Electronic Journal of Statistics
Vol. 7 (2013) 264–291
ISSN: 1935-7524
DOI: 10.1214/13-EJS771

PAC-Bayesian estimation and prediction in sparse additive models

Benjamin Guedj*

Laboratoire de Statistique Théorique et Appliquée
Université Pierre et Marie Curie - UPMC
Tour 25 - 2ème étage, boîte n° 158
4, place Jussieu
75252 Paris Cedex 05, France
e-mail: benjamin.guedj@upmc.fr

and

Pierre Alquier†

School of Mathematical Sciences
University College Dublin
Room 528 - James Joyce Library
Belfield, Dublin 4, Ireland
e-mail: pierre.alquier@ucd.ie

Abstract: The present paper is about estimation and prediction in high-dimensional additive models under a sparsity assumption ($p \gg n$ paradigm). A PAC-Bayesian strategy is investigated, delivering oracle inequalities in probability. The implementation is performed through recent outcomes in high-dimensional MCMC algorithms, and the performance of our method is assessed on simulated data.

AMS 2000 subject classifications: Primary 62G08, 62J02, 65C40.
Keywords and phrases: Additive models, sparsity, regression estimation, PAC-Bayesian bounds, oracle inequality, MCMC, stochastic search.

Received August 2012.

Contents

*Corresponding author.
†Research partially supported by the French "Agence Nationale pour la Recherche" under grant ANR-09-BLAN-0128 "PARCIMONIE".

1. Introduction

Substantial progress has been achieved over the last years in estimating high-dimensional regression models. A thorough introduction to this dynamic field of contemporary statistics is provided by the recent monographs Hastie, Tibshirani and Friedman (2009); Bühlmann and van de Geer (2011). In the popular framework of linear and generalized linear models, the Lasso estimator introduced by Tibshirani (1996) immediately proved successful. Its theoretical properties have been extensively studied and its popularity has never wavered since then, see for example Bunea, Tsybakov and Wegkamp (2006); van de Geer (2008); Bickel, Ritov and Tsybakov (2009); Meinshausen and Yu (2009). However, even though numerous phenomena are well captured within this linear context, restraining high-dimensional statistics to this setting is unsatisfactory. To relax the strong assumptions required in the linear framework, one idea is to investigate a more general class of models, such as nonparametric regression models of the form $Y = f(X) + W$, where Y denotes the response, X the predictor and W a zero-mean noise. A good compromise between complexity and effectiveness is the additive model. It has been extensively studied and formalized for thirty years now. Amongst many other references, the reader is invited to refer to Stone (1985); Hastie and Tibshirani (1986, 1990); Härdle (1990). The core of this model is that the regression function is written as a sum of univariate functions $f = \sum_{i=1}^{p} f_i$, easing its interpretation. Indeed, each covariate's effect is assessed by a unique function. This class of nonparametric models is a popular setting in statistics, despite the fact that classical estimation procedures are known to perform poorly as soon as the number of covariates p exceeds the number of observations n in that setting.

In the present paper, our goal is to investigate a PAC-Bayesian-based prediction strategy in the high-dimensional additive framework ($p \gg n$ paradigm). In that context, estimation is essentially possible at the price of a sparsity assumption, *i.e.*, most of the f_i functions are zero. More precisely, our setting is non-asymptotic. As empirical evidences of sparse representations accumulate, high-dimensional statistics are more and more coupled with a sparsity assumption, namely that the intrinsic dimension p_0 of the data is much smaller than p and n, see e.g. Giraud, Huet and Verzelen (2012). Additive modelling under a sparsity constraint has been essentially studied under the scope of the Lasso in Meier, van de Geer and Bühlmann (2009), Suzuki and Sugiyama (2012) and Koltchinskii and Yuan (2010) or of a combination of functional grouped Lasso and backfitting algorithm in Ravikumar et al. (2009). Those papers inaugurated the study of this problem and contain essential theoretical results consisting in asymptotics (see Meier, van de Geer and Bühlmann (2009); Ravikumar et al. (2009)) and non-asymptotics (see Suzuki and Sugiyama (2012); Koltchinskii and Yuan (2010)) oracle inequalities. The present article should be seen as a constructive contribution towards a deeper understanding of prediction problems in the additive framework. It should also be stressed that our work is to be seen as an attempt to relax as much as possible assumptions made on the model, such as restrictive conditions on the regressors' matrix. We consider

them too much of a non-realistic burden when it comes to prediction problems.

Our *modus operandi* will be based on PAC-Bayesian results, which is original in that context to our knowledge. The PAC-Bayesian theory originates in the two seminal papers Shawe-Taylor and Williamson (1997); McAllester (1999) and has been extensively formalized in the context of classification (see Catoni (2004, 2007)) and regression (see Audibert (2004a,b); Alquier (2006, 2008); Audibert and Catoni (2010, 2011)). However, the methods presented in these references are not explicitly designed to cover the high-dimensional setting under the sparsity assumption. Thus, the PAC-Bayesian theory has been worked out in the sparsity perspective lately, by Dalalyan and Tsybakov (2008, 2012); Alquier and Lounici (2011); Rigollet and Tsybakov (2012). The main message of these studies is that aggregation with a properly chosen prior is able to deal effectively with the sparsity issue. Interesting additional references addressing the aggregation outcomes would be Rigollet (2006); Audibert (2009). The former aggregation procedures rely on an exponential weights approach, achieving good statistical properties. Our method should be seen as an extension of these techniques, and is particularly focused on additive modelling specificities. Contrary to procedures such as the Lasso, the Dantzig selector and other penalized methods which are provably consistent under restrictive assumptions on the Gram matrix associated to the predictors, PAC-Bayesian aggregation requires only minimal assumptions on the model. Our method is supported by oracle inequalities in probability, that are valid in both asymptotic and non-asymptotic settings. We also show that our estimators achieve the optimal rate of convergence over traditional smoothing classes such as Sobolev ellipsoids. It should be stressed that our work is inspired by Alquier and Biau (2011), which addresses the celebrated single-index model with similar tools and philosophy. Let us also mention that although the use of PAC-Bayesian techniques are original in this context, parallel work has been conducted in the deterministic design case by Suzuki (2012).

A major difficulty when considering high-dimensional problems is to achieve a favorable compromise between statistical and computational performances. The recent and thorough monograph Bühlmann and van de Geer (2011) shall provide the reader with valuable insights that address this drawback. As a consequence, the explicit implementation of PAC-Bayesian techniques remains unsatisfactory as existing routines are only put to test with small values of p (typically $p < 100$), contradicting with the high-dimensional framework. In the meantime, as a solution of a convex problem the Lasso proves computable for large values of p in reasonable amounts of time. We therefore focused on improving the computational aspect of our PAC-Bayesian strategy. Monte Carlo Markov Chains (MCMC) techniques proved increasingly popular in the Bayesian community, for they probably are the best way of sampling from potentially complex probability distributions. The reader willing to find a thorough introduction to such techniques is invited to refer to the comprehensive monographs Marin and Robert (2007); Meyn and Tweedie (2009). While Alquier and Biau (2011); Alquier and Lounici (2011) explore versions of the reversible jump MCMC method (RJMCMC) introduced by Green (1995), Dalalyan and Tsybakov (2008, 2012) in-

vestigate a Langevin-Monte Carlo-based method, however only a deterministic design is considered. We shall try to overcome those limitations by considering adaptations of a recent procedure whose comprehensive description is to be found in Petralias (2010); Petralias and Dellaportas (2012). This procedure called Subspace Carlin and Chib algorithm originates in the seminal paper by Carlin and Chib (1995), and has a close philosophy of Hans, Dobra and West (2007), as it favors local moves for the Markov chain. We provide numerical evidence that our method is computationally efficient, on simulated data.

The paper is organized as follows. Section 2 presents our PAC-Bayesian prediction strategy in additive models. In particular, it contains the main theoretical results of this paper which consist in oracle inequalities. Section 3 is devoted to the implementation of our procedure, along with numerical experiments on simulated data, presented in Section 4. Finally, and for the sake of clarity, proofs have been postponed to Section 5.

2. PAC-Bayesian prediction

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space on which we denote by $\{(\mathbf{X}_i, Y_i)\}_{i=1}^n$ a sample of n independent and identically distributed (i.i.d.) random vectors in $(-1,1)^p \times \mathbb{R}$, with $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})$, satisfying

$$Y_i = \psi^\star(\mathbf{X}_i) + \xi_i = \sum_{j=1}^{p} \psi_j^\star(X_{ij}) + \xi_i, \quad i \in \{1, \ldots, p\},$$

where $\psi_1^\star, \ldots, \psi_p^\star$ are p continuous functions $(-1,1) \to \mathbb{R}$ and $\{\xi_i\}_{i=1}^n$ is a set of i.i.d. (conditionaly to $\{(\mathbf{X}_i, Y_i)\}_{i=1}^n$) real-valued random variables. Let $\mathcal{P}$ denote the distribution of the sample $\{(\mathbf{X}_i, Y_i)\}_{i=1}^n$. Denote by $\mathbb{E}$ the expectation computed with respect to $\mathbb{P}$ and let $\|\cdot\|_\infty$ be the supremum norm. We make the two following assumptions.

(**A1**) For any integer k, $\mathbb{E}[|\xi_1|^k] < \infty$, $\mathbb{E}[\xi_1 | \mathbf{X}_1] = 0$ and there exist two positive constants L and σ^2 such that for any integer $k \geq 2$,

$$\mathbb{E}[|\xi_1|^k | \mathbf{X}_1] \leq \frac{k!}{2} \sigma^2 L^{k-2}.$$

(**A2**) There exists a constant $C > \max(1, \sigma)$ such that $\|\psi^\star\|_\infty \leq C$.

Note that A1 implies that $\mathbb{E}\,\xi_1 = 0$ and that the distribution of ξ_1 may depend on $\mathbf{X}_1$. In particular, A1 holds if ξ_1 is a zero-mean gaussian with variance $\gamma^2(\mathbf{X}_1)$ where $x \mapsto \gamma^2(x)$ is bounded.

Further, note that the boundedness assumption A2 plays a key role in our approach, as it allows to use a version of Bernstein's inequality which is one of the two main technical tools we use to state our results. This assumption is not only a technical prerequisite since it proved crucial for critical regimes: indeed, if the intrinsic dimension p_0 of the regression function $\psi^\star$ is still large, the boundedness of the function class allows much faster estimation rates. This point is profusely discussed in Raskutti, Wainwright and Yu (2012).

We are mostly interested in sparse additive models, in which only a few $\{\psi_j^\star\}_{j=1}^p$ are not identically zero. Let $\{\varphi_k\}_{k=1}^\infty$ be a known countable set of continuous functions $\mathbb{R} \to (-1,1)$ called the dictionary. In the sequel, $|\mathcal{H}|$ stands for the cardinality of a set $\mathcal{H}$. For any p-th tuple $\mathbf{m} = (m_1, \dots, m_p) \in \mathbb{N}^p$, denote by $S(\mathbf{m}) \subset \{1, \dots, p\}$ the set of indices of nonzero elements of $\mathbf{m}$, *i.e.*,

$$|S(\mathbf{m})| = \sum_{j=1}^p \mathbb{1}[m_j > 0],$$

and define

$$\Theta_{\mathbf{m}} = \left\{\theta \in \mathbb{R}^{m_1} \times \cdots \times \mathbb{R}^{m_p}\right\},$$

with the convention $\mathbb{R}^0 = \emptyset$. The set $\Theta_{\mathbf{m}}$ is embedded with its canonical Borel field $\mathcal{B}(\Theta_m) = \mathcal{B}(\mathbb{R}^{m_1}) \otimes \cdots \otimes \mathcal{B}(\mathbb{R}^{m_p})$. Denote by

$$\Theta \stackrel{\text{def}}{=} \bigcup_{\mathbf{m} \in \mathcal{M}} \Theta_{\mathbf{m}},$$

which is equipped with the σ-algebra $\mathcal{T} = \sigma\left(\bigvee_{\mathbf{m}\in\mathcal{M}} \mathcal{B}(\Theta_{\mathbf{m}})\right)$, where $\mathcal{M}$ is the collection of models $\mathcal{M} = \{\mathbf{m} = (m_1, \dots, m_p) \in \mathbb{N}^p\}$. Consider the span of the set $\{\varphi_k\}_{k=1}^\infty$, *i.e.*, the set of functions

$$\mathbb{F} = \left\{ \psi_\theta = \sum_{j \in S(\mathbf{m})} \psi_j = \sum_{j \in S(\mathbf{m})} \sum_{k=1}^{m_j} \theta_{jk} \varphi_k \colon \theta \in \Theta_{\mathbf{m}}, \mathbf{m} \in \mathcal{M} \right\},$$

equipped with a countable generated σ-algebra denoted by $\mathcal{F}$. The risk and empirical risk associated to any $\psi_\theta \in \mathbb{F}$ are defined respectively as

$$R(\psi_\theta) = \mathbb{E}[Y_1 - \psi_\theta(\mathbf{X}_1)]^2 \quad \text{and} \quad R_n(\psi_\theta) = r_n(\{\mathbf{X}_i, Y_i\}_{i=1}^n, \psi_\theta),$$

where

$$r_n(\{\mathbf{x}_i, y_i\}_{i=1}^n, \psi_\theta) = \frac{1}{n} \sum_{i=1}^n \left(y_i - \psi_\theta(\mathbf{x}_i)\right)^2.$$

Consider the probability η_α on the set $\mathcal{M}$ defined by

$$\eta_\alpha \colon \mathbf{m} \mapsto \frac{1 - \frac{\alpha}{1-\alpha}}{1 - \left(\frac{\alpha}{1-\alpha}\right)^{p+1}} \binom{p}{|S(\mathbf{m})|}^{-1} \alpha^{\sum_{j=1}^p m_j},$$

for some $\alpha \in (0, 1/2)$. Let us stress the fact that the probability η_α acts as a penalization term over a model $\mathbf{m}$, on the number of its active regressors through the combinatorial term $\binom{p}{|S(\mathbf{m})|}^{-1}$ and on their expansion through $\alpha^{\sum_{j=1}^p m_j}$.

Our procedure relies on the following construction of the probability π, referred to as the prior, in order to promote the sparsity properties of the target regression function $\psi^\star$. For any $\mathbf{m} \in \mathcal{M}$, $\zeta > 0$ and $\mathbf{x} \in \Theta_{\mathbf{m}}$, denote by $\mathcal{B}^1_{\mathbf{m}}(\mathbf{x}, \zeta)$

the ℓ^1-ball centered in $\mathbf{x}$ with radius ζ. For any $\mathbf{m} \in \mathcal{M}$, denote by $\pi_{\mathbf{m}}$ the uniform distribution on $\mathcal{B}^1_{\mathbf{m}}(0, C)$. Define the probability π on $(\Theta, \mathcal{T})$,

$$\pi(A) = \sum_{\mathbf{m}\in\mathcal{M}} \eta_\alpha(\mathbf{m})\pi_{\mathbf{m}}(A), \quad A \in \mathcal{T}.$$

Note that the volume $V_{\mathbf{m}}(C)$ of $\mathcal{B}^1_{\mathbf{m}}(0, C)$ is given by

$$V_{\mathbf{m}}(C) = \frac{(2C)^{\sum_{j\in S(\mathbf{m})} m_j}}{\Gamma\left(\sum_{j\in S(\mathbf{m})} m_j + 1\right)} = \frac{(2C)^{\sum_{j\in S(\mathbf{m})} m_j}}{\left(\sum_{j\in S(\mathbf{m})} m_j\right)!}.$$

Finally, set $\delta > 0$ (which may be interpreted as an inverse temperature parameter) and the posterior Gibbs transition density is

$$\rho_\delta(\{(\mathbf{x}_i, y_i)\}_{i=1}^n, \theta) = \sum_{\mathbf{m}\in\mathcal{M}} \frac{\eta_\alpha(\mathbf{m})}{V_{\mathbf{m}}(C)} \mathbb{1}_{\mathcal{B}^1_{\mathbf{m}}(0,C)}(\theta) \frac{\exp[-\delta r_n(\{\mathbf{x}_i, y_i\}_{i=1}^n, \psi_\theta)]}{\int \exp[-\delta r_n(\{\mathbf{x}_i, y_i\}_{i=1}^n, \psi_\theta)]\pi(\mathrm{d}\theta)}. \quad (2.1)$$

We then consider two competing estimators. The first one is the randomized Gibbs estimator $\hat{\Psi}$, constructed with parameters $\hat{\theta}$ sampled from the posterior Gibbs density, *i.e.*, for any $A \in \mathcal{F}$,

$$\mathbb{P}(\hat{\Psi} \in A | \{\mathbf{X}_i, Y_i\}_{i=1}^n) = \int_A \rho_\delta(\{\mathbf{X}_i, Y_i\}_{i=1}^n, \theta)\pi(\mathrm{d}\theta), \quad (2.2)$$

while the second one is the aggregated Gibbs estimator $\hat{\Psi}^{\mathrm{A}}$ defined as the posterior mean

$$\hat{\Psi}^{\mathrm{A}} = \int \psi_\theta \rho_\delta(\{\mathbf{X}_i, Y_i\}_{i=1}^n, \theta)\pi(\mathrm{d}\theta) = \mathbb{E}[\hat{\Psi} | \{\mathbf{X}_i, Y_i\}_{i=1}^n]. \quad (2.3)$$

These estimators have been introduced in Catoni (2004, 2007) and investigated in further work by Audibert (2004a); Alquier (2006, 2008); Dalalyan and Tsybakov (2008, 2012).

For the sake of clarity, denote by $\mathcal{D}$ a generic numerical constant in the sequel. We are now in a position to write a PAC-Bayesian oracle inequality.

Theorem 2.1. *Let $\hat{\psi}$ and $\hat{\psi}^{\mathrm{A}}$ be realizations of the Gibbs estimators defined by (2.2)–(2.3), respectively. Let A1 and A2 hold. Set $w = 8C\max(L, C)$ and $\delta = n\ell/[w+4(\sigma^2+C^2)]$, for $\ell \in (0, 1)$, and let $\varepsilon \in (0, 1)$. Then with $\mathbb{P}$-probability at least $1 - 2\varepsilon$,*

$$\left.\begin{array}{l} R(\hat{\psi}) - R(\psi^\star) \\ R(\hat{\psi}^{\mathrm{A}}) - R(\psi^\star) \end{array}\right\} \le \mathcal{D} \inf_{\mathbf{m}\in\mathcal{M}} \inf_{\theta\in\mathcal{B}^1_{\mathbf{m}}(0,C)} \left\{ R(\psi_\theta) - R(\psi^\star) + |S(\mathbf{m})| \frac{\log(p/|S(\mathbf{m})|)}{n} + \frac{\log(n)}{n} \sum_{j\in S(\mathbf{m})} m_j + \frac{\log(1/\varepsilon)}{n} \right\}, \quad (2.4)$$

where $\mathcal{D}$ depends upon w, σ, C, ℓ and α defined above.

Under mild assumptions, Theorem 2.1 provides inequalities which admit the following interpretation. If there exists a "small" model in the collection $\mathcal{M}$, *i.e.*, a model $\mathbf{m}$ such that $\sum_{j\in S(\mathbf{m})} m_j$ and $|S(\mathbf{m})|$ are small, such that ψ_θ (with $\theta \in \Theta_{\mathbf{m}}$) is close to $\psi^\star$, then $\hat{\psi}$ and $\hat{\psi}^{\mathrm{A}}$ are also close to $\psi^\star$ up to $\log(n)/n$ and $\log(p)/n$ terms. However, if no such model exists, at least one of the terms $\sum_{j\in S(\mathbf{m})} m_j/n$ and $|S(\mathbf{m})|/n$ starts to emerge, thereby deteriorating the global quality of the bound. A satisfying estimation of $\psi^\star$ is typically possible when $\psi^\star$ admits a sparse representation.

To go further, we derive from Theorem 2.1 an inequality on Sobolev ellipsoids. We show that our procedure achieves the optimal rate of convergence in this setting. For the sake of shortness, we consider Sobolev spaces, however one can easily derive the following results in other functional spaces such as Besov spaces. See Tsybakov (2009) and the references therein.

The notation $\{\varphi_k\}_{k=1}^{\infty}$ now refers to the (non-normalized) trigonometric system, defined as

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

with $j \in \mathbb{N}^*$ and $t \in (-1, 1)$. Let us denote by $S^\star$ the set of indices of non-identically zero regressors. That is, the regression function $\psi^\star$ is

$$\psi^\star = \sum_{j\in S^\star} \psi_j^\star.$$

Assume that for any $j \in S^\star$, $\psi_j^\star$ belongs to the Sobolev ellipsoid $\mathcal{W}(r_j, d_j)$ defined as

$$\mathcal{W}(r_j, d_j) = \left\{ f \in \mathrm{L}^2([-1,1]) : f = \sum_{k=1}^{\infty} \theta_k \varphi_k \quad \text{and} \quad \sum_{i=1}^{\infty} i^{2r_j} \theta_i^2 \le d_j \right\}.$$

with d_j chosen such that $\sum_{j\in S^\star} \sqrt{d_j} \le C\sqrt{6}/\pi$ and for unknown regularity parameters $r_1, \ldots, r_{|S^\star|} \ge 1$. Let us stress the fact that this assumption casts our results onto the adaptive setting. It also implies that $\psi^\star$ belongs to the Sobolev ellipsoid $\mathcal{W}(r, d)$, with $r = \min_{j\in S^\star} r_j$ and $d = \sum_{j\in S^\star} d_j$, *i.e.*,

$$\psi^\star = \sum_{j\in S^\star} \sum_{k=1}^{\infty} \theta_{jk}^\star \varphi_k. \tag{2.5}$$

It is worth pointing out that in that setting, the Sobolev ellipsoid is better approximated by the ℓ^1-ball $\mathcal{B}_{\mathbf{m}}^1(0, C)$ as the dimension of $\mathbf{m}$ grows. Further, make the following assumption.

(**A3**) The distribution of the data $\mathcal{P}$ has a probability density with respect to the corresponding Lebesgue measure, bounded from above by a constant $B > 0$.

Theorem 2.2. *Let $\hat{\psi}$ and $\hat{\psi}^{\mathrm{A}}$ be realizations of the Gibbs estimators defined by (2.2)–(2.3), respectively. Let A1, A2 and A3 hold. Set $w = 8C\max(L, C)$ and $\delta = n\ell/[w+4(\sigma^2+C^2)]$, for $\ell \in (0,1)$, and let $\varepsilon \in (0,1)$. Then with $\mathbb{P}$-probability at least $1-2\varepsilon$,*

$$\left.\begin{array}{l} R(\hat{\psi}) - R(\psi^\star) \\ R(\hat{\psi}^{\mathrm{A}}) - R(\psi^\star) \end{array}\right\} \leq \mathcal{D}\left\{ \sum_{j\in S^\star} d_j^{\frac{1}{2r_j+1}} \left(\frac{\log(n)}{2nr_j}\right)^{\frac{2r_j}{2r_j+1}} + |S^\star|\log(p/|S^\star|)/n + \frac{\log(1/\varepsilon)}{n} \right\},$$

where $\mathcal{D}$ is a constant depending only on w, σ, C, ℓ, α and B.

Theorem 2.2 illustrates that we obtain the minimax rate of convergence over Sobolev classes up to a $\log(n)$ term. Indeed, the minimax rate to estimate a single function with regularity r is $n^{\frac{2r}{2r+1}}$, see for example Tsybakov (2009, Chapter 2). Theorem 2.1 and Theorem 2.2 thus validate our method.

A salient fact about Theorem 2.2 is its links with existing work: assume that all the $\psi_j^\star$ belong to the same Sobolev ellipsoid $\mathcal{W}(r,d)$. The convergence rate is now $\log(n)n^{-\frac{2r}{2r+1}} + \log(p)/n$. This rate (down to a $\log(n)$ term) is the same as the one exhibited by Koltchinskii and Yuan (2010) in the context of multiple kernel learning ($n^{-\frac{2r}{2r+1}} + \log(p)/n$). Suzuki and Sugiyama (2012) even obtain faster rates which correspond to smaller functional spaces. However, the results presented by both Koltchinskii and Yuan (2010) and Suzuki and Sugiyama (2012) are obtained under stringent conditions on the design, which are not necessary to prove Theorem 2.2.

A natural extension is to consider sparsity on both regressors and their expansion, instead of sparse regressors and nested expansion as before. That is, we no longer consider the first m_j dictionary functions for the expansion of regressor j. To this aim, we slightly extend the previous notation. Let $K \in \mathbb{N}^*$ be the length of the dictionary. A model is now denoted by $\mathbf{m} = (\mathbf{m}_1, \ldots, \mathbf{m}_p)$ and for any $j \in \{1, \ldots, p\}$, $\mathbf{m}_j = (m_{j1}, \ldots, m_{jK})$ is a K-sized vector whose entries are 1 whenever the corresponding dictionary function is present in the model and 0 otherwise. Introduce the notation

$$S(\mathbf{m}) = \{\mathbf{m}_j \neq \mathbf{0}, j \in \{1,\ldots,p\}\}, \quad S(\mathbf{m}_j) = \{m_{jk} \neq 0, k \in \{1,\ldots,K\}\}.$$

The prior distribution on the models space $\mathcal{M}$ is now

$$\eta_\alpha \colon \mathbf{m} \mapsto \frac{1 - \alpha\frac{1-\alpha^{K+1}}{1-\alpha}}{1 - \left(\alpha\frac{1-\alpha^{K+1}}{1-\alpha}\right)^{p+1}} \binom{p}{|S(\mathbf{m})|}^{-1} \prod_{j\in S(\mathbf{m})} \binom{K}{|S(\mathbf{m}_j)|}^{-1} \alpha^{|S(\mathbf{m}_j)|},$$

for any $\alpha \in (0, 1/2)$.

Theorem 2.3. *Let $\hat{\psi}$ and $\hat{\psi}^{\mathrm{A}}$ be realizations of the Gibbs estimators defined by (2.2)–(2.3), respectively. Let A1 and A2 hold. Set $w = 8C\max(L, C)$ and $\delta = n\ell/[w+4(\sigma^2+C^2)]$, for $\ell \in (0,1)$, and let $\varepsilon \in (0,1)$. Then with $\mathbb{P}$-probability*

at least $1-2\varepsilon$,

$$\left.\begin{array}{l} R(\hat{\psi})-R(\psi^\star) \\ R(\hat{\psi}^{\mathrm{A}})-R(\psi^\star) \end{array}\right\} \leq \mathcal{D} \inf_{\mathbf{m}\in\mathcal{M}} \inf_{\theta\in\mathcal{B}^1_{\mathbf{m}}(0,C)} \left\{ R(\psi_\theta)-R(\psi^\star) \right.$$
$$\left. +|S(\mathbf{m})|\frac{\log(p/|S(\mathbf{m})|)}{n}+\frac{\log(nK)}{n}\sum_{j\in S(\mathbf{m})}|S(\mathbf{m}_j)|+\frac{\log(1/\varepsilon)}{n} \right\},$$

where $\mathcal{D}$ *depends upon* w, σ, C, ℓ *and* α *defined above.*

3. MCMC implementation

In this section, we describe an implementation of the method outlined in the previous section. Our goal is to sample from the Gibbs posterior distribution ρ_δ. We use a version of the so-called Subspace Carlin and Chib (SCC) developed by Petralias (2010); Petralias and Dellaportas (2012) which originates in the Shotgun Stochastic Search algorithm (see Hans, Dobra and West (2007)). The key idea of the algorithm lies in a stochastic search heuristic that restricts moves in neighborhoods of the visited models. Let $T\in\mathbb{N}^*$ and denote by $\{\theta(t),\mathbf{m}(t)\}_{t=0}^T$ the Markov chain of interest, with $\theta(t)\in\Theta_{\mathbf{m}(t)}$. Define $i\colon t\mapsto\{+,-,=\}$, the three possible moves performed by the algorithm: an addition, a deletion or an adjustment of a regressor. Let $\{\mathbf{e}_1,\dots,\mathbf{e}_p\}$ be the canonical base of $\mathbb{R}^p$. For any model $\mathbf{m}(t)=(m_1(t),\dots,m_p(t))\in\mathcal{M}$, define its neighborhood $\{\mathcal{V}^+[\mathbf{m}(t)],\mathcal{V}^-[\mathbf{m}(t)],\mathcal{V}^=[\mathbf{m}(t)]\}$, where

$$\mathcal{V}^+[\mathbf{m}(t)]=\{\mathbf{k}\in\mathcal{M}\colon \mathbf{k}=\mathbf{m}(t)+x\mathbf{e}_j, x\in\mathbb{N}^*, j\in\{1,\dots,p\}\backslash S[\mathbf{m}(t)]\},$$
$$\mathcal{V}^-[\mathbf{m}(t)]=\{\mathbf{k}\in\mathcal{M}\colon \mathbf{k}=\mathbf{m}(t)-m_j(t)\mathbf{e}_j, j\in S[\mathbf{m}(t)]\},$$

and

$$\mathcal{V}^=[\mathbf{m}(t)]=\{\mathbf{k}\in\mathcal{M}\colon S(\mathbf{k})=S[\mathbf{m}(t)]\}.$$

A move $i(t)$ is chosen with probability $q[i(t)]$. By convention, if $S[\mathbf{m}(t)]=p$ (respectively $S[\mathbf{m}(t)]=1$) the probability of performing an addition move (respectively a deletion move) is zero. Note $\xi\colon\{+,-\}\mapsto\{-,+\}$ and let $D_{\mathbf{m}}$ be the design matrix in model $\mathbf{m}\in\mathcal{M}$. Denote by $\mathrm{LSE}_{\mathbf{m}}$ the least square estimate $\mathrm{LSE}_{\mathbf{m}}=(D'_{\mathbf{m}}D_{\mathbf{m}})^{-1}D'_{\mathbf{m}}\mathbf{Y}$ (with $\mathbf{Y}=(Y_1,\dots,Y_n)$) in model $\mathbf{m}\in\mathcal{M}$. For ease of notation, let $\mathcal{I}$ denote the identity matrix. Finally, denote by $\phi(\cdot;\mu,\Gamma)$ the density of a Gaussian distribution $\mathcal{N}(\mu,\Gamma)$ with mean μ and covariance matrix Γ. A description of the full algorithm is presented in Algorithm 1.

The estimates $\hat{\Psi}$ and $\hat{\Psi}^{\mathrm{A}}$ are obtained as

$$\hat{\Psi}=\sum_{j=1}^p\sum_{k=1}^K\theta_{jk}(T)\varphi_k,$$

and for some burnin $b\in\{1,\dots,T-1\}$,

$$\hat{\Psi}^{\mathrm{A}}=\sum_{j=1}^p\sum_{k=1}^K\left(\frac{1}{T-b}\sum_{\ell=b+1}^T\theta_{jk}(\ell)\right)\varphi_k.$$

Algorithm 1 A Subspace Carlin and Chib-based algorithm

Initialize $(\theta(0), \mathbf{m}(0))$.
for $t = 1$ to T **do**
  Choose a move $i(t)$ with probability $q[i(t)]$.
  For any $\mathbf{k} \in \mathcal{V}^{i(t)}[\mathbf{m}(t-1)]$, generate $\theta_{\mathbf{k}}$ from the proposal density $\phi(\cdot; \mathrm{LSE}_{\mathbf{k}}, \sigma^2 \mathcal{I})$.
  Propose a model $\mathbf{k} \in \mathcal{V}^{i(t)}[\mathbf{m}(t-1)]$ with probability

$$\gamma(\mathbf{m}(t-1), \mathbf{k}) = \frac{A_{\mathbf{k}}}{\sum_{\mathbf{j} \in \mathcal{V}^{i(t)}[\mathbf{m}(t-1)]} A_{\mathbf{j}}},$$

  where

$$A_{\mathbf{j}} = \frac{\rho_\delta(\theta_{\mathbf{j}})}{\phi(\theta_{\mathbf{j}}; \mathrm{LSE}_{\mathbf{j}}, \sigma^2 \mathcal{I})}.$$

  if $i(t) \in \{+, -\}$ **then**
    For any $\mathbf{h} \in \mathcal{V}^{\xi(i(t))}[\mathbf{k}]$, generate $\theta_{\mathbf{h}}$ from the proposal density $\phi(\cdot; \mathrm{LSE}_{\mathbf{h}}, \sigma^2 \mathcal{I})$. Note that $\mathbf{m}(t-1) \in \mathcal{V}^{\xi(i(t))}[\mathbf{k}]$.
    Accept model $\mathbf{k}$, *i.e.*, set $\mathbf{m}(t) = \mathbf{k}$ and $\theta(t) = \theta_{\mathbf{k}}$, with probability

$$\alpha = \min\left(1, \frac{A_{\mathbf{k}} q[i(t)] \gamma(\mathbf{k}, \mathbf{m}(t-1))}{A_{\mathbf{m}(t-1)} q[\xi(i(t))] \gamma(\mathbf{m}(t-1), \mathbf{k})}\right)$$
$$= \min\left(1, \frac{q[i(t)] \sum_{\mathbf{h} \in \mathcal{V}^{i(t)}[\mathbf{m}(t-1)]} A_{\mathbf{h}}}{q[\xi(i(t))] \sum_{\mathbf{h} \in \mathcal{V}^{\xi(i(t))}[\mathbf{k}]} A_{\mathbf{h}}}\right).$$

    Otherwise, set $\mathbf{m}(t) = \mathbf{m}(t-1)$ and $\theta(t) = \theta_{\mathbf{m}(t-1)}$.
  else
    Generate $\theta_{\mathbf{m}(t-1)}$ from the proposal density $\phi(\cdot; \mathrm{LSE}_{\mathbf{m}(t-1)}, \sigma^2 \mathcal{I})$.
    Accept model $\mathbf{k}$, *i.e.*, set $\mathbf{m}(t) = \mathbf{k}$ and $\theta(t) = \theta_{\mathbf{k}}$, with probability

$$\alpha = \min\left(1, \frac{A_{\mathbf{k}} \gamma(\mathbf{k}, \mathbf{m}(t-1))}{A_{\mathbf{m}(t-1)} \gamma(\mathbf{m}(t-1), \mathbf{k})}\right).$$

    Otherwise, set $\mathbf{m}(t) = \mathbf{m}(t-1)$ and $\theta(t) = \theta_{\mathbf{m}(t-1)}$.
  end if
end for

The transition kernel of the chain defined above is reversible with respect to $\rho_\delta \otimes \eta_\alpha$, hence this procedure ensures that $\{\theta(t)\}_{t=1}^T$ is a Markov Chain with stationary distribution ρ_δ.

4. Numerical studies

In this section we validate the effectiveness of our method on simulated data. All our numerical studies have been performed with the software R (see R Core Team (2012)). The method is available on the CRAN website (`http://www.cran.r-project.org/web/packages/pacbpred/index.html`), under the name `pacbpred` (see Guedj (2012)).

Some comments are in order here about how to calibrate the constants C, σ^2, δ and α. Clearly, a too small value for C will stuck the algorithm, preventing the chain to escape from the initial model. Indeed, most proposed models will be discarded since the acceptance ratio will frequently take the

TABLE 1
Each number is the mean (standard deviation) of the RSS over 10 independent runs

MCMC	$p = 50$ 3000 it.	$p = 200$ 10000 it.	$p = 400$ 20000 it.
Model 1	0.0318 (0.0047)	0.0320 (0.0029)	0.0335 (0.0056)
Model 2	0.0411 (0.0061)	0.1746 (0.0639)	0.2201 (0.0992)
Model 3	0.0665 (0.0421)	0.1151 (0.0399)	0.1597 (0.0579)

value 0. Conversely, a large value for C deteriorates the quality of the bound in Theorem 2.1, Theorem 2.2, Theorem 2.3 and Theorem 5.1. However, this only influences the theoretical bound, as its contribution to the acceptance ratio is limited to $\log(2C)$. We thereby proceeded with typically large values of C (such as $C = 10^6$). As the parameter σ^2 is the variance of the proposal distribution ϕ, the practioner should tune it in accordance with the noise level of the data. The parameter requiring the finest calibration is δ: the convergence of the algorithm is sensitive to its choice. Dalalyan and Tsybakov (2008, 2012) exhibit the theoretical value $\delta = n/4\sigma^2$. This value leads to very good numerical performances, as it has been also noticed by Dalalyan and Tsybakov (2008, 2012); Alquier and Biau (2011). The choice for α is guided by a similar reasoning to the one for C. Its contribution to the acceptance ratio is limited to a $\log(1/\alpha)$ term. The value $\alpha = 0.25$ was used in the simulations for its apparent good properties. Although it would be computationally costly, a finer calibration through methods such as cross-validation is possible.

Finally and as a general rule, we strongly encourage practitioners to run several chains of inequal lengths and to adjust the number of iterations needed by observing if the empirical risk is stabilized.

Model 1. $n = 200$ *and* $S^\star = \{1, 2, 3, 4\}$. *This model is similar to Meier, van de Geer and Bühlmann (2009, Section 3, Example 1) and is given by*

$$Y_i = \psi_1^\star(X_{i1}) + \psi_2^\star(X_{i2}) + \psi_3^\star(X_{i3}) + \psi_4^\star(X_{i4}) + \xi_i,$$

with

$$\psi_1^\star \colon x \mapsto -\sin(2x), \quad \psi_2^\star \colon x \mapsto x^3, \quad \psi_3^\star \colon x \mapsto x,$$
$$\psi_4^\star \colon x \mapsto e^{-x} - e/2, \quad \xi_i \sim \mathcal{N}(0, 0.1), \quad i \in \{1, \dots, n\}.$$

The covariates are sampled from independent uniform distributions over $(-1, 1)$.

Model 2. $n = 200$ *and* $S^\star = \{1, 2, 3, 4\}$. *As above but correlated. The covariates are sampled from a multivariate gaussian distribution with covariance matrix* $\Sigma_{ij} = 2^{-|i-j|-2}$, $i, j \in \{1, \dots, p\}$.

Model 3. $n = 200$ *and* $S^\star = \{1, 2, 3, 4\}$. *This model is similar to Meier, van de Geer and Bühlmann (2009, Section 3, Example 3) and is given by*

$$Y_i = 5\psi_1^\star(X_{i1}) + 3\psi_2^\star(X_{i2}) + 4\psi_3^\star(X_{i3}) + 6\psi_4^\star(X_{i4}) + \xi_i,$$

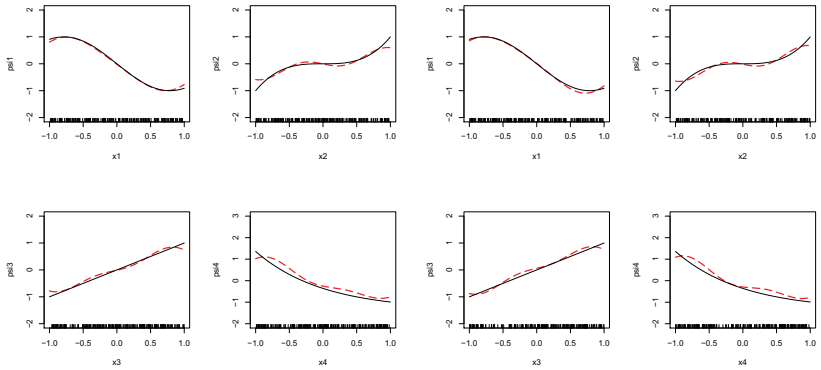
with

$$\psi_1^\star\colon x \mapsto x, \quad \psi_2^\star\colon x \mapsto 4(x^2 - x - 1), \quad \psi_3^\star\colon x \mapsto \frac{\sin(2\pi x)}{2 - \sin(2\pi x)},$$
$$\psi_4^\star\colon x \mapsto 0.1\sin(2\pi x) + 0.2\cos(2\pi x) + 0.3\sin^2(2\pi x) + 0.4\cos^3(2\pi x) + 0.5\sin^3(2\pi x), \quad \xi_i \sim \mathcal{N}(0, 0.5), \quad i \in \{1, \dots, n\}.$$

The covariates are sampled from independent uniform distributions over $(-1, 1)$.

The results of the simulations are summarized in Table 1 and illustrated by Figure 1 and Figure 2. The reconstruction of the true regression function $\psi^\star$ is achieved even in very high-dimensional situations, pulling up our method at the level of the gold standard Lasso.

(a) Model 1, $p = 200$. (b) Model 1, $p = 400$.



(c) Model 2, $p = 50$. (d) Model 3, $p = 50$.

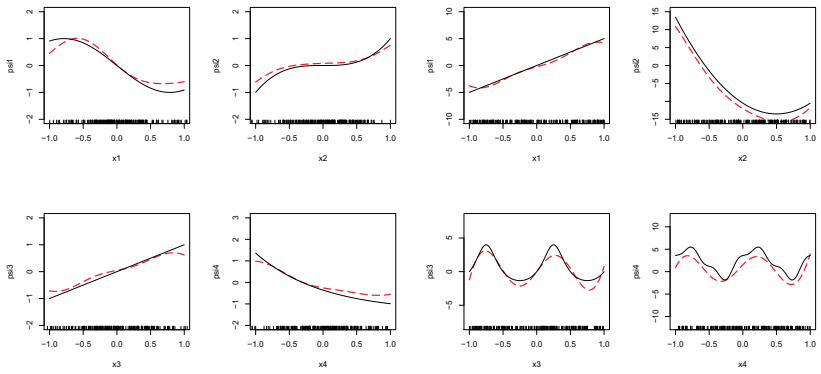
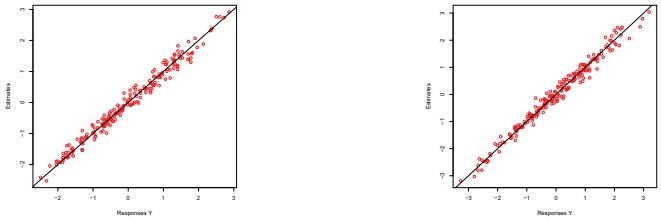


FIG 1. *Estimates (red dashed lines) for* $\psi_1^\star$, $\psi_2^\star$, $\psi_3^\star$ *and* $\psi_4^\star$ *(solid black lines). Other estimates (for* $\psi_j^\star$, $j \notin \{1, 2, 3, 4\}$*) are mostly zero.*

(a) Model 1, $p = 200$. (b) Model 1, $p = 400$.



(c) Model 2, $p = 50$. (d) Model 3, $p = 50$.

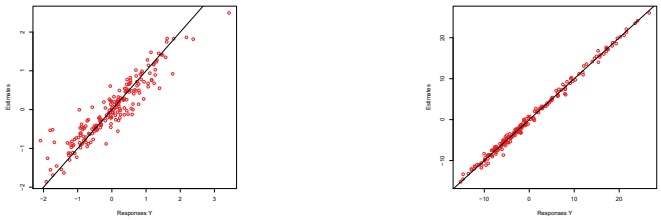


FIG 2. *plot of the responses $Y_1, \dots, Y_n$ against their estimates. The more points on the first bisectrix (solid black line), the better the estimation.*

5. Proofs

To start the chain of proofs leading to Theorem 2.1, Theorem 2.2 and Theorem 2.3, we recall and prove some lemmas to establish Theorem 5.1 which consists in a general PAC-Bayesian inequality in the spirit of Catoni (2004, Theorem 5.5.1) for classification or Catoni (2004, Lemma 5.8.2) for regression. Note also that Dalalyan and Tsybakov (2012, Theorem 1) provides a similar inequality in the deterministic design case. A salient fact on Theorem 5.1 is that the validity of the oracle inequalities only involves the distribution of the noise variable ξ_1, and that distribution is independent of the sample size n.

The proofs of the following two classical results are omitted. Lemma 5.1 is a version of Bernstein's inequality which originates in Massart (2007, Proposition 2.19), whereas Lemma 5.2 appears in Catoni (2004, Equation 5.2.1).

For $x \in \mathbb{R}$, denote $(x)_+ = \max(x, 0)$. Let μ_1, μ_2 be two probabilities. The Kullback-Leibler divergence of μ_1 with respect to μ_2 is denoted $\mathcal{KL}(\mu_1, \mu_2)$ and is

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

Finally, for any measurable space $(A, \mathcal{A})$ and any probability π on $(A, \mathcal{A})$, denote by $\mathcal{M}^1_{+,\pi}(A, \mathcal{A})$ the set of probabilities on $(A, \mathcal{A})$ absolutely continuous with respect to π.

Lemma 5.1. *Let $(T_i)_{i=1}^n$ be independent real-valued variables. Assume that there exist two positive constants v and w such that, for any integer $k \geq 2$,*

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

Then for any $\gamma \in \left(0, \frac{1}{w}\right)$,

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

Lemma 5.2. *Let $(A, \mathcal{A})$ be a measurable space. For any probability μ on $(A, \mathcal{A})$ and any measurable function $h : A \to \mathbb{R}$ such that $\int (\exp \circ h) \mathrm{d}\mu < \infty$,*

$$\log \int (\exp \circ h) \mathrm{d}\mu = \sup_{m \in \mathcal{M}_{+,\pi}^1(A,\mathcal{A})} \int h \mathrm{d}m - \mathcal{KL}(m, \mu),$$

with the 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 is reached for the Gibbs distribution g given by

$$\frac{\mathrm{d}g}{\mathrm{d}\mu}(a) = \frac{\exp(h(a))}{\int (\exp \circ h) \mathrm{d}\mu}, \quad a \in A.$$

Theorem 5.1 is valid in the general regression framework. In the proofs of Lemma 5.3, Lemma 5.4, Lemma 5.5 and Theorem 5.1, we consider a general regression function $\psi^\star$. Denote by $(\Theta, \mathcal{T})$ a space of functions equipped with a countable generated σ-algebra, and let π be a probability on $(\Theta, \mathcal{T})$, referred to as the prior. Lemma 5.3, Lemma 5.4, Lemma 5.5 and Theorem 5.1 follow from the work of Catoni (2004); Dalalyan and Tsybakov (2008, 2012); Alquier (2008); Alquier and Biau (2011). Let $\delta > 0$ and consider the so-called *posterior* Gibbs transition density ρ_δ with respect to π, defined as

$$\rho_\delta(\{\mathbf{x}_i, y_i\}_{i=1}^n, \psi) = \frac{\exp[-\delta r_n(\{\mathbf{x}_i, y_i\}_{i=1}^n, \psi)]}{\int \exp[-\delta r_n(\{\mathbf{x}_i, y_i\}_{i=1}^n, \psi)] \pi(\mathrm{d}\psi)}. \tag{5.1}$$

In the following three lemmas, denote by ρ a *so-called* posterior probability absolutely continuous with respect to π. Let ψ be a realization of a random variable Ψ sampled from ρ.

Lemma 5.3. *Let A1 and A2 hold. Set $w = 8C\max(L, C)$, $\delta \in (0, n/[w + 4(\sigma^2 + C^2)])$ and $\varepsilon \in (0, 1)$. Then with $\mathbb{P}$-probability at least $1 - \varepsilon$*

$$R(\psi) - R(\psi^\star) \leq \frac{1}{1 - \frac{4\delta(\sigma^2+C^2)}{n - w\delta}} \left(R_n(\psi) - R_n(\psi^\star) + \frac{\log \frac{\mathrm{d}\rho}{\mathrm{d}\pi}(\psi) + \log \frac{1}{\varepsilon}}{\delta}\right).$$

Proof. Apply Lemma 5.1 to the variables T_i defined as follow: for any $\psi \in$,

$$T_i = -(Y_i - \psi(\mathbf{X}_i))^2 + (Y_i - \psi^\star(\mathbf{X}_i))^2, \quad i \in \{1, \dots, n\}. \tag{5.2}$$

First, let us note that

$$\begin{aligned} R(\psi) - R(\psi^\star) &= \mathbb{E}[(Y_1 - \psi(\mathbf{X}_1))^2] - \mathbb{E}[(Y_1 - \psi^\star(\mathbf{X}_1))^2] \\ &= \mathbb{E}[(2Y_1 - \psi(\mathbf{X}_1) - \psi^\star(\mathbf{X}_1))(\psi^\star(\mathbf{X}_1) - \psi(\mathbf{X}_1))] \\ &= \mathbb{E}\left[(\psi^\star(\mathbf{X}_1) - \psi(\mathbf{X}_1))\,\mathbb{E}[(2W_1 + \psi^\star(\mathbf{X}_1) - \psi(\mathbf{X}_1))|\,\mathbf{X}_1]\right] \\ &= 2\,\mathbb{E}[(\psi^\star(\mathbf{X}_1) - \psi(\mathbf{X}_1))\,\mathbb{E}[\xi_1|\,\mathbf{X}_1]] + \mathbb{E}[\psi^\star(\mathbf{X}_1) - \psi(\mathbf{X}_1)]^2. \end{aligned}$$

As $\mathbb{E}[\xi_1|\,\mathbf{X}_1] = 0$,

$$R(\psi) - R(\psi^\star) = \mathbb{E}[\psi^\star(\mathbf{X}) - \psi(\mathbf{X})]^2. \tag{5.3}$$

By (5.2), the random variables $(T_i)_{i=1}^n$ are independent. Using Lemma 5.1, we get

$$\begin{aligned} \sum_{i=1}^n \mathbb{E}\,T_i^2 &= \sum_{i=1}^n \mathbb{E}\left[(2Y_i - \psi(\mathbf{X}_i) - \psi^\star(\mathbf{X}_i))^2(\psi(\mathbf{X}_i) - \psi^\star(\mathbf{X}_i))^2\right] \\ &= \sum_{i=1}^n \mathbb{E}\,\mathbb{E}\left[(2W_i + \psi^\star(\mathbf{X}_i) - \psi(\mathbf{X}_i))^2(\psi(\mathbf{X}_i) - \psi^\star(\mathbf{X}_i))^2|\,\mathbf{X}_i\right]. \end{aligned}$$

Next, using that $|a+b|^k \le 2^{k-1}(|a| + |b|)$ for any a, $b \in \mathbb{R}$ and $k \in \mathbb{N}^*$, we get

$$\begin{aligned} \sum_{i=1}^n \mathbb{E}\,T_i^2 &\le 2\sum_{i=1}^n \mathbb{E}\left[(\psi(\mathbf{X}_i) - \psi^\star(\mathbf{X}_i))^2\,\mathbb{E}\left[(4W_i^2 + 4C^2)|\,\mathbf{X}_i\right]\right] \\ &\le 8\left(\sigma^2 + C^2\right)\sum_{i=1}^n \mathbb{E}\left[(\psi(\mathbf{X}_i) - \psi^\star(\mathbf{X}_i))^2\right] \\ &= 8n\left(\sigma^2 + C^2\right)(R(\psi) - R(\psi^\star)) \stackrel{\text{def}}{=} v, \end{aligned} \tag{5.4}$$

where we have used (5.3) in the last equation. It follows that for any integer $k \ge 3$,

$$\begin{aligned} \sum_{i=1}^n \mathbb{E}[(T_i)_+^k] &= \sum_{i=1}^n \mathbb{E}\,\mathbb{E}[(T_i)_+^k|\,\mathbf{X}_i] \\ &\le \sum_{i=1}^n \mathbb{E}\,\mathbb{E}\left[|2Y_i - \psi(\mathbf{X}_i) - \psi^\star(\mathbf{X}_i)|^k|\psi(\mathbf{X}_i) - \psi^\star(\mathbf{X}_i)|^k|\,\mathbf{X}_i\right] \\ &= \sum_{i=1}^n \mathbb{E}\,\mathbb{E}\left[|2W_i + \psi^\star(\mathbf{X}_i) - \psi(\mathbf{X}_i)|^k|\psi(\mathbf{X}_i) - \psi^\star(\mathbf{X}_i)|^k|\,\mathbf{X}_i\right] \\ &\le 2^{k-1}\sum_{i=1}^n \mathbb{E}\,\mathbb{E}\left[\left(2^k|\xi_i|^k + |\psi^\star(\mathbf{X}_i) - \psi(\mathbf{X}_i)|^k\right)|\psi(\mathbf{X}_i) - \psi^\star(\mathbf{X}_i)|^k|\,\mathbf{X}_i\right]. \end{aligned}$$

Using that $|\psi(\mathbf{x}_i) - \psi^\star(\mathbf{x}_i)|^k \le (2C)^{k-2}|\psi(\mathbf{x}_i) - \psi^\star(\mathbf{x}_i)|^2$ and (5.3), we get

$$\sum_{i=1}^n \mathbb{E}[(T_i)_+^k] \le 2^{k-1}\sum_{i=1}^n \left(2^{k-1}k!\sigma^2 L^{k-2} + (2C)^k\right)(2C)^{k-2}[R(\psi) - R(\psi^\star)]$$
$$= \frac{k!}{2} v (2C)^{k-2}\left(\frac{2^{2k-4}\sigma^2 L^{k-2} + \frac{2}{k!}2^{2k-4}C^k}{\sigma^2 + C^2}\right).$$

Recalling that $C > \max(1, \sigma)$ gives

$$\frac{2^{2k-4}\sigma^2 L^{k-2} + \frac{2}{k!}2^{2k-4}C^k}{\sigma^2 + C^2} \le \frac{4^{k-2}\sigma^2 L^{k-2}}{2\sigma^2} + \frac{\frac{2}{k!}4^{k-2}C^k}{C^2}$$
$$\le \frac{1}{2}(4L)^{k-2} + \frac{1}{2}(4C)^{k-2} = [4\max(L, C)]^{k-2}.$$

Hence

$$\sum_{i=1}^n \mathbb{E}[(T_i)_+^k] \le \frac{k!}{2} v w^{k-2}, \quad \text{with} \quad w \stackrel{\text{def}}{=} 8C\max(L, C). \tag{5.5}$$

Applying Lemma 5.1, we obtain, for any real $\delta \in \left(0, \frac{n}{w}\right)$, with $\gamma = \frac{\delta}{n}$,

$$\mathbb{E}\exp[\delta(R_n(\psi^\star) - R_n(\psi) + R(\psi) - R(\psi^\star))] \le \exp\left(\frac{v\delta^2}{2n^2\left(1 - \frac{w\delta}{n}\right)}\right),$$

that is, that for any real number $\varepsilon \in (0, 1)$,

$$\mathbb{E}\exp\left[\delta[R_n(\psi^\star) - R_n(\psi)] + \delta[R(\psi) - R(\psi^\star)]\left(1 - \frac{4\delta(\sigma^2 + C^2)}{n - w\delta}\right) - \log\frac{1}{\varepsilon}\right] \le \varepsilon. \tag{5.6}$$

Next, we use a standard PAC-Bayesian approach (as developed in Audibert (2004a); Catoni (2004, 2007); Alquier (2008)). For any prior probability π on $(\Theta, \mathcal{T})$,

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

By the Fubini-Tonelli theorem

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

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

$$\mathbb{E} \int \exp \left[\delta [R(\psi) - R(\psi^\star)] \left(1 - \frac{4\delta(\sigma^2 + C^2)}{n - w\delta} \right) + \delta [R_n(\psi^\star) - R_n(\psi)] - \log \frac{\mathrm{d}\rho}{\mathrm{d}\pi}(\psi) - \log \frac{1}{\varepsilon} \right] \rho(\mathrm{d}\psi) \leq \varepsilon. \quad (5.7)$$

Recalling that $\mathbb{E}$ stands for the expectation computed with respect to $\mathbb{P}$, the integration symbol may be omitted and we get

$$\mathbb{E} \exp \left[\delta [R(\psi) - R(\psi^\star)] \left(1 - \frac{4\delta(\sigma^2 + C^2)}{n - w\delta} \right) + \delta [R_n(\psi^\star) - R_n(\psi)] - \log \frac{\mathrm{d}\rho}{\mathrm{d}\pi}(\psi) - \log \frac{1}{\varepsilon} \right] \leq \varepsilon.$$

Using the elementary inequality $\exp(\delta x) \geq \mathbb{1}_{\mathbb{R}_+}(x)$, we get, with $\mathbb{P}$-probability at most ε

$$\left(1 - \frac{4\delta(\sigma^2 + C^2)}{n - w\delta} \right) [R(\psi) - R(\psi^\star)] \geq R_n(\psi) - R_n(\psi^\star) + \frac{\log \frac{\mathrm{d}\rho}{\mathrm{d}\pi}(\psi) + \log \frac{1}{\varepsilon}}{\delta}.$$

Taking $\delta < n/[w + 4(\sigma^2 + C^2)]$ implies

$$1 - \frac{4\delta(\sigma^2 + C^2)}{n - w\delta} > 0,$$

and with $\mathbb{P}$-probability at least $1 - \varepsilon$,

$$R(\psi) - R(\psi^\star) \leq \frac{1}{1 - \frac{4\delta(\sigma^2 + C^2)}{n - w\delta}} \left(R_n(\psi) - R_n(\psi^\star) + \frac{\log \frac{\mathrm{d}\rho}{\mathrm{d}\pi}(\psi) + \log \frac{1}{\varepsilon}}{\delta} \right).$$

□

Lemma 5.4. *Let A1 and A2 hold. Set* $w = 8C \max(L, C)$, $\delta \in (0, n/[w + 4(\sigma^2 + C^2)])$ *and* $\varepsilon \in (0, 1)$. *Then with* $\mathbb{P}$*-probability at least* $1 - \varepsilon$

$$\int R_n(\psi)\rho(\mathrm{d}\psi) - R_n(\psi^\star) \leq \left[1 + \frac{4\delta(\sigma^2 + C^2)}{n - w\delta} \right] \left[\int R(\psi)\rho(\mathrm{d}\psi) - R(\psi^\star) \right] + \frac{\mathcal{KL}(\rho, \pi) + \log \frac{1}{\varepsilon}}{\delta}. \quad (5.8)$$

Proof. Set $\psi \in \mathbb{F}$ and $Z_i = (Y_i - \psi(\mathbf{X}_i))^2 - (Y_i - \psi^\star(\mathbf{X}_i))^2$, $i \in \{1, \ldots, n\}$. Since $Z_i = -T_i$ where T_i is defined in (5.2), using the same arguments that lead to

(5.6), we get that for any $\delta \in (0, \in n/w)$ and $\varepsilon \in (0,1)$

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

Using Jensen's inequality, we get

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

Since $\exp(\delta x)\ge \mathbb{1}_{\mathbb{R}_+}(x)$, we obtain with $\mathbb{P}$-probability at most ε

$$\left[-\int R(\psi)\rho(\mathrm{d}\psi)+R(\psi^\star)\right]\left(1+\frac{4\delta(\sigma^2+C^2)}{n-w\delta}\right)+\int R_n(\psi)\rho(\mathrm{d}\psi)$$
$$-R_n(\psi^\star)-\frac{\mathcal{KL}(\rho,\pi)+\log\frac{1}{\varepsilon}}{\delta}\ge 0.$$

Taking $\delta < n/[w+4(\sigma^2+C^2)]$ yields (5.8). □

Lemma 5.5. *Let A1 and A2 hold. Set $w = 8C\max(L,C)$, $\delta \in (0, n/[w+4(\sigma^2+C^2)])$ and $\varepsilon\in(0,1)$. Then with $\mathbb{P}$-probability at least $1-\varepsilon$*

$$\int R(\psi)\rho(\mathrm{d}\psi)-R(\psi^\star)\le\frac{1}{1-\frac{4\delta(\sigma^2+C^2)}{n-w\delta}}\left(\int R_n(\psi)\rho(\mathrm{d}\psi)-R_n(\psi^\star)\right.$$
$$\left.+\frac{\mathcal{KL}(\rho,\pi)+\log\frac{1}{\varepsilon}}{\delta}\right).$$

Proof. Recall (5.7). By Jensen's inequality,

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

Using $\exp(\delta x)\ge\mathbb{1}_{\mathbb{R}_+}(x)$ yields the expected result. □

Theorem 5.1. *Let $\hat\psi$ and $\hat\psi^{\mathrm{A}}$ be realizations of the Gibbs estimators defined by (2.2)–(2.3), respectively. Let A1 and A2 hold. Set $w = 8C\max(L,C)$ and $\delta = n\ell/[w+4(\sigma^2+C^2)]$, for $\ell\in(0,1)$, and let $\varepsilon\in(0,1)$. Then with probability*

at least $1-2\varepsilon$,

$$\left.\begin{array}{l} R(\hat{\psi})-R(\psi^\star) \\ R(\hat{\psi}^{\mathrm{A}})-R(\psi^\star) \end{array}\right\} \leq \mathcal{D} \inf_{\rho\in\mathcal{M}^1_{+,\pi}(\Theta,\mathcal{T})} \left\{ \int R(\psi)\rho(\mathrm{d}\psi) - R(\psi^\star) + \frac{\mathcal{KL}(\rho,\pi)+\log\frac{1}{\varepsilon}}{n} \right\}, \quad (5.9)$$

where $\mathcal{D}$ *is a constant depending only upon* w, σ, C *and* ℓ.

Proof. Recall that the randomized Gibbs estimator $\hat{\Psi}$ is sampled from ρ_δ. Denote by $\hat{\psi}$ a realization of the variable $\hat{\Psi}$. By Lemma 5.3, with $\mathbb{P}$-probability at least $1-\varepsilon$,

$$R(\hat{\psi})-R(\psi^\star) \leq \frac{1}{1-\frac{4\delta(\sigma^2+C^2)}{n-w\delta}} \left(R_n(\hat{\psi})-R_n(\psi^\star) + \frac{\log\frac{\mathrm{d}\rho_\delta}{\mathrm{d}\pi}(\hat{\psi})+\log\frac{1}{\varepsilon}}{\delta} \right).$$

Note that

$$\begin{aligned} \log\frac{\mathrm{d}\rho_\delta}{\mathrm{d}\pi}(\hat{\psi}) &= \log\frac{\exp[-\delta R_n(\hat{\psi})]}{\int\exp[-\delta R_n(\psi)]\pi(\mathrm{d}\psi)} \\ &= -\delta R_n(\hat{\psi}) - \log\int\exp[-\delta R_n(\psi)]\pi(\mathrm{d}\psi). \end{aligned}$$

Thus, with $\mathbb{P}$-probability at least $1-\varepsilon$,

$$R(\hat{\psi})-R(\psi^\star) \leq \frac{1}{1-\frac{4\delta(\sigma^2+C^2)}{n-w\delta}} \left(-R_n(\psi^\star) - \frac{1}{\delta}\log\int\exp[-\delta R_n(\psi)]\pi(\mathrm{d}\psi) + \frac{1}{\delta}\log\frac{1}{\varepsilon} \right).$$

By Lemma 5.2, with $\mathbb{P}$-probability at least $1-\varepsilon$,

$$R(\hat{\psi})-R(\psi^\star) \leq \frac{1}{1-\frac{4\delta(\sigma^2+C^2)}{n-w\delta}} \inf_{\rho\in\mathcal{M}^1_{+,\pi}(\Theta,\mathcal{T})} \left(\int R_n(\psi)\rho(\mathrm{d}\psi) - R_n(\psi^\star) + \frac{\mathcal{KL}(\rho,\pi)+\log\frac{1}{\varepsilon}}{\delta} \right).$$

Finally, by Lemma 5.4, with $\mathbb{P}$-probability at least $1-2\varepsilon$,

$$R(\hat{\psi})-R(\psi^\star) \leq \frac{1+\frac{4\delta(\sigma^2+C^2)}{n-w\delta}}{1-\frac{4\delta(\sigma^2+C^2)}{n-w\delta}} \inf_{\rho\in\mathcal{M}^1_{+,\pi}(\Theta,\mathcal{T})} \left\{ \int R(\psi)\rho(\mathrm{d}\psi) - R(\psi^\star) + \frac{2}{1+\frac{4\delta(\sigma^2+C^2)}{n-w\delta}} \frac{\mathcal{KL}(\rho,\pi)+\log\frac{1}{\varepsilon}}{\delta} \right\}.$$

Apply Lemma 5.5 with the Gibbs posterior probability defined by (5.1). With $\mathbb{P}$-probability at least $1-\varepsilon$,

$$\int R(\psi)\rho_\delta(\mathrm{d}\psi) - R(\psi^\star) \le \frac{1}{1-\frac{4\delta(\sigma^2+C^2)}{n-w\delta}} \left(\int R_n(\psi)\rho_\delta(\mathrm{d}\psi) - R_n(\psi^\star) + \frac{\mathcal{KL}(\rho_\delta,\pi) + \log\frac{1}{\varepsilon}}{\delta} \right).$$

Note that

$$\begin{aligned}\mathcal{KL}(\rho_\delta,\pi) &= \int \log \frac{\exp[-\delta R_n(\psi)]}{\int \exp[-\delta R_n(\psi)]\pi(\mathrm{d}\psi)}\rho_\delta(\mathrm{d}\psi) \\ &= -\delta \int R_n(\psi)\rho_\delta(\mathrm{d}\psi) - \log\left(\int \exp[-\delta R_n(\psi)]\pi(\mathrm{d}\psi)\right).\end{aligned}$$

By Lemma 5.2, with $\mathbb{P}^{\otimes n}$-probability at least $1-\varepsilon$

$$\int R(\psi)\rho_\delta(\mathrm{d}\psi) - R(\psi^\star) \le \frac{1}{1-\frac{4\delta(\sigma^2+C^2)}{n-w\delta}} \inf_{\rho\in\mathcal{M}^1_{+,\pi}(\Theta,\mathcal{T})} \left\{ \int R_n(\psi)\rho(\mathrm{d}\psi) - R_n(\psi^\star) + \frac{\mathcal{KL}(\rho,\pi)+\log\frac{1}{\varepsilon}}{\delta} \right\}.$$

By Lemma 5.4, with $\mathbb{P}^{\otimes n}$-probability at least $1-2\varepsilon$

$$\int R(\psi)\rho_\delta(\mathrm{d}\psi) - R(\psi^\star) \le \frac{1+\frac{4\delta(\sigma^2+C^2)}{n-w\delta}}{1-\frac{4\delta(\sigma^2+C^2)}{n-w\delta}} \inf_{\rho\in\mathcal{M}^1_{+,\pi}(\Theta,\mathcal{T})} \left\{ \int R(\psi)\rho(\mathrm{d}\psi) - R(\psi^\star) + \frac{2}{1+\frac{4\delta(\sigma^2+C^2)}{n-w\delta}} \frac{\mathcal{KL}(\rho,\pi)+\log\frac{1}{\varepsilon}}{\delta} \right\}.$$

As R is a convex function, applying Jensen's inequality gives

$$\int R(\psi)\rho_\delta(\mathrm{d}\psi) \ge R(\hat{\psi}^{\mathrm{A}}).$$

Finally, note that

$$\frac{1+\frac{4\delta(\sigma^2+C^2)}{n-w\delta}}{1-\frac{4\delta(\sigma^2+C^2)}{n-w\delta}} = 1 + \frac{8\ell(\sigma^2+C^2)}{(1-\ell)(w+4\sigma^2+4C^2)}.$$

□

Proof of Theorem 2.1. Let $\rho \in \mathcal{M}^1_{+,\pi}(\Theta,\mathcal{T})$. For any $A \in \mathcal{T}$, note that $\rho(A) = \sum_{\mathbf{m}\in\mathcal{M}} \rho_{\mathbf{m}}(A)$ where $\rho_{\mathbf{m}}(\cdot) = \rho(\cdot \cap \Theta_{\mathbf{m}})$, the trace of ρ on $\Theta_{\mathbf{m}}$. By Theorem 5.1,

with $\mathbb{P}$-probability at least $1-2\varepsilon$

$$R(\hat{\psi}) - R(\psi^\star) \leq \mathcal{D} \inf_{\mathbf{m}\in\mathcal{M}} \inf_{\rho\in\mathcal{M}^1_{+,\pi}(\Theta,\mathcal{T})} \left\{ \int R(\psi)\rho_{\mathbf{m}}(\mathrm{d}\psi) - R(\psi^\star) + \frac{\mathcal{KL}(\rho_{\mathbf{m}},\pi) + \log\frac{1}{\varepsilon}}{n} \right\}. \quad (5.10)$$

Note that for any $\rho \in \mathcal{M}^1_{+,\pi}(\Theta,\mathcal{T})$ and any $\mathbf{m}\in\mathcal{M}$,

$$\mathcal{KL}(\rho_{\mathbf{m}},\pi) = \int \log\left(\frac{\mathrm{d}\rho_{\mathbf{m}}}{\mathrm{d}\pi_{\mathbf{m}}}\right)\mathrm{d}\rho_{\mathbf{m}} + \int \log\left(\frac{\mathrm{d}\pi_{\mathbf{m}}}{\mathrm{d}\pi}\right)\mathrm{d}\rho_{\mathbf{m}}$$

$$= \mathcal{KL}(\rho_{\mathbf{m}},\pi_{\mathbf{m}}) + \log(1/\alpha)\sum_{j\in S(\mathbf{m})} m_j + \log\binom{p}{|S(\mathbf{m})|} + \log\left(\frac{1-\left(\frac{\alpha}{1-\alpha}\right)^{p+1}}{1-\frac{\alpha}{1-\alpha}}\right).$$

Next, using the elementary inequality $\log\binom{n}{k} \leq k\log(ne/k)$ and that $\frac{\alpha}{1-\alpha} < 1$,

$$\mathcal{KL}(\rho_{\mathbf{m}},\pi) \leq \mathcal{KL}(\rho_{\mathbf{m}},\pi_{\mathbf{m}}) + \log(1/\alpha)\sum_{j\in S(\mathbf{m})} m_j + |S(\mathbf{m})|\log\left(\frac{pe}{|S(\mathbf{m})|}\right) + \log\left(\frac{1-\alpha}{1-2\alpha}\right).$$

We restrict the set of all probabilities absolutely continuous with respect to $\pi_{\mathbf{m}}$ to uniform probabilities on the ball $\mathcal{B}^1_{\mathbf{m}}(\mathbf{x},\zeta)$, with $\mathbf{x}\in\mathcal{B}^1_{\mathbf{m}}(0,C)$ and $0<\zeta\leq C-\|\theta\|_1$. Such a probability is denoted by $\mu_{\mathbf{x},\zeta}$. With $\mathbb{P}$-probability at least $1-2\varepsilon$, it yields that

$$R(\hat{\psi}) - R(\psi^\star) \leq \mathcal{D} \inf_{\mathbf{m}\in\mathcal{M}} \inf_{\theta\in\mathcal{B}^1_{\mathbf{m}}(0,C)} \inf_{\mu_{\theta,\zeta},0<\zeta\leq C-\|\theta\|_1} \left\{ \int R(\psi_{\bar{\theta}})\mu_{\theta,\zeta}(\mathrm{d}\bar{\theta}) - R(\psi^\star) + \frac{1}{n}\left[\mathcal{KL}(\mu_{\theta,\zeta},\pi_{\mathbf{m}}) + \log\frac{1}{\varepsilon} + |S(\mathbf{m})|\log\left(\frac{p}{|S(\mathbf{m})|}\right) + \sum_{j\in S(\mathbf{m})} m_j\right]\right\}.$$

Next, note that

$$\mathcal{KL}(\mu_{\theta,\zeta},\pi_{\mathbf{m}}) = \log\left(\frac{V_{\mathbf{m}}(C)}{V_{\mathbf{m}}(\zeta)}\right) = \log\left(\frac{C}{\zeta}\right)\sum_{j\in S(\mathbf{m})} m_j.$$

Note also that

$$\int R(\psi_{\bar{\theta}})\mu_{\theta,\zeta}(\mathrm{d}\bar{\theta}) = \int \mathbb{E}\left[Y_1 - \psi_{\bar{\theta}}(\mathbf{X}_1)\right]^2 \mu_{\theta,\zeta}(\mathrm{d}\bar{\theta})$$

$$= \int \mathbb{E}\left[Y_1 - \psi_\theta(\mathbf{X}_1) + \psi_\theta(\mathbf{X}_1) - \psi_{\bar{\theta}}(\mathbf{X}_1)\right]^2 \mu_{\theta,\zeta}(\mathrm{d}\bar{\theta}),$$

and

$$\int \mathbb{E}\left[Y_1 - \psi_\theta(\mathbf{X}_1) + \psi_\theta(\mathbf{X}_1) - \psi_{\bar\theta}(\mathbf{X}_1)\right]^2 \mu_{\theta,\zeta}(\mathrm{d}\bar\theta)$$
$$= \int R(\psi_\theta)\mu_{\theta,\zeta}(\mathrm{d}\bar\theta) + \int \mathbb{E}\left[\psi_\theta(\mathbf{X}_1) - \psi_{\bar\theta}(\mathbf{X}_1)\right]^2 \mu_{\theta,\zeta}(\mathrm{d}\bar\theta)$$
$$+ 2\int \mathbb{E}\{[Y_1 - \psi_\theta(\mathbf{X}_1)][\psi_\theta(\mathbf{X}_1) - \psi_{\bar\theta}(\mathbf{X}_1)]\}\mu_{\theta,\zeta}(\mathrm{d}\bar\theta).$$

Since $\bar\theta \in \mathcal{B}^1_{\mathbf{m}}(\theta,\zeta)$,

$$\int \mathbb{E}\left[\psi_\theta(\mathbf{X}_1) - \psi_{\bar\theta}(\mathbf{X}_1)\right]^2 \mu_{\theta,\zeta}(\mathrm{d}\bar\theta)$$
$$= \int \mathbb{E}\left[\sum_{j\in S(\mathbf{m})}\sum_{k=1}^{m_j}(\theta_{jk} - \bar\theta_{jk})\varphi_k(X_{1j})\right]^2 \mu_{\theta,\zeta}(\mathrm{d}\bar\theta)$$
$$\le \|\theta - \bar\theta\|_1^2 \max_k \|\varphi_k\|_\infty^2 \le \zeta^2,$$

and by the Fubini-Tonelli theorem,

$$2\int \mathbb{E}\{[Y_1 - \psi_\theta(\mathbf{X}_1)][\psi_\theta(\mathbf{X}_1) - \psi_{\bar\theta}(\mathbf{X}_1)]\}\mu_{\theta,\zeta}(\mathrm{d}\bar\theta)$$
$$= 2\,\mathbb{E}\left[[Y_1 - \psi_\theta(\mathbf{X}_1)]\int [\psi_\theta(\mathbf{X}_1) - \psi_{\bar\theta}(\mathbf{X}_1)]\mu_{\theta,\zeta}(\mathrm{d}\bar\theta)\right] = 0,$$

since $\int \psi_{\bar\theta}(\mathbf{X}_1)\mu_{\theta,\zeta}(\mathrm{d}\bar\theta) = \psi_\theta(\mathbf{X}_1)$. Consequently, as

$$\int R(\psi_\theta)\mu_{\theta,\zeta}(\mathrm{d}\bar\theta) = R(\psi_\theta),$$

we get

$$\int R(\psi_{\bar\theta})\mu_{\theta,\zeta}(\mathrm{d}\bar\theta) \le R(\psi_\theta) + \zeta^2.$$

So with $\mathbb{P}$-probability at least $1 - 2\varepsilon$,

$$R(\hat\psi) - R(\psi^\star) \le \mathcal{D} \inf_{\mathbf{m}\in\mathcal{M}} \inf_{\theta\in\mathcal{B}^1_{\mathbf{m}}(0,C)} \inf_{\mu_{\theta,\zeta},0<\zeta\le C-\|\theta\|_1} \Bigg\{ R(\psi_\theta) + \zeta^2 - R(\psi^\star)$$
$$+ \frac{1}{n}\left[\log(C/\zeta)\sum_{j\in S(\mathbf{m})} m_j + \log\frac{1}{\varepsilon} + |S(\mathbf{m})|\log\left(\frac{p}{|S(\mathbf{m})|}\right) + \sum_{j\in S(\mathbf{m})} m_j\right]\Bigg\}.$$

The function $t \mapsto t^2 + \log(C/t)\sum_{j\in S(\mathbf{m})} m_j/n$ is convex. Its minimum is unique and is reached for $t = [\sum_{j\in S(\mathbf{m})} m_j/(2n)]^{1/2}$. With $\mathbb{P}$-probability at least $1-2\varepsilon$,

$$R(\hat{\psi}) - R(\psi^\star) \leq \mathcal{D} \inf_{\mathbf{m}\in\mathcal{M}} \inf_{\theta\in\mathcal{B}^1_{\mathbf{m}}(0,C)} \left\{ R(\psi_\theta) - R(\psi^\star) \right.$$
$$\left. + |S(\mathbf{m})| \frac{\log(p/|S(\mathbf{m})|)}{n} + \frac{\log(n)}{n} \sum_{j\in S(\mathbf{m})} m_j + \frac{\log(1/\varepsilon)}{n} \right\},$$

where $\mathcal{D}$ is a constant depending only on w, σ, C, ℓ and α. As the same inequality holds for $\hat{\psi}^{\mathrm{A}}$, this concludes the proof. □

Proof of Theorem 2.2. Recall Theorem 2.1. A3 gives

$$R(\psi_\theta) - R(\psi^\star) = \int (\psi_\theta(\mathbf{x}) - \psi^\star(\mathbf{x}))^2 \mathrm{d}\,\mathcal{P}(\mathbf{x}) \leq B \int (\psi_\theta(\mathbf{x}) - \psi^\star(\mathbf{x}))^2 \mathrm{d}\mathbf{x}.$$

For any $\mathbf{m} \in \mathcal{M}$, define

$$\psi^\star_{\mathbf{m}} = \sum_{j\in S^\star} \sum_{k=1}^{m_j} \theta^\star_{jk} \varphi_k.$$

To proceed, we need to check that the projection of $\theta^\star$ onto model $\mathbf{m}$ lies in $\mathcal{B}^1_{\mathbf{m}}(0,C)$, *i.e.*,

$$\sum_{j\in S^\star} \sum_{k=1}^{m_j} |\theta^\star_{jk}| \leq C.$$

Using the Cauchy-Schwarz inequality, we get

$$\sum_{j\in S^\star} \sum_{k=1}^{m_j} |\theta^\star_{jk}| = \sum_{j\in S^\star} \sum_{k=1}^{m_j} k^{r_j} |\theta^\star_{jk}| k^{-r_j}$$
$$\leq \sum_{j\in S^\star} \left[\sqrt{\sum_{k=1}^{m_j} k^{2r_j} (\theta^\star_{jk})^2} \sqrt{\sum_{k=1}^{m_j} k^{-2r_j}} \right].$$

Since for any $t \geq 1$, $\sum_{k=1}^{m_j} k^{-2t} \leq \sum_{k=1}^{\infty} k^{-2t} = \pi^2/6$, the previous inequality yields

$$\sum_{j\in S^\star} \sum_{k=1}^{m_j} |\theta^\star_{jk}| \leq \frac{\pi}{\sqrt{6}} \sum_{j\in S^\star} \sqrt{d_j} \leq C.$$

Recalling (5.3) and A3, for a $\mathbf{m} \in \mathcal{M}$ we may now write that

$$\inf_{\theta\in\Theta_{\mathbf{m}}} R(\psi_\theta) - R(\psi^\star) \leq R(\psi^\star_{\mathbf{m}}) - R(\psi^\star) \leq B \int (\psi^\star(\mathbf{x}) - \psi^\star_{\mathbf{m}}(\mathbf{x}))^2 \mathrm{d}\mathbf{x}$$
$$= B \int \left(\sum_{j\in S^\star} \sum_{k=1+m_j}^{\infty} \theta^\star_{jk} \varphi_k(\mathbf{x}) \right)^2 \mathrm{d}\mathbf{x}.$$

As $\{\varphi_k\}_{k=1}^{\infty}$ forms an orthogonal basis,

$$B\int\left(\sum_{j\in S^\star}\sum_{k=1+m_j}^{\infty}\theta^\star_{jk}\varphi_k(\mathbf{x})\right)^2\mathrm{d}\mathbf{x} = B\sum_{j\in S^\star}\sum_{k=1+m_j}^{\infty}(\theta^\star_{jk})^2 \le B\sum_{j\in S^\star} d_j(1+m_j)^{-2r_j},$$

where the normalizing numerical factors are included in the now generic constant B. As a consequence, with $\mathbb{P}$-probability at least $1-2\varepsilon$,

$$R(\hat{\psi})-R(\psi^\star)\le \mathcal{D}\inf_{\mathbf{m}\in\mathcal{M}}\left\{B\sum_{j\in S^\star}\left\{d_j(1+m_j)^{-2r_j}+\frac{m_j}{n}\log(n)\right\} + |S^\star|\frac{\log(p/|S^\star|)}{n}+\frac{\log(1/\varepsilon)}{n}\right\},$$

where $\mathcal{D}$ is the same constant as in Theorem 2.1. For any $r\ge 2$, the function $t\mapsto d_j(1+t)^{-2r_j}+\frac{\log(n)}{n}t$ is convex and admits a minimum in $\left(\frac{\log(n)}{2r_jd_jn}\right)^{-\frac{1}{2r_j+1}}-1$. Accordingly, choosing $m_j\sim\left(\frac{\log(n)}{2r_jd_jn}\right)^{-\frac{1}{2r_j+1}}-1$ yields that with $\mathbb{P}$-probability at least $1-2\varepsilon$,

$$R(\hat{\psi})-R(\psi^\star)\le\mathcal{D}\left\{\sum_{j\in S^\star}d_j^{\frac{1}{2r_j+1}}\left(\frac{\log(n)}{2nr_j}\right)^{\frac{2r_j}{2r_j+1}}+|S^\star|\frac{\log\left(\frac{p}{|S^\star|}\right)}{n}+\frac{\log(1/\varepsilon)}{n}\right\},$$

where $\mathcal{D}$ is a constant depending only on α, w, σ, C, ℓ and B, and that ends the proof. □

Proof of Theorem 2.3. The proof is similar to the proof of Theorem 2.1. From (5.10) and for any $\rho\in\mathcal{M}^1_{+,\pi}(\Theta,\mathcal{T})$ and any $\mathbf{m}\in\mathcal{M}$,

$$\mathcal{KL}(\rho_{\mathbf{m}},\pi)=\mathcal{KL}(\rho_{\mathbf{m}},\pi_{\mathbf{m}})+\log(1/\alpha)|S(\mathbf{m})|+\log\binom{p}{|S(\mathbf{m})|} + \log\left(\frac{1-\left(\alpha\frac{1-\alpha^{K+1}}{1-\alpha}\right)^{p+1}}{1-\alpha\frac{1-\alpha^{K+1}}{1-\alpha}}\right)+\sum_{j\in S(\mathbf{m})}\log\binom{K}{|S(\mathbf{m}_j)|}.$$

Using the elementary inequality $\log\binom{n}{k}\le k\log(ne/k)$ and that $\alpha\frac{1-\alpha^{K+1}}{1-\alpha}\in(0,1)$ since $\alpha<1/2$,

$$\mathcal{KL}(\rho_{\mathbf{m}},\pi)\le\mathcal{KL}(\rho_{\mathbf{m}},\pi_{\mathbf{m}})+|S(\mathbf{m})|\left[\log(1/\alpha)+\log\left(\frac{pe}{|S(\mathbf{m})|}\right)\right] + \sum_{j\in S(\mathbf{m})}|S(\mathbf{m}_j)|\log\left(\frac{Ke}{|S(\mathbf{m}_j)|}\right)+\log\left(\frac{1-\alpha}{1-2\alpha}\right).$$

Thus with $\mathbb{P}$-probability at least $1-2\varepsilon$,

$$R(\hat{\psi}) - R(\psi^\star) \le \mathcal{D} \inf_{\mathbf{m}\in\mathcal{M}} \inf_{\theta\in\mathcal{B}^1_{\mathbf{m}}(0,C)} \inf_{\mu_{\theta,\zeta}, 0<\zeta\le C-\|\theta\|_1} \left\{ R(\psi_\theta) + \zeta^2 - R(\psi^\star) + \frac{1}{n}\left[[\log(C/\zeta) + \log(K)] \sum_{j\in S(\mathbf{m})} |S(\mathbf{m}_j)| + \log\frac{1}{\varepsilon} + |S(\mathbf{m})| \log\left(\frac{p}{|S(\mathbf{m})|}\right)\right]\right\}.$$

Hence with $\mathbb{P}$-probability at least $1-2\varepsilon$,

$$\left.\begin{array}{l} R(\hat{\psi}) - R(\psi^\star) \\ R(\hat{\psi}^{\mathrm{A}}) - R(\psi^\star) \end{array}\right\} \le \mathcal{D} \inf_{\mathbf{m}\in\mathcal{M}} \inf_{\theta\in\mathcal{B}^1_{\mathbf{m}}(0,C)} \left\{ R(\psi_\theta) - R(\psi^\star) + |S(\mathbf{m})| \frac{\log(p/|S(\mathbf{m})|)}{n} + \frac{\log(nK)}{n} \sum_{j\in S(\mathbf{m})} |S(\mathbf{m}_j)| + \frac{\log(1/\varepsilon)}{n} \right\},$$

where $\mathcal{D}$ is a numerical constant depending upon w, σ, C, ℓ and α. □

Acknowledgements

The authors are grateful to Gérard Biau and Éric Moulines for their constant implication, and to Christophe Giraud and Taiji Suzuki for valuable insights and comments. They also thank an anonymous referee and an associate editor for providing constructive and helpful remarks.

References

Alquier, P. (2006). Transductive and Inductive Adaptive Inference for Regression and Density Estimation PhD thesis, Université Paris 6 - UPMC.

Alquier, P. (2008). PAC-Bayesian Bounds for Randomized Empirical Risk Minimizers. *Mathematical Methods of Statistics* **17** 279–304. arXiv:0712.1698v3 MR2483458

Alquier, P. and Biau, G. (2011). Sparse Single-Index Model. To appear in Journal of Machine Learning Research. arXiv:1101.3229v2

Alquier, P. and Lounici, K. (2011). PAC-Bayesian Theorems for Sparse Regression Estimation with Exponential Weights. *Electronic Journal of Statistics* **5** 127–145. MR2786484

Audibert, J.-Y. (2004a). Aggregated estimators and empirical complexity for least square regression. *Annales de l'Institut Henri Poincaré: Probabilités et Statistiques* **40** 685–736. MR2096215

Audibert, J.-Y. (2004b). Théorie statistique de l'apprentissage: une approche PAC-Bayésienne PhD thesis, Université Paris 6 - UPMC.

Audibert, J.-Y. (2009). Fast learning rates in statistical inference through aggregation. *The Annals of Statistics* **37** 1591–1646. MR2533466

Audibert, J.-Y. and Catoni, O. (2010). Robust linear regression through PAC-Bayesian truncation. Submitted. arXiv:1010.0072v2

Audibert, J.-Y. and Catoni, O. (2011). Robust linear least squares regression. *The Annals of Statistics* **39** 2766–2794. MR2906886

Bickel, P. J., Ritov, Y. and Tsybakov, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics* **37** 1705–1732. arXiv:0801.1095v3 MR2533469

Bunea, F., Tsybakov, A. B. and Wegkamp, M. (2006). Aggregation and sparsity via ℓ_1-penalized least squares. In *Proceedings of the 19th annual conference on Computational Learning Theory* 379–391. Springer-Verlag. MR2280619

Bühlmann, P. and van de Geer, S. A. (2011). *Statistics for High-Dimensional Data*. Springer. MR2807761

Carlin, B. P. and Chib, S. (1995). Bayesian Model choice via Markov Chain Monte Carlo Methods. *Journal of the Royal Statistical Society, Series B* **57** 473–484.

Catoni, O. (2004). *Statistical Learning Theory and Stochastic Optimization. École d'Été de Probabilités de Saint-Flour XXXI – 2001*. Springer. MR2163920

Catoni, O. (2007). *PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning. Lecture notes – Monograph Series* **56**. Institute of Mathematical Statistics. MR2483528

Dalalyan, A. S. and Tsybakov, A. B. (2008). Aggregation by exponential weighting, sharp PAC-Bayesian bounds and sparsity. *Machine Learning* **72** 39–61. arXiv:0803.2839v1

Dalalyan, A. S. and Tsybakov, A. B. (2012). Sparse Regression Learning by Aggregation and Langevin Monte-Carlo. *Journal of Computer and System Sciences* **78** 1423–1443. arXiv:0903.1223v3 MR2926142

Giraud, C., Huet, S. and Verzelen, N. (2012). High-dimensional regression with unknown variance. To appear in Statistical Science. arXiv:1109.5587v2 MR2934907

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82** 711–732. MR1380810

Guedj, B. (2012). pacbpred: PAC-Bayesian Estimation and Prediction in Sparse Additive Models R package version 0.92. `http://cran.r-project.org/web/packages/pacbpred/index.html`

Hans, C., Dobra, A. and West, M. (2007). Shotgun Stochastic Search for "Large p" Regression. *Journal of the American Statistical Association* **102** 507–516. MR2370849

Hastie, T. and Tibshirani, R. (1986). Generalized Additive Models. *Statistical Science* **1** 297–318. MR0858512

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models. Monographs on Statistics and Applied Probability* **43**. Chapman & Hall/CRC. MR1082147

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning – Data mining, Inference, and Prediction*, Second ed. Springer. MR2722294

Härdle, W. K. (1990). *Applied nonparametric regression*. Cambridge University Press. MR1161622

Koltchinskii, V. and Yuan, M. (2010). Sparsity in multiple kernel learning. *The Annals of Statistics* **38** 3660–3695. MR2766864

Marin, J.-M. and Robert, C. P. (2007). *Bayesian Core: A Practical Approach to Computational Bayesian Statistics*. Springer. MR2289769

Massart, P. (2007). *Concentration Inequalities and Model Selection. École d'Été de Probabilités de Saint-Flour XXXIII – 2003*. Springer. MR2319879

McAllester, D. A. (1999). Some PAC-Bayesian Theorems. *Machine Learning* **37** 355–363. MR1811587

Meier, L., van de Geer, S. A. and Bühlmann, P. (2009). High-dimensional additive modeling. *The Annals of Statistics* **37** 3779–3821. arXiv:0806.4115 MR2572443

Meinshausen, N. and Yu, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics* **37** 246–270. arXiv:0806.0145v2 MR2488351

Meyn, S. and Tweedie, R. L. (2009). *Markov Chains and Stochastic Stability*, 2nd ed. Cambridge University Press. MR2509253

Petralias, A. (2010). Bayesian model determination and nonlinear threshold volatility models PhD thesis, Athens University of Economics and Business.

Petralias, A. and Dellaportas, P. (2012). An MCMC model search algorithm for regression problems. *Journal of Statistical Computation and Simulation* **0** 1-19.

R Core Team (2012). R: A Language and Environment for Statistical Computing, Vienna, Austria ISBN 3-900051-07-0. http://www.R-project.org/

Raskutti, G., Wainwright, M. J. and Yu, B. (2012). Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *Journal of Machine Learning Research* **13** 389-427. MR2913704

Ravikumar, P., Lafferty, J., Liu, H. and Wasserman, L. (2009). Sparse additive models. *Journal of the Royal Statistical Society, Series B* **71** 1009–1030. arXiv:0711.4555v2 MR2750255

Rigollet, P. (2006). Inégalités d'oracle, agrégation et adaptation PhD thesis, Université Paris 6 - UPMC.

Rigollet, P. and Tsybakov, A. B. (2012). Sparse estimation by exponential weighting. *Statistical Science* **27** 558-575.

Shawe-Taylor, J. and Williamson, R. C. (1997). A PAC analysis of a Bayes estimator. In *Proceedings of the 10th annual conference on Computational Learning Theory* 2–9. ACM.

Stone, C. J. (1985). Additive regression and other nonparametric models. *The Annals of Statistics* **13** 689–705. MR0790566

Suzuki, T. (2012). PAC-Bayesian Bound for Gaussian Process Regression and Multiple Kernel Additive Model. In *Proceedings of the 25th annual conference on Computational Learning Theory*.

Suzuki, T. and Sugiyama, M. (2012). Fast learning rates of Multiple kernel learning: trade-off between sparsity and smoothness. Submitted. arXiv.org/abs/1203.0565v1

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B* **58** 267–288. MR1379242

Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation. Statistics.* Springer. MR2724359

van de Geer, S. A. (2008). High-dimensional generalized linear models and the Lasso. *The Annals of Statistics* **36** 614–645. MR2396809