Submitted to the Annals of Statistics

ON THE EXPONENTIALLY WEIGHTED AGGREGATE WITH THE LAPLACE PRIOR

By Arnak S. Dalalyan, Edwin Grappin and Quentin Paris

CREST, ENSAE, Université Paris-Saclay, National Research University - Higher School of Economics

In this paper, we study the statistical behaviour of the Exponentially Weighted Aggregate (EWA) in the problem of high-dimensional regression with fixed design. Under the assumption that the underlying regression vector is sparse, it is reasonable to use the Laplace distribution as a prior. The resulting estimator and, specifically, a particular instance of it referred to as the Bayesian lasso, was already used in the statistical literature because of its computational convenience, even though no thorough mathematical analysis of its statistical properties was carried out. The present work fills this gap by establishing sharp oracle inequalities for the EWA with the Laplace prior. These inequalities show that if the temperature parameter is small, the EWA with the Laplace prior satisfies the same type of oracle inequality as the lasso estimator does, as long as the quality of estimation is measured by the prediction loss. Extensions of the proposed methodology to the problem of prediction with low-rank matrices are considered.

1. Introduction. We investigate statistical properties of the Exponentially Weighted Aggregate (EWA) in the context of high-dimensional linear regression with fixed design and under the sparsity scenario. This corresponds to considering data that consist of n random observations $y_1, \dots, y_n \in \mathbb{R}$ and p fixed covariates $\boldsymbol{x}^1, \dots, \boldsymbol{x}^p \in \mathbb{R}^n$. We further assume that there is a vector $\boldsymbol{\beta}^\star \in \mathbb{R}^p$ such that the residuals $\xi_i = y_i - \beta_1^\star \boldsymbol{x}_i^1 - \ldots - \beta_p^\star \boldsymbol{x}_i^p$ are independent, zero mean random variables. In vector notation, this reads as

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta}^\star + \boldsymbol{\xi}, \tag{1}$$

where $\boldsymbol{y} = (y_1, \dots, y_n)^\top$ is the response vector, $\mathbf{X} = (\boldsymbol{x}^1, \dots, \boldsymbol{x}^p) \in \mathbb{R}^{n \times p}$ is the design matrix and $\boldsymbol{\xi}$ is the noise vector. For simplicity, in all mathematical results, the noise vector is assumed to be distributed according to the Gaussian distribution $\mathcal{N}(0, \sigma^2 \mathbf{I}_n)$. We are mainly interested in obtaining mathematical results that cover the high-dimensional setting. This means that our goal is to establish risk bounds that can be small even if the ambient dimension p

MSC 2010 subject classifications: Primary 62J05; secondary 62H12

Keywords and phrases: sparsity, Bayesian lasso, oracle inequality, exponential weights, high-dimensional regression, trace regression, low-rank matrices

is large compared to the sample size. In order to attain this goal, we will consider the, by now, usual sparsity scenario. In other words, the established risk bounds are small if the underlying large vector $\boldsymbol{\beta}^*$ is well approximated by a sparse vector. Note that this setting can be extended to the matrix case, sometimes termed trace-regression (Koltchinskii et al., 2011; Rohde and Tsybakov, 2011). Indeed, if the rows $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ of the design matrix $\mathbf{X}$ are replaced by $m_1 \times m_2$ matrices $\mathbf{X}_1, \dots, \mathbf{X}_n$, then the regression vector $\boldsymbol{\beta}^\star$ is replaced by a $m_1 \times m_2$ matrix $\mathbf{B}^\star$ and the model of trace regression is

$$y_i = \mathrm{Tr}(\mathbf{X}_i^\top \mathbf{B}^\star) + \xi_i, \qquad i = 1, \dots, n.$$

Our focus here is on the statistical properties related to the prediction risk. The important questions of variable selection and estimation in various norms are beyond the scope of the present work.

In the aforementioned vector- and trace-regression models, the most thoroughly studied statistical procedures of estimation and prediction rely on the principle of penalised least squares[1]. In the vector-regression model, assuming that the quadratic loss is used, this corresponds to analysing the properties of the estimator

$$\widehat{\boldsymbol{\beta}}^{\mathrm{PLS}} \in \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \left\{ \frac{1}{2n} \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2 + \lambda \, \mathrm{Pen}(\boldsymbol{\beta}) \right\}, \tag{2}$$

where $\lambda > 0$ is a tuning parameter and $\mathrm{Pen} : \mathbb{R}^p \to \mathbb{R}$ is a sparsity promoting penalty function. The literature on this topic is so rich that it would be impossible to cite here all the relevant papers. We refer the interested reader to the books (Bühlmann and van de Geer, 2011; Giraud, 2015; Koltchinskii, 2011; van de Geer, 2016) and the references therein. Among the sparsity promoting penalties, one can mention the ℓ_0 penalty (which for various choices of λ leads to the BIC (Schwarz, 1978), the AIC (Akaike, 1974) or to Mallows's Cp (Mallows, 1973)), the ℓ_1 penalty or the lasso (Tibshirani, 1996), the ℓ_q (with $0 < q < 1$) or the bridge penalty (Frank and Friedman, 1993; Fu, 1998), the SCAD (Fan and Li, 2001), the minimax concave penalties (Zhang, 2010), the entropy (Koltchinskii, 2009), the SLOPE (Bogdan et al., 2015; Su and Candès, 2016), etc.

The aggregation by exponential weights is an alternative approach to the problems of estimation and prediction that, roughly speaking, replaces the minimisation by the averaging. Assuming that every vector $\boldsymbol{\beta} \in \mathbb{R}^p$ is a candidate for estimating the true vector $\boldsymbol{\beta}^\star$, aggregation (cf., for instance, the survey (Tsybakov, 2014)) consists in computing a weighted average of

[1] Or, more generally, on the penalised empirical risk minimisation

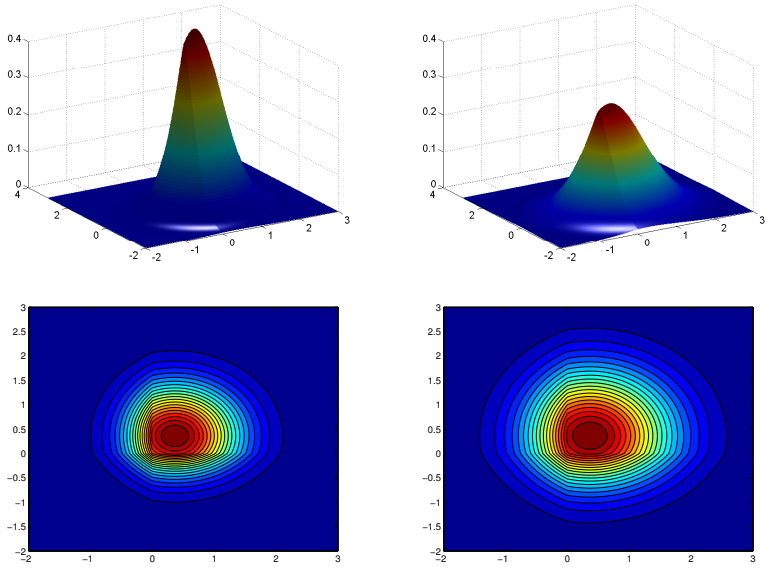
the candidates. Naturally, the weights are to be chosen in a data-driven way. In the case of the exponentially weighted aggregate (EWA), the weight $\widehat{\pi}_n(\boldsymbol{\beta})$ of each candidate vector $\boldsymbol{\beta}$ has the exponential form

$$\widehat{\pi}_n(\boldsymbol{\beta}) \propto \exp\big(-V_n(\boldsymbol{\beta})/\tau\big), \quad \text{where} \quad V_n(\boldsymbol{\beta}) = \frac{1}{2n}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2 + \lambda\,\mathrm{Pen}(\boldsymbol{\beta})$$

is the potential used above for defining the penalised least squares estimator and $\tau > 0$ is an additional tuning parameter referred to as the temperature. Using this notation, the EWA is defined by

$$\widehat{\boldsymbol{\beta}}^{\mathrm{EWA}} = \int_{\mathbb{R}^p} \boldsymbol{\beta}\,\widehat{\pi}_n(\boldsymbol{\beta})\,\mathrm{d}\boldsymbol{\beta}. \tag{3}$$

Exponential weights have been used for a long time in statistical learning theory (cf., for instance, Vovk (1990)). Their use in statistics was initiated by Yuhong Yang in (Yang, 2000a,b,c, 2001) and by Olivier Catoni in a series of preprints, later on included in (Catoni, 2004, 2007). Precise risk bounds for the EWA in the model of regression with fixed design have been established in (Chernousova et al., 2013; Dai et al., 2014; Dalalyan and Salmon, 2012; Dalalyan and Tsybakov, 2007, 2008, 2012b; Golubev and Ostrovski, 2014; Leung and Barron, 2006). In the model of regression with random design, the counterpart of the EWA, often referred to as mirror averaging, has been thoroughly studied in (Audibert, 2009; Chesneau and Lecué, 2009; Dalalyan and Tsybakov, 2012a; Gaïffas and Lecué, 2007; Juditsky et al., 2008; Lecué and Mendelson, 2013; Yuditskiĭ et al., 2005). Note that when the temperature τ equals σ^2/n, the EWA coincides with the Bayesian posterior mean in the regression model with Gaussian noise provided that the prior is defined by $\pi_0(\boldsymbol{\beta}) \propto \exp(-\lambda\,\mathrm{Pen}(\boldsymbol{\beta})/\tau)$. Thanks to this analogy, we will call $\widehat{\pi}_n$ pseudo-posterior density. Let us mention here that, considering the path $\tau \mapsto \widehat{\boldsymbol{\beta}}^{\mathrm{EWA}}$ for $\tau \in (0, \sigma^2/n]$, we get a continuous interpolation between the penalised least squares and the Bayesian posterior mean. Along with these studies, several authors have demonstrated the ability of the EWA to optimally estimate a sparse signal. To this end, various types of priors have been used. For instance, (Alquier and Lounici, 2011; Arias-Castro and Lounici, 2014; Leung and Barron, 2006; Rigollet and Tsybakov, 2011) have employed discrete priors over the set of least-squares estimators with varying supports whereas (Dalalyan and Tsybakov, 2008, 2012b) have used Student-type heavy-tailed priors. In the context of structured sparsity, the EWA has been successfully used in (Alquier and Biau, 2013; Dalalyan et al., 2014; Guedj and Alquier, 2013). Given the close relationship between the EWA and the Bayes estimator, it is worth mentioning here that the problem



$\tau = 0.5$ $\tau = 0.8$

FIG 1. ***Top:*** *the plots of the pseudo-posterior $\widehat{\pi}_n$ with the Laplace prior for the temperature $\tau = 0.5$ (left) and $\tau = 0.8$ (right). One can observe that decreasing the value of τ strengthens the peakedness of the density.* ***Bottom:*** *the level curves of the pseudo-posterior $\widehat{\pi}_n$ with the Laplace prior for the temperature $\tau = 0.5$ (left) and $\tau = 0.8$ (right). One clearly observes the non-differentiability of the density along the axes β_1 and β_2 (caused by the non-differentiability of the ℓ_1-norm).*

of sparse estimation has also received much attention in the literature on Bayesian Statistics (Hans, 2009; Park and Casella, 2008; Wipf et al., 2003). Posterior concentration properties for these methods have been investigated in (Castillo et al., 2015; Castillo and van der Vaart, 2012; Gao et al., 2015; van der Pas et al., 2016).

Despite these efforts, some natural questions remain open. One of them, described in details below, is at the origin of this work. Let us consider the prediction error of a candidate vector $\boldsymbol{\beta}$ with respect to the quadratic loss

$$\ell_n(\boldsymbol{\beta}, \boldsymbol{\beta}^\star) = \frac{1}{n}\|\mathbf{X}(\boldsymbol{\beta} - \boldsymbol{\beta}^\star)\|_2^2 = \frac{1}{n}\sum_{i=1}^n (\boldsymbol{x}_i^\top \boldsymbol{\beta} - \boldsymbol{x}_i^\top \boldsymbol{\beta}^\star)^2. \tag{4}$$

On the one hand, theoretical studies of the lasso (Bellec et al., 2016a,b; Belloni et al., 2014; Bickel et al., 2009; Candes and Tao, 2007; Dalalyan et al., 2017), established[2] sharp upper bounds for the prediction risk of the PLS

[2]Provided that the Gram matrix $\mathbf{X}^\top\mathbf{X}/n$ satisfies suitable assumptions (restricted

estimator (2) for the ℓ_1-penalty $\mathrm{Pen}(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_1$.

Therefore, one could expect the EWA with the Laplace prior $\pi_0(\boldsymbol{\beta}) \propto \exp(-\lambda\|\boldsymbol{\beta}\|_1/\tau)$ to have a high prediction performance. On the other hand, to the best of our knowledge, there is no result in the literature establishing accurate risk bounds for the EWA with Laplace prior. Indeed, a straightforward application of the PAC-Bayesian type risk bounds (McAllester, 1998) for the EWA (such as, for instance, Theorem 1 in (Dalalyan and Tsybakov, 2012b)) to the Laplace prior leads to strongly sub-optimal remainder terms. This raises the following questions:

Q1. Is the EWA with the Laplace prior suitable for prediction under the sparsity scenario?

Q2. If it is, what is the range of temperature τ providing good prediction accuracy?

Q3. How do the statistical properties of the EWA with the Laplace prior compare with those of the lasso?

Related questions are considered in (Castillo et al., 2015). Indeed, for $\boldsymbol{\beta}^\star = \mathbf{0}_p$, $p = n$ and $\mathbf{X}^\top\mathbf{X}/n = \mathbf{I}_n$, Theorem 7 from (Castillo et al., 2015) establishes the following property. For all the reasonable choices[3] of the tuning parameter λ, if the temperature τ in the EWA with the Laplace prior is chosen as $\tau = \sigma^2/n$, then the resulting posterior puts asymptotically no mass on the ball centered at $\boldsymbol{\beta}^\star$ and of radius $\mathrm{Const}(1/\log n)^{1/2}$. This negative result, stated in terms of the posterior contraction rate, can be easily adapted in order to show that, under the previous conditions, the Bayesian posterior mean is sub-optimal.

The present paper completes the picture by establishing some positive results. In particular, it turns out that if the temperature parameter of the EWA with the Laplace prior is of the order $s\sigma^2/(pn)$, where s is the sparsity of $\boldsymbol{\beta}^\star$, then the EWA with the Laplace prior does attain the optimal rate of convergence. Furthermore, it satisfies the same type of sharp sparsity inequality as the lasso does. Interestingly, the proof of this result is based on arguments which differ from those used in the aggregation literature. Indeed, the two previously used techniques for getting oracle inequalities for the EWA and related procedures rely either on the PAC-Bayesian inequality or on the Stein unbiased risk estimate. Instead, the key idea of our proof is to

isometry, restricted eigenvalues, compatibility, etc.).

[3]By "reasonable" we understand here the choice $\lambda = \mathrm{Const}\,\sigma(\frac{\log p}{n})^{1/2}$, for which the lasso is provably rate optimal under the sparsity scenario, provided that the design satisfies a version of the restricted eigenvalue condition.

take advantage of the following relations:

$$\int_{\mathbb{R}^p} \nabla\big(\beta_j^\alpha e^{-V_n(\boldsymbol{\beta})/\tau}\big)\,\mathrm{d}\boldsymbol{\beta} = 0, \qquad j = 1,\dots,p, \quad \alpha = 0, 1.$$

Hence, most of our arguments are independent of the noise distribution and can be extended to other settings (as opposed to the results relying on the Stein formula). Elaborating on this, we prove that the pseudo-posterior $\widehat{\pi}_n$ puts an overwhelming weight on the set of vectors $\boldsymbol{\beta}$ satisfying a sharp oracle inequality with rate-optimal remainder term. In the case of the Gaussian noise, we also obtain the explicit form of the Stein unbiased estimator of the risk of $\widehat{\boldsymbol{\beta}}^{\mathrm{EWA}}$, which can be used for choosing the tuning parameter. Finally, we extend these results to the model of trace regression when the underlying true matrix $\mathbf{B}^\star$ has low rank.

The rest of the paper is organised as follows. The notation used throughout the paper is introduced in the next section. Section 3 analyses the prediction loss of the EWA with the Laplace prior, and Section 4 gathers results characterising the concentration of the pseudo-posterior $\widehat{\pi}_n$. Extensions of these results to the case where the unknown parameter is a (nearly) low-rank matrix are considered in Section 5. A brief summary of the obtained results along with some conclusions is given in Section 6. Finally, the most important proofs are postponed to Section 7.

2. Notation. This paragraph collects notation used throughout the paper. For every integer $k \geq 1$, we write $\mathbf{1}_k$ (resp. $\mathbf{0}_k$) for the vector of $\mathbb{R}^k$ having all coordinates equal to one (resp. zero). We set $[k] = \{1,\dots,k\}$. For every $q \in [0,\infty]$, we denote by $\|\boldsymbol{u}\|_q$ the usual ℓ_q-norm of $\boldsymbol{u} \in \mathbb{R}^k$, that is $\|\boldsymbol{u}\|_q = (\sum_{j\in[k]} |u_j|^q)^{1/q}$ when $0 < q < \infty$, $\|\boldsymbol{u}\|_0 = \mathrm{Card}(\{j : u_j \neq 0\})$ and $\|\boldsymbol{u}\|_\infty = \max_{j\in[k]} |u_j|$. For every integer $k \geq 1$ and any $T \subset [k]$, we denote by T^c and $|T|$ the complementary set $[p] \setminus T$ and the cardinality of T, respectively. For $\boldsymbol{u} \in \mathbb{R}^k$ and $T \subset [k]$, we denote $\boldsymbol{u}_T \in \mathbb{R}^{|T|}$ the vector obtained from $\boldsymbol{u}$ by removing all the coordinates belonging to the set T^c.

In Sections 3 and 4, we recall that $\mathbf{X} \in \mathbb{R}^{n\times p}$ refers to the deterministic design matrix with columns $\boldsymbol{x}^1,\dots,\boldsymbol{x}^p \in \mathbb{R}^n$ and rows $\boldsymbol{x}_1,\dots,\boldsymbol{x}_n \in \mathbb{R}^p$. Finally, our analysis will involve the compatibility factor of the design matrix defined, for any $J \subset [p]$ and $c > 0$, by

$$\kappa_{J,c} = \inf_{\boldsymbol{u}\in\mathbb{R}^p : \|\boldsymbol{u}_{J^c}\|_1 < c\|\boldsymbol{u}_J\|_1} \frac{c^2|J|\|\mathbf{X}\boldsymbol{u}\|_2^2}{n(c\|\boldsymbol{u}_J\|_1 - \|\boldsymbol{u}_{J^c}\|_1)^2}. \tag{5}$$

Note that the compatibility factor, often used for the analysis of the lasso, is slightly larger[4] than the restricted eigenvalue (Bickel et al., 2009). For a

[4]Since this factor appears in the denominator of the risk bound, the larger is the better.

better understanding of these (and related) quantities we refer the reader to (Bickel et al., 2009, Sections 3 and 4) and (van de Geer and Bühlmann, 2009).

Risk bounds established in the present work for the EWA contain a new term, as compared to the analogous risk bounds for the lasso. This term reflects the peakedness of the pseudo-posterior density $\widehat{\pi}_n$ and is defined by

$$H(\tau) = p\tau - \int G(\boldsymbol{u})\widehat{\pi}_n(\boldsymbol{u})\mathrm{d}\boldsymbol{u} + G(\widehat{\boldsymbol{\beta}}^{\mathrm{EWA}}), \tag{6}$$

where $G(\boldsymbol{u}) = {}^1\!/_n\|\mathbf{X}\boldsymbol{u}\|_2^2 + \lambda\|\boldsymbol{u}\|_1$. When the temperature τ is low, close to zero, the pseudo-posterior $\widehat{\pi}_n$ is close to a Dirac measure centred at the lasso, which implies that $H(\tau)$ is close to zero. Furthermore, since the above function G is convex, we have the following bound

$$H(\tau) \le p\tau.$$

In Section 3 and Section 4 we will occasionally use the following matrix notation. For all integers $p \ge 1$, $\mathbf{I}_p$ refers to the identity matrix in $\mathbb{R}^{p\times p}$. For any integers $p \ge 1$ and $q \ge 1$, any matrix $\mathbf{A} \in \mathbb{R}^{p\times q}$ and any subset T of $[q]$, we denote by $\mathbf{A}_T$ the matrix obtained from $\mathbf{A}$ by removing all the columns belonging to T^c. Finally the transpose and the Moore-Penrose pseudoinverse of a matrix $\mathbf{A}$ are denoted by $\mathbf{A}^\top$ and $\mathbf{A}^\dagger$, respectively.

3. Risk bound for the EWA with the Laplace prior. This section is devoted to discussing statistical properties of the EWA with the Laplace prior. Recall that it is defined by (3) as the average with respect to the pseudo-posterior density

$$\widehat{\pi}_n(\boldsymbol{\beta}) \propto \exp(-V_n(\boldsymbol{\beta})/\tau), \quad \text{where } V_n(\boldsymbol{\beta}) = \frac{1}{2n}\sum_{i=1}^n (y_i - \boldsymbol{x}_i^\top\boldsymbol{\beta})^2 + \lambda\,\|\boldsymbol{\beta}\|_1. \tag{7}$$

The emphasis is put on non-asymptotic guarantees in terms of the prediction loss. It is important to mention here that the Laplace prior, $\pi_0(\boldsymbol{\beta}) \propto \exp(-\lambda\|\boldsymbol{\beta}\|_1/\tau)$, makes use of the same scale for all the coordinates of the vector $\boldsymbol{\beta}$. This presumes that the covariates (columns of the matrix $\mathbf{X}$) are already rescaled so that their Euclidean norms are almost equal. An alternative approach (see, for instance, Bickel et al. (2009); Bunea et al. (2007))—that we will not follow here—would consist in replacing the ℓ_1-norm of $\boldsymbol{\beta}$ by the weighted ℓ_1-norm $\sum_{j\in[p]}\|\boldsymbol{x}^j\||\beta_j|$. The next result provides the main risk bound for the EWA.

THEOREM 1. *Assume that data are generated by model* (1) *with* $\boldsymbol{\xi}$ *drawn from the Gaussian distribution* $\mathcal{N}(0, \sigma^2 \mathbf{I}_n)$ *and that the covariates are rescaled so that* $\max_{j\in[p]} {}^1\!/_n \|\boldsymbol{x}^j\|_2^2 \le 1$. *Suppose, in addition, that* $\lambda \ge 2\sigma({}^2\!/_n \log(p/\delta))^{1/2}$, *for some* $\delta \in (0,1)$. *Then, with probability at least* $1-\delta$,

$$\ell_n(\widehat{\boldsymbol{\beta}}^{\mathrm{EWA}}, \boldsymbol{\beta}^\star) \le \inf_{\substack{\bar{\boldsymbol{\beta}}\in\mathbb{R}^p \\ J\subset[p]}} \left\{ \ell_n(\bar{\boldsymbol{\beta}}, \boldsymbol{\beta}^\star) + 4\lambda \|\bar{\boldsymbol{\beta}}_{J^c}\|_1 + \frac{9\lambda^2|J|}{4\kappa_{J,3}} \right\} + 2p\tau, \tag{8}$$

where ℓ_n *is defined in* (4) *and* $\widehat{\boldsymbol{\beta}}^{\mathrm{EWA}}$ *is defined in* (3) *and* (7).

For the lasso estimator, risk bounds of this nature have been developed in (Bellec et al., 2016a; Dalalyan et al., 2017; Koltchinskii et al., 2011; Sun and Zhang, 2012). The risk bound in (8) extends the risk bounds available for the lasso (cf. Theorem 2 in (Dalalyan et al., 2017)) to the EWA with the Laplace prior. Indeed, letting the temperature τ go to zero, the last term in the right-hand side of (8) disappears and we retrieve the risk bound for the lasso. An attractive feature of risk bound (8) is that the factor in front of the term $\ell_n(\bar{\boldsymbol{\beta}}, \boldsymbol{\beta}^\star)$ is equal to one; this is often referred to as a sharp or exact oracle inequality. Furthermore, the other three terms in the right-hand side of (8) are neat and have a simple interpretation. The second term, $4\lambda\|\bar{\boldsymbol{\beta}}_{J^c}\|_1$, accounts for the approximate sparsity; when $\mathbf{X}\boldsymbol{\beta}^\star$ is well approximated by $\mathbf{X}\bar{\boldsymbol{\beta}}$ with a s-sparse vector $\bar{\boldsymbol{\beta}}$, then choosing $J = \{j : \bar{\beta}_j \neq 0\}$ annihilates this term. The third term of the risk bound corresponds to the optimal rate, up to a logarithmic factor, of estimation of a vector $\boldsymbol{\beta}^\star$ concentrated on the known set J. Indeed, if $|J| = s$ and the compatibility factor is bounded away from zero, this term is of order ${}^s\!/_n \log(p)$. Finally, the last term in the above risk bound, $2p\tau$, reflects the influence of the temperature parameter τ. In particular, it shows that if $\tau = \sigma^2/(pn)$ then this term is negligible with respect to the other remainder terms.

The inequality stated in Theorem 1 is a simplified version of the following one (proved in Section 7): for any $\gamma > 1$, in the event $\|\mathbf{X}^\top \boldsymbol{\xi}\|_\infty \le n\lambda/\gamma$, it holds

$$\ell_n(\widehat{\boldsymbol{\beta}}^{\mathrm{EWA}}, \boldsymbol{\beta}^\star) \le \inf_{\substack{\bar{\boldsymbol{\beta}}\in\mathbb{R}^p \\ J\subset[p]}} \left\{ \ell_n(\bar{\boldsymbol{\beta}}, \boldsymbol{\beta}^\star) + 4\lambda \|\bar{\boldsymbol{\beta}}_{J^c}\|_1 + \frac{\lambda^2(\gamma+1)^2|J|}{\gamma^2 \kappa_{J,(\gamma+1)/(\gamma-1)}} \right\} + 2H(\tau), \tag{9}$$

where $H(\tau)$ is defined in (6). On the one hand, one can use this more general result for getting an oracle inequality under more general assumptions on the noise distribution such as those considered, for instance, in (Belloni et al., 2014; Bunea et al., 2007). On the other hand, one can infer from (9) that

the term $H(\tau)$ highlights the difference, in terms of statistical complexity, between the lasso and the EWA with the Laplace prior. It is therefore important to get a precise evaluation of $H(\tau)$ as a function of τ, p and n, and to understand how tight the inequality $H(\tau) \leq p\tau$ is. To answer this question, we restrict our attention to orthonormal designs and show the tightness of the aforementioned inequality. To this end, let us introduce the scaled complementary error function $\Psi_v(t) = e^{t^2/2v} \frac{1}{\sqrt{2\pi v}} \int_t^\infty e^{-u^2/2v} \, \mathrm{d}u$.

PROPOSITION 1. *Let $\widehat{\boldsymbol{\Sigma}}_n = {}^1\!/_n \mathbf{X}^\top \mathbf{X}$ be the Gram matrix and $\widehat{\boldsymbol{\beta}}^{\mathrm{LS}} = {}^1\!/_n \widehat{\boldsymbol{\Sigma}}_n^\dagger \mathbf{X}^\top \boldsymbol{y}$ be the least-squares estimator. Then, we have*

$$H(\tau) = \|\widehat{\boldsymbol{\Sigma}}_n^{1/2} \widehat{\boldsymbol{\beta}}^{\mathrm{EWA}}\|_2^2 + \lambda \|\widehat{\boldsymbol{\beta}}^{\mathrm{EWA}}\|_1 - (\widehat{\boldsymbol{\beta}}^{\mathrm{EWA}})^\top \widehat{\boldsymbol{\Sigma}}_n \widehat{\boldsymbol{\beta}}^{\mathrm{LS}}.$$

Furthermore, when the design is orthonormal, that is $\widehat{\boldsymbol{\Sigma}}_n = \mathbf{I}_p$, then the EWA with the Laplace prior is a thresholding estimator, $\widehat{\beta}_j^{\mathrm{EWA}} = \mathrm{sign}(\widehat{\beta}_j^{\mathrm{LS}}) \big(|\widehat{\beta}_j^{\mathrm{LS}}| - \lambda w(\tau, \lambda, |\widehat{\beta}_j^{\mathrm{LS}}|) \big)$, where

$$w(\tau, \lambda, |\widehat{\beta}_j^{\mathrm{LS}}|) = \frac{\Psi_\tau(\lambda - |\widehat{\beta}_j^{\mathrm{LS}}|) - \Psi_\tau(\lambda + |\widehat{\beta}_j^{\mathrm{LS}}|)}{\Psi_\tau(\lambda - |\widehat{\beta}_j^{\mathrm{LS}}|) + \Psi_\tau(\lambda + |\widehat{\beta}_j^{\mathrm{LS}}|)},$$

and

$$H(\tau) = \sum_{j=1}^p \lambda \big(|\widehat{\beta}_j^{\mathrm{LS}}| - \lambda w(\tau, \lambda, |\widehat{\beta}_j^{\mathrm{LS}}|) \big) \big(1 - w(\tau, \lambda, |\widehat{\beta}_j^{\mathrm{LS}}|) \big).$$

The last expression of $H(\tau)$ provided by the proposition may be used for a numerical evaluation. First, let us note that if we set $\bar{\beta}_j = \widehat{\beta}_j^{\mathrm{LS}}/\sqrt{\tau}$ and $\bar{\lambda} = \lambda/\sqrt{\tau}$, the function $H(\tau)/\tau$ is independent of τ. Indeed, we have $H(\tau)/\tau = \sum_j h(\bar{\lambda}, |\bar{\beta}_j|)$ where

$$h(\bar{\lambda}, z) = \bar{\lambda} \big(z - \bar{\lambda} w(1, \bar{\lambda}, z) \big) \big(1 - w(1, \bar{\lambda}, z) \big), \qquad \forall z > 0.$$

In Figure 2 below, we plot the curves of the functions $z \mapsto h(\bar{\lambda}, z)$ for different values of the parameter $\bar{\lambda}$.

These curves clearly show that the bound $H(\tau) \leq p\tau$, a consequence of $h(\bar{\lambda}, z) \leq 1$, is tight. Another interesting observation is that the function $H(\tau)$ is always nonnegative. This basically implies that the value of τ minimising the right-hand side of (9) is $\tau = 0$. In other terms, the lowest risk bound is obtained for the lasso. This legitimately raises the following question: is there any advantage of using the EWA with the Laplace prior as compared to the lasso? Our firm conviction is that there is an advantage, and will try to explain our viewpoint in the rest of this section.

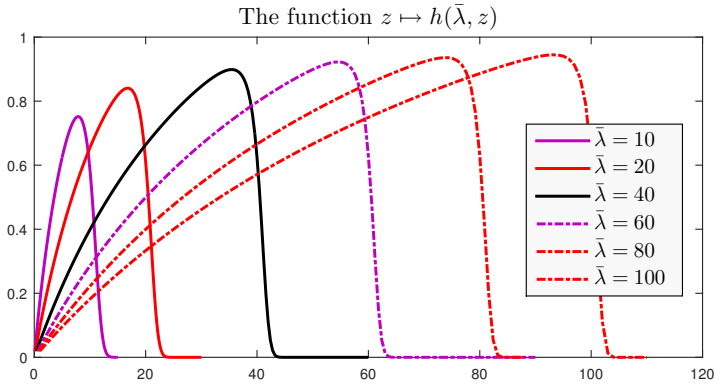


FIG 2. *For different values* $\bar{\lambda} \in \{10, 20, 40, 60, 80, 100\}$*, we plot the function* $z \mapsto h(\bar{\lambda}, z)$.

The point is that the lasso estimator is a nonsmooth function of the data. One of the consequences of this is that the Stein unbiased risk estimate (SURE) for the lasso is a discontinuous function of data. Indeed, as proved in (Tibshirani and Taylor, 2012), The SURE for the lasso (see also the earlier work (Donoho and Johnstone, 1995; Zou et al., 2007)) is given by

$$\widehat{R}^{\text{lasso}}(\lambda) = \frac{1}{n}\|\boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}^{\text{lasso}}(\lambda)\|_2^2 - \sigma^2 + \frac{2\sigma^2}{n}\text{rank}(\mathbf{X}_{\mathcal{A}(\lambda)}),$$

where $\mathcal{A}(\lambda) = \{j \in [p] : \widehat{\beta}_j^{\text{lasso}}(\lambda) \neq 0\}$ is the active set for the lasso estimator with the tuning parameter λ. In theory, this quantity $\widehat{R}^{\text{lasso}}(\lambda)$ can be used for choosing the tuning parameter λ of the lasso. However, in practice, this solution is rarely employed, since $\mathcal{A}(\lambda)$ has a very unstable behaviour as a function of λ and $\boldsymbol{y}$. As a consequence, not only one can get very different "optimal" values of λ for two very close vectors $\boldsymbol{y}$ and $\boldsymbol{y}'$, but is also likely to obtain very different "optimal" values of λ for the same vector $\boldsymbol{y}$ if using two different optimisation algorithms for computing an approximate solution to the lasso problem.

Using Stein's lemma, in the case where $\boldsymbol{\xi}$ is drawn from the Gaussian $\mathcal{N}(0, \sigma^2\mathbf{I}_n)$ distribution, one checks that

$$\widehat{R}^{\text{EWA}}(\lambda, \tau) = \frac{1}{n}\|\boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}^{\text{EWA}}_{\lambda,\tau}\|_2^2 - \sigma^2 + \frac{2\sigma^2}{n^2\tau}\int_{\mathbb{R}^p} \|\mathbf{X}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}^{\text{EWA}}_{\lambda,\tau})\|_2^2\, \widehat{\pi}_{n,\lambda,\tau}(\mathrm{d}\boldsymbol{\beta}) \tag{10}$$

is an unbiased estimator of the risk $\mathbb{E}[\ell_n(\widehat{\boldsymbol{\beta}}^{\text{EWA}}, \boldsymbol{\beta}^\star)]$. Furthermore, the function $(\lambda, \tau) \mapsto \widehat{R}^{\text{EWA}}(\lambda, \tau)$ is clearly continuous on $(0, \infty) \times (0, \infty)$. One can also check that the unbiased risk estimate $\widehat{R}^{\text{EWA}}(\lambda, \tau)$ depends continuously

on the data vector $\boldsymbol{y}$. Therefore, this quantity is arguably more robust to the variation in data and more regular as a function of the tuning parameters as compared to $\widehat{R}^{\text{lasso}}$. This implies that minimising $\widehat{R}^{\text{EWA}}(\lambda, \tau)$ with respect to λ or τ might be a good strategy for choosing these parameters adaptively.

Of course, this requires to be able to numerically compute the right-hand side of (10) or, equivalently, the mean and the covariance matrix of the pseudo-posterior distribution $\widehat{\pi}_n$. For smooth and strongly log-concave densities, the cost of such computations has been recently assessed in (Dalalyan, 2014; Durmus and Moulines, 2016). The adaptation of the approaches developed therein to the pseudo-posterior $\widehat{\pi}_n$, which is neither smooth nor strongly log-concave (but can be approximated by such a function), is an ongoing work.

4. Pseudo-Posterior concentration. Since the EWA estimator has a Bayesian flavour, it is appealing to look at the concentration properties of the pseudo-posterior distribution $\widehat{\pi}_n$. This is particularly important in the light of the results in Castillo et al. (2015) establishing that, for the temperature $\tau = \sigma^2/n$, the pseudo-posterior $\widehat{\pi}_n$ with the Laplace prior puts asymptotically no mass on the set of vectors $\boldsymbol{\beta}$ having a small prediction error. Furthermore, this result is proven for the orthonormal design matrix $\mathbf{X}$, which, intuitively, is a rather favourable situation for the Laplace prior.

The first property that we establish here and that characterises the concentration of the pseudo-posterior around its average is the following upper bound on the variance of the prediction $\mathbf{X}\boldsymbol{\beta}$ when $\boldsymbol{\beta}$ is drawn from $\widehat{\pi}_n$. (Recall that the matrix $\mathbf{X}$ has n rows, so the normalisation by multiplicative factor $1/n$ is natural.)

PROPOSITION 2. *If $\widehat{\pi}_n(\boldsymbol{u}) \propto \exp\left(-V_n(\boldsymbol{u})/\tau\right)$ is the pseudo-posterior with the Laplace prior defined by (7), then, for every $\bar{\boldsymbol{\beta}} \in \mathbb{R}^p$, we have*

$$\int_{\mathbb{R}^p} V_n(\boldsymbol{u})\, \widehat{\pi}_n(\boldsymbol{u})\, \mathrm{d}\boldsymbol{u} \le p\tau + V_n(\bar{\boldsymbol{\beta}}) - \frac{1}{2n} \int_{\mathbb{R}^p} \|\mathbf{X}(\boldsymbol{u} - \bar{\boldsymbol{\beta}})\|_2^2\, \widehat{\pi}_n(\boldsymbol{u})\, \mathrm{d}\boldsymbol{u}. \quad (11)$$

Furthermore, choosing $\bar{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}^{\text{EWA}} = \int_{\mathbb{R}^p} \boldsymbol{u}\, \widehat{\pi}_n(\boldsymbol{u})\, \mathrm{d}\boldsymbol{u}$, we get

$$\frac{1}{n} \int_{\mathbb{R}^p} \|\mathbf{X}(\boldsymbol{u} - \widehat{\boldsymbol{\beta}}^{\text{EWA}})\|_2^2\, \widehat{\pi}_n(\boldsymbol{u})\, \mathrm{d}\boldsymbol{u} \le p\tau. \quad (12)$$

The proof of this result is rather simple and plays an important role in the proof of the oracle inequality stated in Theorem 1. For these reasons, we opted for presenting this proof in this section, instead of postponing it to Section 7.

PROOF. The convexity of the function $\bar{\boldsymbol{\beta}} \mapsto \|\bar{\boldsymbol{\beta}}\|_1$ readily implies that the function $\bar{\boldsymbol{\beta}} \mapsto W_n(\bar{\boldsymbol{\beta}}) = V_n(\bar{\boldsymbol{\beta}}) - 1/2n \|\mathbf{X}(\boldsymbol{u} - \bar{\boldsymbol{\beta}})\|_2^2$ is a convex function, for every fixed $\boldsymbol{u} \in \mathbb{R}^p$. Furthermore, we have $W_n(\boldsymbol{u}) = V_n(\boldsymbol{u})$ and $\nabla W_n(\boldsymbol{u}) = \nabla V_n(\boldsymbol{u})$ at any point $\boldsymbol{u}$ of differentiability of V_n. Therefore,

$$V_n\big(\bar{\boldsymbol{\beta}}\big) \geq V_n(\boldsymbol{u}) + \big(\bar{\boldsymbol{\beta}} - \boldsymbol{u}\big)^\top \nabla V_n(\boldsymbol{u}) + \frac{1}{2n}\big\|\mathbf{X}(\boldsymbol{u} - \bar{\boldsymbol{\beta}})\big\|_2^2, \tag{13}$$

for all $\bar{\boldsymbol{\beta}} \in \mathbb{R}^p$ and for almost all $\boldsymbol{u} \in \mathbb{R}^p$ (those for which V_n is continuously differentiable at $\boldsymbol{u}$). Using the fundamental theorem of calculus, we remark that

$$\int_{\mathbb{R}^p} \nabla V_n(\boldsymbol{u})\, \widehat{\pi}(\boldsymbol{u})\, \mathrm{d}\boldsymbol{u} = -\tau \int_{\mathbb{R}^p} [\nabla \widehat{\pi}_n(\boldsymbol{u})]\, \mathrm{d}\boldsymbol{u} = \mathbf{0}_p \tag{14}$$

and that

$$\begin{aligned}\int_{\mathbb{R}^p} \boldsymbol{u}^\top \nabla V_n(\boldsymbol{u})\, \widehat{\pi}(\boldsymbol{u})\, \mathrm{d}\boldsymbol{u} - p\tau &= \int_{\mathbb{R}^p} \sum_{j=1}^p \left(u_j \frac{\partial V_n}{\partial u_j}(\boldsymbol{u}) - \tau \right) \widehat{\pi}(\boldsymbol{u})\, \mathrm{d}\boldsymbol{u} \\ &= -\tau \int_{\mathbb{R}^p} \sum_{j=1}^p \frac{\partial [u_j \widehat{\pi}_n(\boldsymbol{u})]}{\partial u_j}\, \mathrm{d}\boldsymbol{u} = 0. \end{aligned} \tag{15}$$

Integrating inequality (13) on $\mathbb{R}^p$ with respect to the density $\widehat{\pi}_n$ and using relations (14) and (15), we arrive at

$$V_n\big(\bar{\boldsymbol{\beta}}\big) \geq \int_{\mathbb{R}^p} V_n(\boldsymbol{u})\, \widehat{\pi}_n(\boldsymbol{u})\, \mathrm{d}\boldsymbol{u} - p\tau + \frac{1}{2n} \int_{\mathbb{R}^p} \big\|\mathbf{X}\big(\boldsymbol{u} - \bar{\boldsymbol{\beta}}\big)\big\|_2^2\, \widehat{\pi}_n(\boldsymbol{u})\, \mathrm{d}\boldsymbol{u}. \tag{16}$$

This completes the proof of the first claim of the proposition.

To prove the second claim, we replace $\bar{\boldsymbol{\beta}}$ by $\widehat{\boldsymbol{\beta}}^{\mathrm{EWA}}$ in (16). After rearranging the terms, this yields

$$\frac{1}{2n} \int_{\mathbb{R}^p} \big\|\mathbf{X}\big(\boldsymbol{u} - \widehat{\boldsymbol{\beta}}^{\mathrm{EWA}}\big)\big\|_2^2\, \widehat{\pi}_n(\boldsymbol{u})\, \mathrm{d}\boldsymbol{u} \leq p\tau + V_n\big(\widehat{\boldsymbol{\beta}}^{\mathrm{EWA}}\big) - \int_{\mathbb{R}^p} V_n\, \widehat{\pi}_n. \tag{17}$$

Using once again the fact that $\boldsymbol{u} \mapsto W_n(\boldsymbol{u}) = V_n(\boldsymbol{u}) - 1/2n \|\mathbf{X}(\boldsymbol{u} - \widehat{\boldsymbol{\beta}}^{\mathrm{EWA}})\|_2^2$ is a convex function, we obtain $V_n\big(\widehat{\boldsymbol{\beta}}^{\mathrm{EWA}}\big) = W_n\big(\widehat{\boldsymbol{\beta}}^{\mathrm{EWA}}\big) \leq \int W_n(\boldsymbol{u})\, \widehat{\pi}_n(\boldsymbol{u})\, \mathrm{d}\boldsymbol{u}$, which is equivalent to

$$V_n\big(\widehat{\boldsymbol{\beta}}^{\mathrm{EWA}}\big) - \int_{\mathbb{R}^p} V_n\, \widehat{\pi}_n \leq -\frac{1}{2n} \int_{\mathbb{R}^p} \big\|\mathbf{X}\big(\boldsymbol{u} - \widehat{\boldsymbol{\beta}}^{\mathrm{EWA}}\big)\big\|_2^2\, \widehat{\pi}_n(\boldsymbol{u})\, \mathrm{d}\boldsymbol{u}.$$

This inequality, combined with (17), completes the proof of (12) and of the proposition. □

REMARK 4.1. *A careful inspection of the proof reveals that the claims of the proposition are independent of the precise form of the ℓ_1-penalty. Therefore, the proposition still holds if we replace the ℓ_1-norm by any convex penalty.*

The second claim of the proposition establishes that the dispersion of the distribution $\widehat{\pi}_n$ around its average value $\widehat{\boldsymbol{\beta}}^{\mathrm{EWA}}$ is of the order $(p\tau)^{1/2}$. Interestingly, we show below that the same order of magnitude appears when we determine a region of concentration for the pseudo-posterior $\widehat{\pi}_n$. A key argument in the proof of the latter claim is the following result.

PROPOSITION 3 (Bobkov and Madiman (2011), Theorem 1.1). *Assume that $\widehat{\pi}_n(\boldsymbol{u}) \propto \exp(-V_n(\boldsymbol{u})/\tau)$ is a log-concave probability density*[5] *and let $\boldsymbol{\beta}$ be a random vector drawn from $\widehat{\pi}_n$. Then, for any $t > 0$, the inequality*

$$V_n(\boldsymbol{\beta}) \leq \int_{\mathbb{R}^p} V_n(\boldsymbol{u})\, \widehat{\pi}_n(\boldsymbol{u})\, \mathrm{d}\boldsymbol{u} + \tau\sqrt{p}\, t$$

holds with probability at least $1 - 2e^{-t/16}$.

Using this proposition, we establish the following result (the proof of which is postponed to Section 7) characterising the concentration of $\widehat{\pi}_n$.

THEOREM 2 (Posterior concentration bound). *Assume that data are generated by model (1) with $\boldsymbol{\xi} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_n)$ and rescaled covariates, i.e., $\max_{j\in[p]} {}^1\!/_n \|\boldsymbol{x}^j\|_2^2 \leq 1$. Let the quality of an estimator be measured by the squared prediction loss (4). Assume that the tuning parameter λ satisfies $\lambda \geq 2\sigma\big({}^2\!/_n \log({}^p\!/_\delta)\big)^{1/2}$, for some $\delta \in (0,1)$. Then, with probability at least $1-\delta$, the pseudo-posterior $\widehat{\pi}_n$ with the Laplace prior defined by (7) satisfies*

$$\widehat{\pi}_n\left(\boldsymbol{\beta} : \ell_n(\boldsymbol{\beta}, \boldsymbol{\beta}^\star) \leq \inf_{\substack{\bar{\boldsymbol{\beta}}\in\mathbb{R}^p \\ J\subset[p]}} \left\{ \ell_n(\bar{\boldsymbol{\beta}}, \boldsymbol{\beta}^\star) + 4\lambda \|\bar{\boldsymbol{\beta}}_{J^c}\|_1 + \frac{9\lambda^2|J|}{2\kappa_{J,3}} \right\} + 8p\tau \right) \geq 1 - 2e^{-\sqrt{p}/16}.$$

The latter theorem, in conjunction with Theorem 1, tells us that if we generate a random vector $\boldsymbol{\beta}$ distributed according to the density $\widehat{\pi}_n$, then with high probability it will have a prediction loss almost as small as the one of the EWA, the average with respect to $\widehat{\pi}_n$. This remark might be attractive from the computational point of view, since, at least for some distributions, drawing a random sample is easier than computing the expectation. Note also that by increasing the factor in front of the term $p\tau$ it is possible to

[5] This means that V_n is a convex function.

make the $\widehat{\pi}_n$-probability of the event considered in Theorem 2 even closer to one. The optimality of the term $\sqrt{p}$ in the argument of the exponential present in the last result remains a challenging open question which may be tackled using approaches developed in Hoffmann et al. (2015).

5. Sparsity oracle inequality in the matrix case. In this section, we extend the results of the previous sections to the problem of matrix regression with a low-rankness inducing prior. We first need to introduce additional notations used throughout this section.

5.1. *Specific notation.* For two matrices $\mathbf{A}$ and $\mathbf{B}$ of the same dimension, the scalar product is defined by

$$\langle \mathbf{A}, \mathbf{B} \rangle = \mathrm{Tr}(\mathbf{A}^\top \mathbf{B}).$$

The nuclear norm of a $p \times q$ matrix $\mathbf{A}$ is $\|\mathbf{A}\|_1 = \sum_{k=1}^r s_{\mathbf{A},k}$, where $s_{\mathbf{A},k}$ is the k-th largest singular value of $\mathbf{A}$ and $r = \mathrm{rank}(\mathbf{A})$. The operator norm is $\|\mathbf{A}\| = \sup_{\boldsymbol{x}\in\mathbb{R}^q} \|\mathbf{A}\boldsymbol{x}\|_2/\|\boldsymbol{x}\|_2 = s_{\mathbf{A},1}$. We denote by $\mathcal{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n) \in \mathbb{R}^{n\times m_1\times m_2}$ the three-dimensional tensor playing the role of the design matrix. Besides, let $\|\mathbf{A}\|^2_{L_2(\mathcal{X})} = \langle \mathbf{A}, \mathbf{A}\rangle_{L_2(\mathcal{X})}$ be the prediction loss defined via the "scalar product" $\langle \mathbf{A}, \mathbf{C}\rangle_{L_2(\mathcal{X})} = \frac{1}{n}\sum_{i=1}^n (\langle \mathbf{X}_i, \mathbf{A}\rangle \langle \mathbf{X}_i, \mathbf{C}\rangle)$. We will use the notation $\boldsymbol{u}^\top \mathcal{X} = \sum_{i\in[n]} u_i \mathbf{X}_i \in \mathcal{M}_{m_1,m_2}$ for the product of the tensor $\mathcal{X}$ with the vector $\boldsymbol{u} \in \mathbb{R}^n$.

We now need to define the matrix compatibility factor. Its definition is more involved than in the vector case because of the fact that the left and right singular spaces differ from one matrix to another. Let $\bar{\mathbf{B}}$ be any $m_1 \times m_2$ matrix of rank $r = \mathrm{rank}(\bar{\mathbf{B}})$ having the singular value decomposition $\bar{\mathbf{B}} = \mathbf{V}_1 \boldsymbol{\Sigma} \mathbf{V}_2^\top$. Here, $\boldsymbol{\Sigma}$ is a $r \times r$ diagonal matrix with positive diagonal entries, $\boldsymbol{\Sigma}_{11} \geq \ldots \geq \boldsymbol{\Sigma}_{rr} > 0$, and $\mathbf{V}_j$ is a $m_j \times r$ matrix with orthonormal columns for $j = 1, 2$. For any $J \subset [r]$ and $j = 1, 2$, we define $\mathbf{V}_{j,J}$ as the $m_j \times |J|$ matrix obtained from $\mathbf{V}_j$ by removing the columns with indices lying outside of J. This allows us to introduce the linear operators $\mathcal{P}_{\bar{\mathbf{B}},J^c}$ and $\mathcal{P}^\perp_{\bar{\mathbf{B}},J^c}$ from $\mathcal{M}_{m_1,m_2}$ to $\mathcal{M}_{m_1,m_2}$

$$\mathcal{P}_{\bar{\mathbf{B}},J^c}(\mathbf{U}) = (\mathbf{I}_{m_1} - \mathbf{V}_{1,J}\mathbf{V}_{1,J}^\top)\mathbf{U}(\mathbf{I}_{m_2} - \mathbf{V}_{2,J}\mathbf{V}_{2,J}^\top),$$
$$\mathcal{P}^\perp_{\bar{\mathbf{B}},J^c}(\mathbf{U}) = \mathbf{U} - \mathcal{P}_{\bar{\mathbf{B}},J^c}(\mathbf{U}).$$

We define, for every $\bar{\mathbf{B}} \in \mathcal{M}_{m_1,m_2}$, $J \subset [\mathrm{rank}(\bar{\mathbf{B}})]$ and $c > 0$, the compatibility factor

$$\kappa_{\bar{\mathbf{B}},J,c} = \inf_{\substack{\mathbf{U}\in\mathcal{M}_{m_1,m_2} \\ \|\mathcal{P}_{\bar{\mathbf{B}},J^c}(\mathbf{U})\|_1 < c\|\mathcal{P}^\perp_{\bar{\mathbf{B}},J^c}(\mathbf{U})\|_1}} \frac{c^2|J|\,\|\mathbf{U}\|^2_{L_2(\mathcal{X})}}{\big(c\|\mathcal{P}^\perp_{\bar{\mathbf{B}},J^c}(\mathbf{U})\|_1 - \|\mathcal{P}_{\bar{\mathbf{B}},J^c}(\mathbf{U})\|_1\big)^2}.$$

When $J = [\mathrm{rank}(\bar{\mathbf{B}})]$, we use the notation $\kappa_{\bar{\mathbf{B}},c}$ instead of $\kappa_{\bar{\mathbf{B}},J,c}$. Note that the set $\mathcal{C}(\bar{\mathbf{B}}, J, c) = \{\mathbf{U} \in \mathcal{M}_{m_1,m_2} : \ \|\mathcal{P}_{\bar{\mathbf{B}},J^c}(\mathbf{U})\|_1 < c\|\mathcal{P}^{\perp}_{\bar{\mathbf{B}},J^c}(\mathbf{U})\|_1\}$ defines the cone of dimensionality reduction. It consists of matrices $\mathbf{U}$ that can be written as a sum of two matrices $\mathbf{U}_1$ and $\mathbf{U}_2$ such that $\mathbf{U}_1$ is of small rank and dominates the possibly full-rank matrix $\mathbf{U}_2$, in the sense that $\|\mathbf{U}_2\|_1 \le c\|\mathbf{U}_1\|_1$. Indeed, it suffices to set $\mathbf{U}_1 = \mathcal{P}^{\perp}_{\bar{\mathbf{B}},J^c}(\mathbf{U})$ and to remark that $\mathcal{P}^{\perp}_{\bar{\mathbf{B}},J^c}(\mathbf{U}) = \mathbf{V}_{1,J}\mathbf{V}_{1,J}^{\top}\mathbf{U} + (\mathbf{I}_{m_1} - \mathbf{V}_{1,J}\mathbf{V}_{1,J}^{\top})\mathbf{U}\mathbf{V}_{2,J}\mathbf{V}_{2,J}^{\top}$ is of rank not exceeding $2|J|$.

Similarly to (6), we also define the function

$$H(\tau) = m_1 m_2 \tau - \int_{\mathcal{M}_{m_1,m_2}} G(\mathbf{U})\, \widehat{\pi}_n(\mathbf{U})\, \mathrm{d}\mathbf{U} + G\big(\widehat{\mathbf{B}}\big), \tag{18}$$

where $G(\mathbf{U}) = \|\mathbf{U}\|^2_{L_2(\mathcal{X})} + \lambda\|\mathbf{U}\|_1$. The convexity property of the function G entails that $H(\tau) \le m_1 m_2 \tau$ for every $\tau > 0$.

5.2. *Nuclear-norm prior and the exponential weights.* The observed outcomes are n real random variables $y_1, \dots, y_n \in \mathbb{R}$. Contrary to Sections 3 and 4 where the design points are $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n \in \mathbb{R}^p$, this section studies the situation in which we consider n design matrices $\mathbf{X}_i \in \mathbb{R}^{m_1 \times m_2}$ for $i \in [n]$. We further assume that there is a regression matrix $\mathbf{B}^\star \in \mathcal{M}_{m_1,m_2}$ such that

$$y_i = \mathrm{Tr}(\mathbf{X}_i^{\top}\mathbf{B}^\star) + \xi_i, \qquad i \in [n], \tag{19}$$

where the residuals ξ_i are independent and identically distributed according to a centred Gaussian distribution with variance σ^2. This model is referred to as trace-regression; see, for instance, Rohde and Tsybakov (2011). In this model, the nuclear norm is akin to the ℓ_1 norm in the vector case. Therefore, to some extent, the equivalent of the lasso estimator $\widehat{\mathbf{B}}^{\mathrm{NNP-LS}}_{\lambda}$ with a positive smoothing parameter λ, is defined by

$$\widehat{\mathbf{B}}^{\mathrm{NNP-LS}}_{\lambda} \in \arg\min_{\mathbf{B}\in\mathcal{M}_{m_1,m_2}} \bigg\{ \frac{1}{2n} \sum_{i\in[n]} (\boldsymbol{y}_i - \langle \mathbf{X}_i, \mathbf{B}\rangle)^2 + \lambda\|\mathbf{B}\|_1 \bigg\}.$$

This is the nuclear-norm penalized least-squares estimator. Similarly to the vector case, the above defined estimator $\widehat{\mathbf{B}}^{\mathrm{NNP-LS}}_{\lambda}$ is the maximum a posteriori estimator corresponding to the nuclear-norm prior

$$\pi_0(\mathbf{B}) \propto \exp\bigg\{ - \frac{\lambda\sigma^2\|\mathbf{B}\|_1}{n} \bigg\}.$$

This section investigates the prediction performance of the procedure obtained by replacing the optimisation step by averaging. In the matrix case,

we define the potential function V_n and the pseudo-posterior, respectively, by

$$V_n(\mathbf{B}) = \frac{1}{2n}\sum_{i\in[n]}(\boldsymbol{y}_i - \langle \mathbf{X}_i, \mathbf{B}\rangle)^2 + \lambda\|\mathbf{B}\|_1, \tag{20}$$

and $\widehat{\pi}_n(\mathbf{B}) \propto \exp\{-1/\tau V_n(\mathbf{B})\}$. Using these ingredients, we define the EWA with the nuclear-norm prior by

$$\widehat{\mathbf{B}}^{\mathrm{EWA}} = \int_{\mathcal{M}_{m_1,m_2}} \mathbf{B}\,\widehat{\pi}_n(\mathbf{B})\,\mathrm{d}\mathbf{B}. \tag{21}$$

We aim at studying the performance of this estimator in terms of the in-sample prediction loss

$$\ell_n\big(\widehat{\mathbf{B}}, \mathbf{B}^\star\big) = \|\widehat{\mathbf{B}} - \mathbf{B}^\star\|^2_{L^2(\mathcal{X})} = \frac{1}{n}\sum_{i=1}^n \langle \mathbf{X}_i, \widehat{\mathbf{B}} - \mathbf{B}^\star\rangle^2. \tag{22}$$

5.3. *Oracle Inequality.* The problem of assessing the quality of the nuclear-norm penalised estimators has received a great deal of attention; see, for instance, (Bunea et al., 2011; Candès and Plan, 2011; Candès and Tao, 2010; Gaïffas and Lecué, 2011; Klopp, 2014; Negahban and Wainwright, 2011, 2012; Srebro and Shraibman, 2005). Such an interest in these methods is mainly motivated by the variety of applications in computer vision and image analysis (Harchaoui et al., 2012; Shen and Wu, 2012), recommendation systems (Lim and Teh, 2007; Zhou et al., 2008), and many other areas. Bayesian approaches to the problem of low-rank matrix estimation and prediction has been recently analysed by Alquier (2013); Cottet and Alquier (2016); Mai and Alquier (2015).

Making the parallel with the sparse vector estimation and prediction problem, we can note that the counterpart of the vector sparsity $s = \|\boldsymbol{\beta}^\star\|_0$ in the matrix case is the product $(m_1 + m_2)\mathrm{rank}(\mathbf{B}^\star)$, representing the number of potentially nonzero terms in the singular values decomposition of $\mathbf{B}^\star$. Similarly, the counterpart of the ambient dimension p is the overall number of entries in $\mathbf{B}^\star$ that is $m_1 m_2$. In view of these analogies, the next theorem is a natural extension of Theorem 1 to the model of trace-regression. To state it, we need the following notation:

$$v_{\mathcal{X}} = \left\|\frac{1}{n}\sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i^\top\right\|^{1/2} \bigvee \left\|\frac{1}{n}\sum_{i=1}^n \mathbf{X}_i^\top\mathbf{X}_i\right\|^{1/2}.$$

THEOREM 3. *Assume that data are generated by model (19) with $\boldsymbol{\xi}$ drawn from the Gaussian distribution $\mathcal{N}(\mathbf{0}_n, \sigma^2\mathbf{I}_n)$. Suppose, in addition,*

that $\lambda \geq 2\sigma v_{\mathcal{X}}\{2/n \log((m_1+m_2)/\delta)\}^{1/2}$, *for some* $\delta \in (0,1)$. *Then, with probability at least* $1-\delta$, *the matrix* $\widehat{\mathbf{B}}^{\mathrm{EWA}}$ *defined in* (21) *satisfies*

$$\ell_n(\widehat{\mathbf{B}}^{\mathrm{EWA}}, \mathbf{B}^\star) \leq \inf_{\bar{\mathbf{B}}, J} \left\{ \ell_n(\bar{\mathbf{B}}, \mathbf{B}^\star) + 4\lambda \|\mathcal{P}_{\bar{\mathbf{B}}, J^c}(\bar{\mathbf{B}})\|_1 + \frac{9\lambda^2 |J|}{4\kappa_{\bar{\mathbf{B}}, J, 3}} \right\} + 2m_1 m_2 \tau, \tag{23}$$

where the inf is over all matrices $\bar{\mathbf{B}} \in \mathcal{M}_{m_1,m_2}$ *and all subsets* $J \subset [\mathrm{rank}(\bar{\mathbf{B}})]$.

This result can be seen as an extension of (Koltchinskii et al., 2011, Theorem 2) to the exponentially weighted aggregate with a prior proportional to the exponential of the scaled nuclear norm. Indeed, if we upper bound the infimum over all matrices $\mathbf{B}$ by the infimum over matrices such that $\mathrm{rank}(\mathbf{B}) \leq r$ for some given integer r, we easily see that (23) yields

$$\ell_n(\widehat{\mathbf{B}}^{\mathrm{EWA}}, \mathbf{B}^\star) \leq \inf_{\substack{\bar{\mathbf{B}} \in \mathcal{M}_{m_1,m_2} \\ \mathrm{rank}(\bar{\mathbf{B}}) \leq r}} \left\{ \ell_n(\bar{\mathbf{B}}, \mathbf{B}^\star) + \frac{9\lambda^2 r}{4\kappa_{\bar{\mathbf{B}}, 3}} \right\} + 2m_1 m_2 \tau.$$

An advantage of inequality (23) is that it offers a continuous interpolation between the so called "slow" and "fast" rates. "Slow" rates refer typically to risk bounds that are proportional to λ, whereas "fast" rates are proportional to λ^2. For procedures based on ℓ_1-norm or nuclear-norm penalty, "slow" rates are known to hold without any assumption on the design, while "fast" rates require a kind of compatibility assumption. In (23), taking $J = \varnothing$, the term with λ^2 disappears and we get the "slow" rate proportional to $\lambda\|\bar{\mathbf{B}}\|_1$. The other extreme case corresponding to $J = [\mathrm{rank}(\bar{\mathbf{B}})]$ leads to the "fast" rate proportional to $\lambda^2 \mathrm{rank}(\bar{\mathbf{B}})$, provided that the compatibility factor is bounded away from zero. The risk bound in (23) bridges these two extreme situations by providing the rate $\min_{q \in [r]}\{\lambda(s_{q+1,\bar{\mathbf{B}}} + \ldots + s_{r,\bar{\mathbf{B}}}) + \lambda^2 q\}$, where $r = \mathrm{rank}(\bar{\mathbf{B}})$ and $s_{\ell,\bar{\mathbf{B}}}$ is the ℓ-th largest singular value of $\bar{\mathbf{B}}$. Thus, our risk bound quantifies the quality of prediction in the situations where the true matrix (or the best prediction matrix) is nearly low-rank, but not necessarily exactly low-rank.

Similarly to the vector case, the inequality stated in Theorem 3 is a simplified version of the following one: for any $\gamma > 1$, in the event $\|\boldsymbol{\xi}^\top \mathcal{X}\| \leq n\lambda/\gamma$, the risk $\ell_n(\widehat{\mathbf{B}}^{\mathrm{EWA}}, \mathbf{B}^\star)$ is upper bounded by the expression

$$\inf_{\bar{\mathbf{B}}, J} \left\{ \ell_n(\bar{\mathbf{B}}, \mathbf{B}^\star) + 4\lambda \|\mathcal{P}_{\bar{\mathbf{B}}, J^c}(\bar{\mathbf{B}})\|_1 + \frac{\lambda^2(\gamma+1)^2 |J|}{\gamma^2 \kappa_{\bar{\mathbf{B}}, J, (\gamma+1)/(\gamma-1)}} \right\} + 2H(\tau),$$

where H is defined by (18). This inequality as well as Theorem 3 is proved in Supplementary Material.

5.4. *Pseudo-posterior concentration.* In what follows, we state the result on the pseudo-posterior concentration in the matrix case. Akin to the vector case, one of the main building blocks is (Bobkov and Madiman, 2011, Theorem 1.1), see Proposition 3 above. Since the potential V_n in (20) is convex, the proposition applies and implies that, for every $t > 0$,

$$\widehat{\pi}_n\Big(\mathbf{B} : V_n(\mathbf{B}) \le \int_{\mathcal{M}} V_n(\mathbf{U})\,\widehat{\pi}_n(\mathbf{U})\,\mathrm{d}\mathbf{U} + \tau\sqrt{m_1 m_2}t\Big) \ge 1 - 2e^{-t/16}.$$

After some nontrivial algebra, this allows us to show that a risk bound similar to (3) holds not only for the pseudo-posterior-mean $\widehat{\mathbf{B}}^{\mathrm{EWA}}$, but also for any matrix $\mathbf{B}$ randomly sampled from $\widehat{\pi}_n$.

THEOREM 4. *Let data be generated by model* (19) *with* $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}_n, \sigma^2\mathbf{I}_n)$ *and let the quality of an estimator be measured by the prediction loss* (22). *Assume that* λ *satisfies* $\lambda \ge 2\sigma v_{\mathcal{X}}\{2/n \log((m_1+m_2)/\delta)\}^{1/2}$, *for some* $\delta \in (0,1)$. *Then, with probability at least* $1-\delta$, *the pseudo-posterior* $\widehat{\pi}_n$ *with the nuclear-norm prior defined by* (20) *is such that the* $\widehat{\pi}_n$*-probability of the event*

$$\ell_n(\mathbf{B}, \mathbf{B}^\star) \le \inf_{\substack{\bar{\mathbf{B}} \in \mathcal{M}_{m_1,m_2} \\ J \subset [\mathrm{rank}(\bar{\mathbf{B}})]}} \left\{ \ell_n(\bar{\mathbf{B}}, \mathbf{B}^\star) + 4\lambda \|\mathcal{P}_{\bar{\mathbf{B}},J^c}(\bar{\mathbf{B}})\|_1 + \frac{9\lambda^2|J|}{2\kappa_{\bar{\mathbf{B}},J,3}} \right\} + 8m_1 m_2 \tau$$

is larger than $1 - 2e^{-\sqrt{m_1 m_2}/16}$.

The proof of Theorem 4 is deferred to the Supplementary Material. One can deduce from Theorem 4 that if the temperature parameter τ is sufficiently small, for instance, $\tau \le \lambda^2/(m_1 m_2)$, then a random matrix sampled from the pseudo-posterior $\widehat{\pi}_n$ satisfies nearly the same oracle inequality as the nuclear-norm penalized least-squares estimator. Indeed, the term $8m_1 m_2 \tau$, which is the only difference between the two upper bounds, is in this case negligible with respect to the term involving λ^2.

6. Conclusions. We have considered the model of regression with fixed design and established risk bounds for the exponentially weighted aggregate with the Laplace prior. This class of estimators encompasses important particular cases such as the lasso and the Bayesian lasso. The risk bounds established in the present work exhibit a range of values for the temperature parameter for which the EWA with the Laplace prior has a risk bound of the same order as the lasso. This offers a valuable complement to the negative results by Castillo et al. (2015), which show that the Bayesian lasso is not rate-optimal in the sparsity scenario. Note that the Bayesian

lasso corresponds to the EWA with the Laplace prior for the temperature parameter $\tau = \sigma^2/n$, where σ^2 is the variance of the noise. Our results imply that in order to get rate-optimality in the sparsity scenario, it is sufficient to choose τ smaller than $\sigma^2/(np)$.

We have extended the result outlined in the previous paragraph in two directions. First, we have shown that one can replace the pseudo-posterior mean by any random sample from the pseudo-posterior distribution. This eventually increases the risk by a negligible additional term, but might be useful from a computational point of view. Second, we have established risk bounds of the same flavour in the case of trace-regression, when the unknown parameter is a nearly low-rank large matrix. This result extends those of (Koltchinskii et al., 2011) and unifies risk bounds leading to the "slow" and "fast" rates. Furthermore, our result offers an interpolation between these two extreme cases, see the discussion following Theorem 3.

With some additional work, all the results established in the present work can be extended to the model of regression with random design. Furthermore, the case of a partially labelled sample can be handled by coupling the methodology of the present work with that of (Bellec et al., 2016a). An interesting line of future research is to apply our approach to other priors constructed from convex penalties such as the mixed ℓ_1/ℓ_2-norm used in the group-lasso (Yuan and Lin, 2006), or the weighted ℓ_1-norm of ordered entries used in the slope (Bogdan et al., 2015). Another highly relevant and challenging topic for future work will be to investigate the computational complexity of various methods for approximating the pseudo-posterior mean or for drawing a sample from the pseudo-posterior density.

7. Proofs.

7.1. *Proof of the oracle inequality of Theorem 1.* To ease notation, throughout this section we write $\widehat{\boldsymbol{\beta}}$ instead of $\widehat{\boldsymbol{\beta}}^{\mathrm{EWA}}$. Furthermore, for a function $h : \mathbb{R}^p \to \mathbb{R}$, we often write $\int h\, \widehat{\pi}_n$ instead of $\int_{\mathbb{R}^p} h(\boldsymbol{u})\, \widehat{\pi}_n(\boldsymbol{u})\, \mathrm{d}\boldsymbol{u}$. We split the proof into three steps. The first step, carried out in Lemma 1, consists in deriving an initial upper bound on the prediction loss from the fundamental inequality stated in (11). The second step, performed in Lemma 2, shares many common features with the analogous developments for the lasso and provides a proof of (9). Finally, the third step is a standard bound of the probability of the event $\mathcal{E}_\gamma = \{\|\mathbf{X}^\top \boldsymbol{\xi}\|_\infty \le n\lambda/\gamma\}$ based on the union bound and properties of the Gaussian distribution.

LEMMA 1. *For any $\bar{\boldsymbol{\beta}} \in \mathbb{R}^p$, we have*

$$\ell_n(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}^\star) \le \ell_n(\bar{\boldsymbol{\beta}}, \boldsymbol{\beta}^\star) + 2H(\tau) + \frac{2}{n}\|\mathbf{X}^\top \boldsymbol{\xi}\|_\infty \|\widehat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}\|_1 + 2\lambda(\|\bar{\boldsymbol{\beta}}\|_1 - \|\widehat{\boldsymbol{\beta}}\|_1) - \frac{1}{n}\|\mathbf{X}(\bar{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}})\|_2^2.$$

PROOF. On the one hand, inequality (11) can be rewritten as

$$V_n(\widehat{\boldsymbol{\beta}}) \le V_n(\bar{\boldsymbol{\beta}}) + \underbrace{V_n(\widehat{\boldsymbol{\beta}}) - \int_{\mathbb{R}^p} V_n \, \widehat{\pi}_n + p\tau - \frac{1}{2n}\int_{\mathbb{R}^p} \|\mathbf{X}(\boldsymbol{u} - \bar{\boldsymbol{\beta}})\|_2^2 \, \widehat{\pi}_n(\mathrm{d}\boldsymbol{u})}_{:=A}. \tag{24}$$

On the other hand, one can check that

$$V_n(\widehat{\boldsymbol{\beta}}) - \int V_n \, \widehat{\pi}_n = \frac{1}{2n}\|\mathbf{X}\widehat{\boldsymbol{\beta}}\|_2^2 + \lambda\|\widehat{\boldsymbol{\beta}}\|_1 - \int \Big(\frac{1}{2n}\|\mathbf{X}\boldsymbol{u}\|_2^2 + \lambda\|\boldsymbol{u}\|_1\Big)\, \widehat{\pi}_n(\mathrm{d}\boldsymbol{u}),$$

$$\int_{\mathbb{R}^p} \|\mathbf{X}(\boldsymbol{u} - \bar{\boldsymbol{\beta}})\|_2^2 \, \widehat{\pi}_n(\mathrm{d}\boldsymbol{u}) = \|\mathbf{X}(\bar{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}})\|_2^2 + \int_{\mathbb{R}^p} \|\mathbf{X}\boldsymbol{u}\|_2^2 \, \widehat{\pi}_n(\mathrm{d}\boldsymbol{u}) - \|\mathbf{X}\widehat{\boldsymbol{\beta}}\|_2^2.$$

These inequalities, combined with the definition of H, given in (6), yield

$$A = \frac{1}{n}\|\mathbf{X}\widehat{\boldsymbol{\beta}}\|_2^2 + \lambda\|\widehat{\boldsymbol{\beta}}\|_1 - \int_{\mathbb{R}^p} \Big(\frac{1}{n}\|\mathbf{X}\boldsymbol{u}\|_2^2 + \lambda\|\boldsymbol{u}\|_1\Big)\, \widehat{\pi}_n(\mathrm{d}\boldsymbol{u}) + p\tau - \frac{1}{2n}\|\mathbf{X}(\bar{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}})\|_2^2 = H(\tau) - \frac{1}{2n}\|\mathbf{X}(\bar{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}})\|_2^2. \tag{25}$$

Finally, using the definitions of the prediction loss ℓ_n and the potential V_n, we get that the difference $\ell_n(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}^\star) - \ell_n(\bar{\boldsymbol{\beta}}, \boldsymbol{\beta}^\star)$ equals

$$2\big(V_n(\widehat{\boldsymbol{\beta}}) - V_n(\bar{\boldsymbol{\beta}})\big) + \frac{2}{n}\,\boldsymbol{\xi}^\top \mathbf{X}(\widehat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}) + 2\lambda(\|\bar{\boldsymbol{\beta}}\|_1 - \|\widehat{\boldsymbol{\beta}}\|_1). \tag{26}$$

In view of the duality inequality, the term $\boldsymbol{\xi}^\top \mathbf{X}(\widehat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}})$ is upper bounded in absolute value by $\|\mathbf{X}^\top \boldsymbol{\xi}\|_\infty \|\widehat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}\|_1$. Inserting this inequality and (24) in (26) and using relation (25), we get the claim of the lemma. □

According to Lemma 1, in the event $\mathcal{E}_\gamma = \{\|\mathbf{X}^\top \boldsymbol{\xi}\|_\infty \le n\lambda/\gamma\}$, the loss $\ell_n(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}^\star)$ is upper bounded by

$$\ell_n(\bar{\boldsymbol{\beta}}, \boldsymbol{\beta}^\star) + \frac{2\lambda}{\gamma}(\|\widehat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}\|_1 + \gamma\|\bar{\boldsymbol{\beta}}\|_1 - \gamma\|\widehat{\boldsymbol{\beta}}\|_1) + 2H(\tau) - \frac{1}{n}\|\mathbf{X}(\bar{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}})\|_2^2.$$

LEMMA 2. *For every $J \subset [p]$, we have*

$$\frac{2\lambda}{\gamma}(\|\widehat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}\|_1 + \gamma\|\bar{\boldsymbol{\beta}}\|_1 - \gamma\|\widehat{\boldsymbol{\beta}}\|_1) - \frac{1}{n}\|\mathbf{X}(\bar{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}})\|_2^2 \le 4\lambda\|\bar{\boldsymbol{\beta}}_{J^c}\|_1 + \frac{\lambda^2(\gamma+1)^2|J|}{\gamma^2 \kappa_{J,(\gamma+1)/(\gamma-1)}}.$$

This lemma is essentially a copy of Proposition 2 in (Bellec et al., 2016a). We provide here its proof for the sake of self-containedness.

PROOF. Let us fix a $J \subset \{1, \dots, p\}$ and set $\boldsymbol{u} = \widehat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}$. We have

$$\|\widehat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}\|_1 + \gamma\|\bar{\boldsymbol{\beta}}\|_1 - \gamma\|\widehat{\boldsymbol{\beta}}\|_1 = \|\boldsymbol{u}_J\|_1 + \|\boldsymbol{u}_{J^c}\|_1 + \gamma\|\bar{\boldsymbol{\beta}}_J\|_1 + \gamma\|\bar{\boldsymbol{\beta}}_{J^c}\|_1 - \gamma\|\widehat{\boldsymbol{\beta}}_J\|_1 - \gamma\|\widehat{\boldsymbol{\beta}}_{J^c}\|_1. \tag{27}$$

Using inequalities $\|\bar{\boldsymbol{\beta}}_J\|_1 - \|\widehat{\boldsymbol{\beta}}_J\|_1 \le \|\boldsymbol{u}_J\|_1$ and $\|\widehat{\boldsymbol{\beta}}_{J^c}\|_1 \ge \|\boldsymbol{u}_{J^c}\|_1 - \|\bar{\boldsymbol{\beta}}_{J^c}\|_1$, we deduce from equation (27) that

$$\|\widehat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}\|_1 + \gamma\|\bar{\boldsymbol{\beta}}\|_1 - \gamma\|\widehat{\boldsymbol{\beta}}\|_1 \le (\gamma+1)\|\boldsymbol{u}_J\|_1 - (\gamma-1)\|\boldsymbol{u}_{J^c}\|_1 + 2\gamma\|\bar{\boldsymbol{\beta}}_{J^c}\|_1. \tag{28}$$

Now, from the definition of the compatibility factor $\kappa_{J,c}$ given by equation (5), we infer

$$\|\boldsymbol{u}_J\|_1 - \frac{\gamma-1}{\gamma+1}\|\boldsymbol{u}_{J^c}\|_1 \le \left(\frac{|J|\|\mathbf{X}\boldsymbol{u}\|_2^2}{n\kappa_{J,(\gamma+1)/(\gamma-1)}}\right)^{1/2}. \tag{29}$$

Hence, inequalities (28) end (29) imply that

$$\frac{2\lambda}{\gamma}(\|\widehat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}\|_1 + \gamma\|\bar{\boldsymbol{\beta}}\|_1 - \gamma\|\widehat{\boldsymbol{\beta}}\|_1) - \frac{1}{n}\|\mathbf{X}(\bar{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}})\|_2^2 \le 4\lambda\|\bar{\boldsymbol{\beta}}_{J^c}\|_1 + 2ab - a^2,$$

where we have used the notation $a^2 = \|\mathbf{X}\boldsymbol{u}\|_2^2/n$ and $b^2 = \frac{\lambda^2(\gamma+1)^2|J|}{\gamma^2\kappa_{J,(\gamma+1)/(\gamma-1)}}$. Finally, noticing that

$$2ab - a^2 \le b^2 = \frac{\lambda^2(\gamma+1)^2|J|}{\gamma^2\kappa_{J,(\gamma+1)/(\gamma-1)}},$$

we get the claim of the lemma. □

Combining the claims of the previous lemmas and taking the minimum with respect to J and $\bar{\boldsymbol{\beta}}$, we obtain that the inequality

$$\ell_n(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}^\star) \le \inf_{\substack{\bar{\boldsymbol{\beta}} \in \mathbb{R}^p \\ J \subset [p]}} \left\{ \ell_n(\bar{\boldsymbol{\beta}}, \boldsymbol{\beta}^\star) + 4\lambda\|\bar{\boldsymbol{\beta}}_{J^c}\|_1 + \frac{\lambda^2(\gamma+1)^2|J|}{\gamma^2\kappa_{J,(\gamma+1)/(\gamma-1)}} \right\} + 2H(\tau) \tag{30}$$

holds in the event $\mathcal{E}_\gamma$. The third and the last step of the proof consists in assessing the probability of this event.

LEMMA 3. *If $\mathbf{X} = (\boldsymbol{x}^1, \dots, \boldsymbol{x}^p)$ is a $n \times p$ deterministic matrix with columns $\boldsymbol{x}^j$ satisfying $\|\boldsymbol{x}^j\|_2^2 \le n$ and if $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}_n, \sigma^2\mathbf{I}_n)$, then, for all $\varepsilon > 0$,*

$$\mathbf{P}\big(\|\mathbf{X}^\top\boldsymbol{\xi}\|_\infty > n\varepsilon\big) \le p\,\exp\big(-n\varepsilon^2/(2\sigma^2)\big).$$

PROOF. By the union bound, we get

$$\mathbf{P}\big(\|\mathbf{X}^\top\boldsymbol{\xi}\|_\infty > n\varepsilon\big) = \mathbf{P}\left(\max_{j\in[p]}|\boldsymbol{\xi}^\top\boldsymbol{x}^j| > n\varepsilon\right) \le \sum_{i=1}^{p}\mathbf{P}\big(|\boldsymbol{\xi}^\top\boldsymbol{x}^j| > n\varepsilon\big).$$

Then, noticing that for each $j\in[p]$ the random variable $\boldsymbol{\xi}^\top\boldsymbol{x}^j$ is distributed according to $\mathcal{N}(0,\sigma^2\|\boldsymbol{x}^j\|_2^2)$, we deduce that

$$\mathbf{P}\left(\|\mathbf{X}^\top\boldsymbol{\xi}\|_\infty > n\varepsilon\right) \le 2\sum_{j=1}^{p}\int_{n\varepsilon/(\sigma\|\boldsymbol{x}^j\|_2)}^{+\infty}\phi(u)\,\mathrm{d}u,$$

where ϕ stands for the probability density function of the standard Gaussian distribution. Finally, by using the inequality $\int_x^{+\infty}\phi(u)\,\mathrm{d}u \le 1/2\exp(-x^2/2)$ that holds for every $x>0$, we obtain the result. □

A proof of Theorem 1 can be deduced from the three previous lemmas as follows. Choosing $\gamma=2$ and $\varepsilon=\lambda/2\ge\sigma\sqrt{(2/n)\log(p/\delta)}$ in Lemma 3, we get that the event $\mathcal{E}_\gamma$ has a probability at least $1-\delta$. Furthermore, on this event, we have already established inequality (30). Finally, upper bounding $H(\tau)$ by $p\tau$ leads to the claim of the theorem.

7.2. *Proof of the concentration property of Theorem 2.* Let us introduce the set $\mathcal{B}=\{\boldsymbol{\beta}\in\mathbb{R}^p : V_n(\boldsymbol{\beta})\le\int V_n\,\widehat{\pi}_n + p\tau\}$. Applying Proposition 3 with $t=\sqrt{p}$, we get $\widehat{\pi}_n(\mathcal{B})\ge 1-2e^{-\sqrt{p}/16}$. To prove Theorem 2, it is sufficient to check that in the event $\mathcal{E}_\gamma$ (in particular, with $\gamma=2$), every vector $\boldsymbol{\beta}$ from $\mathcal{B}$ satisfies the inequality

$$\ell_n(\boldsymbol{\beta},\boldsymbol{\beta}^\star) \le \inf_{\substack{\bar{\boldsymbol{\beta}}\in\mathbb{R}^p\\ J\subset[p]}}\left\{\ell_n(\bar{\boldsymbol{\beta}},\boldsymbol{\beta}^\star) + 4\lambda\|\bar{\boldsymbol{\beta}}_{J^c}\|_1 + \frac{9\lambda^2|J|}{2\kappa_{J,3}}\right\} + 8p\tau.$$

In the rest of this proof, $\boldsymbol{\beta}$ is always a vector from $\mathcal{B}$. In view of (11), it satisfies

$$V_n(\boldsymbol{\beta}) \le 2p\tau + V_n(\bar{\boldsymbol{\beta}}) - \frac{1}{2n}\int_{\mathbb{R}^p}\|\mathbf{X}(\boldsymbol{u}-\bar{\boldsymbol{\beta}})\|_2^2\,\widehat{\pi}_n(\boldsymbol{u})\,\mathrm{d}\boldsymbol{u}. \tag{31}$$

Note that (31) holds for every $\bar{\boldsymbol{\beta}}\in\mathbb{R}^p$. Therefore, it also holds for $\bar{\boldsymbol{\beta}}=\boldsymbol{\beta}$ and yields

$$\frac{1}{n}\int_{\mathbb{R}^p}\|\mathbf{X}(\boldsymbol{u}-\boldsymbol{\beta})\|_2^2\,\widehat{\pi}_n(\boldsymbol{u})\,\mathrm{d}\boldsymbol{u} \le 4p\tau. \tag{32}$$

In addition, we have

$$\ell_n(\boldsymbol{\beta},\boldsymbol{\beta}^\star)-\ell_n(\bar{\boldsymbol{\beta}},\boldsymbol{\beta}^\star)=2\big(V_n(\boldsymbol{\beta})-V_n(\bar{\boldsymbol{\beta}})\big)+\frac{2}{n}\,\boldsymbol{\xi}^\top\mathbf{X}(\boldsymbol{\beta}-\bar{\boldsymbol{\beta}})+2\lambda(\|\bar{\boldsymbol{\beta}}\|_1-\|\boldsymbol{\beta}\|_1). \tag{33}$$

Combining (31), (33) and the duality inequality, we get that in $\mathcal{E}_\gamma$,

$$\begin{aligned}\ell_n(\boldsymbol{\beta},\boldsymbol{\beta}^\star)-\ell_n(\bar{\boldsymbol{\beta}},\boldsymbol{\beta}^\star)\le 4p\tau-\frac{1}{n}\int_{\mathbb{R}^p}\|\mathbf{X}(\boldsymbol{u}-\bar{\boldsymbol{\beta}})\|_2^2\,\widehat{\pi}_n(\boldsymbol{u})\,\mathrm{d}\boldsymbol{u}\\ +\frac{2\lambda}{\gamma}\,\|\boldsymbol{\beta}-\bar{\boldsymbol{\beta}}\|_1+2\lambda(\|\bar{\boldsymbol{\beta}}\|_1-\|\boldsymbol{\beta}\|_1).\end{aligned} \tag{34}$$

We use now the inequality $\|\mathbf{X}(\boldsymbol{u}-\bar{\boldsymbol{\beta}})\|_2^2\ge 1/2\|\mathbf{X}(\boldsymbol{\beta}-\bar{\boldsymbol{\beta}})\|_2^2-\|\mathbf{X}(\boldsymbol{u}-\boldsymbol{\beta})\|_2^2$, in conjunction with (32), to deduce from (34) that

$$\begin{aligned}\ell_n(\boldsymbol{\beta},\boldsymbol{\beta}^\star)-\ell_n(\bar{\boldsymbol{\beta}},\boldsymbol{\beta}^\star)\le 8p\tau+\frac{2\lambda}{\gamma}\,\|\boldsymbol{\beta}-\bar{\boldsymbol{\beta}}\|_1+2\lambda(\|\bar{\boldsymbol{\beta}}\|_1-\|\boldsymbol{\beta}\|_1)\\ -\frac{1}{2n}\|\mathbf{X}(\boldsymbol{\beta}-\bar{\boldsymbol{\beta}})\|_2^2.\end{aligned}$$

We can apply now Lemma 2 with $\boldsymbol{\beta}$ instead of $\widehat{\boldsymbol{\beta}}$ and $\mathbf{X}/\sqrt{2}$ instead of $\mathbf{X}$ in order to get the claim of Theorem 2.

7.3. *Proof of Proposition 1.* For the sake of simplicity, we abbreviate $\widehat{\boldsymbol{\beta}}=\widehat{\boldsymbol{\beta}}^{\mathrm{EWA}}$ and $\widehat{\boldsymbol{\beta}}^0=\widehat{\boldsymbol{\beta}}^{\mathrm{LS}}$ throughout the proof. In particular, notation $\widehat{\beta}_j$ (resp. $\widehat{\beta}_j^0$) will refer to the j-th entry of $\widehat{\boldsymbol{\beta}}^{\mathrm{EWA}}$ (resp. $\widehat{\boldsymbol{\beta}}^{\mathrm{LS}}$). First, observe that one can write the posterior density as $\widehat{\pi}(\boldsymbol{u})\propto\exp(-\bar{V}_n(\boldsymbol{u})/\tau)$ with

$$\begin{aligned}\bar{V}_n(\boldsymbol{u})&=V_n(\boldsymbol{u})-\frac{1}{2n}\|\boldsymbol{y}\|^2+\frac{1}{2}\|\widehat{\boldsymbol{\Sigma}}_n^{1/2}\widehat{\boldsymbol{\beta}}^0\|_2^2\\ &=\frac{1}{2}\|\widehat{\boldsymbol{\Sigma}}_n^{1/2}(\boldsymbol{u}-\widehat{\boldsymbol{\beta}}^0)\|_2^2+\lambda\|\boldsymbol{u}\|_1.\end{aligned} \tag{35}$$

On the one hand, the integration by parts formula yields

$$\int_{\mathbb{R}^p}[\boldsymbol{u}^\top\nabla\bar{V}_n(\boldsymbol{u})]\,\widehat{\pi}(\boldsymbol{u})\,\mathrm{d}\boldsymbol{u}=-\tau\int_{\mathbb{R}^p}\boldsymbol{u}^\top\nabla\widehat{\pi}(\boldsymbol{u})\,\mathrm{d}\boldsymbol{u}=p\tau.$$

On the other hand, the expression of $\bar{V}_n(\boldsymbol{u})$ written in (35) leads directly to

$$\int_{\mathbb{R}^p}[\boldsymbol{u}^\top\nabla\bar{V}_n(\boldsymbol{u})]\,\widehat{\pi}(\boldsymbol{u})\,\mathrm{d}\boldsymbol{u}=\int_{\mathbb{R}^p}G(\boldsymbol{u})\,\widehat{\pi}(\boldsymbol{u})\,\mathrm{d}\boldsymbol{u}-\widehat{\boldsymbol{\beta}}^\top\widehat{\boldsymbol{\Sigma}}_n\widehat{\boldsymbol{\beta}}^0,$$

where we recall that $G(\boldsymbol{u})=\|\mathbf{X}\boldsymbol{u}\|_2^2/n+\lambda\|\boldsymbol{u}\|_1=\|\widehat{\boldsymbol{\Sigma}}_n^{1/2}\boldsymbol{u}\|_2^2+\lambda\|\boldsymbol{u}\|_1$. This yields

$$\int_{\mathbb{R}^p}G(\boldsymbol{u})\,\widehat{\pi}(\boldsymbol{u})\,\mathrm{d}\boldsymbol{u}=p\tau+\widehat{\boldsymbol{\beta}}^\top\widehat{\boldsymbol{\Sigma}}_n\widehat{\boldsymbol{\beta}}^0,$$

and, hence,

$$H(\tau) = p\tau - \int_{\mathbb{R}^p} G(\boldsymbol{u})\,\widehat{\pi}(\boldsymbol{u})\,\mathrm{d}\boldsymbol{u} + \|\widehat{\boldsymbol{\Sigma}}_n^{1/2}\widehat{\boldsymbol{\beta}}\|_2^2 + \lambda\|\widehat{\boldsymbol{\beta}}\|_1$$
$$= \|\widehat{\boldsymbol{\Sigma}}_n^{1/2}\widehat{\boldsymbol{\beta}}\|_2^2 + \lambda\|\widehat{\boldsymbol{\beta}}\|_1 - \widehat{\boldsymbol{\beta}}^\top\widehat{\boldsymbol{\Sigma}}_n\widehat{\boldsymbol{\beta}}^0, \tag{36}$$

which proves the first claim of Proposition 1. Let us now consider the case where $\widehat{\boldsymbol{\Sigma}}_n = \mathbf{I}_p$. Then, recalling the definition of $\bar{V}_n(\boldsymbol{u})$ in (35), a straightforward calculation reveals that

$$\bar{V}_n(\boldsymbol{u}) = -\frac{\lambda^2 p}{2} + \sum_{j=1}^p \left[\frac{1}{2}\left(u_j - \widehat{\beta}_j^0 + \lambda\,\mathrm{sign}(u_j)\right)^2 + \lambda\widehat{\beta}_j^0\,\mathrm{sign}(u_j)\right].$$

Hence, we deduce that $\widehat{\pi}(\boldsymbol{u}) = \prod_{j=1}^p \widehat{\pi}_j(u_j)$ where

$$\widehat{\pi}_j(t) \propto \exp\Big(-\frac{1}{2\tau}(t - \widehat{\beta}_j^0 + \lambda\,\mathrm{sign}(t))^2 - \frac{\lambda}{\tau}\widehat{\beta}_j^0\,\mathrm{sign}(t)\Big).$$

Next, let $\varphi(t) = \int_t^{+\infty}\phi(x)\mathrm{d}x$ where ϕ denotes the density function of the standard normal distribution. For a fixed $j \in [p]$, we consider the abbreviations $a = \lambda/\sqrt{\tau}$ and $b = \widehat{\beta}_j^0/\sqrt{\tau}$. Then, the change of variable $u = t/\sqrt{\tau}$ in the first integral below, together with the observation that $\mathrm{sign}(t) = \mathrm{sign}(t/\sqrt{\tau})$ for all real t, leads to

$$\widehat{\beta}_j = \int t\,\widehat{\pi}_j(t)\,\mathrm{d}t = \sqrt{\tau}\,\frac{\int u\exp\{-\frac{1}{2}(u-b+a\,\mathrm{sign}(u))^2 - ab\,\mathrm{sign}(u)\}\,\mathrm{d}u}{\int \exp\{-\frac{1}{2}(u-b+a\,\mathrm{sign}(u))^2 - ab\,\mathrm{sign}(u)\}\,\mathrm{d}u}$$
$$= \sqrt{\tau}\,\frac{(a+b)e^{ab}\varphi(a+b) - (a-b)e^{-ab}\varphi(a-b)}{e^{ab}\varphi(a+b) + e^{-ab}\varphi(a-b)}$$
$$= \sqrt{\tau}\,\mathrm{sign}(b)\frac{(a+|b|)e^{a|b|}\varphi(a+|b|) - (a-|b|)e^{-a|b|}\varphi(a-|b|)}{e^{a|b|}\varphi(a+|b|) + e^{-a|b|}\varphi(a-|b|)}$$
$$= \widehat{\beta}_j^0 + \lambda\,\mathrm{sign}(\widehat{\beta}_j^0)\frac{e^{a|b|}\varphi(a+|b|) - e^{-a|b|}\varphi(a-|b|)}{e^{a|b|}\varphi(a+|b|) + e^{-a|b|}\varphi(a-|b|)}$$
$$= \widehat{\beta}_j^0 + \lambda\,\mathrm{sign}(\widehat{\beta}_j^0)\frac{\Psi(a+|b|) - \Psi(a-|b|)}{\Psi(a+|b|) + \Psi(a-|b|)},$$

where $\Psi(t) = e^{t^2/2}\varphi(t)$. In other terms, noticing that $\Psi_\tau(t) = \Psi(t/\sqrt{\tau})$, we have obtained

$$\widehat{\beta}_j = \mathrm{sign}(\widehat{\beta}_j^0)\left(|\widehat{\beta}_j^0| - \lambda w(\tau,\lambda,|\widehat{\beta}_j^0|)\right), \tag{37}$$

where we have denoted $w(\tau,\lambda,t) = (\Psi_\tau(\lambda-t) - \Psi_\tau(\lambda+t))/(\Psi_\tau(\lambda-t) + \Psi_\tau(\lambda+t))$. Injecting (37) in (36) we arrive at the desired expression for H.

Acknowledgments. The work of Q. Paris was supported by the Russian Academic Excellence Project 5-100. The work of A. D. was partially supported by the grant Investissements d'Avenir (ANR-11-IDEX-0003/Labex Ecodec/ANR-11-LABX-0047) and the chair "LCL/GENES/Fondation du risque, Nouveaux enjeux pour nouvelles données".

8. Supplementary Material. The proofs of Equation (10), as well as the proofs of results of Section 5, have been gathered in the Supplementary Material.

References.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automatic Control*, AC-19:716–723. System identification and time-series analysis.

Alquier, P. (2013). Bayesian methods for low-rank matrix estimation: Short survey and theoretical study. In *24th International Conference, ALT 2013, Singapore. Proceedings*, pages 309–323, Berlin, Heidelberg.

Alquier, P. and Biau, G. (2013). Sparse single-index model. *J. Mach. Learn. Res.*, 14:243–280.

Alquier, P. and Lounici, K. (2011). PAC-Bayesian bounds for sparse regression estimation with exponential weights. *Electron. J. Stat.*, 5:127–145.

Arias-Castro, E. and Lounici, K. (2014). Estimation and variable selection with exponential weights. *Electron. J. Statist.*, 8(1):328–354.

Audibert, J.-Y. (2009). Fast learning rates in statistical inference through aggregation. *Ann. Statist.*, 37(4):1591–1646.

Bellec, P. C., Dalalyan, A. S., Grappin, E., and Paris, Q. (2016a). On the prediction loss of the lasso in the partially labeled setting. Technical report, arXiv:1606.06179.

Bellec, P. C., Lecué, G., and Tsybakov, A. B. (2016b). Slope meets lasso: improved oracle bounds and optimality. Technical report, arXiv:1605.08651.

Belloni, A., Chernozhukov, V., and Wang, L. (2014). Pivotal estimation via square-root lasso in nonparametric regression. *Ann. Statist.*, 42(2):757–788.

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.*, 37(4):1705–1732.

Bobkov, S. and Madiman, M. (2011). Concentration of the information in data with log-concave distributions. *Ann. Probab.*, 39(4):1528–1543.

Bogdan, M., van den Berg, E., Sabatti, C., Su, W., and Candès, E. J. (2015). SLOPE—adaptive variable selection via convex optimization. *Ann. Appl. Stat.*, 9(3):1103–1140.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for high-dimensional data*. Springer Series in Statistics. Springer, Heidelberg. Methods, theory and applications.

Bunea, F., She, Y., and Wegkamp, M. H. (2011). Optimal selection of reduced rank estimators of high-dimensional matrices. *Ann. Statist.*, 39(2):1282–1309.

Bunea, F., Tsybakov, A., and Wegkamp, M. (2007). Sparsity oracle inequalities for the lasso. *Electron. J. Statist.*, 1:169–194.

Candes, E. and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.*, 35(6):2313–2351.

Candès, E. J. and Plan, Y. (2011). Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Trans. Inform. Theory*, 57(4):2342–2359.

Candès, E. J. and Tao, T. (2010). The power of convex relaxation: near-optimal matrix completion. *IEEE Trans. Inform. Theory*, 56(5):2053–2080.

Castillo, I., Schmidt-Hieber, J., and van der Vaart, A. (2015). Bayesian linear regression with sparse priors. *Ann. Statist.*, 43(5):1986–2018.

Castillo, I. and van der Vaart, A. (2012). Needles and straw in a haystack: posterior concentration for possibly sparse sequences. *Ann. Statist.*, 40(4):2069–2101.

Catoni, O. (2004). *Statistical learning theory and stochastic optimization*, volume 1851 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin.

Catoni, O. (2007). *Pac-Bayesian supervised classification: the thermodynamics of statistical learning*. Lecture Notes–Monograph Series, 56. Institute of Mathematical Statistics, Beachwood, OH.

Chernousova, E., Golubev, Y., and Krymova, E. (2013). Ordered smoothers with exponential weighting. *Electron. J. Stat.*, 7:2395–2419.

Chesneau, C. and Lecué, G. (2009). Adapting to unknown smoothness by aggregation of thresholded wavelet estimators. *Statist. Sinica*, 19(4):1407–1417.

Cottet, V. and Alquier, P. (2016). 1-bit Matrix Completion: PAC-Bayesian Analysis of a Variational Approximation. Technical report, arXiv:1604.04191.

Dai, D., Rigollet, P., Xia, L., and Zhang, T. (2014). Aggregation of affine estimators. *Electron. J. Stat.*, 8(1):302–327.

Dalalyan, A., Ingster, Y., and Tsybakov, A. B. (2014). Statistical inference in compound functional models. *Probab. Theory Related Fields*, 158(3-4):513–532.

Dalalyan, A. S. (2014). Theoretical guarantees for approximate sampling from a smooth and log-concave density. to appear in JRSS B , arXiv:1412.7392.

Dalalyan, A. S., Hebiri, M., and Lederer, J. (2017). On the prediction performance of the lasso. *Bernoulli*, 23(1):552–581.

Dalalyan, A. S. and Salmon, J. (2012). Sharp oracle inequalities for aggregation of affine estimators. *Ann. Statist.*, 40(4):2327–2355.

Dalalyan, A. S. and Tsybakov, A. B. (2007). Aggregation by exponential weighting and sharp oracle inequalities. In *Learning theory*, volume 4539 of *Lecture Notes in Comput. Sci.*, pages 97–111. Springer, Berlin.

Dalalyan, A. S. and Tsybakov, A. B. (2008). Aggregation by exponential weighting, sharp pac-bayesian bounds and sparsity. *Machine Learning*, 72(1-2):39–61.

Dalalyan, A. S. and Tsybakov, A. B. (2012a). Mirror averaging with sparsity priors. *Bernoulli*, 18(3):914–944.

Dalalyan, A. S. and Tsybakov, A. B. (2012b). Sparse regression learning by aggregation and Langevin Monte-Carlo. *J. Comput. System Sci.*, 78(5):1423–1443.

Donoho, D. L. and Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, 90(432):1200–1224.

Durmus, A. and Moulines, E. (2016). Sampling from strongly log-concave distributions with the Unadjusted Langevin Algorithm. Technical Report , arXiv:1605.01559.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its

oracle properties. *J. Amer. Statist. Assoc.*, 96(456):1348–1360.

Frank, I. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135.

Fu, W. J. (1998). Penalized regressions: the bridge versus the lasso. *J. Comput. Graph. Statist.*, 7(3):397–416.

Gaïffas, S. and Lecué, G. (2007). Optimal rates and adaptation in the single-index model using aggregation. *Electron. J. Stat.*, 1:538–573.

Gaïffas, S. and Lecué, G. (2011). Sharp oracle inequalities for high-dimensional matrix prediction. *IEEE Trans. Inform. Theory*, 57(10):6942–6957.

Gao, C., van der Vaart, A. W., and Zhou, H. H. (2015). A general framework for bayes structured linear models. Technical report, arXiv:1506.02174.

Giraud, C. (2015). *Introduction to High-Dimensional Statistics*. CRC Press.

Golubev, Y. and Ostrovski, D. (2014). Concentration inequalities for the exponential weighting method. *Math. Methods Statist.*, 23(1):20–37.

Guedj, B. and Alquier, P. (2013). PAC-Bayesian estimation and prediction in sparse additive models. *Electron. J. Stat.*, 7:264–291.

Hans, C. (2009). Bayesian lasso regression. *Biometrika*, 96(4):835–845.

Harchaoui, Z., Douze, M., Paulin, M., Dudik, M., and Malick, J. (2012). Large-scale image classification with trace-norm regularization. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3386–3393.

Hoffmann, M., Rousseau, J., and Schmidt-Hieber, J. (2015). On adaptive posterior concentration rates. *The Annals of Statistics*, 43(5):2259–2295.

Juditsky, A., Rigollet, P., and Tsybakov, A. B. (2008). Learning by mirror averaging. *Ann. Statist.*, 36(5):2183–2206.

Klopp, O. (2014). Noisy low-rank matrix completion with general sampling distribution. *Bernoulli*, 20(1):282–303.

Koltchinskii, V. (2009). Sparse recovery in convex hulls via entropy penalization. *Ann. Statist.*, 37(3):1332–1359.

Koltchinskii, V. (2011). *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: Ecole d'Eté de Probabilités de Saint-Flour XXXVIII-2008*, volume 38. Springer.

Koltchinskii, V., Lounici, K., and Tsybakov, A. B. (2011). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, 39(5):2302–2329.

Lecué, G. and Mendelson, S. (2013). On the optimality of the aggregate with exponential weights for low temperatures. *Bernoulli*, 19(2):646–675.

Leung, G. and Barron, A. R. (2006). Information theory and mixing least-squares regressions. *IEEE Trans. Inform. Theory*, 52(8):3396–3410.

Lim, Y. J. and Teh, Y. W. (2007). Variational Bayesian Approach to Movie Rating Prediction. In *KDD-cup-2007, proceedings*.

Mai, T. T. and Alquier, P. (2015). A Bayesian approach for noisy matrix completion: optimal rate under general sampling distribution. *Electron. J. Stat.*, 9(1):823–841.

Mallows, C. L. (1973). Some comments on c¡sub¿p¡/sub¿. *Technometrics*, 15(4):661–675.

McAllester, D. A. (1998). Some PAC-Bayesian theorems. In *Proceedings of the Eleventh*

Annual Conference on Computational Learning Theory (Madison, WI, 1998), pages 230–234 (electronic). ACM, New York.

Negahban, S. and Wainwright, M. J. (2011). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *Ann. Statist.*, 39(2):1069–1097.

Negahban, S. and Wainwright, M. J. (2012). Restricted strong convexity and weighted matrix completion: optimal bounds with noise. *J. Mach. Learn. Res.*, 13:1665–1697.

Park, T. and Casella, G. (2008). The Bayesian lasso. *J. Amer. Statist. Assoc.*, 103(482):681–686.

Rigollet, P. and Tsybakov, A. (2011). Exponential screening and optimal rates of sparse estimation. *Ann. Statist.*, 39(2):731–771.

Rohde, A. and Tsybakov, A. B. (2011). Estimation of high-dimensional low-rank matrices. *Ann. Statist.*, 39(2):887–930.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.*, 6(2):461–464.

Shen, X. and Wu, Y. (2012). A unified approach to salient object detection via low rank matrix recovery. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 853–860.

Srebro, N. and Shraibman, A. (2005). Rank, trace-norm and max-norm. In Auer, P. and Meir, R., editors, *18th Annual Conference on Learning Theory, COLT 2005. Proceedings*, pages 545–560.

Su, W. and Candès, E. (2016). SLOPE is adaptive to unknown sparsity and asymptotically minimax. *Ann. Statist.*, 44(3):1038–1068.

Sun, T. and Zhang, C.-H. (2012). Scaled sparse linear regression. *Biometrika*, 99(4):879–898.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288.

Tibshirani, R. J. and Taylor, J. (2012). Degrees of freedom in lasso problems. *Ann. Statist.*, 40(2):1198–1232.

Tsybakov, A. B. (2014). Aggregation and minimax optimality in high-dimensional estimation. In *Proceedings of the International Congress of Mathematicians (Seoul, August 2014)*, volume 3, pages 225–246.

van de Geer, S. (2016). *Estimation and Testing Under Sparsity: École d'Été de Probabilités de Saint-Flour XLV – 2015*. Springer International Publishing.

van de Geer, S. and Bühlmann, P. (2009). On the conditions used to prove oracle results for the Lasso. *Electron. J. Stat.*, 3:1360–1392.

van der Pas, S. L., Salomond, J.-B., and Schmidt-Hieber, J. (2016). Conditions for posterior contraction in the sparse normal means problem. *Electron. J. Stat.*, 10(1):976–1000.

Vovk, V. G. (1990). Aggregating strategies. In *Proceedings of the Third Annual Workshop on Computational Learning Theory, COLT 1990*, pages 371–386.

Wipf, D. P., Palmer, J. A., and Rao, B. D. (2003). Perspectives on sparse bayesian learning. In *Advances in Neural Information Processing Systems 16*, pages 249–256.

Yang, Y. (2000a). Adaptive estimation in pattern recognition by combining different procedures. *Statist. Sinica*, 10(4):1069–1089.

Yang, Y. (2000b). Combining different procedures for adaptive regression. *J. Multivariate Anal.*, 74(1):135–161.

Yang, Y. (2000c). Mixing strategies for density estimation. *Ann. Statist.*, 28(1):75–87.

Yang, Y. (2001). Adaptive regression by mixing. *J. Amer. Statist. Assoc.*, 96(454):574–588.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 68(1):49–67.

Yuditskiĭ, A. B., Nazin, A. V., Tsybakov, A. B., and Vayatis, N. (2005). Recursive aggregation of estimators by the mirror descent method with averaging. *Problemy Peredachi Informatsii*, 41(4):78–96.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.*, 38(2):894–942.

Zhou, Y., Wilkinson, D., Schreiber, R., and Pan, R. (2008). Large-scale parallel collaborative filtering for the netflix prize. In *AAIM 2008. Proceedings*, pages 337–348, Berlin, Heidelberg.

Zou, H., Hastie, T., and Tibshirani, R. (2007). On the "degrees of freedom" of the lasso. *Ann. Statist.*, 35(5):2173–2192.

3 avenue Pierre Larousse, 92245 Malakoff, France.
26 Shabolovka street, 119049 Moscow, Russia.