Annales de l'Institut Henri Poincaré - Probabilités et Statistiques
2020, Vol. 56, No. 2, 1465–1483
https://doi.org/10.1214/19-AIHP1010

ANNALES DE L'INSTITUT HENRI POINCARÉ
PROBABILITÉS ET STATISTIQUES
www.imstat.org/aihp

Exponential weights in multivariate regression and a low-rankness favoring prior

Arnak S. Dalalyan

ENSAE/CREST, 5 Av. Le Chatelier, 91120 Palaiseau, France. E-mail: arnak.dalalyan@ensae.fr

Received 12 July 2018; revised 4 March 2019; accepted 6 June 2019

Abstract. We establish theoretical guarantees for the expected prediction error of the exponentially weighted aggregate in the case of multivariate regression that is when the label vector is multidimensional. We consider the regression model with fixed design and bounded noise. The first new feature uncovered by our guarantees is that it is not necessary to require independence of the observations: a symmetry condition on the noise distribution alone suffices to get a sharp risk bound. This result needs the regression vectors to be bounded. A second curious finding concerns the case of unbounded regression vectors but independent noise. It turns out that applying exponential weights to the label vectors perturbed by a uniform noise leads to an estimator satisfying a sharp oracle inequality. The last contribution is the instantiation of the proposed oracle inequalities to problems in which the unknown parameter is a matrix. We propose a low-rankness favoring prior and show that it leads to an estimator that is optimal under weak assumptions.

Résumé. Nous établissons des garanties théoriques pour l'erreur de prédiction de l'agrégat par poids exponentiels dans le cadre de la régression multivariée, c'est-à-dire lorsqu'on souhaite prédire une étiquette multidimensionnelle. Nous considérons le modèle de régression à design fixe et à bruit borné. La première conséquence de nos garanties est qu'il n'est pas nécessaire d'exiger l'indépendance des observations : une condition de symétrie sur la seule distribution du bruit suffit pour obtenir des bornes précises. Ce résultat nécessite que les vecteurs de régression soient bornées. Une seconde conséquence curieuse concerne le cas des vecteurs de régression non bornés mais indépendant de bruit. Il s'avère que l'application des poids exponentiels aux vecteurs d'étiquette perturbés par un bruit uniforme conduit à un estimateur satisfaisant une inégalité d'oracle exacte. La dernière contribution est l'instanciation des inégalités d'oracle proposées à des problèmes dans lesquels le paramètre inconnu est une matrice. Nous proposons une loi a priori favorisant les matrices à faible rang et montrons qu'elle conduit à un estimateur optimal sous des hypothèses faibles.

MSC: Primary 62J05; secondary 62H12

Keywords: Oracle inequality; Exponential weights; Trace regression; Low-rank matrices

1. Introduction

The goal of this paper is to extend the scope of the applications of the exponentially weighted aggregate (EWA) to regression problems with multidimensional labels. Such an extension is important since it makes it possible to cover such problems as the multitask learning, the multiclass classification and the matrix factorization. We consider the regression model with fixed design and additive noise. Our main contributions are mathematical: we establish risk bounds taking the form of PAC-Bayesian type oracle inequalities under various types of assumptions on the noise distribution.

Sharp risk bounds for the exponentially weighted aggregate in the regression with univariate labels have been established in [13,14,16,20,22,24]. These bounds hold under various assumptions on the noise distribution and cover popular examples such as Gaussian, Laplace, uniform and Rademacher noise. One of the important specificities of the setting with multivariate labels is that noise is multivariate as well, and one has to cope with possible correlations within its components. We provide results that not only allow for dependence between noise components corresponding to different labels, but also for dependence between different samples. The corresponding result, stated in Theorem 1, requires, however, some symmetry of the noise distribution. To our knowledge, this is the first oracle inequality that is sharp (*i.e.*, the leading constant is equal to one) and valid under such a general condition on the noise distribution. The remainder term in that inequality is of the order K/n, where K is the number of labels and n is the sample size. This order of magnitude of the remainder term is optimal, in the sense that when all the labels are equal we get the best possible rate.

Nevertheless, one can expect that for weakly correlated labels the remainder term might be of significantly smaller order. This is indeed the case, as shown in Theorem 2, under the additional hypothesis that the n samples are independent. In the obtained sharp oracle inequality, the remainder term is now proportional to $\|\boldsymbol{\Sigma}\|/n$, where $\|\boldsymbol{\Sigma}\|$ is the spectral norm of the noise covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{K\times K}$. Of course, when all the components of the noise vector are highly correlated, the spectral norm $\|\boldsymbol{\Sigma}\|$ is proportional to K and, therefore, the conclusions of Theorem 1 and Theorem 2 are consistent.

The two aforementioned theorems are established under the condition that the aggregated matrices belong to a set having a bounded diameter. The resulting risk bounds scale linearly in that diameter and eventually blow up when the diameter is equal to infinity. However, it has been noticed in that for some distributions this condition can be dropped without deteriorating the remainder term. In particular, this is the case of the Gaussian [24] and the uniform distributions [14]. Furthermore, using an extended version of Stein's lemma, [14] show that the same holds true for any distribution having bounded support and a density bounded away from zero. Corollary 1 in [14] even claims that the same type of bound holds for any symmetric distribution with bounded support. Unfortunately, the proof of this claim is flawed since it relies on Lemma 3 (page 58) that is wrong. In the present work, we have managed to repair this shortcoming and to establish sharp PAC-Bayesian risk bounds for any symmetric distribution with bounded support. This is achieved using a key modification of the aggregation procedure, which consists in adding a suitable defined uniform noise to data vectors before applying the exponential weights. We call the resulting procedure noisy exponentially weighted aggregate. Its statistical properties are presented in Theorem 4.

Finally, we show an application of the obtained PAC-Bayes inequalities to the case of low-rank matrix estimation. We exhibit a well suited prior distribution, termed spectral scaled Student prior, for which the PAC-Bayes inequality leads to optimal remainder term. This prior is the matrix analogue of the scaled Student prior studied in [16,20]. We also provide some hints how this estimator can be implemented using the Langevin Monte Carlo algorithm and present some initial experimental results on the problem of digital image denoising.

Notation

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 $0 < q < \infty$, $\|\boldsymbol{u}\|_0 = \text{Card}(\{j : u_j \neq 0\})$ and $\|\boldsymbol{u}\|_\infty = \max_{j\in[k]} |u_j|$.

For all integers $p \geq 1$, $\mathbf{I}_p$ refers to the identity matrix in $\mathbb{R}^{p\times p}$. Finally the transpose and the Moore–Penrose pseudoinverse of a matrix $\mathbf{A}$ are denoted by $\mathbf{A}^\top$ and $\mathbf{A}^\dagger$, respectively. The spectral norm, the Fobenius norm and the nuclear norm of $\mathbf{A}$ will be respectively denoted by $\|\mathbf{A}\|$, $\|\mathbf{A}\|_F$ and $\|\mathbf{A}\|_1$. For every integer k, t_k and χ^2_k are the Student and the chi-squared distributions with k degrees of freedom.

2. Exponential weights for multivariate regression

In this section we describe the setting of multivariate regression and the main principles of the aggregation by exponential weighting.

2.1. *Multivariate regression model*

We consider the model of multivariate regression with fixed design, in which we observe n feature-label pairs $(\boldsymbol{x}_i, \boldsymbol{Y}_i)$, for $i \in [n]$. The labels $\boldsymbol{Y}_i \in \mathbb{R}^K$ are random vectors with real entries, the features are assumed to be deterministic elements of an arbitrary space $\mathcal{X}$. Note that, unless specified otherwise, we do not assume that the observations are independent.

We introduce the regression function $f^* : \mathcal{X} \to \mathbb{R}^K$ and noise vectors $\boldsymbol{\xi}_i$:

$$\mathbf{F}^*_i = \mathbf{E}[\boldsymbol{Y}_i] = f^*(\boldsymbol{x}_i), \qquad \boldsymbol{\xi}_i = \boldsymbol{Y}_i - \mathbf{F}^*_i, \quad i \in [n].$$

We are interested in estimating the values of f^* at the points $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ only, which amounts to denoising the observed labels $\boldsymbol{Y}_i$. In such a setting, of course, one can forget about the features $\boldsymbol{x}_i$ and the function f^*, since the goal is merely to estimate the $K \times n$ matrix $\mathbf{F}^* = [\mathbf{F}^*_1, \dots, \mathbf{F}^*_n]$. The quality of an estimator $\widehat{\mathbf{F}}$ will be measured using the empirical loss

$$\ell_n(\widehat{\mathbf{F}}, \mathbf{F}^*) = \frac{1}{n}\|\widehat{\mathbf{F}} - \mathbf{F}^*\|_F^2 = \frac{1}{n}\sum_{i\in[n]}\|\widehat{\mathbf{F}}_i - \mathbf{F}^*_i\|_2^2.$$

This quantity is also referred to as in-sample prediction error. Let $\mathcal{F}$ be the set of all $K \times n$ matrices with real entries. The following assumption will be repeatedly used.

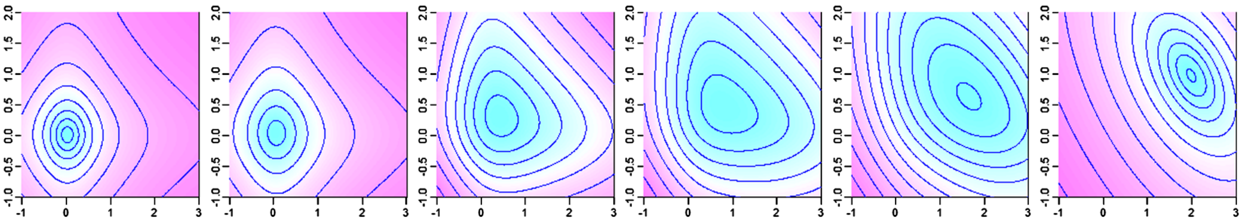

Fig. 1. The contour plots of the log pseudo-posterior for different values of the temperature parameter. The prior is a product of two Student $t(3)$ distributions. For a very large temperature, $\tau = 20$, the first plot from the left, the posterior is very close to the prior. On the other extreme, for $\tau = 0.008$, the utmost right plot, the posterior gets close to a Dirac mass at the observed data $\mathbf{Y}$ (here $\mathbf{Y} = [2, 1]$).

Assumption C(B_ξ, L). For some positive numbers B_ξ and L that, unless otherwise specified, may be equal to $+\infty$, it holds that

$$\max_{i\in[n]} \mathbf{P}\big(\|\boldsymbol{\xi}_i\|_2^2 > KB_\xi^2\big) = 0, \qquad \sup_{\mathbf{F},\mathbf{F}'\in\mathcal{F}} \max_{i\in[n]} \big\|\mathbf{F}_i - \mathbf{F}'_i\big\|_2^2 \le KL^2. \tag{1}$$

Note in (1) the presence of the normalizing factor K in the upper bounds on the Euclidean norms of K-dimensional vectors $\boldsymbol{\xi}_i$ and $(\mathbf{F} - \mathbf{F}')_i$. This allows us to think of the constants B_ξ and L as dimension independent quantities.

2.2. *Exponentially weighted aggregate*

The exponentially weighted aggregate (EWA) is defined as the average with respect to a tempered posterior distribution π_n on $\mathcal{F}$, the set of all $K \times n$ matrices with real entries. To define the tempered posterior π_n, we choose a prior distribution π_0 on $\mathcal{F}$ and a temperature parameter $\tau > 0$, and set

$$\pi_n(d\mathbf{F}) \propto \exp\left\{-\left(\frac{1}{2\tau}\right)\ell_n(\mathbf{F}, \mathbf{Y})\right\}\pi_0(d\mathbf{F}).$$

Figure 1 depicts the form of this posterior density for different values of τ. The EWA is then

$$\widehat{\mathbf{F}}^{\mathrm{EWA}} = \int_{\mathcal{F}} \mathbf{F}\pi_n(d\mathbf{F}). \tag{2}$$

According to the Varadhan–Donsker variational formula, the posterior distribution π_n is the solution of the following optimisation problem:

$$\pi_n \in \arg\min_p \left\{\int_{\mathcal{F}} \frac{1}{2}\ell_n(\mathbf{F}, \mathbf{Y})p(d\mathbf{F}) + \tau D_{\mathrm{KL}}(p \,\|\, \pi_0)\right\},$$

where the inf is taken over all probability measures p on $\mathcal{F}$. We see that the posterior distribution minimises a cost function which contains a term accounting for the fidelity to the observations and a regularisation term proportional to the divergence from the prior distribution. The larger the temperature τ, the closer the posterior π_n is to the prior π_0.

In most situations the integral in (2) cannot be computed in closed form. Even its approximate evaluation using a numerical scheme is often difficult. An appealing alternative is then to use Monte Carlo integration. This corresponds to drawing N samples $\mathbf{F}_1, \dots, \mathbf{F}_N$ from the posterior distribution π_n and to define the Monte Carlo version of the EWA by

$$\widehat{\mathbf{F}}^{\mathrm{MC\text{-}EWA}} = \frac{1}{N}\sum_{\ell=1}^{N} \mathbf{F}_\ell.$$

Of course, the applicability of this method is limited to distributions π_n for which the problem of sampling can be solved at low computational cost. We will see below that this approximation satisfies the same kind of oracle inequality as the original EWA.

3. PAC-Bayes type risk bounds

In this section, we state and discuss several risk bounds for the EWA and related estimators under various conditions. We start with the case of the bounded regression vectors, *i.e.*, the case where the constant L in Assumption C(B_ξ, L) is finite.

Main results of this section, Theorems 1–4, contain two versions of PAC-Bayesian bounds on the expected loss. The second version requires the temperature parameter to be twice larger than the minimal temperature for which the first version holds. However, it has the advantage of containing an additional negative term. The presence of this negative term, proportional to the variance of the posterior π_n, allows us to obtain PAC-Bayesian risk bounds for the Monte-Carlo version of the EWA, as shown in Proposition 2 below. Let us also stress that PAC-Bayes type risk bounds are a powerful theoretical tool, since they may be used to obtain sharp oracle inequalities under various structural assumptions.

3.1. *Bounds without independence assumption but finite L*

We first state the results that hold even when the columns and rows of the noise matrix $\boldsymbol{\xi}$ are dependent. These results, however, require the boundedness of the set of aggregated elements $\mathbf{F}$.

Theorem 1. *Suppose that Assumption $C(B_\xi, L)$ is satisfied and the distribution of $\boldsymbol{\xi}$ is symmetric in the sense that for any sign vector $\boldsymbol{s} \in \{\pm 1\}^n$, the equality in distribution $[s_1\boldsymbol{\xi}_1, \dots, s_n\boldsymbol{\xi}_n] \stackrel{\mathscr{D}}{=} \boldsymbol{\xi}$ holds. Let $\widehat{\mathbf{F}} = \widehat{\mathbf{F}}^{\mathrm{EWA}}$. Then, for every $\tau \geq (\frac{1}{n})(KB_\xi)(2L \vee 3B_\xi)$, we have*

$$\mathbf{E}\big[\ell_n\big(\widehat{\mathbf{F}}, \mathbf{F}^*\big)\big] \leq \inf_p \biggl\{ \int_{\mathcal{F}} \ell_n\big(\mathbf{F}, \mathbf{F}^*\big) p(d\mathbf{F}) + 2\tau D_{\mathrm{KL}}(p \,\|\, \pi_0) \biggr\}, \tag{3}$$

where the inf is taken over all probability measures on $\mathcal{F}$. Furthermore, for larger values of the temperature, $\tau \geq (\frac{1}{n})(KB_\xi)(2L \vee 6B_\xi)$, the following upper bound holds

$$\mathbf{E}\big[\ell_n\big(\widehat{\mathbf{F}}, \mathbf{F}^*\big)\big] \leq \inf_p \biggl\{ \int_{\mathcal{F}} \ell_n\big(\mathbf{F}, \mathbf{F}^*\big) p(d\mathbf{F}) + 2\tau D_{\mathrm{KL}}(p \,\|\, \pi_0) \biggr\} - \frac{1}{2}\int_{\mathcal{F}} \mathbf{E}\big[\ell_n(\widehat{\mathbf{F}}, \mathbf{F})\pi_n(d\mathbf{F})\big].$$

One can remark that the risk bound provided by (3) is an increasing function of the temperature. Therefore, the best risk bound is obtained for the smallest allowed value of temperature, that is

$$\tau = \frac{K}{n} B_\xi (2L \vee 3B_\xi).$$

Assuming B_ξ and L as constants, while $K = K_n$ can grow with n, we see that the remainder term in (3) is of the order K/n. We will see below that using other proof techniques, under somewhat different assumptions on the noise distribution, we can replace K by the spectral norm of the noise covariance matrix $\mathbf{E}[\boldsymbol{\xi}_i\boldsymbol{\xi}_i^\top]$. In the "worst case" when all the entries of $\boldsymbol{\xi}_i$ are equal, these two bounds are of the same order since $\|\mathbf{E}[\boldsymbol{\xi}_i\boldsymbol{\xi}_i^\top]\| = \mathbf{E}[\xi_{i1}^2]\|\mathbf{1}_K\mathbf{1}_K^\top\| = K\mathbf{E}[\xi_{i1}^2]$. Note, however, that the result above does not assume any independence condition on the noise vectors $\boldsymbol{\xi}_i$.

Theorem 2. *We assume that for some $K \times K$ matrix $\boldsymbol{\Sigma} \succeq 0$, we have $\boldsymbol{\xi} = \boldsymbol{\Sigma}^{1/2}\bar{\boldsymbol{\xi}}$ where $\bar{\boldsymbol{\xi}}$ has independent rows $\bar{\boldsymbol{\xi}}_{j\bullet}$ satisfying the following boundedness and symmetry conditions*:

- *for any $(i, j) \in [n] \times [K]$, we have $\mathbf{P}(|\bar{\xi}_{ji}| \leq \bar{B}_\xi) = 1$,*
- *for any sign vector $\boldsymbol{s} \in \{\pm 1\}^n$, the equality in distribution $[s_1\bar{\boldsymbol{\xi}}_{j,1}, \dots, s_n\bar{\boldsymbol{\xi}}_{j,n}] \stackrel{\mathscr{D}}{=} \bar{\boldsymbol{\xi}}_{j\bullet}$ holds.*

In addition, the set $\mathcal{F}$ is such that for some $\bar{L} > 0$, we have $\max_{i\in[n]} \|\boldsymbol{\Sigma}^{1/2}(\mathbf{F}_i - \mathbf{F}_i')\|_\infty \leq \bar{L}$ for every $\mathbf{F}, \mathbf{F}' \in \mathcal{F}$. Let $\widehat{\mathbf{F}} = \widehat{\mathbf{F}}^{\mathrm{EWA}}$. Then, for every $\tau \geq (\frac{1}{n})(\bar{B}_\xi)(2\bar{L} \vee 3\|\boldsymbol{\Sigma}\|\bar{B}_\xi)$, we have

$$\mathbf{E}\big[\ell_n\big(\widehat{\mathbf{F}}, \mathbf{F}^*\big)\big] \leq \inf_p \biggl\{ \int_{\mathcal{F}} \ell_n\big(\mathbf{F}, \mathbf{F}^*\big) p(d\mathbf{F}) + 2\tau D_{\mathrm{KL}}(p \,\|\, \pi_0) \biggr\},$$

where the inf is taken over all probability measures on $\mathcal{F}$. Furthermore, for larger values of the temperature, $\tau \geq (\frac{1}{n})(\bar{B}_\xi)(2\bar{L} \vee 6\|\boldsymbol{\Sigma}\|\bar{B}_\xi)$, the following upper bound holds

$$\mathbf{E}\big[\ell_n\big(\widehat{\mathbf{F}}, \mathbf{F}^*\big)\big] \leq \inf_p \biggl\{ \int_{\mathcal{F}} \ell_n\big(\mathbf{F}, \mathbf{F}^*\big) p(d\mathbf{F}) + 2\tau D_{\mathrm{KL}}(p \,\|\, \pi_0) \biggr\} - \frac{1}{2}\int_{\mathcal{F}} \mathbf{E}\big[\ell_n(\widehat{\mathbf{F}}, \mathbf{F})\pi_n(d\mathbf{F})\big].$$

The strength of this theorem is that it does not require the independence of the observations $\boldsymbol{Y}_i$ corresponding to different values of $i \in [n]$. Only a symmetry condition is required. Furthermore, the resulting risk bound is valid for a temperature parameter which is of order $O(1/n)$ and, hence, is independent of the dimension K of label vectors $\boldsymbol{Y}_i$.

The proofs of Theorem 1 and Theorem 2, postponed to Section 7, rely on the following interesting construction related to the Skorokhod embedding. If $\gamma > 0$ is a fixed number and ξ is a random variable having a symmetric distribution, then one can devise a new random variable ζ such that $\xi + \gamma\zeta$ has the same distribution as $(1+\gamma)\xi$ and $\mathbf{E}[\zeta|\xi] = 0$. The construction of the pair (ξ, ζ) is as follows. We first draw a random variable R of the same distribution as $|\xi|$ and a Brownian motion $(B_t : t \geq 0)$ independent of R. We then define the two stopping times T and T_γ by

$$T = \inf\{t \geq 0 : |B_t| = R\}, \qquad T_\gamma = \inf\{t \geq 0 : |B_t| = (1+\gamma)R\}.$$

One can easily check that the random variable B_T has the same distribution as ξ whereas B_{T_γ} has the same distribution as $(1+\gamma)\xi$. Furthermore, since conditionally to $B_T = x$, the process $(B_{T+t} - x : t \geq 0)$ is a Brownian motion, we have $\mathbf{E}[B_{T_\gamma} - B_T | B_T] = 0$. Therefore, the pair $\xi := B_T$ and $\zeta := (B_{T_\gamma} - B_T)/\gamma$ satisfies the aforementioned conditions. If we set $\eta = \zeta/\xi$, we can check that

$$\eta = \begin{cases} 1, & \text{with probability } 1 - \frac{\gamma}{1+2\gamma}, \\ -1 - \frac{1}{\gamma}, & \text{with probability } \frac{\gamma}{1+2\gamma}. \end{cases}$$

This is exactly the formula used in Lemma 3 below. This particular example of the Skorokhod embedding relies heavily on the symmetry of the distribution of ξ. There are other constructions that do not need this condition. We believe that some of them can be used to further relax the assumptions of Theorem 1 and Theorem 2. This is, however, out of scope of the present work.

3.2. *Bounds under independence with infinite L*

The previous two theorems require from the set $\mathcal{F}$ of aggregated elements to have a finite diameter L (or $\bar{L}$) and this diameter enters (linearly) in the risk bound through the temperature. The presence of this condition is dictated by the techniques of the proofs; we see no reason for the established oracle inequalities to fail in the case of infinite L. In the present section, we state some results that are proved using another technique, building on the celebrated Stein lemma, which do not need L to be finite.

Theorem 3. *Assume that for some $K \times K$ positive semidefinite matrix $\mathbf{\Sigma}$, the noise matrix $\boldsymbol{\xi} = \mathbf{\Sigma}^{1/2}\bar{\boldsymbol{\xi}}$ with $\bar{\boldsymbol{\xi}}$ satisfying the following conditions*:

C1. *all the random variables $\bar{\xi}_{j,i}$ are iid with zero mean and unit variance,*
C2. *the measure $m_{\bar{\xi}}(x)\,dx$, where $m_{\bar{\xi}}(x) = -\mathbf{E}[\bar{\xi}_{j,i}\mathbb{1}(\bar{\xi}_{j,i} \leq x)]$, is absolutely continuous with respect to the distribution of $\bar{\xi}_{j,i}$ with a Radon–Nikodym derivative*[1] $g_{\bar{\xi}}$,
C3. *$g_{\bar{\xi}}$ is bounded by some constant $G_{\bar{\xi}} < \infty$.*

Let $\widehat{\mathbf{F}} = \widehat{\mathbf{F}}^{\mathrm{EWA}}$. For any $\tau \geq (\|\mathbf{\Sigma}\| G_{\bar{\xi}})/n$, we have

$$\mathbf{E}\big[\ell_n\big(\widehat{\mathbf{F}}, \mathbf{F}^*\big)\big] \leq \inf_p \left\{ \int_{\mathcal{F}} \ell_n\big(\mathbf{F}, \mathbf{F}^*\big) p(d\mathbf{F}) + 2\tau D_{\mathrm{KL}}(p\|\pi_0) \right\}. \tag{4}$$

Furthermore, if $\tau \geq 2(\|\mathbf{\Sigma}\| G_{\bar{\xi}})/n$, then

$$\mathbf{E}\big[\ell_n\big(\widehat{\mathbf{F}}, \mathbf{F}^*\big)\big] = \inf_p \left\{ \int_{\mathcal{F}} \ell_n\big(\mathbf{F}, \mathbf{F}^*\big) p(d\mathbf{F}) + 2\tau D_{\mathrm{KL}}(p\|\pi_0) \right\} - \frac{1}{2}\int_{\mathcal{F}} \mathbf{E}\big[\ell_n(\mathbf{F}, \widehat{\mathbf{F}})\pi_n(d\mathbf{F})\big].$$

As mentioned in [14, pp. 43–44], many distributions satisfy assumptions C2 and C3. For instance, for the Gaussian distribution $\mathcal{N}(0, \sigma^2)$ and for the uniform in $[-b, b]$ distributions these assumptions are fulfilled with $G_{\bar{\xi}} = \sigma^2$ and $G_{\bar{\xi}} = b^2/2$, respectively. More generally, if $\bar{\xi}_{j,i}$ has a density $p_{\bar{\xi}}$ with bounded support $[-b, b]$, then the assumptions are satisfied with $G_{\bar{\xi}} = \mathbf{E}[|\bar{\xi}_{j,i}|]/\min_{|x|\leq b} p_{\bar{\xi}}(x)$. Furthermore, if a random variable $\bar{\xi}$ satisfies these assumptions, then for every $\gamma \in \mathbb{R}$, the scaled random variable $\gamma\bar{\xi}$ satisfies these assumptions with $G_{\gamma\bar{\xi}} = \gamma^2 G_{\bar{\xi}}$. Here, we add another class to the family of distributions satisfying C2 and C3: unimodal distributions with compact support.

---

[1] This means that for any bounded and measurable function h, we have $\int_{\mathbb{R}} h(x) m_{\bar{\xi}}(x)\,dx = \mathbf{E}[h(\bar{\xi}_{j,i}) g_{\bar{\xi}}(\bar{\xi}_{j,i})]$.

Proposition 1. *Assume that $\bar\xi_{j,i}$ has a density $p_{\bar\xi}$ with respect to the Lebesgue measure such that $p_{\bar\xi}(x) = 0$ for every $x \notin [-b, b]$ and, for some $a \in [-b, b]$, $p_{\bar\xi}$ is increasing on $[-b, a]$ and decreasing on $[a, b]$. Then, $\bar\xi_{j,i}$ satisfies C2 and C3 with $G_{\bar\xi} = (\frac{b^2}{2})$.*

Perhaps the most important shortcoming of the last theorem is that it cannot be applied to the discrete distributions of noise. In fact, if the distribution of $\bar\xi_{j,i}$ is discrete, then there is no chance condition C2 to be satisfied. This is due to the fact that the measure $m_{\bar\xi}\,dx$, being absolutely continuous with respect to the Lebesgue measure, cannot be absolutely continuous with respect to a counting measure. On the other hand, Theorem 1 and Theorem 2 can be applied to discrete noise distributions, but they require boundedness of the family $\mathcal{F}$. At this stage, we do not know whether it is possible to extend PAC-Bayesian type risk bound (4) to discrete distributions and unbounded sets $\mathcal{F}$. However, in the case of a bounded discrete noise, we propose a simple modification of the EWA for which (4) is valid.

The modification mentioned in the previous paragraph consists in adding a uniform noise to the entries of the observed labels $\mathbf{Y}_i$. Thus, we define the noisy exponential weighting aggregate, nEWA, by

$$\widehat{\mathbf{F}}^{\mathrm{nEWA}} = \int_{\mathcal{F}} \mathbf{F}\bar\pi_n(d\mathbf{F}), \qquad \bar\pi_n(d\mathbf{F}) \propto \exp\left\{-\left(\frac{1}{2\tau}\right)\ell_n(\mathbf{F}, \bar{\mathbf{Y}})\right\}\pi_0(d\mathbf{F}), \tag{5}$$

where $\bar\pi_n$ is defined in the same way as π_n but for the perturbed matrix $\bar{\mathbf{Y}} = \mathbf{Y} + \boldsymbol{\zeta}$, with $\boldsymbol{\zeta}$ a $K \times n$ random perturbation matrix.

Theorem 4. *Let $\widehat{\mathbf{F}}^{\mathrm{nEWA}}$ be the noisy EWA defined by (5). Assume that*

C4. *entries $\xi_{j,i}$ of the noise matrix $\boldsymbol{\xi}$ are iid with zero mean and bounded by some constant $B_\xi > 0$,*
C5. *entries $\zeta_{j,i}$ of the perturbation matrix are iid uniformly distributed in $[-B_\xi, B_\xi]$.*

Then, for any $\tau \ge 2B_\xi^2/n$, we have

$$\mathbf{E}\big[\ell_n\big(\widehat{\mathbf{F}}^{\mathrm{nEWA}}, \mathbf{F}^*\big)\big] \le \inf_p\left\{\int_{\mathcal{F}} \ell_n\big(\mathbf{F}, \mathbf{F}^*\big)p(d\mathbf{F}) + 2\tau D_{\mathrm{KL}}(p\|\pi_0)\right\}. \tag{6}$$

Furthermore, if $\tau \ge 4B_\xi^2/n$, then for $\widehat{\mathbf{F}} = \widehat{\mathbf{F}}^{\mathrm{nEWA}}$

$$\mathbf{E}\big[\ell_n\big(\widehat{\mathbf{F}}, \mathbf{F}^*\big)\big] = \inf_p\left\{\int_{\mathcal{F}} \ell_n\big(\mathbf{F}, \mathbf{F}^*\big)p(d\mathbf{F}) + 2\tau D_{\mathrm{KL}}(p\|\pi_0)\right\} - \frac{1}{2}\int_{\mathcal{F}} \mathbf{E}\big[\ell_n(\mathbf{F}, \widehat{\mathbf{F}})\bar\pi_n(d\mathbf{F})\big].$$

Proof of Theorem 4. Let us check that the matrix of perturbed labels $\bar{\mathbf{Y}}$ satisfies the conditions of Theorem 3 with $\boldsymbol{\Sigma} = \mathbf{I}_K$. To this end, we set $\widetilde\xi_{j,i} = \xi_{j,i} + \zeta_{j,i}$. We will check that the distribution of $\widetilde\xi_{j,i}$ satisfies conditions C2 and C3 (condition C1 is straightforward). Since the distribution of $\widetilde\xi_{j,i}$ is the convolution of that of $\xi_{j,i}$ and a uniform distribution, it admits a density with respect to the Lebesgue measure which is given by

$$\widetilde p(x) = \frac{1}{2B_\xi}\mathbf{P}\big(|\xi_{j,i} - x| \le B_\xi\big) = \frac{1}{2B_\xi}\mathbf{P}\big(\xi_{j,i} \in [x - B_\xi, x + B_\xi] \cap [-B_\xi, B_\xi]\big).$$

The set $A_x := [x - B_\xi, x + B_\xi] \cap [-B_\xi, B_\xi]$ is empty if $|x| > 2B_\xi$. When x runs from $-2B_\xi$, the set A_x increases, while it decreases when x runs from 0 to $2B_\xi$. This implies that the density $\widetilde p$ is zero outside the interval $[-2B_\xi, 2B_\xi]$ and unimodal in this interval. Therefore, it satisfies Proposition 1 with $b = 2B_\xi$ and $a = 0$. This implies that conditions C2 and C3 are fulfilled with $G_{\bar\xi} = 2B_\xi^2$ and $\|\boldsymbol{\Sigma}\| = 1$. Thus, the conclusion of Theorem 3 applies and yields the claims of Theorem 4. □

We can replace in Theorem 4 the condition C4 by $\boldsymbol{\xi} = \boldsymbol{\Sigma}^{1/2}\bar{\boldsymbol{\xi}}$, where $\bar\xi_{j,i}$ are iid and bounded. In this case, the contamination added to the labels is of the form $\boldsymbol{\Sigma}^{1/2}\bar{\boldsymbol{\zeta}}$, where $\bar\zeta_{j,i}$'s are iid uniform. The claims of Theorem 4 remain valid, but they are of limited interest, since it is not likely to find a situation in which the matrix $\boldsymbol{\Sigma}$ is known.

3.3. *Risk bounds for the Monte Carlo EWA*

All the four theorems of the previous sections contain two risk bounds. The first bound is, in each case, more elegant and valid for a smaller value of the temperature than the second bound. However, the latter appears to be more useful for getting guarantees for the Monte Carlo version of the EWA. This is due to the fact that the additional term in the second risk bounds is proportional to the difference of the risks between the MC-EWA and the EWA.

Proposition 2. *If $\widehat{\mathbf{F}}^{\text{MC-EWA}}$ is the MC-EWA with N Monte Carlo samples, then*

$$\mathbf{E}\big[\ell_n\big(\widehat{\mathbf{F}}^{\text{MC-EWA}}, \mathbf{F}^*\big)\big] = \mathbf{E}\big[\ell_n\big(\widehat{\mathbf{F}}^{\text{EWA}}, \mathbf{F}^*\big)\big] + \frac{1}{N}\int_{\mathcal{F}} \mathbf{E}\big[\ell_n\big(\mathbf{F}, \widehat{\mathbf{F}}^{\text{EWA}}\big)\pi_n(d\mathbf{F})\big].$$

Therefore, if the conditions of one of the four foregoing theorems are satisfied and τ is chosen accordingly then, as soon as $N \geq 2$,

$$\mathbf{E}\big[\ell_n\big(\widehat{\mathbf{F}}^{\text{MC-EWA}}, \mathbf{F}^*\big)\big] \leq \inf_p \bigg\{\int_{\mathcal{F}} \ell_n\big(\mathbf{F}, \mathbf{F}^*\big) p(d\mathbf{F}) + 2\tau D_{\text{KL}}(p\|\pi_0)\bigg\}.$$

The proof of this result is straightforward. For the sake of completeness, it is provided in Section 7. Note that this result bounds only the expected error, where the expectation is taken with respect to both the noise matrix $\boldsymbol{\xi}$ and the Monte Carlo sample. Using standard concentration inequalities, this bound can be complemented by an evaluation of the deviation between the Monte Carlo average $\widehat{\mathbf{F}}^{\text{MC-EWA}}$ and its "expectation" $\widehat{\mathbf{F}}^{\text{EWA}}$.

3.4. *Relation to previous work*

To the best of our knowledge, the first result in the spirit of the oracle inequalities presented in foregoing sections was established by [24], using a technique based on Stein's unbiased risk estimate for regression with Gaussian noise developed in [18,19]. Extensions to more general noise distributions were presented in [13,14] and later on refined in [16]. These papers studied the problem of aggregating "frozen" (that is independent of the data used for the aggregation) estimators, except [24] who considered the case of projection estimators as well. In his PhD thesis, [23] proved that analogous oracle bounds hold for the problem of aggregation of shrinkage estimators. The case of more general linear estimators was explored by [5,10,12]. In the context of sparsity, statistical properties of exponential weights were studied in [3,26,27]. [20] proved that analogous results hold true in the sequential setting, see also [29].

There is also extensive literature on the exponential weights for problems with iid obsrvations, such as the density model, the regression with random design, etc. We refer the interested reader to [4,9,15,22,31,32] and the references therein. It is useful to note here that the proof techniques used in the iid setting and in the setting with deterministic design considered in the present work are very different. Furthermore, the version exponential of the exponential weights used in the iid setting involves an additional averaging step and is therefore referred to as progressive mixture or mirror averaging.

4. EWA with low-rank favoring priors

To give a concrete example of application of the results established in previous section, let us consider the so called reduced rank regression model. An asymptotic analysis of this model goes back to [21], whereas more recent results can be found in [6,7] and the references therein. It corresponds to assuming that the matrix $\mathbf{F}^* = \mathbf{E}[\mathbf{Y}]$ has a small rank, as compared to its maximal possible value $K \wedge n$. Equivalently, this means that the observed K dimensional vectors $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ belong, up to a noise term, to a low dimensional subspace. Such problems arise, for instance, in subspace clustering or in multi-index problems. Of course, one can estimate the matrix $\mathbf{F}^*$ by the PCA, but it requires rather precise knowledge of the rank.

In order to get an estimator that takes advantage of the (nearly) low-rank property of the matrix $\mathbf{F}^*$, we suggest to use the following prior

$$\pi_0(d\mathbf{F}) \propto \det\big(\lambda^2 \mathbf{I}_K + \mathbf{F}\mathbf{F}^\top\big)^{-(n+K+2)/2} d\mathbf{F}, \tag{7}$$

where $\lambda > 0$ is a tuning parameter. From now on, with a slight abuse of notation, we will denote by $\pi_0(\mathbf{F})$ the probability density function of the measure $\pi_0(d\mathbf{F})$. The same will be done for $p(d\mathbf{F})$ and $\pi_n(d\mathbf{F})$. We will refer to π_0 as the spectral scaled Student prior, since one easily checks that

$$\pi_0(\mathbf{F}) \propto \prod_{j=1}^{K} \big(\lambda^2 + s_j(\mathbf{F})^2\big)^{-(n+K+2)/2},$$

where $s_j(\mathbf{F})$ denotes the jth largest singular value of $\mathbf{F}$. We can recognize in the last display the density function of the scaled Student t evaluated at $s_j(\mathbf{F})$. Thus, the scaled spectral Student prior operates on the spectrum of $\mathbf{F}$ as the sparsity favoring prior introduced in [16] on the vectors. Another interesting property of this prior, is that if $\mathbf{F} \sim \pi_0$, then the marginal distributions of the columns of $\mathbf{F}$ are scaled multivariate Student t_3.

Lemma 1. *If* $\mathbf{F}$ *is a random matrix having as density the function* π_0, *then the random vectors* $\mathbf{F}_i$ *are all drawn from the* K*-variate scaled Student distribution* $(\lambda/\sqrt{3})t_{3,K}$. *As a consequence, we have* $\int_{\mathcal{F}} \|\mathbf{F}_i\|_2^2 \pi_0(\mathbf{F})\,d\mathbf{F} = \lambda^2 K$.

From a mathematical point of view, the nice feature of the aforementioned prior is that the Kullback–Leibler divergence between π_0 and its shifted by a matrix $\bar{\mathbf{F}}$ version grows proportionally to the rank of $\bar{\mathbf{F}}$, when all the other parameters remain fixed. This is formalized in the next result.

Lemma 2. *Let* $\bar{p}$ *be the probability density function obtained from the prior* π_0 *by a translation,* $\bar{p}(\mathbf{F}) = \pi_0(\mathbf{F} - \bar{\mathbf{F}})$. *Then, for any matrix* $\bar{\mathbf{F}}$ *of at most rank* r, *we have*

$$D_{\mathrm{KL}}(\bar{p}\|\pi_0) \le 2r(n+K+2)\log\left(1 + \frac{\|\bar{\mathbf{F}}\|_F}{\sqrt{2r}\lambda}\right) \le 2r(n+K+2)\log\bigl(1 + \|\bar{\mathbf{F}}\|/\lambda\bigr).$$

The proof of this result is deferred to the Appendix. Applying this lemma, in conjunction with Theorem 3, we get a risk bound in the reduced rank regression problem which illustrates the power of the exponential weights.

Theorem 5. *Let the noise matrix* $\boldsymbol{\xi}$ *and the artificial perturbation matrix* $\boldsymbol{\zeta}$ *satisfy the assumptions of Theorem 4. Let* π_0 *be the scaled spectral Student prior (7) with some tuning parameter* $\lambda > 0$. *Then, for every* $\tau \ge 2B_\xi^2/n$, *we have*

$$\mathbf{E}\bigl[\ell_n\bigl(\widehat{\mathbf{F}}^{\mathrm{nEWA}}, \mathbf{F}^*\bigr)\bigr] \le \inf_{\bar{\mathbf{F}}}\left\{\ell_n\bigl(\bar{\mathbf{F}}, \mathbf{F}^*\bigr) + 4r(\bar{\mathbf{F}})(n+K+2)\tau\log\left(1 + \frac{\|\bar{\mathbf{F}}\|_F}{\sqrt{2r(\bar{\mathbf{F}})}\lambda}\right)\right\} + \lambda^2 K,$$

where $r(\bar{\mathbf{F}}) = \operatorname{rank}(\bar{\mathbf{F}})$ *and the inf is taken over all* $K \times n$ *matrices* $\bar{\mathbf{F}}$.

Proof of Theorem 5. Let us fix an arbitrary matrix $\bar{\mathbf{F}}$ and denote its rank by r. We apply Theorem 4 and upper bound the inf with respect to all probability distributions p by the right hand side of (6) evaluated at the distribution $\bar{p}$ defined in Lemma 2. This yields

$$\mathbf{E}\bigl[\ell_n\bigl(\widehat{\mathbf{F}}^{\mathrm{nEWA}}, \mathbf{F}^*\bigr)\bigr] \le \int_{\mathcal{F}} \ell_n\bigl(\mathbf{F}, \mathbf{F}^*\bigr)\pi_0(\mathbf{F} - \bar{\mathbf{F}})\,d\mathbf{F} + 4r(\bar{\mathbf{F}})(n+K+2)\tau\log\left(1 + \frac{\|\bar{\mathbf{F}}\|_F}{\sqrt{2r}\lambda}\right). \tag{8}$$

Using the translation invariance of the Lebesgue measure and the fact that $\int \mathbf{F}\pi_0(\mathbf{F})\,d\mathbf{F} = \mathbf{0}$, we get

$$\int_{\mathcal{F}} \ell_n\bigl(\mathbf{F}, \mathbf{F}^*\bigr)\pi_0(\mathbf{F} - \bar{\mathbf{F}})\,d\mathbf{F} = \ell_n\bigl(\bar{\mathbf{F}}, \mathbf{F}^*\bigr) + \frac{1}{n}\int_{\mathcal{F}} \|\mathbf{F}\|_F^2 \pi_0(\mathbf{F})\,d\mathbf{F}. \tag{9}$$

Using Lemma 1 and the fact that $\|\mathbf{F}\|_F^2 = \sum_{i\in[n]} \|\mathbf{F}_i\|_2^2$, one checks that $\int_{\mathcal{F}} \|\mathbf{F}\|_F^2 \pi_0(\mathbf{F})\,d\mathbf{F} \le \lambda^2 K n$. Combining this inequality with (9) and (8), we get the claim of the theorem. □

There are many papers using a Bayesian approach to the problem of prediction with low rank matrices, see [1] and the references therein. All the methods we are aware of put a prior on $\mathbf{F}$ which is defined as follows: first choose one prior π_0^U on the space of $K \times (K \wedge n)$ matrices, a prior π_0^V on the space of $n \times (K \wedge n)$ matrices and a third prior π^γ on $\mathbb{R}^{K\wedge n}$, then define π_0 as the distribution of $\mathbf{U}\mathbf{D}_{\boldsymbol{\gamma}}^2\mathbf{V}^\top$ when the triplet $(\mathbf{U}, \mathbf{V}, \boldsymbol{\gamma})$ is drawn from the product distribution $\pi_0^U \times \pi_0^V \times \pi_0^\gamma$ (see, for instance, [2]).

To our knowledge, [30] is the only work dealing with prior (7) in a context related to low rank matrix estimation and completion. It proposes variational approximations to the Bayes estimator and demonstrates their good performance on various data sets. In a sense, Theorem 5 provides theoretical justification for the empirically observed good statistical properties of the prior defined in (7).

Let us briefly discuss the inequality of Theorem 5. If we choose $\tau = 2B_\xi^2/n$ and $\lambda^2 = B_\xi^2(n+K)/(nK)$, we get the following upper bound on the risk $\mathbf{E}[\ell_n(\widehat{\mathbf{F}}^{\mathrm{nEWA}}, \mathbf{F}^*)]$:

$$\inf_{\bar{\mathbf{F}}}\left\{\ell_n\bigl(\bar{\mathbf{F}}, \mathbf{F}^*\bigr) + \frac{8B_\xi^2 r(\bar{\mathbf{F}})(n+K+2)}{n}\log\left(1 + \frac{\|\bar{\mathbf{F}}\|_F\sqrt{K}}{\sqrt{2r(\bar{\mathbf{F}})}B_\xi}\right)\right\} + \frac{B_\xi^2(n+K)}{n}. \tag{10}$$

We see that (10) handles optimally mis-specification, since it is an oracle inequality with a leading constant 1, and the remainder term is of optimal order $r(n+K)/n$, up to a logarithmic factor. The extra logarithmic factor in the last display can be further simplified using the inequality $\|\bar{\mathbf{F}}\|_F \leq \|\bar{\mathbf{F}}\|\sqrt{r(\bar{\mathbf{F}})}$. This implies that the term inside the logarithm is the signal-to-noise ratio $\sqrt{K}\|\bar{\mathbf{F}}\|/B_\xi$. Other oracle inequalities with nearly optimal remainder terms in the context of low-rank matrix estimation and completion are exposed in [1,2]. However, those results are not sharp oracle inequalities since the factor in front of the leading term in the upper bound is larger than 1. All the oracle inequalities we are aware of, for the model under consideration, contain analogous (and often much more complicated) logarithmic factors.

Using the properties of the scaled Student prior exposed in Lemma 1 and Lemma 2, one can establish oracle inequalities in other statistical problems in which the unknown parameter is a matrix, such as matrix completion, trace regression or multiclass classification, see [6,8,25,28]. This is left to future work.

5. Implementation and a few numerical experiments

In this section, we report the results of some proof of concept numerical experiments. We propose to compute an approximation of the EWA with the scaled multivariate Student prior by a suitable version of the Langevin Monte Carlo algorithm. To describe the letter, let us first remark that

$$\log \pi_n(\mathbf{F}) = -\frac{1}{2\tau}\ell_n(\mathbf{F},\mathbf{Y}) - \frac{(n+K+2)}{2}\log\det\bigl(\lambda^2\mathbf{I}_K + \mathbf{F}\mathbf{F}^\top\bigr). \tag{11}$$

From this relation, we can infer that

$$-\nabla \log \pi_n(\mathbf{F}) = \frac{1}{n\tau}(\mathbf{F}-\mathbf{Y}) + (n+K+2)\bigl(\lambda^2\mathbf{I}_K + \mathbf{F}\mathbf{F}^\top\bigr)^{-1}\mathbf{F}.$$

The (constant-step) Langevin MC is defined by choosing an initial matrix $\mathbf{F}_0$ and then by using the recursion

$$\mathbf{F}_{k+1} = \mathbf{F}_k + h\nabla \log \pi_n(\mathbf{F}) + \sqrt{2h}\mathbf{W}_k, \quad k = 0, 1, \dots,$$

where $h > 0$ is the step-size and $\mathbf{W}_0, \mathbf{W}_1, \dots$ are independent Gaussian random matrices with iid standard Gaussian entries. For (strongly) log-concave densities π, nonasymptotic guarantees for the LMC have been recently established in [11,17], but they do not carry over the present case since the right hand side of (11) is not concave. Our numerical experiments show that (despite the absence of theoretical guarantees) the LMC converges and leads to relevant results.

Note that a direct application of the Langevin MC algorithm involves a $K \times K$ matrix inversion at each iteration. This might be costly and can slow down significantly the algorithm. We suggest to replace this matrix inversion by a few steps of gradient descent for a suitably chosen optimization problem. Indeed, one can easily check that the matrix $\mathbf{M} = (\lambda^2\mathbf{I}_K + \mathbf{F}\mathbf{F}^\top)^{-1}\mathbf{F}$ is nothing else but the solution to the convex optimization problem

$$\min\bigl\{\bigl\|\mathbf{I}_n - \mathbf{F}^\top\mathbf{M}\bigr\|_F^2 + 2\lambda^2\|\mathbf{M}\|_F^2\bigr\}.$$

We use ten gradient descent steps for approximating the solution of this optimization problem. This does not require neither matrix inversion nor svd or other costly operation.

We applied this algorithm to the problem of image denoising (see Figure 2). We chose an RGB image of resolution 120×160 and applied to it an additive Gaussian white noise of standard deviation $\sigma \in \{10, 30, 50\}$. In order to make use of the denoising algorithm based on the aforementioned Langevin MC, we transformed the noisy image into a matrix of size 192×300. Each row of this transformed matrix corresponds to a patch of size $10 \times 10 \times 3$ of the noisy image. The patches are chosen to be non-overlapping in order to get a reasonable dimensionality. We expect the result to be better for overlapping patches, but the computational cost will also be high. The parameters were chosen as follows:

$$\tau = 2\sigma^2/n; \qquad \lambda = 10\sigma\sqrt{(n+K)/(nK)}; \qquad h = 10; \qquad k_{\max} = 4000.$$

Note that the values of τ and λ are suggested by our theoretical results, while the step-size h and the number of iterations of the LMC, $k_{\max}$, were chosen experimentally. The LMC after $k_{\max}$ iterations provides one sample that is approximately drawn from the pseudo-posterior π_n. We did $N = 400$ repetitions and averaged the obtained results for approximating the posterior mean.

6. Conclusion

We have studied the expected in-sample prediction error of the Exponentially Weighted Algorithm (EWA) in the context of multivariate regression with possible dependent noise. We have shown that under boundedness assumptions on the noise

Fig. 2. The result of the experiment on image densoising. Left: the original $120 \times 160 \times 3$ image. Middle: the noisy image for different values of σ. Right: the denoised image.

and the aggregated elements, the EWA satisfies a PAC-Bayes type sharp oracle inequality, provided that the temperature parameter is sufficiently large. The remainder term in these oracle inequalities is of arguably optimal order of magnitude and is consistent with the corresponding results obtained in the model of univariate regression. An interesting observation is that if we apply the EWA to the data matrix artificially contaminated by a uniform noise, the resulting procedure satisfies a sharp oracle inequality under a much weaker assumption on the noise distribution. In particular, this allows to cover any distribution with bounded support. We have also included the results of a small numerical experiment on image denoising, that shows the applicability of the EWA.

7. Proofs of the main results

The proofs of all the main theorems stated in the previous sections are gathered in this section. The proofs of some technical lemmas are deferred to Appendix A.

Proof of Theorem 1. We wish to upper bound $\ell_n(\widehat{\mathbf{F}}^{\mathrm{EWA}}, \mathbf{F}^*)$. Let $\boldsymbol{\zeta}$ be a random matrix such that $\mathbf{E}[\boldsymbol{\zeta}|\mathbf{Y}] = \mathbf{0}$ and define

$$\ell_n\big(\mathbf{F}, \mathbf{F}^*, \boldsymbol{\zeta}\big) = \ell_n\big(\mathbf{F}, \mathbf{F}^*\big) + \frac{2}{n}\big\langle \boldsymbol{\zeta}, \mathbf{F} - \mathbf{F}^*\big\rangle. \tag{12}$$

In what follows, we use the short notation $\widehat{\mathbf{F}}$ instead of $\widehat{\mathbf{F}}^{\mathrm{EWA}}$. We have, for every $\alpha > 0$,

$$\begin{aligned}
\ell_n\big(\widehat{\mathbf{F}}, \mathbf{F}^*, \boldsymbol{\zeta}\big) &= \frac{1}{\alpha}\log\exp\big\{\alpha\ell_n\big(\widehat{\mathbf{F}}, \mathbf{F}^*, \boldsymbol{\zeta}\big)\big\} \\
&= \underbrace{\frac{1}{\alpha}\log\int_{\mathcal{F}} e^{\alpha(\ell_n(\widehat{\mathbf{F}}, \mathbf{F}^*, \boldsymbol{\zeta}) - \ell_n(\mathbf{F}, \mathbf{F}^*, \boldsymbol{\zeta}))}\pi_n(d\mathbf{F})}_{:=S_1(\alpha)} - \underbrace{\frac{1}{\alpha}\log\int_{\mathcal{F}} e^{-\alpha\ell_n(\mathbf{F}, \mathbf{F}^*, \boldsymbol{\zeta})}\pi_n(d\mathbf{F})}_{:=S(\alpha)}.
\end{aligned}$$

The next two lemmas provide suitable upper bounds on the magnitude of the terms $S(\alpha)$ and $S_1(\alpha)$.

Lemma 3. *Let $\boldsymbol{\xi} = [\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n]$ be a $K \times n$ random matrix with real entries having a symmetric distribution (see the statement of Theorem 1). Let $\boldsymbol{\zeta}_i$ be defined as $\boldsymbol{\zeta}_i = -\boldsymbol{\xi}_i\eta_i$, where η_i are iid random variables independent of $\boldsymbol{\xi}$ and satisfying*

$$\eta_i = \begin{cases} 1, & \text{with probability } 1 - \frac{\alpha\tau}{1+2\alpha\tau}, \\ -1 - \frac{1}{\alpha\tau}, & \text{with probability } \frac{\alpha\tau}{1+2\alpha\tau}. \end{cases}$$

Then, the expectation of the random variable S can be bounded as follows:

$$-\left(\frac{1}{\alpha}\right)\mathbf{E}\big[S(\alpha)\big] \le \inf_p\left\{\int_{\mathcal{F}} \ell_n\big(\mathbf{F}, \mathbf{F}^*\big)p(d\mathbf{F}) + 2\tau D_{\mathrm{KL}}(p \,\|\, \pi_0)\right\},$$

where the inf is taken over all probability measures on $\mathcal{F}$.

Lemma 4. *Let the random vectors $\boldsymbol{\zeta}_i$, $i \in [n]$ be as defined in Lemma 3. Then, we have*

$$\lim_{\alpha\to 0}\frac{1}{\alpha}\mathbf{E}\big[S_1(\alpha)|\boldsymbol{\xi}\big] \le \sum_{i\in[n]} \tau \log\int_{\mathcal{F}} e^{(2/n\tau)\boldsymbol{\xi}_i^\top(\widehat{\mathbf{F}}_i - \mathbf{F}_i)}\pi_n(d\mathbf{F}) - \int_{\mathcal{F}} \ell_n(\widehat{\mathbf{F}}, \mathbf{F})\pi_n(d\mathbf{F}).$$

Applying these two lemmas, we get

$$\begin{aligned}
\mathbf{E}\big[\ell_n\big(\widehat{\mathbf{F}}, \mathbf{F}^*\big)\big] &= \mathbf{E}\big[\ell_n\big(\widehat{\mathbf{F}}, \mathbf{F}^*, \boldsymbol{\zeta}\big)\big] \\
&= \lim_{\alpha\to 0}\frac{\mathbf{E}(\mathbf{E}[S_1(\alpha)|\boldsymbol{\xi}]) - \mathbf{E}[S(\alpha)]}{\alpha} \\
&\le \inf_p\left\{\int_{\mathcal{F}} \ell_n\big(\mathbf{F}, \mathbf{F}^*\big)p(d\mathbf{F}) + 2\tau D_{\mathrm{KL}}(p \,\|\, \pi_0)\right\} - \int_{\mathcal{F}} \ell_n(\widehat{\mathbf{F}}, \mathbf{F})\pi_n(d\mathbf{F}) \\
&\quad + \sum_{i\in[n]} \tau\mathbf{E}\left\{\log\int_{\mathcal{F}} e^{(2/n\tau)\boldsymbol{\xi}_i^\top(\widehat{\mathbf{F}}_i - \mathbf{F}_i)}\pi_n(d\mathbf{F})\right\}.
\end{aligned} \tag{13}$$

Then, for every $\tau \ge (2K/n)(B_\xi L)$, we have

$$\begin{aligned}
e^{(2/n\tau)\boldsymbol{\xi}_i^\top(\widehat{\mathbf{F}}_i - \mathbf{F}_i)} &\le 1 + \frac{2\boldsymbol{\xi}_i^\top(\widehat{\mathbf{F}}_i - \mathbf{F}_i)}{n\tau} + \frac{3(\boldsymbol{\xi}_i^\top(\widehat{\mathbf{F}}_i - \mathbf{F}_i))^2}{(n\tau)^2} \\
&\le 1 + \frac{2\boldsymbol{\xi}_i^\top(\widehat{\mathbf{F}}_i - \mathbf{F}_i)}{n\tau} + \frac{3KB_\xi^2\|\widehat{\mathbf{F}}_i - \mathbf{F}_i\|_2^2}{(n\tau)^2}.
\end{aligned}$$

This implies that

$$\int_{\mathcal{F}} e^{(2/n\tau)\boldsymbol{\xi}_i^\top(\widehat{\mathbf{F}}_i-\mathbf{F}_i)}\pi_n(d\mathbf{F}) \le 1+\frac{3KB_\xi^2}{(n\tau)^2}\int_{\mathcal{F}}\|\widehat{\mathbf{F}}_i-\mathbf{F}_i\|_2^2\pi_n(d\mathbf{F}).$$

Combining the last display with (13) and using the inequality $\log(1+x)\le x$, we arrive at

$$\begin{aligned}\mathbf{E}\big[\ell_n\big(\widehat{\mathbf{F}},\mathbf{F}^*\big)\big] &\le \inf_p\bigg\{\int_{\mathcal{F}}\ell_n\big(\mathbf{F},\mathbf{F}^*\big)p(d\mathbf{F})+2\tau D_{\mathrm{KL}}(p\,\|\,\pi_0)\bigg\}\\ &\quad -\bigg(1-\frac{3KB_\xi^2}{n\tau}\bigg)\mathbf{E}\bigg\{\int_{\mathcal{F}}\ell_n(\widehat{\mathbf{F}},\mathbf{F})\pi_n(d\mathbf{F})\bigg\}.\end{aligned}$$

This completes the proof of the theorem. □

Proof of Theorem 2. The proof follows the same arguments as those used in the proof of Theorem 1. That is why we will skip some technical details. The main difference is in the definition of the matrix $\boldsymbol{\zeta}$ and the subsequent computations related to the evaluation of the term $S_2(\alpha)$. Thus, for any random matrix $\boldsymbol{\zeta}$ such that $\mathbf{E}[\boldsymbol{\zeta}|\mathbf{Y}]=\mathbf{0}$ and for $\ell_n(\mathbf{F},\mathbf{F}^*,\boldsymbol{\zeta})=\ell_n(\mathbf{F},\mathbf{F}^*)+\frac{2}{n}\langle\boldsymbol{\zeta},\mathbf{F}-\mathbf{F}^*\rangle$, we have

$$\mathbf{E}\big[\ell_n\big(\widehat{\mathbf{F}},\mathbf{F}^*\big)\big]=\mathbf{E}\big[\ell_n\big(\widehat{\mathbf{F}},\mathbf{F}^*,\boldsymbol{\zeta}\big)\big]=\lim_{\alpha\to 0}\frac{\mathbf{E}[S_1(\alpha)]-\mathbf{E}[S(\alpha)]}{\alpha},$$

where S and S_1 are the same as in the proof of Theorem 1. We instantiate the matrix $\boldsymbol{\zeta}$ as follows: $\boldsymbol{\zeta}=\boldsymbol{\Sigma}^{1/2}\bar{\boldsymbol{\zeta}}$ where the entries of $\bar{\boldsymbol{\zeta}}$ are given by $\bar\zeta_{j,i}=-\bar\xi_{j,i}\eta_{j,i}$, with $\eta_{j,i}$ being iid random variables independent of $\boldsymbol{\xi}$ and satisfying

$$\eta_{j,i}=\begin{cases}1, & \text{with probability } 1-\frac{\alpha\tau}{1+2\alpha\tau},\\ -1-\frac{1}{\alpha\tau}, & \text{with probability } \frac{\alpha\tau}{1+2\alpha\tau}.\end{cases}$$

One easily checks that the resulting vector $\bar{\boldsymbol{\xi}}_{j\bullet}-2\alpha\tau\bar{\boldsymbol{\zeta}}_{j,\bullet}$ has the same distribution as the vector $(1+2\alpha\tau)\bar{\boldsymbol{\xi}}_{j,\bullet}$, for every $j\in[K]$. Furthermore, for different values of j, these vectors are independent. This implies that the matrix $\bar{\boldsymbol{\xi}}-2\alpha\tau\bar{\boldsymbol{\zeta}}$ has the same distribution as the matrix $(1+2\alpha\tau)\bar{\boldsymbol{\xi}}$, which is sufficient for getting the conclusion of Lemma 3. That is

$$-\bigg(\frac{1}{\alpha}\bigg)\mathbf{E}\big[S(\alpha)\big]\le\inf_p\bigg\{\int_{\mathcal{F}}\ell_n\big(\mathbf{F},\mathbf{F}^*\big)p(d\mathbf{F})+2\tau D_{\mathrm{KL}}(p\,\|\,\pi_0)\bigg\},$$

where the inf is taken over all probability measures on $\mathcal{F}$. To bound the term S_1, we use a result similar to that of Lemma 4.

Lemma 5. *Let the random matrix $\boldsymbol{\zeta}$ be defined as above. Set $\mathbf{H}(\mathbf{F})=\boldsymbol{\Sigma}^{1/2}(\widehat{\mathbf{F}}-\mathbf{F})$. Then, we have*

$$\lim_{\alpha\to 0}\frac{1}{\alpha}\mathbf{E}\big[S_1(\alpha)|\boldsymbol{\xi}\big]\le\sum_{\substack{i\in[n]\\ j\in[K]}}\tau\log\int_{\mathcal{F}}e^{(2/n\tau)\bar\xi_{j,i}\mathbf{H}_{j,i}(\mathbf{F})}\pi_n(d\mathbf{F})-\int_{\mathcal{F}}\ell_n(\widehat{\mathbf{F}},\mathbf{F})\pi_n(d\mathbf{F}).$$

Then, for every $\tau\ge(2/n)(\bar B_\xi\bar L)$, we have

$$\begin{aligned}e^{(2/n\tau)\bar\xi_{j,i}\mathbf{H}_{j,i}(\mathbf{F})}&\le 1+\frac{2\bar\xi_{j,i}\mathbf{H}_{j,i}(\mathbf{F})}{n\tau}+\frac{3\bar\xi_{j,i}^2\mathbf{H}_{j,i}(\mathbf{F})^2}{(n\tau)^2}\\ &\le 1+\frac{2\bar\xi_{j,i}\mathbf{H}_{j,i}(\mathbf{F})}{n\tau}+\frac{3\bar B_\xi^2\mathbf{H}_{j,i}^2(\mathbf{F})}{(n\tau)^2}.\end{aligned}$$

Using the fact that $\int_{\mathcal{F}}\mathbf{H}(\mathbf{F})\pi_n(d\mathbf{F})=0$, the last display implies that

$$\int_{\mathcal{F}}e^{(2/n\tau)\bar\xi_{j,i}\mathbf{H}_{j,i}(\mathbf{F})}\pi_n(d\mathbf{F})\le 1+\frac{3\bar B_\xi^2}{(n\tau)^2}\int_{\mathcal{F}}\mathbf{H}_{j,i}^2(\mathbf{F})\pi_n(d\mathbf{F}).$$

Combining the last display with Lemma 5 and using the inequality $\log(1+x) \le x$, we arrive at

$$\begin{aligned}
\lim_{\alpha\to 0}\frac{1}{\alpha}\mathbf{E}\big[S_1(\alpha)|\boldsymbol{\xi}\big] &\le \sum_{i,j}\frac{3\bar{B}_\xi^2}{n^2\tau}\int_{\mathcal{F}}\mathbf{H}_{j,i}^2(\mathbf{F})\pi_n(d\mathbf{F}) - \int_{\mathcal{F}}\ell_n(\widehat{\mathbf{F}},\mathbf{F})\pi_n(d\mathbf{F}) \\
&= \frac{3\bar{B}_\xi^2}{n^2\tau}\int_{\mathcal{F}}\big\|\boldsymbol{\Sigma}^{1/2}(\widehat{\mathbf{F}}-\mathbf{F})\big\|_F^2\pi_n(d\mathbf{F}) - \int_{\mathcal{F}}\ell_n(\widehat{\mathbf{F}},\mathbf{F})\pi_n(d\mathbf{F}) \\
&\le \bigg(\frac{3\bar{B}_\xi^2\|\boldsymbol{\Sigma}\|}{n\tau}-1\bigg)\int_{\mathcal{F}}\ell_n(\widehat{\mathbf{F}},\mathbf{F})\pi_n(d\mathbf{F}).
\end{aligned}$$

This completes the proof of the theorem. □

Proof of Theorem 3. We outline here only the main steps of the proof, without going too much into the details. One can extend the Stein lemma from the Gaussian distribution to that of $\bar{\xi}_{j,i}$, provided the conditions of Theorem 3 are satisfied (see Lemma 1 in [14] for a similar result). The resulting claim is that the random variable

$$\widehat{r} := \ell_n(\widehat{\mathbf{F}},\mathbf{Y}) - \operatorname{tr}(\boldsymbol{\Sigma}) + \frac{2}{n}\sum_{i=1}^n\sum_{j=1}^K g_{\bar{\xi}}(\bar{\xi}_{j,i})\partial_{\bar{\xi}_{j,i}}\big(\boldsymbol{\Sigma}^{1/2}\widehat{\mathbf{F}}\big)_{j,i} \tag{14}$$

satisfies $\mathbf{E}[\widehat{r}] = \mathbf{E}[\ell_n(\widehat{\mathbf{F}},\mathbf{F}^*)]$. On the one hand, using Varadhan–Donsker's variational formula, we get

$$\begin{aligned}
\mathbf{E}\big[\ell_n(\widehat{\mathbf{F}},\mathbf{Y})\big] &\le \mathbf{E}\big[\ell_n(\widehat{\mathbf{F}},\mathbf{Y}) + 2\tau D_{\mathrm{KL}}(\pi_n\|\pi_0)\big] \\
&= \mathbf{E}\bigg[\int_{\mathcal{F}}\ell_n(\mathbf{F},\mathbf{Y})\pi_n(d\mathbf{F}) + 2\tau D_{\mathrm{KL}}(\pi_n\|\pi_0)\bigg] - \int_{\mathcal{F}}\mathbf{E}\big[\ell_n(\mathbf{F},\widehat{\mathbf{F}})\pi_n(d\mathbf{F})\big] \\
&\le \mathbf{E}\bigg[\inf_p\bigg(\int_{\mathcal{F}}\ell_n(\mathbf{F},\mathbf{Y})p(d\mathbf{F}) + 2\tau D_{\mathrm{KL}}(p\|\pi_0)\bigg)\bigg] - \int_{\mathcal{F}}\mathbf{E}\big[\ell_n(\mathbf{F},\widehat{\mathbf{F}})\pi_n(d\mathbf{F})\big] \\
&\le \inf_p\mathbf{E}\bigg[\bigg(\int_{\mathcal{F}}\ell_n(\mathbf{F},\mathbf{Y})p(d\mathbf{F}) + 2\tau D_{\mathrm{KL}}(p\|\pi_0)\bigg)\bigg] - \int_{\mathcal{F}}\mathbf{E}\big[\ell_n(\mathbf{F},\widehat{\mathbf{F}})\pi_n(d\mathbf{F})\big] \\
&= \inf_p\bigg(\int_{\mathcal{F}}\ell_n\big(\mathbf{F},\mathbf{F}^*\big)p(d\mathbf{F}) + \operatorname{tr}(\boldsymbol{\Sigma}) + 2\tau D_{\mathrm{KL}}(p\|\pi_0)\bigg) - \int_{\mathcal{F}}\mathbf{E}\big[\ell_n(\mathbf{F},\widehat{\mathbf{F}})\pi_n(d\mathbf{F})\big].
\end{aligned}$$

On the other hand, computing the partial derivative $\partial_{Y_{j,i}}(\boldsymbol{\Sigma}^{1/2}\widehat{\mathbf{F}})_{j,i}$, we get

$$\begin{aligned}
\partial_{\bar{\xi}_{j,i}}\big(\boldsymbol{\Sigma}^{1/2}\widehat{\mathbf{F}}\big)_{j,i} &= \boldsymbol{e}_j^\top\boldsymbol{\Sigma}^{1/2}(\partial_{\bar{\boldsymbol{\xi}}_i}\widehat{\mathbf{F}}_i)\boldsymbol{e}_j \\
&= \frac{1}{2n\tau}\boldsymbol{e}_j^\top\boldsymbol{\Sigma}^{1/2}\bigg(\int_{\mathcal{F}}\mathbf{F}_i(\mathbf{F}_i-\mathbf{Y}_i)^\top\pi_n(d\mathbf{F}) - \widehat{\mathbf{F}}_i(\widehat{\mathbf{F}}_i-\mathbf{Y}_i)^\top\bigg)\boldsymbol{\Sigma}^{1/2}\boldsymbol{e}_j \\
&= \frac{1}{2n\tau}\int_{\mathcal{F}}\big\{\boldsymbol{e}_j^\top\boldsymbol{\Sigma}^{1/2}(\mathbf{F}-\widehat{\mathbf{F}})_i\big\}^2\pi_n(d\mathbf{F}).
\end{aligned}$$

From this relation, we infer that

$$\begin{aligned}
\sum_{j=1}^K g_{\bar{\xi}}(\bar{\xi}_{j,i})\partial_{\bar{\xi}_{j,i}}\big(\boldsymbol{\Sigma}^{1/2}\widehat{\mathbf{F}}\big)_{j,i} &= \frac{1}{2n\tau}\sum_{j=1}^K g_{\bar{\xi}}(\bar{\xi}_{j,i})\int_{\mathcal{F}}\big\{\boldsymbol{e}_j^\top\boldsymbol{\Sigma}^{1/2}(\mathbf{F}-\widehat{\mathbf{F}})_i\big\}^2\pi_n(d\mathbf{F}) \\
&\le \frac{G_{\bar{\xi}}}{2n\tau}\int_{\mathcal{F}}\big\|\boldsymbol{\Sigma}^{1/2}(\mathbf{F}_i-\widehat{\mathbf{F}}_i)\big\|_2^2\pi_n(d\mathbf{F}) \\
&\le \frac{\|\boldsymbol{\Sigma}\|G_{\bar{\xi}}}{2n\tau}\int_{\mathcal{F}}\|\mathbf{F}_i-\widehat{\mathbf{F}}_i\|_2^2\pi_n(d\mathbf{F}).
\end{aligned}\tag{15}$$

Combining (14)–(15), we arrive at

$$\mathbf{E}\big[\ell_n\big(\widehat{\mathbf{F}},\mathbf{F}^*\big)\big] = \mathbf{E}[\widehat{r}]$$

$$
\begin{aligned}
&\le \inf_p\big(\ell_n\big(\mathbf{F},\mathbf{F}^*\big)+2\tau D_{\mathrm{KL}}(p\|\pi_0)\big)-\int_{\mathcal{F}}\mathbf{E}\big[\ell_n(\mathbf{F},\widehat{\mathbf{F}})\pi_n(d\mathbf{F})\big]\\
&\quad+\frac{\|\mathbf{\Sigma}\|G_{\bar\xi}}{n^2\tau}\sum_{i=1}^n\int_{\mathcal{F}}\mathbf{E}\big[\|\widehat{\mathbf{F}}_i-\mathbf{F}_i\|_2^2\pi_n(d\mathbf{F})\big]\\
&=\inf_p\big(\ell_n\big(\mathbf{F},\mathbf{F}^*\big)+2\tau D_{\mathrm{KL}}(p\|\pi_0)\big)-\bigg(1-\frac{\|\mathbf{\Sigma}\|G_{\bar\xi}}{n\tau}\bigg)\int_{\mathcal{F}}\mathbf{E}\big[\ell_n(\mathbf{F},\widehat{\mathbf{F}})\pi_n(d\mathbf{F})\big].
\end{aligned}
$$

This completes the proof. □

Proof of Proposition 1. Without loss of generality, we assume that $a\ge 0$. We have, for every $x\in[a,b]$,

$$
m_{\bar\xi}(x)=\int_x^b yp_{\bar\xi}(y)\,dy\le p_{\bar\xi}(x)\int_x^b y\,dy\le\bigg(\frac{b^2}{2}\bigg)p_{\bar\xi}(x).
$$

Similarly, for every $x\in[-b,0]$, we have $x\le a$ and, therefore,

$$
m_{\bar\xi}(x)=-\int_{-b}^x yp_{\bar\xi}(y)\,dy\le p_{\bar\xi}(x)\int_{-b}^x(-y)\,dy\le\bigg(\frac{b^2}{2}\bigg)p_{\bar\xi}(x).
$$

Finally, for $x\in[0,a]$, we have

$$
m_{\bar\xi}(x)=\int_{-b}^x(-y)p_{\bar\xi}(y)\,dy\le\int_{-b}^0(-y)p_{\bar\xi}(y)\,dy\le\bigg(\frac{b^2}{2}\bigg)p_{\bar\xi}(0)\le\bigg(\frac{b^2}{2}\bigg)p_{\bar\xi}(x)
$$

and the claim of the lemma follows. □

Appendix A: Proofs of postponed technical results

Proof of Lemma 3. Using the fact that $\pi_n(d\mathbf{F})\propto\exp\{-(\frac{1}{2\tau})\ell_n(\mathbf{F},\mathbf{Y})\}\pi_0(d\mathbf{F})$ and that $\ell_n(\mathbf{F},\mathbf{Y})=\ell_n(\mathbf{F},\mathbf{F}^*)+(\frac{2}{n})\langle\boldsymbol{\xi},\mathbf{F}^*-\mathbf{F}\rangle+(\frac{1}{n})\|\boldsymbol{\xi}\|_F^2$, we arrive at

$$
\begin{aligned}
-S(\alpha)&=\log\int_{\mathcal{F}}e^{-(\frac{1}{2\tau})\ell_n(\mathbf{F},\mathbf{F}^*)-(\frac{1}{n\tau})\langle\boldsymbol{\xi},\mathbf{F}^*-\mathbf{F}\rangle}\pi_0(d\mathbf{F})\\
&\quad-\log\int_{\mathcal{F}}e^{-(\alpha+\frac{1}{2\tau})\ell_n(\mathbf{F},\mathbf{F}^*)-(\frac{1}{n\tau})\langle\boldsymbol{\xi}-2\alpha\tau\boldsymbol{\zeta},\mathbf{F}^*-\mathbf{F}\rangle}\pi_0(d\mathbf{F})
\end{aligned}
$$

One easily checks that the random matrix $\boldsymbol{\xi}-2\alpha\tau\boldsymbol{\zeta}$ has the same distribution as the matrix $(1+2\alpha\tau)\boldsymbol{\xi}$ and, therefore,

$$
\begin{aligned}
-\mathbf{E}\big[S(\alpha)\big]&=\mathbf{E}\bigg[\log\int_{\mathcal{F}}e^{-(\frac{1}{2\tau})\ell_n(\mathbf{F},\mathbf{F}^*)-(\frac{1}{n\tau})\langle\boldsymbol{\xi},\mathbf{F}^*-\mathbf{F}\rangle}\pi_0(d\mathbf{F})\bigg]\\
&\quad-\mathbf{E}\bigg[\log\int_{\mathcal{F}}e^{-(2\alpha\tau+1)(\frac{1}{2\tau}\ell_n(\mathbf{F},\mathbf{F}^*)+(\frac{1}{n\tau})\langle\boldsymbol{\xi},\mathbf{F}^*-\mathbf{F}\rangle)}\pi_0(d\mathbf{F})\bigg].
\end{aligned}
$$

Applying the Hölder inequality $\int_{\mathcal{F}}G\,d\pi_0\le(\int_{\mathcal{F}}G^{2\alpha\tau+1}\,d\pi_0)^{1/(2\alpha\tau+1)}$ to the first expectation of the right hand side, we get

$$
-\mathbf{E}\big[S(\alpha)\big]\le-\frac{2\alpha\tau}{2\alpha\tau+1}\mathbf{E}\bigg[\log\int_{\mathcal{F}}e^{-(2\alpha\tau+1)(\frac{1}{2\tau}\ell_n(\mathbf{F},\mathbf{F}^*)+(\frac{1}{n\tau})\langle\boldsymbol{\xi},\mathbf{F}^*-\mathbf{F}\rangle)}\pi_0(d\mathbf{F})\bigg].
$$

Donsker–Varadhan's variational inequality implies that

$$
-\frac{1}{\alpha}\mathbf{E}\big[S(\alpha)\big]\le\mathbf{E}\bigg[\inf_p\bigg\{\int_{\mathcal{F}}\bigg(\ell_n\big(\mathbf{F},\mathbf{F}^*\big)+\bigg(\frac{2}{n}\bigg)\langle\boldsymbol{\xi},\mathbf{F}^*-\mathbf{F}\rangle\bigg)p(d\mathbf{F})+\frac{2\tau}{2\alpha\tau+1}D_{\mathrm{KL}}(p\parallel\pi_0)\bigg\}\bigg]
$$

$$\le \inf_p \left\{ \int_{\mathcal{F}} \mathbf{E}\left[\ell_n(\mathbf{F}, \mathbf{F}^*) + \left(\frac{2}{n}\right)\langle \boldsymbol{\xi}, \mathbf{F}^* - \mathbf{F}\rangle \right] p(d\mathbf{F}) + \frac{2\tau}{2\alpha\tau + 1} D_{\mathrm{KL}}(p \parallel \pi_0) \right\}$$
$$\le \inf_p \left\{ \int_{\mathcal{F}} \mathbf{E}\left[\ell_n(\mathbf{F}, \mathbf{F}^*) \right] p(d\mathbf{F}) + \frac{2\tau}{2\alpha\tau + 1} D_{\mathrm{KL}}(p \parallel \pi_0) \right\}.$$

The desired result follows from the last display using the inequality $2\alpha\tau + 1 \ge 1$. □

Proof of Lemma 4. In view of (12), we have

$$\ell_n(\widehat{\mathbf{F}}, \mathbf{F}^*, \boldsymbol{\zeta}) - \ell_n(\mathbf{F}, \mathbf{F}^*, \boldsymbol{\zeta}) = \ell_n(\widehat{\mathbf{F}}, \mathbf{F}^*) - \ell_n(\mathbf{F}, \mathbf{F}^*) + \frac{2}{n}\langle \boldsymbol{\zeta}, \widehat{\mathbf{F}} - \mathbf{F}\rangle$$

which implies that,

$$S_1(\alpha) = \log \int_{\mathcal{F}} e^{\alpha(\ell_n(\widehat{\mathbf{F}}, \mathbf{F}^*) - \ell_n(\mathbf{F}, \mathbf{F}^*) + \frac{2}{n}\langle \boldsymbol{\zeta}, \widehat{\mathbf{F}} - \mathbf{F}\rangle)} \pi_n(d\mathbf{F}).$$

Using the definition of the expectation, we get

$$\Psi(\alpha) := \mathbf{E}\left[S_1(\alpha) | \boldsymbol{\xi}\right] = \sum_{\boldsymbol{s} \in \{0,1\}^n} \frac{(\alpha\tau)^{\|\boldsymbol{s}\|_1}(1 + \alpha\tau)^{n - \|\boldsymbol{s}\|_1}}{(1 + 2\alpha\tau)^n} \log \int_{\mathcal{F}} e^{\Phi(\alpha, \boldsymbol{s}, \mathbf{F})} \pi_n(d\mathbf{F}),$$

where

$$\Phi(\alpha, \boldsymbol{s}, \mathbf{F}) := \alpha\big(\ell_n(\widehat{\mathbf{F}}, \mathbf{F}^*) - \ell_n(\mathbf{F}, \mathbf{F}^*)\big) - \frac{2}{n}\sum_{i=1}^n \left\{ \alpha(1 - s_i) - s_i\left(\alpha + \frac{1}{\tau}\right) \right\} \langle \boldsymbol{\xi}_i, \widehat{\mathbf{F}}_i - \mathbf{F}_i\rangle.$$

One easily checks that the function $\Psi(\alpha)$ is differentiable in $(0, \infty)$ and that $\Psi(0) = 0$. Therefore,

$$\lim_{\alpha \to 0} \frac{\Psi(\alpha)}{\alpha} = \lim_{\alpha \to 0} \frac{\Psi(\alpha) - \Psi(0)}{\alpha} = \Psi'(0)$$
$$= \frac{d}{d\alpha}\bigg|_{\alpha = 0} \sum_{\substack{\boldsymbol{s} \in \{0,1\}^n \\ \|\boldsymbol{s}\|_1 \le 1}} \frac{(\alpha\tau)^{\|\boldsymbol{s}\|_1}(1 + \alpha\tau)^{n - \|\boldsymbol{s}\|_1}}{(1 + 2\alpha\tau)^n} \log \int_{\mathcal{F}} e^{\Phi(\alpha, \boldsymbol{s}, \mathbf{F})} \pi_n(d\mathbf{F}). \tag{16}$$

Let us now compute the derivatives with respect to α of the terms of the last sum. First of all, remark that all the terms corresponding to $\boldsymbol{s}$ with $\|\boldsymbol{s}\|_1 \ge 2$ have derivatives vanishing at $\alpha = 0$. This is due to the fact that these terms can be written as $\alpha^2 \varrho(\alpha)$ with a function ϱ which is well defined and differentiable in $[0, 1]$. Therefore, it suffices to consider the terms with $\|\boldsymbol{s}\|_1 \le 1$. For the term corresponding to $\boldsymbol{s} = \mathbf{0}$, since $\Phi(0, \mathbf{0}, \mathbf{F}) = 0$, we have

$$\frac{d}{d\alpha}\bigg|_{\alpha=0} \left\{ \frac{(1 + \alpha\tau)^n}{(1 + 2\alpha\tau)^n} \log \int_{\mathcal{F}} e^{\Phi(\alpha, \mathbf{0}, \mathbf{F})} \pi_n(d\mathbf{F}) \right\} = \frac{d}{d\alpha}\bigg|_{\alpha=0} \left\{ \log \int_{\mathcal{F}} e^{\Phi(\alpha, \mathbf{0}, \mathbf{F})} \pi_n(d\mathbf{F}) \right\}$$
$$= \frac{d}{d\alpha}\bigg|_{\alpha=0} \left\{ \int_{\mathcal{F}} e^{\Phi(\alpha, \mathbf{0}, \mathbf{F})} \pi_n(d\mathbf{F}) \right\}$$
$$\overset{(a)}{=} \int_{\mathcal{F}} \frac{d}{d\alpha}\bigg|_{\alpha=0} \Phi(\alpha, \mathbf{0}, \mathbf{F}) \pi_n(d\mathbf{F}).$$

In the above equality (a), the permutation of the integral and the derivative is justified by the fact that the function $\mathbf{F} \mapsto e^{\Phi(\alpha, \mathbf{0}, \mathbf{F})}$ and its derivative with respect to α are π_n-integrable in a neighborhood of 0. Using that $\Phi(\alpha, \mathbf{0}, \mathbf{F})$ is a linear function of α, as well as the fact that $\int \mathbf{F}\pi_n(d\mathbf{F}) = \widehat{\mathbf{F}}$, we arrive at

$$\frac{d}{d\alpha}\bigg|_{\alpha=0} \left\{ \frac{(1 + \alpha\tau)^n}{(1 + 2\alpha\tau)^n} \log \int_{\mathcal{F}} e^{\Phi(\alpha, \mathbf{0}, \mathbf{F})} \pi_n(d\mathbf{F}) \right\} = -\int_{\mathcal{F}} \ell_n(\widehat{\mathbf{F}}, \mathbf{F}) \pi_n(d\mathbf{F}).$$

Let us go back to (16) and evaluate the terms corresponding to vectors $\boldsymbol{s}$ such that $\|\boldsymbol{s}\|_1 = 1$. This means that only one entry of $\boldsymbol{s}$ is equal to one, all the others being equal to zero. Hence, if we denote by $\boldsymbol{e}_i$ the ith element of the canonical basis of $\mathbb{R}^n$, we get

$$\begin{aligned}
&\frac{d}{d\alpha}\bigg|_{\alpha=0}\sum_{\substack{s\in\{0,1\}^n\\ \|s\|_1=1}}\frac{(\alpha\tau)^{\|s\|_1}(1+\alpha\tau)^{n-\|s\|_1}}{(1+2\alpha\tau)^n}\log\int_{\mathcal{F}}e^{\Phi(\alpha,s,\mathbf{F})}\pi_n(d\mathbf{F})\\
&\quad=\sum_{i\in[n]}\frac{d}{d\alpha}\bigg|_{\alpha=0}\frac{(\alpha\tau)(1+\alpha\tau)^{n-1}}{(1+2\alpha\tau)^n}\log\int_{\mathcal{F}}e^{\Phi(\alpha,e_i,\mathbf{F})}\pi_n(d\mathbf{F})\\
&\quad=\sum_{i\in[n]}\tau\log\int_{\mathcal{F}}e^{\Phi(0,e_i,\mathbf{F})}\pi_n(d\mathbf{F})=\sum_{i\in[n]}\tau\log\int_{\mathcal{F}}e^{(2/n\tau)\boldsymbol{\xi}_i^\top(\widehat{\mathbf{F}}_i-\mathbf{F}_i)}\pi_n(d\mathbf{F}).
\end{aligned}$$

This completes the proof of the lemma. □

Proof of Proposition 2. Recall that $\widehat{\mathbf{F}}^{\text{MC-EWA}}$ is the average of N independent random matrices $\mathbf{F}_1,\dots,\mathbf{F}_N$ such that their conditional distribution given $\mathbf{Y}$ is π_n. This implies that $\mathbf{E}[\mathbf{F}_i]=\widehat{\mathbf{F}}^{\text{EWA}}=\widehat{\mathbf{F}}$. Since ℓ_n is the square loss, the bias-variance decomposition yields

$$\begin{aligned}
\mathbf{E}\big[\ell_n\big(\widehat{\mathbf{F}}^{\text{MC-EWA}},\mathbf{F}^*\big)|\mathbf{Y}\big]&=\frac{1}{n}\big\|\mathbf{E}\big[\widehat{\mathbf{F}}^{\text{MC-EWA}}|\mathbf{Y}\big]-\mathbf{F}^*\big\|_F^2+\frac{1}{n}\mathbf{E}\big[\big\|\widehat{\mathbf{F}}^{\text{MC-EWA}}-\mathbf{E}\big[\widehat{\mathbf{F}}^{\text{MC-EWA}}|\mathbf{Y}\big]\big\|_F^2|\mathbf{Y}\big]\\
&=\frac{1}{n}\big\|\widehat{\mathbf{F}}-\mathbf{F}^*\big\|_F^2+\frac{1}{n}\mathbf{E}\big[\big\|\widehat{\mathbf{F}}^{\text{MC-EWA}}-\widehat{\mathbf{F}}\big\|_F^2|\mathbf{Y}\big]\\
&=\ell_n\big(\widehat{\mathbf{F}},\mathbf{F}^*\big)+\frac{1}{nN}\int_{\mathcal{F}}\|\mathbf{F}-\widehat{\mathbf{F}}\|_F^2\pi_n(d\mathbf{F})\\
&=\ell_n\big(\widehat{\mathbf{F}},\mathbf{F}^*\big)+\frac{1}{N}\int_{\mathcal{F}}\ell_n(\mathbf{F},\widehat{\mathbf{F}})\pi_n(d\mathbf{F}).
\end{aligned}$$

Taking the expectation of both sides of the above inequality, we get the first claim of the distribution. To prove the second claim, let us assume that conditions of Theorem 3 are fulfilled and that $\tau\ge 2(\|\boldsymbol{\Sigma}\|G_{\bar{\xi}})/n$. Then, for $\widehat{\mathbf{F}}=\widehat{\mathbf{F}}^{\text{EWA}}$ we know that

$$\mathbf{E}\big[\ell_n\big(\widehat{\mathbf{F}},\mathbf{F}^*\big)\big]=\inf_p\bigg\{\int_{\mathcal{F}}\ell_n\big(\mathbf{F},\mathbf{F}^*\big)p(d\mathbf{F})+2\tau D_{\text{KL}}(p\|\pi_0)\bigg\}-\frac{1}{2}\int_{\mathcal{F}}\mathbf{E}\big[\ell_n(\mathbf{F},\widehat{\mathbf{F}})\pi_n(d\mathbf{F})\big].$$

Combining with the first claim of the proposition, this leads to

$$\begin{aligned}
\mathbf{E}\big[\ell_n\big(\widehat{\mathbf{F}}^{\text{MC-EWA}},\mathbf{F}^*\big)\big]&\le\mathbf{E}\big[\ell_n\big(\widehat{\mathbf{F}},\mathbf{F}^*\big)\big]+\frac{1}{N}\int_{\mathcal{F}}\mathbf{E}\big[\ell_n(\mathbf{F},\widehat{\mathbf{F}})\pi_n(d\mathbf{F})\big]\\
&\le\inf_p\bigg\{\int_{\mathcal{F}}\ell_n\big(\mathbf{F},\mathbf{F}^*\big)p(d\mathbf{F})+2\tau D_{\text{KL}}(p\|\pi_0)\bigg\}\\
&\quad-\frac{1}{2}\int_{\mathcal{F}}\mathbf{E}\big[\ell_n(\mathbf{F},\widehat{\mathbf{F}})\pi_n(d\mathbf{F})\big]+\frac{1}{N}\int_{\mathcal{F}}\mathbf{E}\big[\ell_n(\mathbf{F},\widehat{\mathbf{F}})\pi_n(d\mathbf{F})\big].
\end{aligned}$$

This readily implies that, as soon as $N\ge 2$,

$$\mathbf{E}\big[\ell_n\big(\widehat{\mathbf{F}}^{\text{MC-EWA}},\mathbf{F}^*\big)\big]\le\inf_p\bigg\{\int_{\mathcal{F}}\ell_n\big(\mathbf{F},\mathbf{F}^*\big)p(d\mathbf{F})+2\tau D_{\text{KL}}(p\|\pi_0)\bigg\}.$$

This completes the proof. □

Proof of Lemma 1. For any bounded and measurable function $h:\mathbb{R}^K\to\mathbb{R}$, we have

$$\begin{aligned}
\int_{\mathcal{F}}h(\mathbf{F}_1)\pi_0(\mathbf{F})\,d\mathbf{F}&\overset{(a)}{=}\frac{1}{C_\lambda}\int_{\mathcal{F}}\frac{h(\mathbf{F}_1)}{\det(\lambda^2\mathbf{I}_K+\mathbf{F}\mathbf{F}^\top)^{(n+K+2)/2}}\,d\mathbf{F}\\
&\overset{(b)}{=}\frac{1}{C_1}\int_{\mathcal{F}}\frac{h(\lambda\mathbf{M}_1)}{\det(\mathbf{I}_K+\mathbf{M}\mathbf{M}^\top)^{(n+K+2)/2}}\,d\mathbf{M}
\end{aligned}$$

where in (a) we have used the notation $C_\lambda = \int_{\mathcal{F}} \det(\lambda^2 \mathbf{I}_K + \mathbf{F}\mathbf{F}^\top)^{-(n+K+2)/2}\, d\mathbf{F}$, whereas in (b) we have made the change of variable $\mathbf{F} = \lambda \mathbf{M}$. In order to compute the last integral, we make another change of variable, $\mathbf{M} \rightsquigarrow \bar{\mathbf{M}}$, given by $\mathbf{M} = [\bar{\mathbf{M}}_1, (\mathbf{I} + \bar{\mathbf{M}}_1\bar{\mathbf{M}}_1^\top)^{1/2}\bar{\mathbf{M}}_{2:n}]$. This yields

$$\begin{aligned} d\mathbf{M} = d\mathbf{M}_1\, d\mathbf{M}_{2:n} &= d\bar{\mathbf{M}}_1 \det\bigl(\mathbf{I} + \bar{\mathbf{M}}_1\bar{\mathbf{M}}_1^\top\bigr)^{(n-1)/2}\, d\bar{\mathbf{M}}_{2:n} \\ &\overset{(c)}{=} \bigl(1 + \|\bar{\mathbf{M}}_1\|_2^2\bigr)^{(n-1)/2}\, d\bar{\mathbf{M}}_1\, d\bar{\mathbf{M}}_{2:n} \end{aligned}$$

and

$$\begin{aligned} \det\bigl(\mathbf{I} + \mathbf{M}\mathbf{M}^\top\bigr) &= \det\bigl(\mathbf{I} + \mathbf{M}_1\mathbf{M}_1^\top + \mathbf{M}_{2:n}\mathbf{M}_{2:n}^\top\bigr) \\ &= \det\bigl(\mathbf{I} + \bar{\mathbf{M}}_1\bar{\mathbf{M}}_1^\top + \bigl(\mathbf{I} + \bar{\mathbf{M}}_1\bar{\mathbf{M}}_1^\top\bigr)^{1/2}\bar{\mathbf{M}}_{2:n}\bar{\mathbf{M}}_{2:n}^\top\bigl(\mathbf{I} + \bar{\mathbf{M}}_1\bar{\mathbf{M}}_1^\top\bigr)^{1/2}\bigr) \\ &= \det\bigl(\bigl(\mathbf{I} + \bar{\mathbf{M}}_1\bar{\mathbf{M}}_1^\top\bigr)^{1/2}\bigl(\mathbf{I} + \bar{\mathbf{M}}_{2:n}\bar{\mathbf{M}}_{2:n}^\top\bigr)\bigl(\mathbf{I} + \bar{\mathbf{M}}_1\bar{\mathbf{M}}_1^\top\bigr)^{1/2}\bigr) \\ &= \det\bigl(\mathbf{I} + \bar{\mathbf{M}}_1\bar{\mathbf{M}}_1^\top\bigr)\det\bigl(\mathbf{I} + \bar{\mathbf{M}}_{2:n}\bar{\mathbf{M}}_{2:n}^\top\bigr) \\ &\overset{(c')}{=} \bigl(1 + \|\bar{\mathbf{M}}_1\|_2^2\bigr)\det\bigl(\mathbf{I} + \bar{\mathbf{M}}_{2:n}\bar{\mathbf{M}}_{2:n}^\top\bigr), \end{aligned}$$

where in (c) and (c') we have used the fact that the matrix $\mathbf{I} + \bar{\mathbf{M}}_1\bar{\mathbf{M}}_1^\top$ has all its eigenvalues equal to 1 except the largest one, corresponding to the eigenvector $\bar{\mathbf{M}}_1$, which is equal to $1 + \|\bar{\mathbf{M}}_1\|_2^2$. Using the same change of variable in C_1, and replacing $\bar{\mathbf{M}}_1$ by $\boldsymbol{x}$ for convenience, we get

$$\begin{aligned} \int_{\mathcal{F}} h(\mathbf{F}_1)\pi_0(\mathbf{F})\, d\mathbf{F} &= \frac{\int_{\mathbb{R}^K} h(\lambda\boldsymbol{x})(1 + \|\boldsymbol{x}\|_2^2)^{-(K+3)/2}\, d\boldsymbol{x}}{\int_{\mathbb{R}^K}(1 + \|\boldsymbol{x}\|_2^2)^{-(K+3)/2}\, d\boldsymbol{x}} \\ &= \frac{\int_{\mathbb{R}^K} h(\lambda\boldsymbol{y}/\sqrt{3})(1 + \|\boldsymbol{y}\|_2^2/3)^{-(K+3)/2}\, d\boldsymbol{y}}{3\int_{\mathbb{R}^K}(1 + \|\boldsymbol{y}\|_2^2/3)^{-(K+3)/2}\, d\boldsymbol{y}}. \end{aligned}$$

In the last expression, we recognize the probability density function of the multivariate t_3-distribution. Since the covariance matrix of a K-variate t_ν distribution is equal to $\frac{\nu}{\nu-2}\mathbf{I}_K$, we infer that

$$\int_{\mathcal{F}} \|\mathbf{F}\|_F^2 \pi_0(\mathbf{F})\, d\mathbf{F} = nK\lambda^2.$$

This completes the proof of the lemma. □

Proof of Lemma 2. It holds that

$$\begin{aligned} D_{\mathrm{KL}}(\bar{p}\|\pi_0) &= \int_{\mathcal{F}} \log\biggl(\frac{\bar{p}(\mathbf{F})}{\pi_0(\mathbf{F})}\biggr)\bar{p}(\mathbf{F})\, d\mathbf{F} \\ &= \int_{\mathcal{F}} \log\biggl(\frac{\pi_0(\mathbf{F} - \bar{\mathbf{F}})}{\pi_0(\mathbf{F})}\biggr)\pi_0(\mathbf{F} - \bar{\mathbf{F}})\, d\mathbf{F} \\ &\overset{(1)}{=} \int_{\mathcal{F}} \log\biggl(\frac{\pi_0(\mathbf{U})}{\pi_0(\bar{\mathbf{F}} + \mathbf{U})}\biggr)\pi_0(\mathbf{U})\, d\mathbf{U} \\ &\overset{(2)}{=} \int_{\mathcal{F}} \log\biggl(\frac{\pi_0(\mathbf{F})}{\pi_0(\bar{\mathbf{F}} - \mathbf{F})}\biggr)\pi_0(\mathbf{F})\, d\mathbf{F}. \end{aligned}$$

In (1) above we have used the change of variable $\mathbf{U} = \mathbf{F} - \bar{\mathbf{F}}$, while in (2) we have set $\mathbf{F} = -\mathbf{U}$ and have used the fact that the density π_0 is symmetric. To ease notation, we set $\mathbf{A} = (\lambda^2\mathbf{I}_K + \mathbf{F}\mathbf{F}^\top)^{-1/2}$ and $\mathbf{B} = \lambda^2\mathbf{I}_K + (\mathbf{F} - \bar{\mathbf{F}})(\mathbf{F} - \bar{\mathbf{F}})^\top$. We have

$$2\log\biggl(\frac{\pi_0(\mathbf{F})}{\pi_0(\bar{\mathbf{F}} - \mathbf{F})}\biggr) = (n + K + 2)\log\biggl(\frac{\det(\mathbf{B})}{\det(\mathbf{A}^{-2})}\biggr)$$

$$= (n + K + 2)\log\bigl(\det(\mathbf{ABA})\bigr)$$
$$= (n + K + 2)\sum_{j=1}^{K}\log s_j(\mathbf{ABA}), \tag{17}$$

where $s_j(\mathbf{ABA})$ is the jth largest eigenvalue of the symmetric matrix $\mathbf{ABA}$. Let r be the rank of $\bar{\mathbf{F}}$. The first claim is that the matrix $\mathbf{ABA}$ has at most $2r$ singular values different from one. Indeed, one can check that

$$\mathbf{ABA} - \mathbf{I}_K = \mathbf{A}\bar{\mathbf{F}}\bar{\mathbf{F}}^\top\mathbf{A} - \mathbf{A}\bar{\mathbf{F}}\mathbf{F}^\top\mathbf{A} - \mathbf{A}\mathbf{F}\bar{\mathbf{F}}^\top\mathbf{A}.$$

The matrix at the right hand side is at most of rank $2r$. This implies that $\mathbf{ABA} - \mathbf{I}_K$ has at most $2r$ nonzero eigenvalues. Therefore, the number of eigenvalues of $\mathbf{ABA}$ different from 1 is not larger than $2r$, which implies that the sum at the right hand side of (17) has at most $2r$ nonzero entries.

Let $\boldsymbol{u}_j$ be the unit eigenvector corresponding to the eigenvalue s_j. We know that, for every $j \in [2r]$, $s_j = \boldsymbol{u}_j^\top\mathbf{ABA}\boldsymbol{u}_j$. Using , we get

$$\begin{aligned}
s_j &= 1 + \boldsymbol{u}_j^\top\bigl(\mathbf{A}\bar{\mathbf{F}}\bar{\mathbf{F}}^\top\mathbf{A} - \mathbf{A}\bar{\mathbf{F}}\mathbf{F}^\top\mathbf{A} - \mathbf{A}\mathbf{F}\bar{\mathbf{F}}^\top\mathbf{A}\bigr)\boldsymbol{u}_j \\
&= 1 + \bigl\|\bar{\mathbf{F}}^\top\mathbf{A}\boldsymbol{u}_j\bigr\|_2^2 - 2\bigl(\bar{\mathbf{F}}^\top\mathbf{A}\boldsymbol{u}_j\bigr)^\top\bigl(\mathbf{F}^\top\mathbf{A}\boldsymbol{u}_j\bigr) \\
&\le 1 + \bigl\|\bar{\mathbf{F}}^\top\mathbf{A}\boldsymbol{u}_j\bigr\|_2^2 + 2\bigl\|\bar{\mathbf{F}}^\top\mathbf{A}\boldsymbol{u}_j\bigr\|_2\bigl\|\mathbf{F}^\top\mathbf{A}\boldsymbol{u}_j\bigr\|_2 \\
&\overset{(\star)}{\le} \bigl(1 + \bigl\|\bar{\mathbf{F}}^\top\mathbf{A}\boldsymbol{u}_j\bigr\|_2\bigr)^2,
\end{aligned}$$

where $(\star)$ follows from the fact that $\|\mathbf{F}^\top\mathbf{A}\boldsymbol{u}_j\|_2 \le \|\mathbf{F}^\top\mathbf{A}\| \le 1$. Using the concavity of the function $\log(1 + x^{1/2})$ over $(0, +\infty)$, we arrive at

$$\begin{aligned}
2\log\biggl(\frac{\pi_0(\mathbf{F})}{\pi_0(\bar{\mathbf{F}} - \mathbf{F})}\biggr) &= (n + K + 2)\sum_{j=1}^{2r}\log s_j(\mathbf{ABA}) \\
&\le 2(n + K + 2)\sum_{j=1}^{2r}\log\bigl(1 + \bigl\{\bigl\|\bar{\mathbf{F}}^\top\mathbf{A}\boldsymbol{u}_j\bigr\|_2^2\bigr\}^{1/2}\bigr) \\
&\le 4(n + K + 2)r\log\left(1 + \left\{\frac{1}{2r}\sum_{j=1}^{2r}\bigl\|\bar{\mathbf{F}}^\top\mathbf{A}\boldsymbol{u}_j\bigr\|_2^2\right\}^{1/2}\right).
\end{aligned}$$

Finally, since $\boldsymbol{u}_j$'s are orthonormal and $\mathbf{A} \preceq \lambda^{-1}\mathbf{I}_K$, we get the claim of the lemma. □

Appendix B: Flaw in Corollary 1 of [14]

As mentioned in the Introduction, Corollary 1 in [14] relies heavily on Lemma 3 of the same paper, that claims that

$$x + \log\biggl(1 + \frac{1}{\alpha_0}\bigl(e^{-x\alpha_0} - 1\bigr)\biggr) \le \frac{x^2\alpha_0}{2}, \qquad \forall x \in \mathbb{R}, \forall \alpha_0 > 0.$$

Unfortunately, this inequality is not always true. In particular, the argument of the logarithm is not always positive, which implies that the left hand side is not always well defined. For instance, one can check that if $\alpha = 0.5$ and $x \ge 2$, we have

$$1 + \frac{1}{\alpha_0}\bigl(e^{-x\alpha_0} - 1\bigr) = 2e^{-0.5x} - 1 \le (2/e) - 1 \le 0.$$

Acknowledgement

This research is supported by a grant of the French National Research Agency (ANR), "Investissements d'Avenir" (LabEx Ecodec/ANR-11-LABX-0047) and the chair "LCL/GENES/Fondation du risque, Nouveaux enjeux pour nouvelles données".

References

[1] P. Alquier. Bayesian methods for low-rank matrix estimation: Short survey and theoretical study. In *Algorithmic Learning Theory* 309–323. *Berlin, Heidelberg*, 2013. MR3133074 https://doi.org/10.1007/978-3-642-40935-6_22

[2] P. Alquier and B. Guedj. An oracle inequality for quasi-Bayesian nonnegative matrix factorization. *Math. Methods Statist.* **26** (1) (2017) 55–67. MR3656072 https://doi.org/10.3103/S1066530717010045

[3] P. Alquier and K. Lounici. PAC-Bayesian bounds for sparse regression estimation with exponential weights. *Electron. J. Stat.* **5** (2011) 127–145. MR2786484 https://doi.org/10.1214/11-EJS601

[4] J.-Y. Audibert. Fast learning rates in statistical inference through aggregation. *Ann. Statist.* **37** (4) (2009) 1591–1646. MR2533466 https://doi.org/10.1214/08-AOS623

[5] P. C. Bellec. Optimal bounds for aggregation of affine estimators. *Ann. Statist.* **46** (1) (2018) 30–59. MR3766945 https://doi.org/10.1214/17-AOS1540

[6] F. Bunea, Y. She and M. H. Wegkamp. Optimal selection of reduced rank estimators of high-dimensional matrices. *Ann. Statist.* **39** (2) (2011) 1282–1309. MR2816355 https://doi.org/10.1214/11-AOS876

[7] F. Bunea, Y. She and M. H. Wegkamp. Joint variable and rank selection for parsimonious estimation of high-dimensional matrices. *Ann. Statist.* **40** (5) (2012) 2359–2388. MR3097606 https://doi.org/10.1214/12-AOS1039

[8] E. J. Candès and T. Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Trans. Inform. Theory* **56** (5) (2010) 2053–2080. MR2723472 https://doi.org/10.1109/TIT.2010.2044061

[9] O. Catoni. *Pac-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*. *Lecture Notes – Monograph Series* **56**. Institute of Mathematical Statistics, Beachwood, OH, 2007. MR2483528

[10] D. Dai, P. Rigollet, L. Xia and T. Zhang. Aggregation of affine estimators. *Electron. J. Stat.* **8** (1) (2014) 302–327. MR3192554 https://doi.org/10.1214/14-EJS886

[11] A. S. Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** (3) (2017) 651–676. MR3641401 https://doi.org/10.1111/rssb.12183

[12] A. S. Dalalyan and J. Salmon. Sharp oracle inequalities for aggregation of affine estimators. *Ann. Statist.* **40** (4) (2012) 2327–2355. MR3059085 https://doi.org/10.1214/12-AOS1038

[13] A. S. Dalalyan and A. B. Tsybakov. Aggregation by exponential weighting and sharp oracle inequalities. In *Learning Theory* 97–111. *LNCS* **4539**, 2007. MR2397581 https://doi.org/10.1007/978-3-540-72927-3_9

[14] A. S. Dalalyan and A. B. Tsybakov. Aggregation by exponential weighting, sharp pac-Bayesian bounds and sparsity. *Mach. Learn.* **72** (1–2) (2008) 39–61.

[15] A. S. Dalalyan and A. B. Tsybakov. Mirror averaging with sparsity priors. *Bernoulli* **18** (3) (2012) 914–944. MR2948907 https://doi.org/10.3150/11-BEJ361

[16] A. S. Dalalyan and A. B. Tsybakov. Sparse regression learning by aggregation and Langevin Monte-Carlo. *J. Comput. System Sci.* **78** (5) (2012) 1423–1443. MR2926142 https://doi.org/10.1016/j.jcss.2011.12.023

[17] A. Durmus and E. Moulines. Sampling from strongly log-concave distributions with the Unadjusted Langevin Algorithm. Technical Report, 2016. Available at arXiv:1605.01559.

[18] E. I. George. Combining minimax shrinkage estimators. *J. Amer. Statist. Assoc.* **81** (394) (1986) 437–445. MR0845881

[19] E. I. George. Minimax multiple shrinkage estimation. *Ann. Statist.* **14** (1) (1986) 188–205. MR0829562 https://doi.org/10.1214/aos/1176349849

[20] S. Gerchinovitz. Sparsity regret bounds for individual sequences in online linear regression. *J. Mach. Learn. Res.* **14** (1) (2013) 729–769. MR3049487

[21] A. J. Izenman. Reduced-rank regression for the multivariate linear model. *J. Multivariate Anal.* **5** (2) (1975) 248–264. MR0373179 https://doi.org/10.1016/0047-259X(75)90042-1

[22] A. Juditsky, P. Rigollet and A. B. Tsybakov. Learning by mirror averaging. *Ann. Statist.* **36** (5) (2008) 2183–2206. MR2458184 https://doi.org/10.1214/07-AOS546

[23] G. Leung. Information Theory and Mixing Least Squares Regression. PhD thesis, Yale University, 2004.

[24] G. Leung and A. R. Barron. Information theory and mixing least-squares regressions. *IEEE Trans. Inform. Theory* **52** (8) (2006) 3396–3410. MR2242356 https://doi.org/10.1109/TIT.2006.878172

[25] S. Negahban and M. J. Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *Ann. Statist.* **39** (2) (2011) 1069–1097. MR2816348 https://doi.org/10.1214/10-AOS850

[26] P. Rigollet and A. Tsybakov. Exponential screening and optimal rates of sparse estimation. *Ann. Statist.* **39** (2) (2011) 731–771. MR2816337 https://doi.org/10.1214/10-AOS854

[27] P. Rigollet and A. B. Tsybakov. Sparse estimation by exponential weighting. *Statist. Sci.* **27** (4) (2012) 558–575. MR3025134 https://doi.org/10.1214/12-STS393

[28] A. Rohde and A. B. Tsybakov. Estimation of high-dimensional low-rank matrices. *Ann. Statist.* **39** (2) (2011) 887–930. MR2816342 https://doi.org/10.1214/10-AOS860

[29] O. Wintenberger. Optimal learning with Bernstein online aggregation. *Mach. Learn.* **106** (1) (2017) 119–141. MR3592152 https://doi.org/10.1007/s10994-016-5592-6

[30] L. Yang, J. Fang, H. Duan, H. Li and B. Zeng. Fast low-rank Bayesian matrix completion with hierarchical Gaussian prior models. *IEEE Trans. Signal Process.* **66** (11) (2018) 2804–2817. MR3816906 https://doi.org/10.1109/tsp.2018.2816575

[31] Y. Yang. Adaptive estimation in pattern recognition by combining different procedures. *Statist. Sinica* **10** (4) (2000) 1069–1089. MR1804557

[32] Y. Yang. Combining different procedures for adaptive regression. *J. Multivariate Anal.* **74** (1) (2000) 135–161. MR1790617 https://doi.org/10.1006/jmva.1999.1884