PAC-Bayesian Bounds for Sparse Regression Estimation with Exponential Weights

Pierre Alquier, Karim Lounici

PAC-BAYESIAN BOUNDS FOR SPARSE REGRESSION ESTIMATION WITH EXPONENTIAL WEIGHTS

PIERRE ALQUIER, KARIM LOUNICI

Abstract. We consider the sparse regression model where the number of parameters p is larger than the sample size n. The difficulty when considering high-dimensional problems is to propose estimators achieving a good compromise between statistical and computational performances. The BIC estimator for instance performs well from the statistical point of view [7] but can be computed for values of p of at most a few tens. The Lasso estimator is solution of a convex minimization problem. Hence it can be computed for large value of p. However stringent conditions on the design are required to establish the statistical properties of this estimator. Dalalyan and Tsybakov [14] propose a method achieving a good compromise between the statistical and computational aspects of the problem. Their estimator can be computed for reasonably large p and satisfies nice statistical properties under weak assumptions on the design. However, [14] concerns only the empirical risk and proposes only results in expectation. In this paper, we propose an aggregation procedure similar to that of [14] but with improved statistical performances. Our main result concerns the expected risk and is given in probability. We also propose a MCMC method to compute our estimator for reasonably large values of p.

MSC 2000 subject classification: Primary: 62J07; Secondary: 62J05, 62G08, 62F15, 62B10, 68T05.

Key words and phrases: Sparsity Oracle Inequality, High-dimensional Regression, Exponential Weights, PAC-bayesian, RJMCMC

1. Introduction

We observe n independent pairs $(X_1, Y_1), ..., (X_n, Y_n) \in \mathcal{X} \times \mathbb{R}$ (where $\mathcal{X}$ is any measurable set) such that

$$Y_i = f(X_i) + W_i,\ 1 \leqslant i \leqslant n, \tag{1.1}$$

where $f : \mathcal{X} \to \mathbb{R}$ is the unknown regression function and the noise variables $W_1, \dots, W_n$ are independent of $X_1, \dots, X_n$ and such that $\mathbb{E}W_i = 0$ and $\mathbb{E}W_i^2 \leqslant \sigma^2$ for some $\sigma^2 > 0$ and any $1 \leqslant i \leqslant n$. The distribution of the sample is denoted by $\mathbb{P}$, the corresponding expectation is denoted by $\mathbb{E}$. For any function $g : \mathcal{X} \to \mathbb{R}$ define $\|g\|_n = \left(\sum_{i=1}^n g(X_i)^2/n\right)^{1/2}$ and $\|g\| = \left(\mathbb{E}\|g\|_n^2\right)^{1/2}$. Let $\mathcal{F} = \{\phi_1, \dots, \phi_p\}$ be a set of functions $\phi_j : \mathcal{X} \to \mathbb{R}$ such that w.l.o.g. $\|\phi_j\| = 1$ for any j (this assumption can be relaxed). For any $\theta \in \mathbb{R}^p$ define $f_\theta = \sum_{j=1}^p \theta_j \phi_j$ and the risk

$$R(\theta) = \mathbb{E}[r(\theta)]$$

where

$$r(\theta) = \frac{1}{n}\sum_{i=1}^n \Big(Y_i - f_\theta(X_i)\Big)^2 = \|Y - f_\theta\|_n^2$$

Date: March 21, 2010.

and $Y = (Y_1, ..., Y_n)^T$. Let us choose $\overline{\theta} \in \arg\min_{\mathbb{R}^p} R$. Note that the minimum may not be unique but this does not matter since we consider here the prediction problem, that is the estimation of $\min_{\mathbb{R}^p} R$.

It is a known fact that the least-square estimator $\hat{\theta}_{LSE} \in \arg\min_{\Theta} r$ performs poorly in high-dimension $p > n$ since it satisfies

$$\mathbb{E}\left(R(\hat{\theta}_{LSE}) - R(\overline{\theta})\right) \leqslant C\sigma^2,$$

for some $C > 0$. Assume now there exists a vector $\overline{\theta} \in \arg\min_{\mathbb{R}^p} R$ with a number of nonzero coordinates $p_0 \leq n$. If the indexes of these coordinates are known, then we can construct an estimator $\hat{\theta}_n^0$ such that

$$\mathbb{E}\left(R(\hat{\theta}_n^0) - R(\overline{\theta})\right) \leqslant C\sigma^2 \frac{p_0}{n}.$$

The estimator $\hat{\theta}_n^0$ is called oracle estimator since the set of indexes of the nonzero coordinates of $\overline{\theta}$ is unknown in practice. The problematic is now to build an estimator, when the set of nonzero coordinates of $\overline{\theta}$ is unknown, with statistical performances closed to that of the oracle estimator $\hat{\theta}_n^0$.

A possible approach is to consider solutions of penalized empirical risk minimization problems:

$$\hat{\theta}_{pen} \in \arg\min_{\theta\in\Theta} \left[r(\theta) + pen(\theta)\right],$$

where the penalization $pen(\theta)$ is proportional to the number of nonzero components of θ as for instance C_p, AIC and BIC criteria [29, 1, 34]. Bunea, Tsybakov and Wegkamp [7] established for the BIC estimator $\hat{\theta}_n^{BIC}$ the following result

$$\mathbb{E}\left(R(\hat{\theta}_n^{BIC}) - R(\overline{\theta})\right) \leqslant C\sigma^2 \frac{p_0}{n} \log p,$$

for some $C > 0$. The above inequality is called sparsity oracle inequality because the upper bound is linear in p_0 the number of nonzero coordinates of $\overline{\theta}$ and logarithmic in p the total number of coordinates. Despite some good statistical properties, these estimators can be computed in practice for p of the order at most a few tens since they are solutions of nonconvex optimization problems. Considering convex penalty function leads to computationally feasible optimization problems. A popular example is the Lasso estimator (cf. Frank and Friedman [19] and Tibshirani [36]) with the penalty term $pen(\theta) = \lambda \sum_{j=1}^p |\theta_j|$ where $\lambda > 0$ is some regularization parameter. However, the Lasso estimator requires strong assumptions on the dictionary $\mathcal{D}$ and the design to establish the statistical properties. Bunea, Tsybakov and Wegkamp [6] assume a mutual coherence condition on the dictionary, Bickel, Ritov and Tsybakov [6] and Koltchinskii [25] established their results under a restricted eigenvalue condition. An alternative to penalization, the LOL method, proposed by Kerkyacharian, Mougeot, Picard and Tribouley [24], also requires the mutual coherence assumption.

Simultaneously, the PAC-Bayesian approach for regression estimation was developed by Audibert [4, 5] and Alquier [2, 3], based on previous works in the classification context by Catoni [10, 11, 12], Mc Allester [31], Shawe-Taylor and Williamson [35]. This framework is very interesting to study the excess risk $R(\cdot) - R(\overline{\theta})$ in the regression context since it requires very weak conditions on the dictionary. However, the methods of these papers are not computationally feasible when p becomes large. Dalalyan and Tsybakov [14, 15, 16, 17] propose an exponential weights procedure related to the PAC-Bayesian approach with good statistical and computational performances. However they consider deterministic design, thus they established their statistical result only for the empirical excess risk instead of the true excess risk $R(\cdot) - R(\overline{\theta})$.

In this paper, we propose two exponential weights estimation procedures. The first one is an exponential weights combination of the least squares estimators in all the possible sub-models. Note that in the literature on aggregation, the elements of the dictionary are often preliminary estimators computed with a frozen fraction of the initial sample so that these estimators are considered deterministic functions. Next the aggregate is computed using this dictionary and the remaining data. This scheme is referred to as 'data splitting'. Experiments were carried out on the necessity of data splitting. They tend to show that data splitting may not be necessary in practice in some situations. However, to our knowledge, these observations lack theoretical confirmation. In this paper, we do not use data splitting scheme. The same sample is used to build the preliminary estimators and to construct the aggregate and we establish for this procedure a sparsity oracle inequality with optimal bounds in the deterministic design case in the spirit of [14]. For the second procedure, the design is assumed to be random. We use the PAC-bayesian techniques of Catoni [12] to produce an estimator satisfying a sparsity oracle inequality for the true excess risk. Then we propose efficient Monte Carlo computation algorithms to compute both estimators. Our algorithms are inspired from the computational bayesian theory for model selection, see George and McCulloch [21], Casella and Moreno [8] or Cui and George [13] among others. A review on Monte Carlo algorithms applied to bayesian analysis and variable selection can be found in Casella and Robert [9] or George [20]. Both estimators are computed using the Hastings-Metropolis algorithm. For the second one, we use a particular form of the RJMCMC ("Reversible Jump Markov Chain Monte Carlo") method proposed by Green [22]. Note that in a work parallel to ours, Rigollet and Tsybakov [33] consider exponentially weighted aggregates with discrete priors and suggest another version of the Metropolis-Hastings algorithm to compute them.

The paper is organized as follows. In Section 2 we define a general aggregation procedure and derive a sparsity oracle inequality in the deterministic design case. In Section 3 the design can be either deterministic or random. We propose a modification of the first aggregation procedure for which we can establish a sparsity oracle inequality in probability for the true excess risk. Section 4 is devoted to the RJMCMC algorithm used to effectively implement our estimators. In Section 5 we carry out a simulation study and compare the performances of our methods with the Lasso. Finally, Section 6 contains the proofs of our results.

2. Sparsity Oracle Inequality in Expectation

Throughout this section, we assume that the design is deterministic and the noise variables $W_1, \dots, W_n$ are i.i.d. gaussian $N(0, \sigma^2)$. For any integer $d \geq 2$, real $p > 0$ and vector $z \in \mathbb{R}^d$ define $|z|_p = (\sum_{j=1}^M |\theta_j^p|)^{1/p}$ and $|z|_\infty = \max_{1 \leq j \leq M} |\theta_j|$.

For any $J \subset \{1, ..., p\}$ and $K > 0$ define

$$\Theta(J) = \left\{ \theta \in \mathbb{R}^p : \quad \theta_j = 0 \forall j \notin J \right\}, \tag{2.1}$$

and

$$\Theta_K(J) = \left\{ \theta \in \mathbb{R}^p : \quad |\theta|_1 \leq K \quad \text{and} \quad \theta_j = 0 \forall j \notin J \right\}. \tag{2.2}$$

For the sake of simplicity we will write $\Theta_K = \Theta_K(\{1, ..., p\})$.

For any subset $J \subset \{1, ..., p\}$ define

$$\hat{\theta}_J \in \arg \min_{\theta \in \Theta(J)} r(\theta).$$

The aggregate $\hat{f}_n$ is defined as follows

$$\hat{f}_n = f_{\hat{\theta}_n}, \quad \hat{\theta}_n = \hat{\theta}_n(\lambda, \pi) \stackrel{\triangle}{=} \frac{\sum_{k=0}^{n} \sum_{\substack{J \subset \{1,\dots,p\} \\ |J|=k}} \pi_J e^{-\lambda\left(r(\hat{\theta}_J)+\frac{2\sigma^2|J|}{n}\right)} \hat{\theta}_J}{\sum_{k=0}^{n} \sum_{\substack{J \subset \{1,\dots,p\} \\ |J|=k}} \pi_J e^{-\lambda r(\hat{\theta}_J)+\frac{2\sigma^2|J|}{n}}} \tag{2.3}$$

where $\lambda > 0$ is the temperature parameter, π is the prior probability distribution on $\mathcal{P}(\{1,\dots,p\})$ (the set of all subset of $\{1,...,p\}$), that is, for any $J \in \{1,...,p\}$, $\pi_J \geq 0$ and $\sum_{J\in\mathcal{P}(\{1,\dots,p\})} \pi_J = 1$). In Section 4 we show that the estimator $\hat{\theta}_n$ can be computed in reasonable time even when p is large using a MCMC scheme, namely, Hastings-Metropolis algorithm. The parameters π and λ must be tuned in a suitable way. The choice of π is discussed below. The choice of the temperature parameter λ is discussed in Section 5.

We now state the main results of this section.

Proposition 1. *Assume that the noise variables $W_1,\dots,W_n$ are i.i.d. $N(0,\sigma^2)$. Then the aggregate $\hat{\theta}_n$ defined by (2.3) with $0 < \lambda \leqslant \frac{n}{4\sigma^2}$ satisfies*

$$\mathbb{E}\left[r(\hat{\theta}_n)\right] \leqslant \min_{\substack{J \in \mathcal{P}(\{1,\dots,p\}) \\ |J| \leqslant n}} \left\{\mathbb{E}[r(\hat{\theta}_J)] + \frac{1}{\lambda}\log\left(\frac{1}{\pi_J}\right)\right\}. \tag{2.4}$$

Proposition 1 can be compared with [14]. We emphasize that contrary to [14] which considered the pure aggregation problem with the dictionary $\mathcal{F}$, here we consider the dictionary $\mathcal{D} = \{\hat{f}_{\hat{\theta}_J},\ J \in \mathcal{P}(\{1,\dots,p\}),\ |J| \leqslant n\}$ which depends on the data. Note also that the same data set is used to build the dictionary $\mathcal{D}$ and the aggregate $\hat{f}_n$, which shows in this case that data splitting is not necessary.

Proposition 1 holds true for any prior π. We now define a prior yielding an interesting sparsity oracle inequality. Fix $\alpha \in (0,1)$. Define π as follows

$$\pi_J = \frac{\alpha^{|J|}}{1-\alpha}\binom{p}{|J|}^{-1}, \quad \forall J \in \mathcal{P}(\{1,...,p\}). \tag{2.5}$$

We have the following Theorem

Theorem 1. *Assume that the noise variables $W_1,\dots,W_n$ are i.i.d. $N(0,\sigma^2)$. Then the aggregate $\hat{f}_n = f_{\hat{\theta}_n}$, with $\lambda = \frac{n}{4\sigma^2}$ and π taken as in (2.5), satisfies*

$$\begin{aligned}\mathbb{E}\left[\|\hat{f}_n - f\|_n^2\right] \\ \leqslant \min_{\theta\in\mathbb{R}^p}\left\{\|f_\theta - f\|_n^2 + \frac{\sigma^2|J(\theta)|}{n}\left(4\log\left(\frac{pe}{|J(\theta)|\alpha}\right)+1\right) + \frac{4\sigma^2\log(1-\alpha)}{n}\right\},\end{aligned} \tag{2.6}$$

where for any $\theta \in \mathbb{R}^p$ $J(\theta) = \{j : \theta_j \neq 0\}$.

Tsybakov [37] introduced the notion of optimal rate of aggregation adapting existing tools from the minimax theory. Note that the rate we derive in Theorem 1 is the optimal rate of sparse aggregation if $M \geqslant n$ and $|J(\theta)| \leqslant \sqrt{n}$ (this is a straightforward consequence of Theorem 5 in [28]). Theorem 1 improves upon [14] where a similar sparsity oracle inequality is derived with sub-optimal rate of aggregation and the restriction $\theta \in \Theta_K$ for some $K > 0$ in the right-hand-side whereas our result holds for any $\theta \in \mathbb{R}^p$.

3. Sparsity Oracle Inequality in Probability

From now on, the design can be either deterministic or random. We make the mild assumption that $L := \max_{1\leq j\leq M} \|\phi_j\|_\infty < \infty$.

We also make the following assumption on the noise in this section.

Assumption 1. The noise variables $W_1, \dots, W_n$ are independent and independent from $X_1, \dots, X_n$. We assume also that there exists two known constants $\sigma > 0$ and $\xi > 0$ such that

$$\mathbb{E}(W_i^2) \leq \sigma^2$$

$$\forall k \geq 3, \quad \mathbb{E}(|W_i|^k) \leq \sigma^2 k! \xi^{k-2}.$$

The estimation method is a version of the Gibbs estimator introduced by Catoni [11, 12]. Fix $K \geq 1$ and $c > 0$. First we define the prior probability distribution as follows. For any $J \subset \{1, ..., p\}$ let u_J denote the uniform measure on $\Theta_{K+c}(J)$. We define

$$\mathrm{m}(d\theta) = \sum_{J \subset \{1,\dots,p\}} \pi_J \mathrm{u}_J(d\theta)$$

with π taken as in (2.5).

We are now ready to define our estimator. For any $\lambda > 0$ we consider the probability measure $\tilde{\rho}$ admitting the following density w.r.t. the probability measure m

$$\frac{d\tilde{\rho}}{d\mathrm{m}}(\theta) = \frac{e^{-\lambda r(\theta)}}{\int_{\Theta_K} e^{-\lambda r} d\mathrm{m}}. \tag{3.1}$$

The aggregate $\tilde{f}_n$ is defined as follows

$$\tilde{f}_n = f_{\tilde{\theta}_n}, \quad \tilde{\theta}_n = \tilde{\theta}_n(\lambda, \mathrm{m}) = \int_{\Theta_K} \theta \tilde{\rho}_\lambda(d\theta). \tag{3.2}$$

The practical computation of $\tilde{\theta}_n$ is discussed in Section 4, using Green's [20] RJMCMC method.

Define

$$\mathcal{C}_1 = \left[8\sigma^2 + (2\|f\|_\infty + L(2K + c))^2\right] \vee \left[8[\xi + (2\|f\|_\infty + L(2K + c))]L(2K + c)\right],$$

and

$$\mathcal{C}_2 = L[2\|f\|_\infty + L(2K + c) + 2\sigma].$$

We can now state the main result of this section.

Theorem 2. *Let Assumption 1 be satisfied. Take $K > 1$, $c = n^{-1}$ and $\lambda = \lambda^* = \frac{n}{2\mathcal{C}_1}$. Then we have, for any $\varepsilon \in (0, 1)$, with probability at least $1 - \epsilon$,*

$$R(\tilde{\theta}_\lambda) \leq \min_{\theta \in \Theta_K} \left\{ R(\theta) + \frac{3\mathcal{C}_2}{n} + \frac{8\mathcal{C}_1}{n} \left[|J(\theta)| \log(K + c) + \left(|J(\theta)| \log\left(\frac{enp}{\alpha |J(\theta)|}\right) + \log\left(\frac{2\sqrt{1-\alpha}}{\varepsilon}\right)\right)\right]\right\}.$$

The choice $\lambda = \lambda^*$ comes from the optimization of a (rather pessimistic) upper bound on the risk R (see Inequality (6.5) below). However this choice is not necessarily the best choice in practice even though it gives the good order of magnitude for λ. Section 5 illustrates this point. The practitioner may use cross-validation to properly tune the temperature parameter.

"Localization" techniques introduced by Catoni [11] would allow us to replace the $\log np$ term by $\log p$. We refrain from implementing these techniques here for the sake of simplicity.

4. Practical computation of the estimator

Practical computation of $\tilde{\theta}_n$ and $\hat{\theta}_n$, for a given temperature $\lambda > 0$, is delicate. Indeed, exact computation of these estimators requires considering all subsets $\Theta_K(J)$ $\forall J \subset \{1, ..., p\}$. Since there are 2^p such subsets, exact computation of $\hat{\theta}_n$ or $\tilde{\theta}_n$ is not feasible for large p. However, since our estimators are defined as expectations of posterior distributions, we can approximate them via a Monte Carlo procedure. There is an extensive literature on Monte Carlo computational methods, especially in bayesian statistics where estimators can sometimes be expressed as expectations of posterior distribution, see Casella and Robert [9] for example. Markov Chains Monte Carlo (MCMC) algorithms such as Hastings-Metropolis or Gibbs sampling are classical in the case where we integrate w.r.t a discrete distribution, or to an absolutely continuous probability distribution w.r.t. the Lebesgue measure on $\mathbb{R}^p$. We use this procedure to compute $\hat{\theta}_n$. The computation of $\tilde{\theta}_n$ is more delicate. Indeed $\tilde{\rho}_\lambda$ is absolutely continuous w.r.t. the measure $\mathrm{m}(d\theta)$ that involves a mixture of Lebesgue measures on spaces of different dimensions. A way to proceed with such measures was proposed by Green [22] under the name "Reversible Jump Markov Chain Monte Carlo", RJMCMC, and applied successfully in various problems of model selection like multiple change-point problems, image segmentation and partition models in [22] or selection of the number of components in a mixture model in Green and Richardson [23]. We propose to adapt this procedure to our setting to compute $\tilde{\theta}_n$.

4.1. Computation of $\hat{\theta}_n$ via Hastings-Metropolis sampling.

In this subsection, we write a particular form of Hastings-Metropolis algorithm that will allow to compute any estimator of the form

$$\hat{\theta}^{(w)} = \sum_{\substack{J \subset \{1, ..., p\} \\ |J| \leq n}} w_J \hat{\theta}_J$$

where we already defined

$$\hat{\theta}_J \in \arg\min_{\theta \in \Theta(J)} r(\theta),$$

and

$$w_J \geq 0 \ \ \forall J \quad \text{and} \quad \sum_{\substack{J \subset \{1, ..., p\} \\ |J| \leq n}} w_J = 1.$$

Hastings-Metropolis algorithm starts from an arbitrary value (say $J^{(0)} = \emptyset$), a simple transition kernel $k(\cdot, \cdot)$ on the set $\{J \subset \{1, ..., p\}, |J| \leq n\}$ and updates $J^{(t)}$ to $J^{(t+1)}$ using the following scheme:

- draw $I^{(t)}$ from $k(J^{(t)}, \cdot)$;
- take

$$J^{(t+1)} = \begin{cases} I^{(t)} & \text{with proba. } \alpha(J^{(t)}, I^{(t)}) = \min\left(1, \frac{w_{I^{(t)}} k(I^{(t)}, J^{(t)})}{w_{J^{(t)}} k(J^{(t)}, I^{(t)})}\right) \\ J^{(t)} & \text{with proba. } 1 - \alpha(J^{(t)}, I^{(t)}). \end{cases}$$

We stop after T steps and compute

$$\hat{\theta}^{w,T,bo} = \frac{1}{T - bo + 1} \sum_{t=bo}^{T} \hat{\theta}_{J^{(t)}}.$$

This algorithm ensures that $(J^{(t)})_t$ is a Markov chain with invariant probability distribution $(w_J)_J$ (see Casella and Robert [9]). Here, the set $\{J \subset \{1, ..., p\}, |J| \leq n\}$ being finite, we just have to check that the chain is irreducible and aperiodic to obtain the convergence of $\hat{\theta}^{w,T,bo}$ to $\hat{\theta}^{(w)}$ (in probability).

In practice, we use the following kernel k:

$$k(J,\cdot) = k_+(J,\cdot)\mathbb{1}_{\{|J|=0\}} + \frac{k_+(J,\cdot)+k_-(J,\cdot)}{2}\mathbb{1}_{\{0<|J|<n\}} + k_-(J,\cdot)\mathbb{1}_{\{|J|=n\}}$$

where $k_+(\cdot,\cdot)$ and $k_-(\cdot,\cdot)$ are two kernels that we define now. The kernel k_+ adds an element to J whereas k_- removes one element from J. When we try to add an element, it is reasonable to consider first features that are the most correlated with the current residual. Similarly when we try to remove one element, we give priority the feature with the smallest coefficients in absolute value.

Formally, we choose some parameter $\zeta > 0$. We put, for $j \notin J$:

$$k_+(J, J\cup\{j\}) = \frac{e^{\zeta|c_j|}}{\sum_{h\notin J} e^{\zeta|c_h|}}$$

where c_j is the coefficient of linear correlation between $\left(Y_i - f_{\hat\theta_J}(X_i)\right)_{1\le i\le n}$ and $(\phi_j(X_i))_{1\le i\le n}$. And, for $j \in J$:

$$k_-(J, J\setminus\{j\}) = \frac{e^{-\zeta\left|(\hat\theta_J)_j\right|}}{\sum_{h\in J} e^{-\zeta\left|(\hat\theta_J)_h\right|}}.$$

4.2. **RJMCMC algorithm and computation of $\tilde\theta_n$.** The RJMCMC algorithm proposed by Green [22] is an application of the Hastings-Metropolis to the case of a measure absolutely continuous with relation to a more sophisticated distribution (in our case m). We use here this method to compute $\tilde\theta_n = \int_{\Theta_K} \theta\tilde\rho_\lambda(d\theta)$. We start from $\theta^{(0)} = 0$ and then, at each step, update $\theta^{(t)}$ to $\theta^{(t+1)}$ using the transition kernel (be careful, this time k has a density w. r. t. the measure, say, m, and, for the sake of simplicity, we let $\tilde\rho_\lambda(\cdot)$ denote indifferently the measure $\tilde\rho_\lambda$ and its density with respect to m, this is a standard notation in the MCMC literature):

- draw $\tau^{(t)}$ from $k(\theta^{(t)},\cdot)$;
- take

$$\vartheta^{(t)} = \begin{cases} \tau^{(t)} & \text{with proba. } \alpha(\theta^{(t)},\tau^{(t)}) = \min\left(1, \frac{\tilde\rho_\lambda(\tau^{(t)})k(\tau^{(t)},\theta^{(t)})}{\tilde\rho_\lambda(\tau^{(t)})k(\theta^{(t)},\tau^{(t)})}\right) \\ \theta^{(t)} & \text{with proba. } 1-\alpha(\theta^{(t)},\tau^{(t)}) \end{cases};$$

- draw $\theta^{(t+1)}$ from the distribution $\tilde\rho_\lambda(d\theta|\theta\in\Theta(J(\vartheta^{(t+1)})))$.

This algorithm ensures that $(\theta^{(t)})_t$ is a Markov chain with invariant probability distribution ρ_λ, see [9]. The last step ensures that we can make a move inside the current model even if the model change was rejected by the Hastings-Metropolis step. Note that is is straightforward to draw $\theta^{(t+1)}$ as, for any J, the distribution

$$\tilde\rho_\lambda(d\theta|\theta\in\Theta(J)) \propto e^{-\lambda r(\theta)}\mathrm{u}_J(d\theta)$$

is a truncated gaussian distribution. Moreover, this step does not affect the fact that $\tilde\rho_\lambda$ is an invariant distribution of the simulated Markov chain

Here, we use the following kernel k:

$$k(\theta,\cdot) = k_+(\theta,\cdot)\mathbb{1}_{\{J(\theta)=\emptyset\}} + \frac{k_+(\theta,\cdot)+k_-(\theta,\cdot)}{2}\mathbb{1}_{\{0<|J(\theta)|<n\}} + k_-(\theta,\cdot)\mathbb{1}_{\{|J(\theta)|=n\}\}}$$

where, for some $\zeta > 0$,

$$k_+(\theta,d\theta') = \sum_{j\notin J(\theta)} \frac{e^{\zeta|c_j(\theta)|}}{\sum_{h\notin J(\theta)} e^{\zeta|c_h(\theta)|}}\tilde\rho_\lambda(d\theta'|\theta'\in\Theta(J(\theta)\cup\{j\}))$$

with $c_j(\theta)$ is the coefficient of linear correlation between $(Y_i - f_\theta(X_i))_{1\leq i\leq n}$ and $(\phi_j(X_i))_{1\leq i\leq n}$ and,

$$k_-(\theta, d\theta') = \sum_{j\in J(\theta)} \frac{e^{-\zeta|\theta_j|}}{\sum_{h\in J(\theta)} e^{-\zeta|\theta_h|}} \tilde{\rho}_\lambda(d\theta'|\theta' \in \Theta(J(\theta)\setminus\{j\})).$$

5. Simulations

We compare in this section the exponential weights estimators $\hat{\theta}_n$ and $\tilde{\theta}_n$ to the LASSO on a toy example introduced by Tibshirani [36].

5.1. Description of the experiments.

We consider the toy example of [36]:

$$\forall i \in \{1, \dots, 20\}, \quad Y_i = \langle \beta, X_i \rangle + \varepsilon_i$$

with $X_i \in \mathcal{X} = \mathbb{R}^p$, $\beta \in \mathbb{R}^p$ and the ε_i are i. i. d. from a gaussian distribution with mean 0 and standard deviation σ.

The X_i's are i.i.d. and drawn from the gaussian distribution with zero mean and variance matrix:

$$\Sigma(\rho) = \left(\rho^{|i-j|}\right)_{\substack{i\in\{1,\dots,p\}\\ j\in\{1,\dots,p\}}}$$

for some $\rho \in [0, 1)$. Note that Tibshirani's toy example is set with $p = 8$ whereas we consider here $p \geq 8$.

Define the regression vector as follows

$$\beta = (3, 1.5, 0, 0, 2, 0, 0, 0, ...),$$

corresponding to a "sparse situation".

We will take σ respectively equal to 1 ("low noise") and 3 ("noisy case"); the value of ρ is fixed to 0.5.

5.2. Estimation results for small p.

In this subsection, we carry out experiments for $p = 8$ and $p = 30$.

We compute for our exponential weights estimator the minimum quadratic risk $\inf_{\lambda\in\Lambda} |X(\hat{\theta}_n(\lambda, \pi) - \beta)|_2$ where the temperature parameter λ is taken in a grid Λ. We proceed similarly for $\tilde{\theta}_n = \tilde{\theta}_n(\lambda, \mathrm{m})$ and for the LASSO estimator

$$\theta_n^L = \theta_n^L(\mu) = \arg\min_\theta \left[r(\theta) + \mu|\theta|_1\right]$$

for μ taken in a grid Λ'. We take

$$\Lambda = \{2, 3, 4, ..., 25\}\frac{1}{20}\frac{n}{s^2}$$

(motivated by our theoretical results) and

$$\Lambda' = \{1, 2, 3, ..., 70\}\frac{1}{10}\sqrt{\frac{\sigma^2 \log(p)}{n}}$$

(motivated, for example, by [6]). We fix $\zeta = 2$ in the algorithm and $\alpha = 10$. The MCMC algorithms are implemented with $T = 12000$ and $bo = 2000$.

We perform every experiment 20 times and give the results in Table 1. Convergence of the estimator can be checked, see for example Figure 1.

We can see on these experiments that the exponential weights estimators outperforms the LASSO in the low noise model $\sigma = 1$. When σ grows, our procedures seem to become less stable, especially the estimator $\hat{\theta}_n$. However for $p = 30$, we observe that the estimator $\tilde{\theta}_n$ performs better than the LASSO for both value of the noise.

TABLE 1. Results for the estimation of β, with small p. For each possible combination of σ and p, we report the median, mean and the standard deviation values of respectively $\inf_{\lambda\in\Lambda} |X(\hat{\theta}_n - \beta)|_2^2$, $\inf_{\lambda\in\Lambda} |X(\tilde{\theta}_n - \beta)|_2^2$ and $\inf_{\mu\in\Lambda'} |X(\hat{\theta}_n^L - \beta)|_2^2$.

σ	p	what?	$\hat{\theta}_\mu^{LASSO}$	$\hat{\theta}_n$	$\tilde{\theta}_n$
3	8	median	**0.70**	0.99	0.94
		mean	**0.75**	0.94	0.83
		s.d.	0.40	0.47	0.37
1	8	median	0.17	0.14	**0.14**
		mean	0.23	0.20	**0.19**
		s.d.	0.17	0.16	0.15
3	30	median	**0.82**	1.47	0.88
		mean	1.03	1.81	**1.02**
		s.d.	0.66	1.00	0.57
1	30	median	0.34	0.20	**0.19**
		mean	0.40	0.24	**0.20**
		s.d.	0.22	0.16	0.12

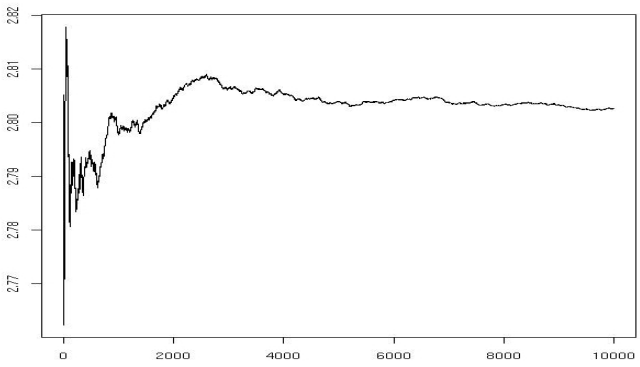

FIGURE 1. Convergence of the first coefficient in an experiment with $\sigma^2 = 1$, $p = 8$. We represent the first coordinate of $\frac{1}{N-bo+1}\sum_{t=bo}^{N} \hat{\theta}_{I^{(t)}}$ as a function of $N = b_0, ..., T$.

5.3. **Estimation results for large p.** In this subsection, we consider the cases $p = 100$ and $p = 1000$. As $\hat{\theta}_n$ becomes unstable when the dimension grows so that we simply remove it from the simulation study and focus on the comparison between $\tilde{\theta}_n$ and the LASSO. The results are given in Table 2.

We observe on these simulations that $\tilde{\theta}_n$ outperforms the LASSO in all cases.

5.4. **Some comments on model selection.** It is a known fact that the regularization parameter should be chosen differently for the LASSO depending on the problem at hand: prediction (minimization of $|X(\hat{\theta}_n^L(\mu) - \beta)|_2^2$) or variable selection. Denote for the Lasso estimator by $\mu(S)$ the set of regularization parameters yielding a good result for the variables selection problem and by $\mu(P)$ the set of regularization parameters yielding a good result for the prediction problem. We consider below the estimators $\hat{\theta}_n^L(\mu(P))$ and $\hat{\theta}_n^L(\mu(S))$. Note that if the sets $\mu(S)$ and $\mu(P)$ contain more than one element, we consider the value of the regularization parameter μ yielding the best possible performance depending on the criterion of interest, prediction risk or variable selection. We define the selected number of

TABLE 2. Results for the estimation of β, with large p. For each possible combination of σ and p, we report the median, mean and the standard deviation values of $\inf_{\lambda\in\Lambda} |X(\tilde{\theta}_n - \beta)|_2^2$ and $\inf_{\mu\in\Lambda'} |X(\hat{\theta}_n^L - \beta)|_2^2$.

σ	p	what?	$\hat{\theta}_\mu^{LASSO}$	$\tilde{\theta}_n$
3	100	median	1.55	**1.46**
		mean	1.69	**1.58**
		s.d.	0.86	0.92
1	100	median	0.44	**0.14**
		mean	0.47	**0.22**
		s.d.	0.23	0.20
3	1000	median	2.13	**1.96**
		mean	2.11	**2.03**
		s.d.	0.62	0.56
1	1000	median	0.72	**0.40**
		mean	0.75	**0.50**
		s.d.	0.32	0.31

TABLE 3. Additional results for the case $p = 8$, $\sigma = 1$. First array: prediction result. Second array: number of selected components.

	$\hat{\theta}_n^L(\mu(S))$	$\hat{\theta}_n^L(\mu(P))$	$\tilde{\theta}_n(\lambda(S), \mathrm{m})$	$\tilde{\theta}_n(\lambda(P), \mathrm{m})$
median	1.30	0.17	0.24	0.14
mean	2.03	0.23	0.27	0.19
s.d.	1.77	0.17	0.16	0.15
	$\hat{\theta}_n^L(\mu(S))$	$\hat{\theta}_n^L(\mu(P))$	$\hat{\theta}_n(\lambda(S), \mathrm{m})$	$\tilde{\theta}_n(\lambda(P), \mathrm{m})$
median	3.00	6.00	3.00	2.50
mean	3.00	5.65	3.00	2.40
s.d.	0.00	1.04	0.00	1.19

components as the bayesian MAP, maximum *a posteriori*). We define similarly for the exponential weights estimator $\tilde{\theta}_n(\lambda, m)$ the regularization parameters $\lambda(S)$ and $\lambda(P)$.

We propose some additional results on the first experiment ($p = 8$, $\sigma = 1$) in Table 3. We compare the performances of $\hat{\theta}_n^L(\mu(P))$, $\hat{\theta}_n^L(\mu(S)$, $\tilde{\theta}_n(\lambda(P), \mathrm{m})$ and $\tilde{\theta}_n(\lambda(S), \mathrm{m})$.

The results show that our method can perform variables selection and estimation simultaneously while the LASSO cannot.

5.5. **Some comments on computation time.** We can roughly analyze the computational complexity of our MCMC algorithm $\hat{\theta}_n$:

(1) First, there are T MCMC steps.
(2) At each step $t \leq T$, we have to choose which new component we want to add (or remove) from the current model. There are at most p possible choices, and for each choice j we have to compute the correlation between $(Y_i - f_{\hat{\theta}_{J^{(t)}}})_i^n$ and $(\phi_j(X_i))_i^n$, so this takes $\mathcal{O}(np)$ operations.
(3) Finally, at each step t we have to compute $\hat{\theta}_{J^{(t)}}$, this takes at most $\mathcal{O}(|J|^3)$ operations.

So finally, the number of operations is $\mathcal{O}\left(T(np + E_\lambda(|J|^3))\right)$ where $E_\lambda(|J|)$ is the expectation of $|J|$ under the aggregation distribution with temperature parameter

λ

$$E_\lambda(|J|) = \frac{\sum_{k=0}^{n} \sum_{\substack{J \subset \{1,\dots,p\} \\ |J|=k}} \pi_J e^{-\lambda\left(r(\hat{\theta}_J) + \frac{2\sigma^2|J|}{n}\right)} |J|}{\sum_{k=0}^{n} \sum_{\substack{J \subset \{1,\dots,p\} \\ |J|=k}} \pi_J e^{-\lambda\left(r(\hat{\theta}_J) + \frac{2\sigma^2|J|}{n}\right)}}.$$

For properly tuned λ, we observe $\mathbb{E}(|J|) \simeq |J(\overline{\theta})|$. We understand why the sparsity of the parameter has an important influence on the computation time. Consider for example the case $p = 100$ and $n = 50$. If $|J(\overline{\theta})| = 10$ then $n * p = 5000 > 1000 = |J(\overline{\theta})|^3$ whereas if $|J(\overline{\theta})| = 25$ then $n * p = 5000 < 15625 = |J(\overline{\theta})|^3$.

All the simulations were performed with the R software [32]. The code are available on request by e-mail.

6. Proofs

6.1. **Proofs of Section 2.** This proof uses an argument from Leung and Barron [27].

Proof of Proposition 1. The mapping $Y \to \hat{f}_n(Y) \stackrel{\triangle}{=} (\hat{f}_n(X_1, Y), \dots, \hat{f}_n(X_n, Y))^T$ is clearly continuously differentiable by composition of elementary differentiable functions. For any subset $J \subset \{1, \dots, p\}$ define $A_J = (\phi_j(X_i))_{1 \le i \le n, j \in J}$, $\Sigma_J = \frac{1}{n} A_J^T A_J$, $\Phi_J(\cdot) = (\phi_j(\cdot))_{j \in J}$ and

$$g_J = e^{-\lambda\left(\|Y - f_J\|_n^2 + \frac{2\sigma^2|J|}{n}\right)}$$

where

$$f_J(x, Y) = \frac{1}{n} Y^T A_J \Sigma_J^+ \Phi_J(x)^T,$$

and Σ_J^+ denotes the pseudo-inverse of Σ_J. Denote by ∂_i the derivative w.r.t. Y_i. Simple computations give

$$\partial_i f_J(x, Y) = \frac{1}{n} \Phi_J(X_i) \Sigma_J^{-1} \Phi_J(x)^T,$$

$$(\partial_i f_J(X_1, Y), \dots, \partial_i f_J(X_n, Y)) Y = f_J(X_i, Y),$$

and

$$\sum_{l=1}^{n} f_J(X_l, Y) \partial_i f_J(X_l, Y) \quad = \quad f_J(X_i, Y).$$

Thus we have

$$\begin{aligned} \partial_i(g_J) &= -\lambda \partial_i \left(\|Y - f_J\|_n^2\right) g_J \\ &= -\frac{2\lambda}{n} \left((Y_i - f_J(X_i, Y)) - \sum_{l=1}^{n} \partial_i f_J(X_l, Y)(Y_l - f_J(X_l, Y)) \right) g_J \\ &= -\frac{2\lambda}{n} (Y_i - f_J(X_i, Y)) g_J, \end{aligned}$$

Recall that

$$\hat{f}_n = \frac{\sum_{k=0}^{n} \sum_{J \in \mathcal{P}(\{1,\dots,p\}), |J|=k} \pi_J g_J f_J}{\sum_{k=0}^{n} \sum_{J \in \mathcal{P}(\{1,\dots,p\}), |J|=k} \pi_J g_J}.$$

We have

$$\partial_i \hat{f}_n(X_i, Y) \quad = \quad \frac{\sum_{k=0}^{n} \sum_{J \in \mathcal{P}(\{1,\dots,p\}), |J|=k} \pi_J \left(\partial_i(g_J) f_J(X_i, Y) + g_J \partial_i(f_J(X_i, Y))\right)}{\sum_{k=0}^{n} \sum_{J \in \mathcal{P}(\{1,\dots,p\}), |J|=k} \pi_J g_J}$$

$$
\begin{aligned}
& - \frac{\left(\sum_{k=0}^n \sum_{J\in\mathcal{P}(\{1,\dots,p\}),|J|=k} \pi_J g_J f_J\right)\left(\sum_{k=0}^n \sum_{J\in\mathcal{P}(\{1,\dots,p\}),|J|=k} \pi_J \partial_i(g_J)\right)}{(\sum_{k=0}^n \sum_{J\in\mathcal{P}(\{1,\dots,p\}),|J|=k} \pi_J g_J)^2} \\
= & -\frac{2\lambda}{n} Y_i \hat{f}_n + \frac{2\lambda}{n} \frac{\sum_{k=0}^n \sum_{J\in\mathcal{P}(\{1,\dots,p\}),|J|=k} f_J(X_i,Y)^2 \pi_J g_J}{\sum_{k=0}^n \sum_{J\in\mathcal{P}(\{1,\dots,p\}),|J|=k} \pi_J g_J} \\
& + \frac{1}{n} \frac{\sum_{k=0}^n \sum_{J\in\mathcal{P}(\{1,\dots,p\}),|J|=k} \Phi_J(X_i)\Sigma_J^{-1}\Phi_J(X_i)^T \pi_J g_J f_J}{\sum_{k=0}^n \sum_{J\in\mathcal{P}(\{1,\dots,p\}),|J|=k} \pi_J g_J} \\
+ & \frac{2\lambda}{n} Y_i \hat{f}_n - \frac{2\lambda}{n} \hat{f}_n^2 \\
= & \frac{2\lambda}{n} \frac{\sum_{k=0}^n \sum_{J\in\mathcal{P}(\{1,\dots,p\}),|J|=k} (f_J(X_i,Y) - \hat{f}_n(X_i,Y))^2 \pi_J g_J}{\sum_{k=0}^n \sum_{J\in\mathcal{P}(\{1,\dots,p\}),|J|=k} \pi_J g_J} \\
& + \frac{1}{n} \frac{\sum_{k=0}^n \sum_{J\in\mathcal{P}(\{1,\dots,p\}),|J|=k} \Phi_J(X_i)\Sigma_J^{-1}\Phi_J(X_i)^T \pi_J g_J}{\sum_{k=0}^n \sum_{J\in\mathcal{P}(\{1,\dots,p\}),|J|=k} \pi_J g_J} \geq 0. \qquad (6.1)
\end{aligned}
$$

Consider the following estimator of the risk

$$
\hat{r}_n(Y) = \|\hat{f}_n(Y) - Y\|_n^2 + \frac{2\sigma^2}{n} \sum_{i=1}^n \partial_i \hat{f}_n(X_i, Y) - \sigma^2. \qquad (6.2)
$$

Using an argument based on Stein's identity as in [26] we now prove that

$$
\mathbb{E}(\hat{r}_n(Y)) = \mathbb{E}\left(\|\hat{f}_n(Y) - f\|_n^2\right).
$$

We have

$$
\begin{aligned}
\mathbb{E}\left(\|\hat{f}_n(Y) - f\|_n^2\right) &= \mathbb{E}\left(\|\hat{f}_n(Y) - Y\|_n^2 + \frac{2}{n}\sum_{i=1}^n W_i(\hat{f}_n(X_i,Y) - f(X_i))\right) - \sigma^2 \\
&= \mathbb{E}\left(\|\hat{f}_n(Y) - Y\|_n^2 + \frac{2}{n}\sum_{i=1}^n W_i \hat{f}_n(X_i,Y)\right) - \sigma^2. \qquad (6.3)
\end{aligned}
$$

For $\mathbf{z} = (z_1, \dots, z_n)^T \in \mathbb{R}^n$ write $F_{W,i}(\mathbf{z}) = \prod_{j\neq i} F_{W,i}(z_j)$, where F_W denotes the c.d.f. of the random variable W_1. Since $\mathbb{E}(W_i) = 0$ we have

$$
\begin{aligned}
\mathbb{E}\left(W_i \hat{f}_n(X_i,Y)\right) &= \mathbb{E}\left(W_i \int_0^{W_i} \partial_i \hat{f}_n(X_i, Y_1, \dots, Y_{i-1}, f(X_i) + z, Y_{i+1}, \dots, Y_n) dz\right) \\
&= \int_{\mathbb{R}^{n-1}} \left(\int_{\mathbb{R}} y \int_0^y \partial_i \hat{f}_n(X_i, f + \mathbf{z}) dz_i dF_W(y)\right) dF_{W,i}(\mathbf{z}). \qquad (6.4)
\end{aligned}
$$

In view of (6.1) we can apply Fubini's Theorem to the right-hand-side of (6.4). We obtain under the assumption $W \sim \mathcal{N}(0, \sigma^2)$ that

$$
\begin{aligned}
\int_{\mathbb{R}^+} \int_0^y \partial_i \hat{f}_n(X_i, f + \mathbf{z}) dz_i dF_W(y) &= \int_{\mathbb{R}^+} \int_{z_i}^\infty y dF_W(y) \partial_i \hat{f}_n(X_i, f + \mathbf{z}) dz_i \\
&= \int_{\mathbb{R}^+} \sigma^2 \partial_i \hat{f}_n(X_i, f + \mathbf{z}) dF_W(z_i),
\end{aligned}
$$

A Similar equality holds for the integral over $\mathbb{R}^-$. Thus we obtain

$$
\mathbb{E}\left(W_i \hat{f}_n(X_i,Y)\right) = \sigma^2 \mathbb{E}\left(\partial_i \hat{f}_n(X_i,Y)\right).
$$

Combining (6.2), (6.3) and the above display gives

$$
\mathbb{E}\left(\hat{r}_n(Y)\right) = \mathbb{E}\left(\|\hat{f}_n(Y) - f\|_n^2\right).
$$

Since $\hat{f}_n(\cdot,Y)$ is the expectation of $f_J(\cdot,Y)$ w.r.t. the probability distribution $\propto g\cdot\pi$, we have

$$\|\hat{f}_n(\cdot,Y)-Y\|_n^2 = \frac{\sum_{k=0}^n \sum_{J\in\mathcal{P}(\{1,\dots,p\}),\,|J|=k} \left(\|f_J(\cdot,Y)-Y\|_n^2 - \|f_J(\cdot,Y)-\hat{f}_n(Y)\|_n^2\right) g_J\pi_J}{\sum_{k=0}^n \sum_{J\in\mathcal{P}(\{1,\dots,p\}),\,|J|=k} g_J\pi_J}.$$

For the sake of simplicity set $f_J = f_J(\cdot,Y)$ and $\hat{f}_n = \hat{f}_n(\cdot,Y)$. Combining (6.2), the above display and $\lambda \leqslant \frac{n}{4\sigma^2}$ yields

$$\begin{aligned}
\hat{r}_n(Y) &= \frac{\sum_{k=0}^n \sum_{J\in\mathcal{P}(\{1,\dots,p\}),\,|J|=k} \left(\|f_J-Y\|_n^2 - \sum_{i=1}^n \left(\frac{4\lambda\sigma^2}{n}-1\right)\|f_J-\hat{f}_n\|_n^2\right) g_J\pi_J}{\sum_{k=0}^n \sum_{J\in\mathcal{P}(\{1,\dots,p\}),\,|J|=k} \pi_J g_J} \\
&\quad + \frac{2\sigma^2}{n^2}\sum_{i=1}^n \frac{\sum_{k=0}^n \sum_{J\in\mathcal{P}(\{1,\dots,p\}),|J|=k} \Phi_J(X_i)\Sigma_J^{-1}\Phi_J(X_i)^T\pi_J g_J}{\sum_{k=0}^n \sum_{J\in\mathcal{P}(\{1,\dots,p\}),|J|=k} \pi_J g_J} - \sigma^2 \\
&\leqslant \sum_{k=0}^n \sum_{J\in\mathcal{P}(\{1,\dots,p\}),|J|=k} \left(\|f_J-Y\|_n^2 + \frac{2\sigma^2}{n}|J|\right) g_J\pi_J - \sigma^2.
\end{aligned}$$

By definition of g_J we have

$$\begin{aligned}
\|f_J-Y\|_n^2 + \frac{2\sigma^2|J|}{n} &= -\frac{1}{\lambda}\log\left(\frac{g_J}{\sum_{k=0}^n \sum_{J\in\mathcal{P}(\{1,\dots,p\}),|J|=k} g_J\pi_J}\right) \\
&\quad + \frac{1}{\lambda}\log\left(\sum_{k=0}^n \sum_{J\subset\{1,\dots,p\},\,|J|=k} g_J\pi_J\right).
\end{aligned}$$

Summing the above inequality w.r.t. the probability distribution $g\cdot\pi$ (with the suitable normalization) and using the fact that

$$\sum_{k=0}^n \sum_{J\subset\{1,\dots,p\},\,|J|=k} \theta_J(Y)\log\left(\frac{g_J}{\sum_{k=0}^n \sum_{J\in\mathcal{P}(\{1,\dots,p\}),|J|=k} g_J\pi_J}\right) = K(g\cdot\pi,\pi) \geqslant 0$$

as well as a convex duality argument (cf., e.g., [18], p. 264) we get

$$\hat{r}_n(Y) \leqslant \sum_{k=0}^n \sum_{J\in\mathcal{P}(\{1,\dots,p\}),|J|=k} \left(\|Y-f_J\|_n^2 + \frac{2\sigma^2}{n}|J|\right)\pi'_J + \frac{1}{\lambda}K(\pi',\pi) - \sigma^2,$$

for all probability measure π' on $\mathcal{P}(\{1,\dots,p\})$. Taking the expectation in the last inequality we get for any π'

$$\begin{aligned}
\mathbb{E}\left(\|\hat{f}_n-f\|_n^2\right) &= \mathbb{E}(\hat{r}_n(Y)) \\
&\leqslant \sum_{k=0}^n \sum_{J\in\mathcal{P}(\{1,\dots,p\}),|J|=k} \left(\mathbb{E}(\|f_J-Y\|_n^2) + \frac{2\sigma^2}{n}|J|\right)\pi'_J + \frac{1}{\lambda}K(\pi',\pi) - \sigma^2 \\
&\leqslant \sum_{k=0}^n \sum_{J\in\mathcal{P}(\{1,\dots,p\}),|J|=k} \left(\mathbb{E}(\|f_J-f\|_n^2) + \frac{2}{n}\sum_{i=1}^n \mathbb{E}(W_i f_J(X_i,Y)) + \frac{2\sigma^2}{n}|J|\right)\pi'_J \\
&\qquad + \frac{1}{\lambda}K(\pi',\pi) \\
&\leqslant \sum_{k=0}^n \sum_{J\in\mathcal{P}(\{1,\dots,p\}),|J|=k} \left(\mathbb{E}(\|f_J-f\|_n^2) + \frac{4\sigma^2}{n}|J|\right)\pi'_J + \frac{1}{\lambda}K(\pi',\pi),
\end{aligned}$$

where we have used Stein's argument $\mathbb{E}(W_i f_J(X_i,Y)) = \sigma^2\mathbb{E}\left(\partial_i f_J(X_i,Y)\right)$ and the fact that $\sum_{i=1}^n \partial_i f_J(X_i,Y) = 1$ in the last line. Finally taking π' in the set of Dirac distributions on the subset J of $\{1,\dots,p\}$ yields the theorem. □

Proof of Theorem 1. First note that any minimizer $\theta \in \mathbb{R}^p$ of the right-hand-side in (2.6) is such that $|J(\theta)| \leqslant \operatorname{rank}(A) \leqslant n$ where we recall that $A = (\phi_j(X_i))_{1 \leqslant i \leqslant n, 1 \leqslant j \leqslant p}$. Indeed, for any $\theta \in \mathbb{R}^p$ such that $|J(\theta)| > \operatorname{rank}(A)$ we can construct a vector $\theta' \in \mathbb{R}^p$ such that $f_\theta = f_{\theta'}$ and $|J(\theta')| \leqslant \operatorname{rank}(A)$ and the mapping $x \to x \log\left(\frac{ep\alpha}{x}\right)$ is non-decreasing on $(0, p]$.

Next for any $J \in \mathcal{P}(\{1, \dots, p\})$ we have

$$\mathbb{E}[\|f_J - f\|_n^2] = \min_{\theta \in \Theta(J)} \|f_\theta - f\|_n^2 + \frac{\sigma^2 |J|}{n} = \min_{\theta \in \Theta(J)} \left\{ \|f_\theta - f\|_n^2 + \frac{\sigma^2 |J(\theta)|}{n} \right\}.$$

Thus

$$\begin{aligned}
&\min_{J \in \mathcal{P}(\{1,\dots,p\}), |J| \leqslant n} \left(\mathbb{E}[\|f_J - f\|_n^2] + \frac{1}{\lambda} \log\left(\frac{1}{\pi_J}\right) + \frac{\sigma^2 J}{n} \right) \\
&\qquad = \min_{J \in \mathcal{P}(\{1,\dots,p\}) : |J| \leqslant n} \min_{\theta \in \Theta(J)} \left(\|f_\theta - f\|_n^2 + \frac{1}{\lambda} \log\left(\frac{1}{\pi_{J(\theta)}}\right) + \frac{\sigma^2 |J(\theta)|}{n} \right) \\
&\qquad = \min_{\theta \in \mathbb{R}^p} \left(\|f_\theta - f\|_n^2 + \frac{1}{\lambda} \log\left(\frac{1}{\pi_{J(\theta)}}\right) + \frac{\sigma^2 |J(\theta)|}{n} \right).
\end{aligned}$$

Combining the above display with Proposition 1 and our definition of the prior π gives the result. □

6.2. **Proof of Theorem 2.** We state below a version of Bernstein's inequality useful in the proof of Theorem 2. See Proposition 2.9 page 24 in [30], more precisely Inequality (2.21).

Lemma 1. *Let T_1, ..., T_n be independent real valued random variables. Let us assume that there is two constants v and w such that*

$$\sum_{i=1}^n \mathbb{E}(T_i^2) \leq v$$

and for all integers $k \geq 3$,

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

Then, for any $\zeta \in (0, 1/w)$,

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

Proof of Theorem 2. For any $\theta \in \Theta_{K+c}$ and $\theta' \in \Theta_K$ define the random variables

$$T_i = -\left(Y_i - f_\theta(X_i)\right)^2 + \left(Y_i - f_{\theta'}(X_i)\right)^2.$$

Note that these variables are independent. We have

$$\begin{aligned}
\sum_{i=1}^n \mathbb{E}(T_i^2) &= \sum_{i=1}^n \mathbb{E}\left\{ [2Y_i - f_{\theta'}(X_i) - f_\theta(X_i)]^2 [f_{\theta'}(X_i) - f_\theta(X_i)]^2 \right\} \\
&= \sum_{i=1}^n \mathbb{E}\left\{ [2W_i + 2f(X_i) + f_{\theta'}(X_i) - f_\theta(X_i)]^2 [f_{\theta'}(X_i) - f_\theta(X_i)]^2 \right\} \\
&\leq \sum_{i=1}^n \mathbb{E}\left\{ \left[8W_i^2 + (2\|f\|_\infty + L(2K + c))^2\right] [f_{\theta'}(X_i) - f_\theta(X_i)]^2 \right\} \\
&= \sum_{i=1}^n \mathbb{E}\left[8W_i^2 + (2\|f\|_\infty + L(2K + c))^2\right] \mathbb{E}\left\{ [f_{\theta'}(X_i) - f_\theta(X_i)]^2 \right\}
\end{aligned}$$

$$\leq n\left[8\sigma^2+(2\|f\|_\infty+L(2K+c))^2\right]\left[R(\theta)-R(\theta')\right]=:v(\theta,\bar\theta)=v.$$

For any integer $k\geq 3$ we have

$$\sum_{i=1}^n\mathbb{E}\left[(T_i)_+^k\right]\leq\sum_{i=1}^n\mathbb{E}\left[|2Y_i-f_{\theta'}(X_i)-f_\theta(X_i)|^k\,|f_{\theta'}(X_i)-f_\theta(X_i)|^k\right]$$

$$\leq\sum_{i=1}^n\mathbb{E}\left[2^{2k-1}\left[|W_i|^k+(\|f\|_\infty+L(K+c/2)^k\right]|f_{\theta'}(X_i)-f_\theta(X_i)|^k\right]$$

$$\leq\sum_{i=1}^n\mathbb{E}\left[2^{2k-1}\left[|W_i|^k+(\|f\|_\infty+L(K+c/2)^k\right][L(2K+c)]^{k-2}\left[f_{\theta'}(X_i)-f_\theta(X_i)\right]^2\right]$$

$$\leq 2^{2k-1}\left[\sigma^2k!\xi^{k-2}+(\|f\|_\infty+L(K+c/2))^k\right][L(2K+c)]^{k-2}\sum_{i=1}^n\mathbb{E}\left\{\left[f_{\theta'}(X_i)-f_\theta(X_i)\right]^2\right\}$$

$$\leq\frac{8(\sigma^2k!\xi^{k-2}+(\|f\|_\infty+L(K+c/2))^k)(4L(2K+c))^{k-2}}{8\sigma^2+(2\|f\|_\infty+L(2K+c))^2}v\leq v\frac{k!w^{k-2}}{2}$$

with $w:=8[\xi+2(\|f\|_\infty+L(K+c/2))]L(2K+c)$.

Next, for any $\lambda\in]0,n/w[$ and $\theta\in\Theta_{K+c}$, applying Lemma 1 with $\zeta=\lambda/n$ gives

$$\mathbb{E}\exp\left[\lambda\Big(R(\theta)-R(\theta')-r(\theta)+r(\theta')\Big)\right]\leq\exp\left[\frac{v\lambda^2}{2n^2(1-\frac{w\lambda}{n})}\right].$$

Set $C=8\sigma^2+(2\|f\|_\infty+L(2K+c))^2$. For any $\epsilon>0$ the last display yields

$$\mathbb{E}\exp\left[\left(\lambda-\frac{\lambda^2C}{2n(1-\frac{w\lambda}{n})}\right)\Big(R(\theta)-R(\theta')\Big)+\lambda\Big(-r(\theta)+r(\theta')\Big)-\log\frac{1}{\varepsilon}\right]\leq\varepsilon.$$

Integrating w.r.t. the probability distribution $\pi(\cdot)$ we get

$$\int\mathbb{E}\exp\left[\left(\lambda-\frac{\lambda^2C}{2n(1-\frac{w\lambda}{n})}\right)\Big(R(\theta)-R(\theta')\Big)+\lambda\Big(-r(\theta)+r(\theta')\Big)-\log\frac{1}{\varepsilon}\right]\pi(d\theta)\leq\varepsilon.$$

Next, Fubini's theorem gives

$$\mathbb{E}\int\exp\left[\left(\lambda-\frac{\lambda^2C}{2n(1-\frac{w\lambda}{n})}\right)\Big(R(\theta)-R(\theta')\Big)+\lambda\Big(-r(\theta)+r(\theta')\Big)-\log\frac{1}{\varepsilon}\right]\pi(d\theta)\leq\varepsilon.$$

$$\mathbb{E}\int\exp\left[\left(\lambda-\frac{\lambda^2C}{2n(1-\frac{w\lambda}{n})}\right)\Big(R(\theta)-R(\theta')\Big)+\lambda\Big(-r(\theta)+r(\theta')\Big)-\log\left[\frac{d\tilde\rho_\lambda}{d\pi}(\theta)\right]-\log\frac{1}{\varepsilon}\right]\tilde\rho_\lambda(d\theta)\leq\varepsilon.$$

Jensen's inequality yields

$$\mathbb{E}\exp\left[\left(\lambda-\frac{\lambda^2C}{2n(1-\frac{w\lambda}{n})}\right)\left(\int Rd\tilde\rho_\lambda-R(\theta')\right)\right.$$

$$+\lambda\left(-\int r d\tilde{\rho}_\lambda + r(\theta')\right) - \mathcal{K}(\tilde{\rho}_\lambda, \pi) - \log\frac{1}{\varepsilon}\Bigg] \leq \varepsilon.$$

Now, using the basic inequality $\exp(x) \geq \mathbb{1}_{\mathbb{R}_+}(x)$ we get

$$\mathbb{P}\Bigg\{\left(\lambda - \frac{\lambda^2 C}{2n(1-\frac{w\lambda}{n})}\right)\left(\int R d\tilde{\rho}_\lambda - R(\theta')\right) + \lambda\left(-\int r d\tilde{\rho}_\lambda + r(\theta')\right) - \mathcal{K}(\tilde{\rho}_\lambda, \pi) - \log\frac{1}{\varepsilon}\Bigg] \geq 0\Bigg\} \leq \varepsilon.$$

Using Jensen's inequality again gives

$$\int R d\tilde{\rho}_\lambda \geq R\left(\int \theta \tilde{\rho}_\lambda(d\theta)\right) = R(\tilde{\theta}_\lambda).$$

Combining the last two displays we obtain

$$\mathbb{P}\Bigg\{R(\tilde{\theta}_\lambda) - R(\theta') \leq \frac{\int r d\tilde{\rho}_\lambda - r(\theta') + \frac{1}{\lambda}\left[\mathcal{K}(\tilde{\rho}_\lambda, \pi) + \log\frac{1}{\varepsilon}\right]}{1 - \frac{\lambda C}{2(n-w\lambda)}}\Bigg\} \geq 1 - \varepsilon.$$

Now, using Lemma 1.1.3 (page 4) in Catoni [12] we obtain that

$$\mathbb{P}\Bigg\{R(\tilde{\theta}_\lambda) - R(\theta') \leq \inf_{\rho \in \mathcal{M}_+^1(\Theta_{K+c})} \frac{\int r d\rho - r(\theta') + \frac{1}{\lambda}\left[\mathcal{K}(\rho, \pi) + \log\frac{1}{\varepsilon}\right]}{1 - \frac{\lambda C}{2(n-w\lambda)}}\Bigg\} \geq 1 - \varepsilon$$

where $\mathcal{M}_+^1(\Theta_K)$ is the set of all probability measures over Θ_K (with the borel σ-algebra). A similar argument to upper bound r by R combined with the union bound yield

$$\mathbb{P}\Bigg\{R(\tilde{\theta}_\lambda) - R(\theta') \leq \inf_{\rho \in \mathcal{M}_+^1(\Theta_{K+c})} \frac{\left(1 + \frac{\lambda C}{2(n-w\lambda)}\right)\left(\int R d\rho - R(\theta')\right) + \frac{2}{\lambda}\left[\mathcal{K}(\rho, \pi) + \log\frac{2}{\varepsilon}\right]}{1 - \frac{\lambda C}{2(n-w\lambda)}}\Bigg\} \geq 1 - \varepsilon.$$

Now for any $\theta' \in \Theta_K$ and any $\delta \in]0, c]$ taking ρ as the uniform probability measure on the set $\{t \in \Theta(J(\theta')) : |t - \theta'|_1 \leq \delta\} \subset \Theta_{K+c}(J(\theta'))$ gives

$$\mathbb{P}\Bigg\{R(\tilde{\theta}_\lambda) \leq \min_{\theta' \in \Theta_K}\Bigg(R(\theta') + \frac{1}{1 - \frac{\lambda C}{2(n-w\lambda)}}\Bigg[\left(1 + \frac{\lambda C}{2(n-w\lambda)}\right)\mathcal{C}_2\delta + \frac{2}{\lambda}\left(|J(\theta')|\log\frac{K+c}{\delta} + |J(\theta')|\log\frac{1}{\alpha} + \log(1-\alpha) + \log\binom{p}{|J(\theta')|} + \log\frac{2}{\varepsilon}\right)\Bigg]\Bigg)\Bigg\} \geq 1 - \varepsilon.$$

Taking $\delta = c = n^{-1}$ and the inequality $\log\binom{p}{|J(\theta')|} \leq |J(\theta')|\log\frac{pe}{|J(\theta')|}$ give

$$\mathbb{P}\Bigg\{R(\tilde{\theta}_\lambda) \leq \min_{\theta' \in \Theta_K}\Bigg(R(\theta') + \frac{1}{1 - \frac{\lambda C}{2(n-w\lambda)}}\Bigg[\left(1 + \frac{\lambda C}{2(n-w\lambda)}\right)\frac{\mathcal{C}_2}{n} + \frac{2}{\lambda}\left(|J(\theta')|\log(K+c) + |J(\theta')|\log\left(\frac{epn}{\alpha|J(\theta')|}\right) + \log\left(\frac{2\sqrt{1-\alpha}}{\varepsilon}\right)\right)\Bigg]\Bigg)\Bigg\} \geq 1 - \varepsilon \tag{6.5}$$

Taking now $\lambda = n/(2\mathcal{C}_1)$ (where we recall that $\mathcal{C}_1 = C \vee w$) in (6.5) gives

$$\mathbb{P}\Bigg\{ R(\tilde{\theta}_\lambda) \leq \min_{\theta' \in \Theta_K} \Bigg(R(\theta') + \frac{3\mathcal{C}_2}{n} + \frac{8\mathcal{C}_1}{n} \Bigg[|J(\theta')| \log\left(K + c\right) + \left(|J(\theta')| \log\left(\frac{enp}{\alpha |J(\theta')|} \right) + \log\left(\frac{2\sqrt{1-\alpha}}{\varepsilon} \right) \right) \Bigg] \Bigg) \Bigg\} \geq 1 - \varepsilon,$$

where we have used that $1 - \frac{\lambda C}{2(n - w\lambda)} \geq 1/2$ and $1 + \frac{\lambda C}{2(n-w\lambda)} \leq 3/2$. □

References

[1] H. Akaike. Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Csaki, editors, *2nd International Symposium on Information Theory*, pages 267–281. Budapest: Akademia Kiado, 1973.

[2] P. Alquier. *Transductive and Inductive Adaptative Inference for Regression and Density Estimation*. PhD thesis, University Paris 6, 2006.

[3] P. Alquier. Pac-bayesian bounds for randomized empirical risk minimizers. *Mathematical Methods of Statistics*, 17(4):279–304, 2008.

[4] J.-Y. Audibert. Aggregated estimators and empirical complexity for least square regression. *Annales de l'Institut Henri Poincaré: Probability and Statistics*, 40(6):685–736, 2004.

[5] J.-Y. Audibert. *PAC-Bayesian Statistical Learning Theory*. PhD thesis, University Paris 6, 2004.

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

[7] F. Bunea, A.B. Tsybakov, and M.H. Wegkamp. Aggregation for Gaussian regression. *Annals of Statistics*, 35:1674–1697, 2007.

[8] G. Casella and E. Moreno. Objective bayesian variable selection. *Journal of the American Statistical Association*, 101:157–167, 2006.

[9] G. Casella and C. Robert. *Monte Carlo Statistical Methods*. Springer-Verlag, 2nd Edition 2004.

[10] O. Catoni. A pac-bayesian approach to adaptative classification. *Preprint Laboratoire de Probabilités et Modèles Aléatoires*, 2003.

[11] O. Catoni. *Statistical Learning Theory and Stochastic Optimization, Lecture Notes in Mathematics (Saint-Flour Summer School on Probability Theory 2001, ed. J. Picard)*. Springer, 2004.

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

[13] Wen Cui and I. E. George. Empirical bayes vs. fully bayes variable selection. *Journal of Statistical Planning and Inference*, 138(4):888–900, 2008.

[14] A. Dalalyan and A. Tsybakov. Aggregation by exponential weighting, sharp oracle inequalities and sparsity. *Machine Learning*, 72(1-2):39–61, 2008.

[15] A.S. Dalalyan and A.B. Tsybakov. Pac-bayesian bounds for the expected error of aggregation by exponential weights. Technical report, Université Paris 6, CREST and CERTIS, Ecole des Ponts ParisTech, 2009. personal communication.

[16] A.S. Dalalyan and A.B. Tsybakov. Mirror averaging with sparsity priors. arXiv:1003:1189, 2010.

[17] A.S. Dalalyan and A.B. Tsybakov. Sparse regression learning by aggregation and langevin monte-carlo. arXiv:09.1223v3, 2010.

[18] A. Dembo and O. Zeitouni. *Large Deviation Techniques and Applications*. Springer, 1998.

[19] L. Frank and J. Friedman. A statistical view on some chemometrics regression tools. *Technometrics*, 16:499–511, 1993.

[20] I. E. George. The variable selection problem. *Journal of the American Statistician Association*, 95(452):1304–1308, 2000.

[21] I. E. George and R. E. McCulloch. Approaches for bayesian model selection. *Statistica Sinica*, 7:339–373, 1997.

[22] P. J. Green. Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

[23] P. J. Green and S. Richardson. On bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society: series B (statistical methodology)*, 59(4):731–792, 1997.

[24] G. Kerkyacharian, M. Mougeot, D. Picard, and K. Tribouley. Learning Ouf of Leaders. Preprint, arXiv:1001.1919, 2010.

[25] V. Koltchinskii. Sparsity in empirical risk minimization. Annales de l'Institut Henri Poincaré, Probability and Statistics (to appear).
[26] E. L. Lehmann and G. Casella. *Theory of Point Estimation*. Springer, 1998.
[27] G. Leung and A.R. Barron. Information theory and mixing least-squares regressions. *IEEE Trans. Inform. Theory*, 52(8):3396–3410, 2006.
[28] K. Lounici. Generalized mirror averaging and d-convex aggregation. *Math. Methods Statist.*, 16(3), 2007.
[29] C. L. Mallows. Some comments on c_p. *Technometrics*, 15:661–676, 1973.
[30] P. Massart. *Concentration Inequalities and Model Selection (Saint-Flour Summer School on Probability Theory 2003, ed. J. Picard)*. Springer, 2007.
[31] D. A. McAllester. Some pac-bayesian theorems. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory (Madison, WI, 1998)*, pages 230–234. ACM, 1998.
[32] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. ISBN 3-900051-07-0.
[33] P. Rigollet and A.B. Tsybakov. Exponential screening and optimal rates of sparse estimation. manuscript. manuscript, 2010.
[34] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464, 1978.
[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, COLT'97*, pages 2–9. ACM, 1997.
[36] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B*, 58(1):267–288, 1996.
[37] A.B. Tsybakov. Optimal rates of aggregation. In *Computationnal Learning theory and Kernel Machines (COLT)*, volume 2777 of *Lecture Notes in Artificial Intelligence*, pages 303–313. Springer, heidelberg, 2003.

CREST and LPMA -Université Paris 7, University of Cambridge
E-mail address: alquier@ensae.fr, k.lounici@statslab.cam.ac.uk