation, inequality (4.4) leads to

$$
\begin{aligned}
&R(\hat{\theta}_\lambda, \hat{f}_\lambda) - R(\theta^\star, f^\star) \\
&\leq \frac{1}{1 - \frac{\lambda[(2C+1)^2+4\sigma^2]}{n-w\lambda}} \inf_{\substack{I \subset \{1,\dots,p\} \\ 1 \leq M \leq n}} \inf_{\delta,\gamma>0} \Bigg\{ \\
&\qquad \left(1 + \frac{\lambda\left[(2C+1)^2+4\sigma^2\right]}{n-w\lambda}\right)\left(\int R(\theta,f)\mathrm{d}\rho_{I,M,\delta,\gamma}(\theta,f) - R(\theta^\star,f^\star)\right) \\
&\qquad + 2\frac{\mathcal{K}(\rho_{I,M,\delta,\gamma}, \mu_I \otimes \nu_M) + \log\left[\frac{\binom{p}{|I|}}{2^{-|I|}10^{-M}}\right] + \log\left(\frac{1}{\varepsilon}\right)}{\lambda}\Bigg\}.
\end{aligned} \tag{4.5}
$$

To finish the proof, we have to control the different terms in (4.5). Note first that

$$
\log\binom{p}{|I|} \leq |I|\log\left(\frac{pe}{|I|}\right)
$$

and, consequently,

$$
\begin{aligned}
\log\left[\frac{\binom{p}{|I|}}{2^{-|I|}10^{-M}}\right] &\leq |I|\log\left(\frac{pe}{|I|}\right) + |I|\log 2 + M\log 10 \\
&= |I|\log\left(\frac{2pe}{|I|}\right) + M\log 10.
\end{aligned} \tag{4.6}
$$

Next,

$$
\begin{aligned}
\mathcal{K}(\rho_{I,M,\delta,\gamma}, \mu_I \otimes \nu_M) &= \mathcal{K}(\rho^1_{I,M,\delta} \otimes \rho^2_{I,M,\gamma}, \mu_I \otimes \nu_M) \\
&= \mathcal{K}(\rho^1_{I,M,\delta}, \mu_I) + \mathcal{K}(\rho^2_{I,M,\gamma}, \nu_M).
\end{aligned}
$$

Observe that

$$
\mathcal{K}(\rho^2_{I,M,\gamma}, \nu_M) = \log\left(\frac{\int \mathbf{1}_{[\sum_{j=1}^M j|\beta_j| \leq C+1]}\mathrm{d}\beta}{\int \mathbf{1}_{[\sum_{j=1}^M j|\beta_j - (\beta^\star_{I,M})_j| \leq \gamma]}\mathrm{d}\beta}\right) = M\log\left(\frac{C+1}{\gamma}\right)
$$

and, similarly,

$$
\mathcal{K}(\rho^1_{I,M,\delta}, \mu_I) = \log\left(\frac{1}{\int \mathbf{1}_{[\|\theta-\theta^\star_{I,M}\|_1 \leq \delta]}\mathrm{d}\mu_I(\theta)}\right) \leq |I|\log\left(\frac{2}{\delta}\right).
$$

The last inequality is obtained by observing that, in the worst case, the set $\{\theta : \|\theta - \theta^{\star}_{I,M}\|_1 \leq \delta\}$ intersects $2^{|I|}$ sides of $\mathcal{S}^p_{1,+}(I)$. Putting all the pieces together, we are led to

$$\mathcal{K}(\rho_{I,M,\delta,\gamma}, \mu_I \otimes \nu_M) \leq |I| \log\left(\frac{2}{\delta}\right) + M \log\left(\frac{C}{\gamma}\right). \tag{4.7}$$

Finally, it remains ton control the term

$$\int R(\theta, f) \mathrm{d}\rho_{I,M,\delta,\gamma}(\theta, f).$$

To this aim, we may write

$$\begin{aligned}
&\int R(\theta, f) \mathrm{d}\rho_{I,M,\delta,\gamma}(\theta, f) \\
&\quad = \int \mathbb{E}\left[\left(Y - f(\theta^T \mathbf{X})\right)^2\right] \mathrm{d}\rho_{I,M,\delta,\gamma}(\theta, f) \\
&\quad = \int \mathbb{E}\Big[\big(Y - f^{\star}_{I,M}(\theta^{\star T}_{I,M}\mathbf{X}) + f^{\star}_{I,M}(\theta^{\star T}_{I,M}\mathbf{X}) - f(\theta^{\star T}_{I,M}\mathbf{X}) \\
&\qquad\quad + f(\theta^{\star T}_{I,M}\mathbf{X}) - f(\theta^T\mathbf{X})\big)^2\Big] \mathrm{d}\rho_{I,M,\delta,\gamma}(\theta, f) \\
&\quad = R(\theta^{\star}_{I,M}, f^{\star}_{I,M}) \\
&\qquad + \int \mathbb{E}\Big[\big(f^{\star}_{I,M}(\theta^{\star T}_{I,M}\mathbf{X}) - f(\theta^{\star T}_{I,M}\mathbf{X})\big)^2 \\
&\qquad + \left(f(\theta^{\star T}_{I,M}\mathbf{X}) - f(\theta^T\mathbf{X})\right)^2 \\
&\qquad + 2\left(Y - f^{\star}_{I,M}(\theta^{\star T}_{I,M}\mathbf{X})\right)\left(f^{\star}_{I,M}(\theta^{\star T}_{I,M}\mathbf{X}) - f(\theta^{\star T}_{I,M}\mathbf{X})\right) \\
&\qquad + 2\left(Y - f^{\star}_{I,M}(\theta^{\star T}_{I,M}\mathbf{X})\right)\left(f(\theta^{\star T}_{I,M}\mathbf{X}) - f(\theta^T\mathbf{X})\right) \\
&\qquad + 2\left(f^{\star}_{I,M}(\theta^{\star T}_{I,M}\mathbf{X}) - f(\theta^{\star T}_{I,M}\mathbf{X})\right)\left(f(\theta^{\star T}_{I,M}\mathbf{X}) - f(\theta^T\mathbf{X})\right)\Big] \mathrm{d}\rho_{I,M,\delta,\gamma}(\theta, f) \\
&\quad := R(\theta^{\star}_{I,M}, f^{\star}_{I,M}) + \mathbf{A} + \mathbf{B} + \mathbf{C} + \mathbf{D} + \mathbf{E}.
\end{aligned}$$

Computation of C By Fubini's theorem,

$$\begin{aligned}
\mathbf{C} &= \mathbb{E}\left[\int 2\left(Y - f^{\star}_{I,M}(\theta^{\star T}_{I,M}\mathbf{X})\right)\left(f^{\star}_{I,M}(\theta^{\star T}_{I,M}\mathbf{X}) - f(\theta^{\star T}_{I,M}\mathbf{X})\right) \mathrm{d}\rho_{I,M,\delta,\gamma}(\theta, f)\right] \\
&= \mathbb{E}\left\{\int \left[2\left(Y - f^{\star}_{I,M}(\theta^{\star T}_{I,M}\mathbf{X})\right)\right.\right. \\
&\qquad\qquad \left.\left. \times \int \left(f^{\star}_{I,M}(\theta^{\star T}_{I,M}\mathbf{X}) - f(\theta^{\star T}_{I,M}\mathbf{X})\right) \mathrm{d}\rho^2_{I,M,\gamma}(f)\right] \mathrm{d}\rho^1_{I,M,\delta}(\theta)\right\}.
\end{aligned}$$

But

$$\int \left(f^{\star}_{I,M}(\theta^{\star T}_{I,M}\mathbf{X}) - f(\theta^{\star T}_{I,M}\mathbf{X})\right) \mathrm{d}\rho^2_{I,M,\gamma}(f) = 0$$

since $\rho^2_{I,M,\gamma}$ is centered at $f^{\star}_{I,M}$. This proves that $C = 0$.

Control of A Clearly,

$$\mathbf{A} \leq \int \sup_{y\in\mathbb{R}} \left((f^{\star}_{I,M}(y) - f(y)\right)^2 \mathrm{d}\rho^2_{I,M,\gamma}(f) \leq \gamma^2.$$

Control of B We have

$$\begin{aligned}
\mathbf{B} &= \int \mathbb{E}\left[\left(f(\theta^{\star T}_{I,M}\mathbf{X}) - f(\theta^T\mathbf{X})\right)^2\right] \mathrm{d}\rho_{I,M,\delta,\gamma}(\theta, f) \\
&\leq \int \mathbb{E}\left[\left(\ell(C+1)(\theta^{\star T}_{I,M} - \theta^T)\mathbf{X}\right)^2\right] \mathrm{d}\rho^1_{I,M,\delta}(\theta) \\
&\quad \text{(by the mean value theorem)} \\
&\leq \ell^2(C+1)^2\mathbb{E}\left[\|\mathbf{X}\|^2_\infty\right] \int \|\theta^{\star}_{I,M} - \theta\|^2_1 \mathrm{d}\rho^1_{I,M,\delta}(\theta) \\
&\leq \ell^2(C+1)^2\delta^2 \\
&\quad \text{(by Assumption } \mathbf{D}\text{).}
\end{aligned}$$

Control of E Write

$$\begin{aligned}
|\mathbf{E}| &\leq 2\int \mathbb{E}\Big[\left|f^{\star}_{I,M}(\theta^{\star T}_{I,M}\mathbf{X}) - f(\theta^{\star T}_{I,M}\mathbf{X})\right| \\
&\qquad \times \left|f(\theta^{\star T}_{I,M}\mathbf{X}) - f(\theta^T\mathbf{X})\right|\Big] \mathrm{d}\rho_{I,M,\delta,\gamma}(\theta, f) \\
&\leq 2\int \mathbb{E}\Big[\left|f^{\star}_{I,M}(\theta^{\star T}_{I,M}\mathbf{X}) - f(\theta^{\star T}_{I,M}\mathbf{X})\right| \\
&\qquad \times \ell(C+1)\left|(\theta^{\star T}_{I,M} - \theta^T)\mathbf{X}\right|\Big] \mathrm{d}\rho_{I,M,\delta,\gamma}(\theta, f) \\
&\leq 2\left(\int \mathbb{E}\left[\left(f^{\star}_{I,M}(\theta^{\star T}_{I,M}\mathbf{X}) - f(\theta^{\star T}_{I,M}\mathbf{X})\right)^2\right] \mathrm{d}\rho_{I,M,\delta,\gamma}(\theta, f)\right)^{\frac{1}{2}} \\
&\qquad \left(\int \mathbb{E}\left[\left(\ell(C+1)(\theta^{\star T}_{I,M} - \theta^T)\mathbf{X}\right)^2\right] \mathrm{d}\rho_{I,M,\delta,\gamma}(\theta, f)\right)^{\frac{1}{2}} \\
&\quad \text{(by the Cauchy-Schwarz inequality)} \\
&\leq 2\left(\gamma^2\right)^{\frac{1}{2}}\left(\ell^2(C+1)^2\delta^2\right)^{\frac{1}{2}} \\
&= 2\ell(C+1)\gamma\delta.
\end{aligned}$$

Control of D Finally,

$$
\begin{aligned}
\mathbf{D} &= 2\int \mathbb{E}\left[\left(Y - f^\star_{I,M}(\theta^{\star T}_{I,M}\mathbf{X})\right)\left(f(\theta^{\star T}_{I,M}\mathbf{X}) - f(\theta^T\mathbf{X})\right)\right] \mathrm{d}\rho_{I,M,\delta,\gamma}(\theta, f) \\
&= 2\int \mathbb{E}\left[\left(Y - f^\star_{I,M}(\theta^{\star T}_{I,M}\mathbf{X})\right)\left(f^\star_{I,M}(\theta^{\star T}_{I,M}\mathbf{X}) - f^\star_{I,M}(\theta^T\mathbf{X})\right)\right] \mathrm{d}\rho^1_{I,M,\delta}(\theta) \\
&\quad \left(\text{since } \int f \mathrm{d}\rho^2_{I,M,\delta}(f) = f^\star_{I,M}\right) \\
&= 2\mathbb{E}\left[\left(Y - f^\star_{I,M}(\theta^{\star T}_{I,M}\mathbf{X})\right)\int \left(f^\star_{I,M}(\theta^{\star T}_{I,M}\mathbf{X}) - f^\star_{I,M}(\theta^T\mathbf{X})\right) \mathrm{d}\rho^1_{I,M,\delta}(\theta)\right] \\
&\le 2\sqrt{\mathbb{E}\left[\left(Y - f^\star_{I,M}(\theta^{\star T}_{I,M}\mathbf{X})\right)^2\right]} \\
&\quad \times \sqrt{\mathbb{E}\left[\int \left(f^\star_{I,M}(\theta^{\star T}_{I,M}\mathbf{X}) - f^\star_{I,M}(\theta^T\mathbf{X})\right) \mathrm{d}\rho^1_{I,M,\delta}(\theta)\right]^2} \\
&\quad \text{(by the Cauchy-Schwarz inequality)} \\
&= 2\sqrt{R(\theta^\star_{I,M}, f^\star_{I,M})}\sqrt{\mathbb{E}\left[\int \left(f^\star_{I,M}(\theta^{\star T}_{I,M}\mathbf{X}) - f^\star_{I,M}(\theta^T\mathbf{X})\right) \mathrm{d}\rho^1_{I,M,\delta}(\theta)\right]^2}.
\end{aligned}
$$

The inequality

$$
\begin{aligned}
\left|f^\star_{I,M}(\theta^{\star T}_{I,M}\mathbf{X}) - f^\star_{I,M}(\theta^T\mathbf{X})\right| &\le \ell(C+1)\left|(\theta^{\star T}_{I,M} - \theta^T)\mathbf{X}\right| \\
&\le \ell(C+1)\|\theta^\star_{I,M} - \theta\|_1
\end{aligned}
$$

leads to

$$
\begin{aligned}
&\left[\int \left(f^\star_{I,M}(\theta^{\star T}_{I,M}\mathbf{X}) - f^\star_{I,M}(\theta^T\mathbf{X})\right) \mathrm{d}\rho^1_{I,M,\delta}(\theta)\right]^2 \\
&\qquad \le \ell^2(C+1)^2\left[\int \|\theta^\star_{I,M} - \theta\|_1 \mathrm{d}\rho^1_{I,M,\delta}(\theta)\right]^2.
\end{aligned}
$$

Consequently,

$$
\left[\int \left(f^\star_{I,M}(\theta^{\star T}_{I,M}\mathbf{X}) - f^\star_{I,M}(\theta^T\mathbf{X})\right) \mathrm{d}\rho^1_{I,M,\delta}(\theta)\right]^2 \le \ell^2(C+1)^2\delta^2,
$$

and therefore

$$
\begin{aligned}
D &\le 2\ell(C+1)\delta\sqrt{R(0,0)/2} \\
&\le \sqrt{2}\ell(C+1)\delta\sqrt{C^2+\sigma^2}.
\end{aligned}
$$

Thus, taking $\delta = \gamma = 1/n$ and putting all the pieces together, we obtain

$$A + B + C + D + E \leq \frac{\mathcal{C}_1}{n},$$

where $\mathcal{C}_1$ is a positive constant function of C, σ and ℓ. Combining this inequality with (4.5)-(4.7) yields, with probability larger than $1 - 2\varepsilon$,

$$\begin{aligned} R(\hat{\theta}_\lambda, \hat{f}_\lambda) &- R(\theta^\star, f^\star) \\ &\leq \frac{1}{1 - \frac{\lambda[(2C+1)^2+4\sigma^2]}{n-w\lambda}} \inf_{\substack{I \subset \{1,\ldots,p\} \\ 1 \leq M \leq n}} \left\{ \left(1 + \frac{\lambda\left[(2C+1)^2 + 4\sigma^2\right]}{n - w\lambda}\right) \left(R(\theta^\star_{I,M}, f^\star_{I,M}) \right.\right. \\ &\qquad \left.\left. - R(\theta^\star, f^\star) + \frac{\mathcal{C}_1}{n} \right) + 2\frac{M\log(10Cn) + |I|\log\left(\frac{4epn}{|I|}\right) + \log\left(\frac{1}{\varepsilon}\right)}{\lambda} \right\}. \end{aligned}$$

Choosing finally

$$\lambda = \frac{n}{w + 2\left[(2C+1)^2 + 4\sigma^2\right]},$$

we obtain that there exists a positive constant $\mathcal{C}_2$ function of L, C, σ and ℓ such that, with probability at least $1 - 2\varepsilon$,

$$\begin{aligned} R(\hat{\theta}_\lambda, \hat{f}_\lambda) - R(\theta^\star, f^\star) \leq \mathcal{C}_2 \inf_{\substack{I \subset \{1,\ldots,p\} \\ 1 \leq M \leq n}} &\left\{ R(\theta^\star_{I,M}, f^\star_{I,M}) - R(\theta^\star, f^\star) \right. \\ &\left. + \frac{M\log(10Cn) + |I|\log\left(\frac{4epn}{|I|}\right) + \log\left(\frac{1}{\varepsilon}\right)}{n} \right\}. \end{aligned}$$

This concludes the proof of Theorem 2.1.

4.3 Proof of Corollary 2.1

We already know, by Theorem 2.1, that with probability at least $1 - 2\varepsilon$,

$$\begin{aligned} R(\hat{\theta}_\lambda, \hat{f}_\lambda) - R(\theta^\star, f^\star) \leq \mathcal{C} \inf_{\substack{I \subset \{1,\ldots,p\} \\ 1 \leq M \leq n}} &\left\{ R(\theta^\star_{I,M}, f^\star_{I,M}) - R(\theta^\star, f^\star) \right. \\ &\left. + \frac{M\log(Cn) + |I|\log\left(\frac{pn}{|I|}\right) + \log\left(\frac{1}{\varepsilon}\right)}{n} \right\}. \end{aligned}$$

By definition, for all $(\theta, f) \in \mathcal{S}^p_{1,+}(I) \times \mathcal{F}_M(C)$,

$$R(\theta^\star_{I,M}, f^\star_{I,M}) \leq R(\theta, f).$$

In particular, if there exists $I^\star$ such that $\theta^\star \in \mathcal{S}^p_{1,+}(I^\star)$, then

$$R(\hat{\theta}_\lambda, \hat{f}_\lambda) - R(\theta^\star, f^\star) \leq \mathcal{C} \inf_{\substack{1 \leq M \leq n \\ f \in \mathcal{F}_M(C)}} \left\{ R(\theta^\star, f) - R(\theta^\star, f^\star) \right.$$
$$\left. + \frac{M \log(Cn) + |I^\star| \log\left(\frac{pn}{|I^\star|}\right) + \log\left(\frac{1}{\varepsilon}\right)}{n} \right\}. \tag{4.8}$$

Observe that, for any $f \in \mathcal{F}_M(C)$,

$$\begin{aligned} R(\theta^\star, f) - R(\theta^\star, f^\star) &= \int_{\mathbb{R}^p} \left[f\left(\theta^{\star T}\mathbf{x}\right) - f^\star\left(\theta^{\star T}\mathbf{x}\right)\right]^2 \mathrm{d}\mathbf{P}(\mathbf{x}, y) \\ &\leq B^2 \int_{-1}^{1} \left[f(t) - f^\star(t)\right]^2 \mathrm{d}t. \end{aligned}$$

Since $f^\star \in L_2([-1, 1])$, we may write

$$f^\star = \sum_{j=1}^{\infty} \beta_j^\star \varphi_j$$

and apply (4.8) with

$$f = \sum_{j=1}^{M} \beta_j^\star \varphi_j.$$

In order to do so, we just need to check that $f \in \mathcal{F}_M(C)$, that is

$$\sum_{j=1}^{M} j|\beta_j^\star| \leq C + 1.$$

But, by the Cauchy-Schwarz inequality,

$$\begin{aligned} \sum_{j=1}^{M} j|\beta_j^\star| &= \sum_{j=1}^{M} j^k |\beta_j^\star| j^{1-k} \\ &\leq \sqrt{\sum_{j=1}^{M} j^{2k}(\beta_j^\star)^2} \sqrt{\sum_{j=1}^{M} j^{2-2k}}. \end{aligned}$$

Thus,

$$\sum_{j=1}^{M} j|\beta_j^\star| \leq \frac{\pi}{\sqrt{6}}\sqrt{\sum_{j=1}^{M} j^{2k}(\beta_j^\star)^2}$$
$$\text{(since, by assumption, } k \geq 2)$$
$$\leq C+1$$
$$\text{(since } f^\star \in \mathcal{W}(k, 6(C+1)^2/\pi^2)).$$

Next, with this choice of f,

$$\int_{-1}^{1} [f(t) - f^\star(t)]^2 \,\mathrm{d}t \leq \Lambda M^{-2k}$$

for some positive constant Λ depending only on k and C (see for instance Tsybakov [38]). Therefore, inequality (4.8) takes the form

$$R(\hat{\theta}_\lambda, \hat{f}_\lambda) - R(\theta^\star, f^\star) \leq \mathcal{C} \inf_{1\leq M\leq n} \left\{ \Lambda M^{-2k} + \frac{M\log(Cn) + |I^\star|\log\left(\frac{pn}{|I^\star|}\right) + \log\left(\frac{1}{\varepsilon}\right)}{n} \right\}. \tag{4.9}$$

Letting $\lceil . \rceil$ be the ceiling function and choosing $M = \lceil (n/\log(Cn))^{\frac{1}{2\beta+1}} \rceil$ in (4.9) concludes the proof.

5 Annex: Description of the MCMC algorithm

This annex is intended to make thoroughly clear the specification of the proposal conditional densities k_1 and k_2 introduced in Section 3.

5.1 Notation

To provide explicit formulas for the conditional densities $k_1((\tau, h)|(\theta, f))$ and $k_2((\tau, h)|(\theta, f))$, we first set

$$f = \sum_{j=1}^{m_f} \beta_{f,j}\varphi_j \quad \text{and} \quad h = \sum_{j=1}^{m_h} \beta_{h,j}\varphi_j,$$

where it is recalled that $\{\varphi_j\}_{j=1}^{\infty}$ denotes the (non-normalized) trigonometric system. We let I (respectively, J) be the set of nonzero coordinates of the vector θ (respectively, τ), and denote finally by θ_I (respectively, τ_J) the vector of dimension $|I|$ (respectively, $|J|$) which contains the nonzero coordinates of θ (respectively, $|\tau|$). Recall that all densities are defined with respect to the prior π, which is made explicit in Subsection 2.2.

For a generic $h \in \mathcal{F}_{m_h}(C)$, given $\tau \in \mathcal{S}_{1,+}^p$ and $s > 0$, we let the density $\mathrm{dens}_s(h|\tau, m_h)$ with respect to π be defined as

$$\begin{aligned}
&\mathrm{dens}_s(h|\tau, m_h) \\
&\quad \propto \exp\left[-\frac{1}{2s^2}\sum_{j=1}^{m_h}\left(\beta_{h,j} - \tilde{\beta}_j(\tau, m_h)\right)\right]\mathbf{1}\left[\sum_{j=1}^{m_h} j^2|\beta_{h,j}| \le C+1\right],
\end{aligned}$$

where the $\tilde{\beta}_j(\tau, m_h)$ are the empirical least square coefficients given by

$$\left\{\tilde{\beta}_j(\tau, m_h)\right\}_{j=1,\ldots,m_h} \in \arg\min_{b\in\mathbb{R}^{m_h}} \sum_{i=1}^{n}\left(Y_i - \sum_{j=1}^{m_h} b_j\varphi_j(\tau^T\mathbf{X}_i)\right)^2.$$

In the experiments, we fixed $s = 0.1$. Note that simulating with respect to $\mathrm{dens}_s(h|\tau, m_h)$ is an easy task, since one just needs to compute a least square estimate and then draw from a truncated Gaussian distribution.

5.2 Description of k_1

We take

$$\begin{aligned}
k_1(\cdot|(\theta, f)) = &\frac{2k_{1,=}(\cdot|(\theta, f)) + k_{1,+}(\cdot|(\theta, f))}{3}\mathbf{1}_{[|I|=1]} \\
&+ \frac{k_{1,-}(\cdot|(\theta, f)) + 2k_{1,=}(\cdot|(\theta, f)) + k_{1,+}(\cdot|(\theta, f))}{4}\mathbf{1}_{[1<|I|<p]} \\
&+ \frac{k_{1,-}(\cdot|(\theta, f)) + 2k_{1,=}(\cdot|(\theta, f))}{3}\mathbf{1}_{[|I|=p]}.
\end{aligned}$$

Roughly, the idea is that $k_{1,-}$ tries to remove one component in θ, $k_{1,=}$ keeps the same number of components, whereas $k_{1,+}$ adds one component. The density $k_{1,=}$ takes the form

$$k_{1,=}((\tau, h)|(\theta, f)) = k_{1,=}(\tau|\theta)\mathrm{dens}_s(h|\tau, m_f).$$

The density $k_{1,=}(.|\theta)$ is the density of τ when $J = I$ and

$$\tau_I = \frac{\theta_I + E}{\|\theta_I + E\|_1}\mathrm{sgn}\left((\theta_I + E)_{j(\theta_I+E)}\right),$$

where $E = (E_1, \ldots, E_{|I|})$ and the E_i are independent random variables uniformly distributed in $[-\delta, \delta]$. Throughout, the value of δ was fixed at 0.5. It is noteworthy that when we change the parameter from θ to τ, then we also change the function from f to h. Thus, with this procedure, the link function h is more "adapted" to τ and the subsequent move is more likely to be accepted in the Hastings-Metropolis algorithm.

In the case where we are to remove one component, $k_{1,-}$ is given by

$$k_{1,-}\left((\tau, h)|(\theta, f)\right) = \sum_{j \in I} c_j \mathbf{1}_{[\tau = \theta_{-j}]} \text{dens}_s(h|\tau, m_f),$$

where θ_{-j} is just obtained from θ by setting the j-th component to 0 and by renormalizing the parameter in order to have $\|\theta_{-j}\|_1 = 1$. We set

$$c_j = \frac{\exp\left(-|\theta_j|\right) \mathbf{1}_{[|\theta_j| < \delta]}}{\sum_{\ell \in I} \exp\left(-|\theta_\ell|\right) \mathbf{1}_{[|\theta_\ell| < \delta]}}.$$

The idea is that smaller components are more likely to be removed than larger ones. Finally, the density $k_{1,+}$ takes the form

$$k_{1,+}\left((\tau, h)|(\theta, f)\right) = \sum_{j \notin I} c'_j \mathbf{1}_{[\tau_{-j} = \theta]} \frac{\mathbf{1}_{[|\tau_j| < \delta]}}{2\delta} \text{dens}_s(h|\tau, m_f).$$

We set

$$c'_j = \frac{\exp\left(\left|\sum_{i=1}^n \left(Y_i - f(\theta^T \mathbf{X}_i)\right)(\mathbf{X}_i)_j\right|\right)}{\sum_{\ell \notin I} \exp\left(\left|\sum_{i=1}^n \left(Y_i - f(\theta^T \mathbf{X}_i)\right)(\mathbf{X}_i)_\ell\right|\right)}$$

where $(\mathbf{X}_i)_j$ denotes the j-th component of $\mathbf{X}_i$. In words, the idea is that a new nonzero coordinate in θ is more likely to be interesting in the model if the corresponding feature is correlated with the current residual.

5.3 Description of k_2

In the same spirit, we let the conditional density k_2 be defined as

$$\begin{aligned} k_2\left(\cdot|(\theta, f)\right) = & \frac{2k_{2,=}\left(\cdot|(\theta, f)\right) + k_{2,+}\left(\cdot|(\theta, f)\right)}{3} \mathbf{1}_{[m_f = 1]} \\ & + \frac{k_{2,-}\left(\cdot|(\theta, f)\right) + 2k_{2,=}\left(\cdot|(\theta, f)\right) + k_{2,+}\left(\cdot|(\theta, f)\right)}{4} \mathbf{1}_{[1 < m_f < n]} \\ & + \frac{k_{2,-}\left(\cdot|(\theta, f)\right) + 2k_{2,=}\left(\cdot|(\theta, f)\right)}{3} \mathbf{1}_{[m_f = n]}. \end{aligned}$$

We choose

$$k_{2,=}\left((\tau, h)|(\theta, f)\right) = \mathbf{1}_{[\tau = \theta]} \text{dens}_s(h|\tau, m_f)$$

and

$$k_{2,+}\left((\tau,h)|(\theta,f)\right) = \mathbf{1}_{[\tau=\theta]}\mathrm{dens}_s(h|\tau,m_f+1).$$

With this choice, $m_h = m_f+1$, which means that the proposal density tries to add one coefficient in the expansion of h, while leaving θ unchanged. Finally

$$k_{2,-}\left((\tau,h)|(\theta,f)\right) = \mathbf{1}_{[\tau=\theta]}\mathrm{dens}_s(h|\tau,m_f-1),$$

and the proposal tries to remove one coefficient in h.

References

[1] P. Alquier. PAC-Bayesian bounds for randomized empirical risk minimizers. *Mathematical Methods of Statistics*, 17:279–304, 2008.

[2] P. Alquier and K. Lounici. *PAC-Bayesian bounds for sparse regression estimation with exponential weights*. Preprint PMA-1346, arXiv:1009.2707, 2010.

[3] A. Antoniadis, G. Grégoire, and I.W. McKeague. Bayesian estimation in single-index models. *Statistica Sinica*, 14:1147–1164, 2004.

[4] J.-Y. Audibert. Aggregated estimators and empirical complexity for least square regression. *Annales de l'Institut Henri Poincaré: Probability and Statistics*, 40:685–736, 2004.

[5] R.E. Bellman. *Adaptive Control Processes: A Guided Tour*. Princeton University Press, 1961.

[6] P.J. Bickel, Y. Ritov, and A.B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37:1705–1732, 2009.

[7] A.M. Bruckstein, D.L. Donoho, and M. Elad. From sparse solutions of systems of equations to sparse modeling of signals and images. *SIAM Review*, 51:34–81, 2009.

[8] F. Bunea, A. Tsybakov, and M. Wegkamp. Sparsity oracle inequalities for the Lasso. *Electronic Journal of Statistics*, 1:169–194, 2007.

[9] E.J. Candès and T. Tao. The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35:2313–2351, 2005.

[10] O. Catoni. *Statistical Learning Theory and Stochastic Optimization*. Springer-Verlag, 2004.

[11] O. Catoni. *PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*, volume 56 of *Lecture Notes-Monograph Series*. IMS, 2007.

[12] J.M. Chambers, W.S. Cleveland, B. Kleiner, and P.A. Tukey. *Graphical Methods for Data Analysis*. Wadsworth & Brooks, Belmont, 1983.

[13] X. Chen, C. Zou, and R.D. Cook. Coordinate-independent sparse sufficient dimension reduction and variable selection. *The Annals of Statistics*, 38:3696–3723, 2010.

[14] P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis. Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 47:547–553, 2009.

[15] A.S. Dalalyan, A. Juditsky, and V. Spokoiny. A new algorithm for estimating the effective dimension-reduction subspace. *Journal of Machine Learning Research*, 9:1647–1678, 2008.

[16] A.S. Dalalyan and A.B. Tsybakov. Aggregation by exponential weighting, sharp PAC-Bayesian bounds and sparsity. *Machine Learning*, 72:39–61, 2008.

[17] A.S. Dalalyan and A.B. Tsybakov. *Sparse regression learning by aggregation and Langevin Monte-Carlo*. Preprint PMA, arXiv:0903.1223, 2010.

[18] M. Delecroix, M. Hristache, and V. Patilea. On semiparametric M-estimation in single-index regression. *Journal of Statistical Planning and Inference*, 136:730–769, 2006.

[19] S. Gaïffas and G. Lecué. Optimal rates and adaptation in the single-index model using aggregation. *Electronic Journal of Statistics*, 1:538–573, 2007.

[20] P.J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian Model determination. *Biometrika*, 82:711–732, 1995.

[21] L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer-Verlag, New York, 2002.

[22] W. Härdle, P. Hall, and H. Ichimura. Optimal smoothing in single-index models. *The Annals of Statistics*, 21:157–178, 1993.

[23] D. Jr. Harrison and D.L. Rubinfeld. Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, 5:81–102, 1978.

[24] J.L. Horowitz. *Semiparametric Methods in Econometrics*. Springer-Verlag, 1998.

[25] H. Ichimura. Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *Journal of Econometrics*, 58:71–120, 1993.

[26] O. Lopez. Single-index regression models with right-censored responses. *Journal of Statistical Planning and Inference*, 139:1082–1097, 2009.

[27] J.-M. Marin and C. Robert. *Bayesian Core: A practical approach to computational Bayesian analysis*. Springer, New York, 2007.

[28] P. Massart. *Concentration Inequalities and Model Selection*. Springer-Verlag, Berlin, 2007.

[29] D.A. McAllester. Some PAC-Bayesian theorems. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, pages 230–234, New York, 1998. ACM.

[30] P. McCullagh and J.A. Nelder. *Generalized Linear Models*. Chapman and Hall, 1983.

[31] E.A. Nadaraya. On estimating regression. *Theory of Probability and its Applications*, 9:141–142, 1964.

[32] E.A. Nadaraya. Remarks on nonparametric estimates for density functions and regression curves. *Theory of Probability and its Applications*, 15:134–137, 1970.

[33] J.R. Quinlan. Combining instance-based and model-based learning. In *Proceedings of the Tenth International Conference on Machine Learning*, pages 236–243, Amherst, 1993. Morgan Kaufmann.

[34] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, 2008.

[35] J. Shawe-Taylor and R. Williamson. A PAC analysis of a Bayes estimator. In *Proceedings of the Tenth Annual Conference on Computational Learning Theory*, pages 2–9, New York, 1997. ACM.

[36] C.J. Stone. Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, 10:1040–1053, 1982.

[37] R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1996.

[38] A.B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.

[39] L.J. van't Veer, H. Dai, M.J. van de Vijver, Y.D. He, A.A.M. Hart, M. Mao, H.L. Peterse, K. van der Kooy, M.J. Marton, A.T. Witteveen, G.J. Schreiber, R.M. Kerkhoven, C. Roberts, P.S. Linsley, R. Bernards, and S.H. Friend. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415:530–536, 2002.

[40] H.B. Wang. Bayesian estimation and variable selection for single index models. *Computational Statistics and Data Analysis*, 53:2617–2627, 2009.

[41] G.S. Watson. Smooth regression analysis. *Sankhyā Series A*, 26:359–372, 1964.

[42] I.-C. Yeh. Modeling of strength of high-performance concrete using artificial neural networks. *Cement and Concrete Research*, 28:1797–1808, 1998.

[43] I.-C. Yeh. Modeling slump flow of concrete using second-order regressions and artificial neural networks. *Cement and Concrete Composites*, 29:474–480, 2007.