Dimension-free PAC-Bayesian bounds for matrices, vectors, and linear least squares regression.

Olivier Catoni [*] and Ilaria Giulini [†]

December 31, 2017

Abstract: This paper is focused on dimension-free PAC-Bayesian bounds, under weak polynomial moment assumptions, allowing for heavy tailed sample distributions. It covers the estimation of the mean of a vector or a matrix, with applications to least squares linear regression. Special efforts are devoted to the estimation of the Gram matrix, due to its prominent role in high-dimension data analysis.

Key words: PAC-Bayesian bounds, sub-Gaussian mean estimator, random vector, random matrix, least squares regression, dimension-free bounds

MSC2010: 62J10, 62J05, 62H12, 62H20, 62F35, 15B52.

1. Introduction The subject of this paper is to discuss dimension-free PAC-Bayesian bounds for matrices and vectors. It comes after Catoni (2016) and Giulini (2017a), the first paper discussing dimension dependent bounds and the second one dimension-free bounds, under a kurtosis like assumption about the data distribution. Here, in contrast, we envision even weaker assumptions, and focus on dimension-free bounds only.

Our main objective is the estimation of the mean of a random vector and of a random matrix. Finding sub-Gaussian estimators for the mean of a non necessarily sub-Gaussian random vector has been the subject of much research in the last few years, with important contributions from Joly, Lugosi and Oliveira (2017), Lugosi and Mendelson (2017) and Minsker (2015). While in Joly, Lugosi and Oliveira (2017) the statistical error bound still has a residual dependence on the dimension of the ambient space, in Lugosi and Mendelson (2017) this dependence is removed, for an estimator of the median of means type. However, this estimator is not easy to compute and the bound contains large constants. We propose here another type of estimator, that can be seen as a multidimensional extension of Catoni (2012). It provides a nonasymptotic confidence region with the same diameter (including the values of the constants) as the Gaussian concentration inequality stated in equation (1.1) of Lugosi and Mendelson (2017), although in our case, the confidence region is not necessarily a ball, but still a convex set. The Gaussian bound concerns the

---
[*]CREST – CNRS, UMR 9194, Université Paris Saclay, France; e-mail: olivier.catoni@ensae.fr

[†] Laboratoire de Probabilités et Modèles Aléatoires, Université Paris Diderot, France; e-mail: giulini@math.univ-paris-diderot.fr

estimation of the expectation of a Gaussian random vector by the mean of an i.i.d. sample, whereas in our case, we only assume that the variance is finite, a much weaker hypothesis.

In Minsker (2016) the question of estimating the mean of a random matrix is addressed. The author uses exponential matrix inequalities in order to extend Catoni (2012) to matrices and to control the operator norm of the error. In the bounds at confidence level $1 - \delta$, the complexity term is multiplied by $\log(\delta^{-1})$. Here, we extend Catoni (2012) using PAC-Bayesian bounds to measure complexity, and define an estimator with a bound where the term $\log(\delta^{-1})$ is multiplied by some directional variance term only, and not the complexity factor, that is larger.

After recalling in Section 2 the PAC-Bayesian inequality that will be at the heart of many of our proofs, we deal successively with the estimation of a random vector (Section 3) and of a random matrix (Section 4). Section 6 is devoted to the estimation of the Gram matrix, due to its prominent role in multidimensional data analysis. In Section 7 we introduce some applications to least squares regression.

2. Some well known PAC-Bayesian inequality   This is a preliminary section, where we state the PAC-Bayesian inequality that we will use throughout this paper to obtain deviation inequalities holding uniformly with respect to some parameter.

Consider a random variable $X \in \mathcal{X}$ and a measurable parameter space Θ. Let $\mu \in \mathcal{M}_+^1(\Theta)$ be a probability measure on Θ and $f : \Theta \times \mathcal{X} \to \mathbb{R}$ a bounded measurable function. For any other probability measure ρ on Θ, define the Kullback divergence function $\mathcal{K}(\rho, \mu)$ as usual by the formula

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

Let $(X_1, \ldots, X_n)$ be n independent copies of X.

PROPOSITION 2.1.   *For any $\delta \in]0, 1[$, with probability at least $1 - \delta$, for any probability measure $\rho \in \mathcal{M}_+^1(\Theta)$,*

$$
\frac{1}{n} \sum_{i=1}^{n} \int f(\theta, X_i) \mathrm{d}\rho(\theta) \leq \int \log\left[\mathbb{E}\left(\exp(f(\theta, X))\right)\right] \mathrm{d}\rho(\theta) + \frac{\mathcal{K}(\rho, \mu) + \log(\delta^{-1})}{n}.
$$

PROOF.   It is a consequence of equation (5.2.1) page 159 of Catoni (2004). Indeed, let us recall the identity

$$
\log\left(\int \exp(h(\theta)) \, \mathrm{d}\mu(\theta)\right) = \sup_{\rho}\left\{\int h(\theta) \, \mathrm{d}\rho(\theta) - \mathcal{K}(\rho, \mu)\right\},
$$

where h may be any bounded measurable function (extensions to unbounded h are possible but will not be required in this paper), and where the supremum in ρ is taken on all probability measures on the measurable parameter space Θ. The proof may be found in (Catoni, 2004, page 159). Combined with Fubini's lemma, it yields

$$\mathbb{E}\left\{\exp\sup_{\rho}\left[\int\left(\sum_{i=1}^{n} f(\theta, X_i) - n\log\left[\mathbb{E}\left(\exp(f(\theta, X))\right)\right]\right)\mathrm{d}\rho(\theta) - \mathcal{K}(\rho, \mu)\right]\right\}$$

$$= \mathbb{E}\left\{\int \exp\left(\sum_{i=1}^{n} f(\theta, X_i) - n\log\left[\mathbb{E}\left(\exp(f(\theta, X))\right)\right]\right)\mathrm{d}\mu(\theta)\right\}$$

$$= \int \mathbb{E}\exp\left(\sum_{i=1}^{n} f(\theta, X_i) - n\log\left[\mathbb{E}\left(\exp(f(\theta, X))\right)\right]\right)\mathrm{d}\mu(\theta)$$

$$= \int \prod_{i=1}^{n}\left[\frac{\mathbb{E}\left(\exp(f(\theta, X_i))\right)}{\mathbb{E}\left(\exp(f(\theta, X))\right)}\right]\mathrm{d}\mu(\theta) = 1.$$

Since $\mathbb{E}(\exp(W)) \leq 1$ implies that

$$\mathbb{P}(W \geq \log(\delta^{-1})) = \mathbb{E}\left(\mathbb{1}\left[\delta\exp(W) \geq 1\right]\right) \leq \mathbb{E}(\delta\exp(W)) \leq \delta,$$

we obtain the desired result, considering

$$W = \sup_{\rho}\left[\int\left(\sum_{i=1}^{n} f(\theta, X_i) - n\log\left[\mathbb{E}\left(\exp(f(\theta, X))\right)\right]\right)\mathrm{d}\rho(\theta) - \mathcal{K}(\rho, \mu)\right].$$

$\square$

3. Estimation of the mean of a random vector   Let $X \in \mathbb{R}^d$ be a random vector and let $(X_1, \ldots, X_n)$ be n independent copies of X. In this section, we will estimate the mean $\mathbb{E}(X)$ and obtain dimension-free non-asymptotic bounds for the estimation error.

Let $\mathbb{S}_d = \{\theta \in \mathbb{R}^d : \|\theta\| = 1\}$ be the unit sphere of $\mathbb{R}^d$ and let I_d be the identity matrix of size $d \times d$. Let $\rho_\theta = \mathcal{N}(\theta, \beta^{-1} I_d)$ be the normal distribution centered at $\theta \in \mathbb{R}^d$, whose covariance matrix is $\beta^{-1} I_d$, where β is a positive real parameter.

Instead of estimating directly the mean vector $\mathbb{E}(X)$, our strategy will be rather to estimate its component $\langle\theta, \mathbb{E}(X)\rangle$ in each direction $\theta \in \mathbb{S}_d$ of the unit sphere. For this, we introduce the estimator of $\langle\theta, \mathbb{E}(X)\rangle$ defined as

$$\mathcal{E}(\theta) = \frac{1}{n\lambda}\sum_{i=1}^{n}\int \psi\left(\lambda\langle\theta', X_i\rangle\right)\mathrm{d}\rho_\theta(\theta'), \qquad \theta \in \mathbb{S}_d, \lambda > 0,$$

3

where ψ is the symmetric influence function

$$(1) \qquad \psi(t) = \begin{cases} t - t^3/6, & -\sqrt{2} \le t \le \sqrt{2} \\ 2\sqrt{2}/3, & t > \sqrt{2} \\ -2\sqrt{2}/3, & t < -\sqrt{2} \end{cases}$$

and where the positive constants λ and β will be chosen afterward.

As stated in the following lemma, we chose this influence function because it is close to the identity in a neighborhood of zero and is such that $\exp(\psi(t))$ is bounded by polynomial functions.

LEMMA 3.1. *For any $t \in \mathbb{R}$,*

$$-\log(1 - t + t^2/2) \le \psi(t) \le \log(1 + t + t^2/2).$$

PROOF. Put $f(t) = \log(1 + t + t^2/2)$. Remark that $f'(t) = \dfrac{1+t}{1+t+t^2/2}$ for $t \in \mathbb{R}$ and that $\psi'(t) = 1 - t^2/2$ for $t \in [-\sqrt{2}, \sqrt{2}]$. As $\psi(0) = f(0) = 0$ and

$$[f'(t) - \psi'(t)](1 + t + t^2/2) = \frac{t^3(2-t)}{4},$$

$$\psi'(t) \le f'(t), \qquad 0 \le t \le \sqrt{2},$$

$$\psi'(t) \ge f'(t), \quad -\sqrt{2} \le t \le 0,$$

proving that

$$\psi(t) \le f(t), \qquad -\sqrt{2} \le t \le \sqrt{2}.$$

Since f is increasing on $[\sqrt{2}, +\infty[$ and decreasing on $]-\infty, -\sqrt{2}]$, while ψ is constant on these two intervals, the above inequality can be extended to all $t \in \mathbb{R}$. From the symmetry $\psi(-t) = -\psi(t)$, we deduce the converse inequality

$$-f(-t) \le \psi(t), \qquad t \in \mathbb{R}$$

that ends the proof. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad$ □

Since $\lambda\langle\theta', X_i\rangle$ follows a normal distribution with mean $\lambda\langle\theta, X_i\rangle$ and standard deviation $\lambda\beta^{-1/2}\|X_i\|$, and since the influence function ψ is piecewise polynomial, the estimator $\mathcal{E}$ can be computed explicitly in terms of the standard normal distribution function. This is done in the following lemma.

LEMMA 3.2. *Let $W \sim \mathcal{N}(0,1)$ be a standard Gaussian real valued random variable. For any $m \in \mathbb{R}$ and any $\sigma \in \mathbb{R}_+$, define*

$$\varphi(m, \sigma) = \mathbb{E}[\psi(m + \sigma W)].$$

4

The function φ can be computed as

$$\varphi(m,\sigma) = m(1 - \sigma^2/2) - m^3/6 + r(m,\sigma),$$

where, introducing $F(a) = \mathbb{P}(W \le a)$, $a \in \mathbb{R}$, the correction term r is

$$
\begin{aligned}
r(m,\sigma) = {}& \frac{2\sqrt{2}}{3}\left[F\left(\frac{-\sqrt{2}+m}{\sigma}\right) - F\left(\frac{-\sqrt{2}-m}{\sigma}\right)\right] \\
& - (m - m^3/6)\left[F\left(\frac{-\sqrt{2}+m}{\sigma}\right) + F\left(\frac{-\sqrt{2}-m}{\sigma}\right)\right] \\
& + \sigma\frac{(1-m^2/2)}{\sqrt{2\pi}}\left[\exp\left(-\frac{1}{2}\left(\frac{\sqrt{2}+m}{\sigma}\right)^2\right) - \exp\left(-\frac{1}{2}\left(\frac{\sqrt{2}-m}{\sigma}\right)^2\right)\right] \\
& + \frac{m\sigma^2}{2}\left\{F\left(\frac{-\sqrt{2}-m}{\sigma}\right) + F\left(\frac{-\sqrt{2}+m}{\sigma}\right)\right. \\
& + \frac{1}{\sqrt{2\pi}}\left[\frac{(\sqrt{2}+m)}{\sigma}\exp\left[-\frac{1}{2}\left(\frac{\sqrt{2}+m}{\sigma}\right)^2\right] + \frac{(\sqrt{2}-m)}{\sigma}\exp\left[-\frac{1}{2}\left(\frac{\sqrt{2}-m}{\sigma}\right)^2\right]\right]\right\} \\
& + \frac{\sigma^3}{6\sqrt{2\pi}}\left\{\left[\left(\frac{\sqrt{2}-m}{\sigma}\right)^2 + 2\right]\exp\left[-\frac{1}{2}\left(\frac{\sqrt{2}-m}{\sigma}\right)^2\right]\right. \\
& \left. - \left[\left(\frac{\sqrt{2}+m}{\sigma}\right)^2 + 2\right]\exp\left[-\frac{1}{2}\left(\frac{\sqrt{2}+m}{\sigma}\right)^2\right]\right\}.
\end{aligned}
$$

Remark that the correction term is small when $|m|$ is small and σ is small, since

$$F(-t) \le \min\left\{\frac{1}{t\sqrt{2\pi}}, \frac{1}{2}\right\}\exp\left(-\frac{t^2}{2}\right), \qquad t \in \mathbb{R}_+.$$

PROOF. The proof of this lemma is a simple computation, based on the expression

$$
\begin{aligned}
\psi(t) = {}& \left(t - \frac{t^3}{6}\right)[\mathbb{1}(t \le \sqrt{2}) - \mathbb{1}(t \le -\sqrt{2})] \\
& + \frac{2\sqrt{2}}{3}[1 - \mathbb{1}(t \le \sqrt{2}) - \mathbb{1}(t \le -\sqrt{2})], \qquad t \in \mathbb{R},
\end{aligned}
$$

on the identities

$$\mathbb{E}[\mathbb{1}(W \le a)] = F(a),$$

$$\mathbb{E}[\mathbb{1}(W \le a)W] = -\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{a^2}{2}\right),$$

$$\mathbb{E}[\mathbb{1}(W \le a)W^2] = F(a) - \frac{a}{\sqrt{2\pi}} \exp\left(-\frac{a^2}{2}\right),$$

$$\mathbb{E}[\mathbb{1}(W \le a)W^3] = -\frac{(a^2 + 2)}{\sqrt{2\pi}} \exp\left(-\frac{a^2}{2}\right),$$

and on the fact that $F(-t) = 1 - F(t)$. $\qquad\square$

Accordingly, the estimator $\mathcal{E}$ can be computed as

$$\mathcal{E}(\theta) = \frac{1}{n\lambda} \sum_{i=1}^{n} \varphi\left(\lambda\langle\theta, X_i\rangle, \lambda\beta^{-1/2}\|X_i\|\right)$$

$$= \frac{1}{n} \sum_{i=1}^{n} \langle\theta, X_i\rangle\left(1 - \frac{\lambda^2\|X_i\|^2}{2\beta}\right) - \frac{\lambda^2\langle\theta, X_i\rangle^3}{6} + r\left(\lambda\langle\theta, X_i\rangle, \lambda\beta^{-1/2}\|X_i\|\right).$$

3.1. Estimation without centering

PROPOSITION 3.3. *Assume that*

$$\mathbb{E}(\|X\|^2) = \mathbf{Tr}[\mathbb{E}(XX^\top)] \le T < \infty$$

and $\qquad \sup_{\theta\in\mathcal{S}} \mathbb{E}(\langle\theta, X\rangle^2) \le v \le T < \infty,$

where T and v are two known constants and where $\mathcal{S} \subset \mathbb{S}_d$ is an arbitrary symmetric subset of the unit sphere, meaning that if $\theta \in \mathcal{S}$ then $-\theta \in \mathcal{S}$. Choose any confidence parameter $\delta \in]0, 1[$ and set the constants λ and β used in the definition of the estimator $\mathcal{E}$ to

$$\lambda = \sqrt{\frac{2\log(\delta^{-1})}{nv}},$$

$$\beta = \sqrt{nT}\lambda = \sqrt{\frac{2T\log(\delta^{-1})}{v}}.$$

NON ASYMPTOTIC CONFIDENCE REGION: *With probability at least $1 - \delta$,*

$$\sup_{\theta\in\mathcal{S}}|\mathcal{E}(\theta) - \langle\theta, \mathbb{E}(X)\rangle| \le \sqrt{\frac{T}{n}} + \sqrt{\frac{2v\log(\delta^{-1})}{n}}.$$

Consider an estimator $\widehat{m} \in \mathbb{R}^d$ of $\mathbb{E}(X)$ satisfying

$$\sup_{\theta\in\mathcal{S}}|\mathcal{E}(\theta) - \langle\theta, \widehat{m}\rangle| \le \sqrt{\frac{T}{n}} + \sqrt{\frac{2v\log(\delta^{-1})}{n}}.$$

With probability at least $1 - \delta$, such a vector exists and

$$\sup_{\theta \in \mathcal{S}} |\langle \theta, \widehat{m} - \mathbb{E}(X) \rangle| \leq \sup_{\theta \in \mathcal{S}} |\mathcal{E}(\theta) - \langle \theta, \widehat{m} \rangle| + \sqrt{\frac{T}{n}} + \sqrt{\frac{2v \log(\delta^{-1})}{n}}$$

$$\leq 2\sqrt{\frac{T}{n}} + 2\sqrt{\frac{2v \log(\delta^{-1})}{n}}.$$

REMARK 3.1.   In particular in the case when $\mathcal{S} = \mathbb{S}_d$ is the whole unit sphere, we obtain with probability at least $1 - \delta$ the bound

$$\|\widehat{m} - \mathbb{E}(X)\| = \sup_{\theta \in \mathbb{S}_d} \langle \theta, \widehat{m} - \mathbb{E}(X) \rangle \leq 2 \left(\sqrt{\frac{T}{n}} + \sqrt{\frac{2v \log(\delta^{-1})}{n}} \right).$$

By choosing $\widehat{m}$ as the middle of a diameter of the confidence region, we could do a little better and replace the factor 2 in this bound by a factor $\sqrt{3}$.

PROOF.   According to the PAC-Bayesian inequality of Proposition 2.1 on page 2, with probability at least $1 - \delta$, for any $\theta \in \mathcal{S}$,

$$\mathcal{E}(\theta) \leq \frac{1}{\lambda} \int \log \left[\mathbb{E} \left(\exp \psi \left(\lambda \langle \theta', X \rangle \right) \right) \right] d\rho_\theta(\theta') + \frac{\mathcal{K}(\rho_\theta, \rho_0) + \log(\delta^{-1})}{n\lambda}.$$

We can then use the polynomial approximation of $\exp(\psi(t))$ given by Lemma 3.1 on page 4, remarking that $\mathcal{K}(\rho_\theta, \rho_0) = \beta/2$ and that $\log(1 + z) \leq z$, to deduce that

$$\mathcal{E}(\theta) \leq \mathbb{E}(\langle \theta, X \rangle) + \frac{\lambda}{2} \int \mathbb{E}(\langle \theta', X \rangle^2) \, d\rho_\theta(\theta') + \frac{\beta + 2\log(\delta^{-1})}{2n\lambda}$$

$$= \mathbb{E}(\langle \theta, X \rangle) + \frac{\lambda}{2} \left[\mathbb{E}(\langle \theta, X \rangle^2) + \frac{\mathbb{E}(\|X\|^2)}{\beta} \right] + \frac{\beta + 2\log(\delta^{-1})}{2n\lambda}$$

$$\leq \mathbb{E}(\langle \theta, X \rangle) + \frac{\lambda}{2}(v + T/\beta) + \frac{\beta + 2\log(\delta^{-1})}{2n\lambda}$$

$$= \langle \theta, \mathbb{E}(X) \rangle + \sqrt{\frac{T}{n}} + \sqrt{\frac{2v \log(\delta^{-1})}{n}}.$$

We conclude by considering both $\theta \in \mathcal{S}$ and $-\theta \in \mathcal{S}$ to get the reverse inequality, using the assumption that $\mathcal{S}$ is symmetric and remarking that $\mathcal{E}(-\theta) = -\mathcal{E}(\theta)$.
The existence with probability $1 - \delta$ of $\widehat{m}$ satisfying the required inequality is granted by the fact that on the event defined by the above PAC-Bayesian inequality, the expectation $\mathbb{E}(X)$ belongs to the confidence region that, as a result, cannot be empty. $\square$

3.2. *Centered estimate* The bounds in the previous section are simple, but they are stated in terms of uncentered moments of order two where we would have expected a variance. In this section, we explain how to deduce centered bounds from the uncentered bounds of the previous section, through the use of a sample splitting scheme.

Assume that

$$\mathbb{E}(\|X - \mathbb{E}(X)\|^2) \leq \overline{T} < \infty,$$

$$\text{and} \sup_{\theta \in \mathbb{S}_d} \mathbb{E}(\langle \theta, X - \mathbb{E}(X) \rangle^2) \leq \overline{v} \leq \overline{T} < \infty,$$

where $\overline{v}$ and $\overline{T}$ are known constants. Remark that when these bounds hold, the bounds

(2) $$v = \overline{v} + \|\mathbb{E}(X)\|^2 \text{ and } T = \overline{T} + \|\mathbb{E}(X)\|^2$$

hold in the previous section. Assume that we know also some bound b such that

$$\|\mathbb{E}(X)\|^2 \leq b.$$

Split the sample in two parts $(X_1, \ldots, X_k)$ and $(X_{k+1}, \ldots, X_n)$. Use the first part to construct an estimator $\widetilde{m}$ of $\mathbb{E}(X)$ as described in Proposition 3.3 on page 6, choosing $\mathbb{S} = \mathbb{S}_d$. According to this proposition and by equation (2), with probability at least $1 - \delta$,

$$\|\widetilde{m} - \mathbb{E}(X)\| \leq 2 \sqrt{\frac{\overline{T} + b}{k}} + 2 \sqrt{\frac{2(\overline{v} + b) \log(\delta^{-1})}{k}} = \sqrt{\frac{A}{k}},$$

where we have put $A = 4 \left(\sqrt{\overline{T} + b} + \sqrt{2(\overline{v} + b) \log(\delta^{-1})} \right)^2$.

We then construct an estimator $\mathcal{E}(\theta)$ of $\langle \theta, \mathbb{E}(X) \rangle$, $\theta \in \mathbb{S}_d$, built as described in Proposition 3.3, based on the sample $(X_{k+1} - \widetilde{m}, \ldots, X_n - \widetilde{m})$ and on the constants $\overline{T} + A/k$ and $\overline{v} + A/k$. With probability at least $1 - 2\delta$,

$$\sup_{\theta \in \mathbb{S}_d} |\mathcal{E}(\theta) - \langle \theta, \mathbb{E}(X) \rangle| \leq B_{n,k} = \sqrt{\frac{\overline{T} + A/k}{n - k}} + \sqrt{\frac{2(\overline{v} + A/k) \log(\delta^{-1})}{n - k}},$$

and we can, if needed, deduce from $\mathcal{E}(\theta)$ an estimator $\widehat{m}$ such that with probability at least $1 - 2\delta$,

$$\|\widehat{m} - \mathbb{E}(X)\| \leq 2B_{n,k}.$$

If we want the correction term A/k to behave as a second order term when n tends to ∞, we can for example take $k = \sqrt{n}$, in which case $n - k$ is equivalent to n at infinity, so that $B_{n,\sqrt{n}}$ is equivalent to

$$\sqrt{\frac{\overline{T}}{n}} + \sqrt{\frac{2\,\overline{v}\log(\delta^{-1})}{n}}.$$

Let us also mention that a simpler estimator, obtained by shrinking the norm of X_i, is also possible. It comes with a sub-Gaussian deviation bound under the slightly stronger hypothesis that $\mathbb{E}(\|X\|^p) < \infty$ for some (non necessarily integer) exponent $p > 2$, and is described in Catoni and Giulini (2017).

4. Mean matrix estimate Let $M \in \mathbb{R}^{p \times q}$ be a random matrix and let $M_1, \ldots, M_n$ be n independent copies of M. In this section, we will provide an estimator for $\mathbb{E}(M)$.

From the previous section, we already have an estimator $\widehat{m}$ of $\mathbb{E}(M)$ with a bounded Hilbert-Schmidt norm $\|\widehat{m} - \mathbb{E}(M)\|_{\text{HS}}$, since from the point of view of the Hilbert-Schmidt norm, M is nothing but a random vector of size pq. Here, we will be interested in another natural norm, the operator norm

$$\|M\|_\infty = \sup_{\theta \in \mathbb{S}_q} \|M\theta\|.$$

Indeed, recalling that

$$\|M\|_\infty = \sup_{\theta \in \mathbb{S}_q, \xi \in \mathbb{S}_p} \langle \xi, M\theta \rangle = \sup_{\xi \in \mathbb{S}_p} \|M^\top \xi\| = \sup_{\theta \in \mathbb{S}_q, \xi \in \mathbb{S}_p} \mathbf{Tr}(\theta \xi^\top M),$$

we see that we can deduce results from the previous section on vectors, considering the scalar product between matrices

$$\langle M, N \rangle = \mathbf{Tr}(M^\top N), \quad M, N \in \mathbb{R}^{p \times q},$$

and the part of the unit sphere defined as

$$\mathbb{S} = \{\xi \theta^\top \ : \ \xi \in \mathbb{S}_p, \theta \in \mathbb{S}_q\}.$$

Doing so, we obtain in the uncentered case a bound of the form

$$\|\widehat{m} - \mathbb{E}(M)\| \le 2\sqrt{\frac{\mathbb{E}(\|M\|_{\text{HS}}^2)}{n}} + 2\sqrt{\frac{2}{n}\sup_{\xi \in \mathbb{S}_p, \theta \in \mathbb{S}_q} \mathbb{E}(\langle \xi, M\theta \rangle^2) \log(\delta^{-1})}.$$

We will show in the next section that the second δ-dependent term is satisfactory whereas the first δ-independent term can be improved.

9

4.1. *Estimation without centering*   Consider the influence function ψ defined by equation (1) on page 4.

For any $\xi \in \mathbb{R}^p$, let $\nu_\xi = \mathcal{N}(\xi, \beta^{-1} I_p)$, where I_p is the identity matrix of size $p \times p$. In the same way, let $\rho_\theta = \mathcal{N}(\theta, \gamma^{-1} I_q)$, $\theta \in \mathbb{R}^q$. Consider the estimator of $\langle \xi, \mathbb{E}(M) \theta \rangle$ defined as

$$\mathcal{E}(\xi, \theta) = \frac{1}{\lambda n} \sum_{i=1}^{n} \int \psi\left(\lambda \langle \xi', M_i \theta' \rangle\right) \mathrm{d}\nu_\xi(\xi') \, \mathrm{d}\rho_\theta(\theta'), \qquad \xi \in \mathbb{R}^p, \theta \in \mathbb{R}^q.$$

PROPOSITION 4.1.   *For any parameters $\delta \in]0, 1[$, $\lambda, \beta, \gamma \in]0, \infty[$, with probability at least $1 - \delta$, for any $\xi \in \mathbb{R}^p$ and any $\theta \in \mathbb{R}^q$,*

$$|\mathcal{E}(\xi, \theta) - \mathbb{E}(\langle \xi, M\theta \rangle)|$$
$$\leq \frac{\lambda}{2} \left[\mathbb{E}(\langle \xi, M\theta \rangle^2) + \frac{\mathbb{E}(\|M\theta\|^2)}{\beta} + \frac{\mathbb{E}(\|M^\top \xi\|^2)}{\gamma} + \frac{\mathbb{E}(\|M\|_{\mathrm{HS}}^2)}{\beta \gamma} \right.$$
$$\left. + \frac{\beta + \gamma + 2 \log(\delta^{-1})}{2n\lambda}. \right.$$

PROOF.   The PAC-Bayesian inequality of Proposition 2.1 on page 2 tells us that with probability at least $1 - \delta$, for any $\xi \in \mathbb{R}^p$ and any $\theta \in \mathbb{R}^q$,

$$\mathcal{E}(\xi, \theta) \leq \lambda^{-1} \int \log\left\{ \mathbb{E}\left[\exp\left(\psi(\lambda \langle \xi', M\theta' \rangle) \right) \right] \right\} \mathrm{d}\nu_\xi(\xi') \mathrm{d}\rho_\theta(\theta')$$
$$+ \frac{\mathcal{K}(\nu_\xi, \nu_0)}{n\lambda} + \frac{\mathcal{K}(\rho_\theta, \rho_0)}{n\lambda} + \frac{\log(\delta^{-1})}{n\lambda},$$

Using the properties of ψ (Lemma 3.1 on page 4) and Fubini's lemma, we get

$$\mathcal{E}(\xi, \theta) \leq \langle \xi, \mathbb{E}(M)\theta \rangle + \frac{\lambda}{2} \mathbb{E}\left(\int \langle \xi', M\theta' \rangle^2 \, \mathrm{d}\nu_\xi(\xi') \mathrm{d}\rho_\theta(\theta') \right) + \frac{\beta + \gamma + 2 \log(\delta^{-1})}{2n\lambda}.$$

As

$$\int \langle \xi', M\theta' \rangle^2 \, \mathrm{d}\nu_\xi(\xi') \mathrm{d}\rho_\theta(\theta') = \langle \xi, M\theta \rangle^2 + \frac{\|M\theta\|^2}{\beta} + \frac{\|M^\top \xi\|^2}{\gamma} + \frac{\|M\|_{\mathrm{HS}}^2}{\beta \gamma},$$

this concludes the proof.   □

Let us now discuss the question of computing $\mathcal{E}(\xi, \theta)$. Remark that, according to Lemma 3.2 on page 4, for any $x \in \mathbb{R}^p$,

$$\int \psi(\langle \xi', x \rangle) \, \mathrm{d}\nu_\xi(\xi') = \varphi(\langle \xi, x \rangle, \beta^{-1/2} \|x\|)$$

10

$$= \langle \xi, x \rangle - \frac{\langle \xi, x \rangle \|x\|^2}{2\beta} - \frac{\langle \xi, x \rangle^3}{6} + r\left(\langle \xi, x \rangle, \beta^{-1/2} \|x\|\right).$$

It is also easy to check that

$$\int \|M_i \theta'\|^2 \, \mathrm{d}\rho_\theta(\theta') = \|M_i \theta\|^2 + \frac{\|M_i\|_{\mathrm{HS}}^2}{\gamma},$$

$$\int \langle \xi, M_i \theta' \rangle \|M_i \theta'\|^2 \, \mathrm{d}\rho_\theta(\theta') = \langle \xi, M_i \theta \rangle \|M_i \theta\|^2 + \frac{1}{\gamma} \langle \xi, M_i \theta \rangle \|M_i\|_{\mathrm{HS}}^2$$
$$+ \frac{2}{\gamma} \langle \xi, M_i M_i^\top M_i \theta \rangle,$$

and $\displaystyle \int \langle \xi, M_i \theta' \rangle^3 \, \mathrm{d}\rho_\theta(\theta') = \langle \xi, M_i \theta \rangle^3 + \frac{3}{\gamma} \langle \xi, M_i \theta \rangle \|M_i^\top \xi\|^2.$

Consider a standard random vector $W_q \sim \mathcal{N}(0, I_q)$. We obtain that

$$\int \psi\left(\lambda \langle \xi', M_i \theta' \rangle\right) \, \mathrm{d}\nu_\xi(\xi') \mathrm{d}\rho_\theta(\theta') =$$
$$\int \left(\lambda \langle \xi, M_i \theta' \rangle - \frac{\lambda^3 \langle \xi, M_i \theta' \rangle \|M_i \theta'\|^2}{2\beta} - \frac{\lambda^3}{6} \langle \xi, M_i \theta' \rangle^3 \right.$$
$$\left. + r\left(\lambda \langle \xi, M_i \theta' \rangle, \lambda \beta^{-1/2} \|M_i \theta'\|\right) \right) \mathrm{d}\rho_\theta(\theta'),$$

so that

$$\mathcal{E}(\xi, \theta) = \frac{1}{n} \sum_{i=1}^{n} \langle \xi, M_i \theta \rangle - \frac{\lambda^2}{6} \langle \xi, M_i \theta \rangle^3$$
$$- \frac{\lambda^2}{2\beta} \langle \xi, M_i \theta \rangle \|M_i \theta\|^2 - \frac{\lambda^2}{2\gamma} \langle \xi, M_i \theta \rangle \|M_i^\top \xi\|^2$$
$$- \frac{\lambda^2}{2\beta\gamma} \langle \xi, M_i \theta \rangle \|M_i\|_{\mathrm{HS}}^2 - \frac{\lambda^2}{\beta\gamma} \langle \xi, M_i M_i^\top M_i \theta \rangle$$
$$+ \frac{1}{\lambda} \mathbb{E}\left[r\left(\lambda \langle M_i^\top \xi, \theta + \gamma^{-1/2} W_q \rangle, \lambda \beta^{-1/2} \|M_i(\theta + \gamma^{-1/2} W_q)\| \right) \right].$$

The last term is not explicit, since it contains an expectation, but should be most of the time a small reminder and can be evaluated using a Monte-Carlo numerical scheme. This gives a more explicit and efficient method than evaluating directly $\mathcal{E}(\xi, \theta)$ using a Monte-Carlo simulation for the couple of random variables $(\xi', \theta') \sim \nu_\xi \otimes \rho_\theta$.

PROPOSITION 4.2. *Assume that the following finite bounds are known*

$$v \geq \sup_{\xi \in \mathbb{S}_p, \theta \in \mathbb{S}_q} \mathbb{E}(\langle \xi, M\theta \rangle^2) = \sup_{\xi \in \mathbb{S}_p, \theta \in \mathbb{S}_q} (\xi^\top \otimes \xi^\top) \mathbb{E}(M \otimes M)(\theta \otimes \theta),$$

$$t \geq \sup_{\theta \in \mathbb{S}_q} \mathbb{E}(\|M\theta\|^2) = \sup_{\theta \in \mathbb{S}_q} \langle \theta, \mathbb{E}(M^\top M)\theta \rangle = \|\mathbb{E}(M^\top M)\|_\infty,$$

$$u \geq \sup_{\xi \in \mathbb{S}_p} \mathbb{E}(\|M^\top \xi\|^2) = \sup_{\xi \in \mathbb{S}_p} \langle \xi, \mathbb{E}(MM^\top)\xi \rangle = \|\mathbb{E}(M^\top M)\|_\infty,$$

$$T \geq \mathbb{E}(\|M\|_{\mathrm{HS}}^2),$$

and choose

$$\lambda = \sqrt{\frac{\beta + \gamma + 2\log(\delta^{-1})}{n(v + t/\beta + u/\gamma + T/(\beta\gamma))}}.$$

For any values of $\delta \in {]0,1[}$, $\beta, \gamma \in {]0,\infty[}$, with probability at least $1 - \delta$, for any $\xi \in \mathbb{S}_p$, any $\theta \in \mathbb{S}_q$,

$$|\mathcal{E}(\xi,\theta) - \langle \xi, \mathbb{E}(M)\theta \rangle| \leq B_n = \sqrt{\left(v + \frac{t}{\beta} + \frac{u}{\gamma} + \frac{T}{\beta\gamma}\right) \frac{\beta + \gamma + 2\log(\delta^{-1})}{n}}.$$

Consider now any estimator $\widehat{m}$ of $\mathbb{E}(M)$. With probability at least $1 - \delta$,

$$\|\widehat{m} - \mathbb{E}(M)\|_\infty \leq \sup_{\xi \in \mathbb{S}_p, \theta \in \mathbb{S}_q} |\mathcal{E}(\xi,\theta) - \langle \xi, \widehat{m}\,\theta \rangle| + B_n.$$

In particular, if we choose $\widehat{m}$ such that,

$$\sup_{\xi \in \mathbb{S}_p, \theta \in \mathbb{S}_q} |\mathcal{E}(\xi,\theta) - \langle \xi, \widehat{m}\,\theta \rangle| \leq B_n,$$

with probability at least $1 - \delta$, this choice is possible and

$$\|\widehat{m} - \mathbb{E}(M)\|_\infty \leq 2B_n.$$

REMARK 4.1. In particular, choosing $\beta = \gamma = 2\max\left\{\dfrac{(t+u)}{v}, \sqrt{\dfrac{T}{v}}\right\}$, we get

$$B_n \leq \sqrt{\frac{2v}{n}\left(2\log(\delta^{-1}) + 4\max\left\{\frac{t+u}{v}, \sqrt{\frac{T}{v}}\right\}\right)}.$$

The bound B_n is of the type $\sqrt{\dfrac{2v\,[\mathcal{C} + \log(\delta^{-1})]}{n}}$, with a complexity (or dimension) term $\mathcal{C}$ equal to

$$\mathcal{C} = 4\max\left\{\frac{t+u}{v}, \sqrt{\frac{T}{v}}\right\} + \log(\delta^{-1}).$$

12

REMARK 4.2. Let us envision a simple case to compare the precision of the bounds in a setting where dimension-free and dimension-dependent bounds coincide. Assume more specifically that the entries of the matrix M,

$$M_{i,j} \quad 1 \leq i \leq p, 1 \leq j \leq q,$$

are centered and i.i.d.. Assume that $\sigma = \sqrt{\mathbb{E}(M_{i,j}^2)}$ is known, and take

$$v = \sup_{\xi \in \mathbb{S}_p, \theta \in \mathbb{S}_q} \mathbb{E}(\langle \xi, M\theta \rangle^2) = \sigma^2,$$

$$t = \sup_{\theta \in \mathbb{S}_q} \mathbb{E}(\|M\theta\|^2) = p\sigma^2,$$

$$u = \sup_{\xi \in \mathbb{S}_p} \mathbb{E}(\|M^\top \xi\|^2) = q\sigma^2,$$

$$T = \mathbb{E}(\|M\|_{\text{HS}}^2) = pq\sigma^2.$$

Choosing $\beta = \gamma = 2(p + q)$, we get a complexity term equal to

$$\mathcal{C} = 4(p + q) + \log(\delta^{-1}),$$

whereas the bound of the previous section made for vectors has a complexity factor equal to pq.

4.2. *Controlling both the operator norm error and the Hilbert-Schmidt error*
There are situations where it is desirable to control both $\|\widehat{m} - \mathbb{E}(M)\|_\infty$ and $\|\widehat{m} - \mathbb{E}(M)\|_{\text{HS}}$. To do so we can very easily combine Propositions 3.3 on page 6 and Proposition 4.2 on page 11, since these two propositions are based on the construction of confidence regions.

More precisely, first consider $M \in \mathbb{R}^{p \times q}$ as a vector and use the scalar product

$$\langle \theta, M \rangle_{\text{HS}} = \mathbf{Tr}(\theta^\top M), \quad \theta \in \mathbb{R}^{p \times q}.$$

Applying Proposition 3.3 on page 6, we can build an estimator $\mathcal{E}_{\text{HS}}(\theta)$ such that with probability at least $1 - \delta$,

$$\sup_{\theta \in \mathbb{R}^{p \times q}, \|\theta\|_{\text{HS}} = 1} |\mathcal{E}_{\text{HS}}(\theta) - \mathbf{Tr}(\theta^\top \mathbb{E}(M))| \leq A_n = \sqrt{\frac{T}{n}} + \sqrt{\frac{2v \log(\delta^{-1})}{n}}.$$

On the other hand, we can also apply Proposition 4.2 on page 11 and build an estimator $\mathcal{E}(\xi, \theta), \xi \in \mathbb{S}_p, \theta \in \mathbb{S}_q$, such that with probability at least $1 - \delta$,

$$\sup_{\xi \in \mathbb{S}_p, \theta \in \mathbb{S}_q} |\mathcal{E}(\xi, \theta) - \langle \xi, \mathbb{E}(M)\theta \rangle| \leq B_n = \sqrt{\frac{2v}{n}\left(2\log(\delta^{-1}) + 4\max\left\{\frac{t+u}{v}, \sqrt{\frac{T}{v}}\right\}\right)}.$$

13

PROPOSITION 4.3. *Consider a matrix $\widehat{m}$ such that*

$$\sup_{\theta \in \mathbb{R}^{p \times q}, \|\theta\|_{\mathrm{HS}}=1} |\mathcal{E}_{\mathrm{HS}}(\theta) - \mathbf{Tr}(\theta^\top \widehat{m})| \leq A_n$$

$$and \quad \sup_{\xi \in \mathbb{S}_p, \theta \in \mathbb{S}_q} |\mathcal{E}(\xi, \theta) - \langle \xi, \widehat{m}\, \theta \rangle| \leq B_n.$$

Combining Propositions 3.3 and 4.2 shows that, with probability at least $1 - 2\delta$, such a matrix $\widehat{m}$ exists and satisfies both

$$\|\widehat{m} - \mathbb{E}(M)\|_{\mathrm{HS}} \leq 2A_n \quad and \quad \|\widehat{m} - \mathbb{E}(M)\|_\infty \leq 2B_n.$$

Remark that B_n is typically smaller than A_n as expected in interesting large dimension situations.

4.3. *Centered estimator* As already done in the case of the estimation of the mean of a random vector, we deduce in this section centered bounds from the uncentered bounds of the previous sections, using sample splitting.

Put $m = \mathbb{E}(M)$ and $\overline{M} = M - m$. Assume that we know finite constants $\overline{v}, \overline{t}, \overline{u}, \overline{T}$ such that

$$\sup_{\xi \in \mathbb{S}_p, \theta \in \mathbb{S}_q} \mathbb{E}(\langle \xi, \overline{M}\theta \rangle^2) \leq \overline{v} < \infty,$$

$$\sup_{\theta \in \mathbb{S}_q} \mathbb{E}(\|\overline{M}\theta\|^2) \leq \overline{t} < \infty,$$

$$\sup_{\xi \in \mathbb{S}_p} \mathbb{E}(\|\overline{M}^\top \xi\|^2) \leq \overline{u} < \infty,$$

$$\mathbb{E}(\|\overline{M}\|_{\mathrm{HS}}^2) \leq \overline{T} < \infty.$$

When this is true, we can take for the previous uncentered constants

$$v = \overline{v} + \|m\|_\infty^2, \quad t = \overline{t} + \|m\|_\infty^2, \quad u = \overline{u} + \|m\|_\infty^2, \quad T = \overline{T} + \|m\|_{\mathrm{HS}}^2.$$

In view of this, it is suitable to assume that we also know some finite constants b and c such that

$$\|m\|_\infty^2 \leq b \quad and \quad \|m\|_{\mathrm{HS}}^2 \leq c.$$

As we see that the Hilbert-Schmidt norm $\|m\|_{\mathrm{HS}}$ comes into play, we will use the combined preliminary estimate provided by Proposition 4.3.

Given an i.i.d. matrix sample $(M_1, \ldots, M_n)$, first use $(M_1, \ldots, M_k)$ to build a preliminary estimator $\widetilde{m}$ as described in Proposition 4.3. With probability at least $1 - \delta/2$,

$$\|\widetilde{m} - m\|_{\mathrm{HS}} \leq \sqrt{\frac{A}{k}} \quad and \quad \|\widetilde{m} - m\|_\infty \leq \sqrt{\frac{B}{k}},$$

14

$$\text{where } A = 4\left(\sqrt{2(\bar{v} + b) \log(4/\delta)} + \sqrt{\overline{T} + c} \right)^2$$

$$\text{and } B = 8(\bar{v} + b)\left(2 \log(4/\delta) + 4 \max\left\{ \frac{\bar{t} + \bar{u} + 2b}{\bar{v} + b}, \left(\frac{\overline{T} + c}{\bar{v} + b}\right)^{1/2} \right\} \right).$$

Then use the sample $(M_{k+1} - \widetilde{m}, \ldots, M_n - \widetilde{m})$ to build an estimator $\mathcal{E}(\xi, \theta)$, $\xi \in \mathbb{S}_p$, $\theta \in \mathbb{S}_q$, based on the construction described in Proposition 4.2 on page 11, at confidence level $1 - \delta/2$. It is such that with probability at least $1 - \delta$,

$$|\mathcal{E}(\xi, \theta) - \langle \xi, m\, \theta \rangle| \leq C_{n,k}$$

$$= \sqrt{\frac{2(\bar{v} + B/k)}{n - k}\left(2 \log(2/\delta) + 4 \max\left\{ \frac{\bar{t} + \bar{u} + 2B/k}{\bar{v} + B/k}, \left(\frac{\overline{T} + A/k}{\bar{v} + B/k}\right)^{1/2} \right\} \right)}.$$

If we choose for instance $k = \sqrt{n}$, we obtain that

$$C_{n, \sqrt{n}} \underset{n \to \infty}{\sim} \sqrt{\frac{2\bar{v}}{n}\left(2 \log(2/\delta) + 4 \max\left\{ \frac{\bar{t} + \bar{u}}{\bar{v}}, \left(\frac{\overline{T}}{\bar{v}}\right)^{1/2} \right\} \right)}.$$

5. Adaptive estimators The results presented in the previous sections assume that there exist known upper bounds for some quantities as $\mathbb{E}(\|X\|^2)$ in the case of a mean vector estimate or $\mathbb{E}(\|M\|_{\text{HS}}^2)$ in the matrix case. Here we would like to adapt to these quantities, in the case when those bounds are not known.

To do so, we will use an asymmetric influence function $\psi : \mathbb{R}_+ \longrightarrow \mathbb{R}_+$ defined on the positive real line only as

(3)
$$\psi(t) = \begin{cases} t - t^2/2, & 0 \leq t \leq 1, \\ 1/2, & 1 \leq t. \end{cases}$$

LEMMA 5.1. *For any $t \in \mathbb{R}_+$,*

$$-\log(1 - t + t^2) \leq \psi(t) \leq \log(1 + t).$$

PROOF. Let us put $f(t) = -\log(1 - t + t^2)$ and $g(t) = \log(1 + t)$. Remark that $f(0) = g(0) = \psi(0) = 0$. Remark also that for any $t \in [0, 1]$,

$$f'(t) = \frac{1 - 2t}{1 - t + t^2}, \quad \psi'(t) - f'(t) = \frac{t^2(2 - t)}{1 - t + t^2} \leq 0,$$

$$\text{and} \quad g'(t) = 1 - \frac{t}{1 + t} \geq \psi'(t) = 1 - t.$$

As on the interval $[1, \infty[$, f is decreasing, g is increasing and ψ is constant, this proves the lemma. $\square$

15

Similarly to the previous case, considering a standard Gaussian real valued random dom variable $W \sim \mathcal{N}(0,1)$, we can introduce the function

$$\varphi(m,\sigma) = \mathbb{E}\{\psi[(m+\sigma W)_+]\},$$

where $t_+ = \max\{t, 0\}$, and explicitly compute φ as

$$\varphi(m,\sigma) = \left(m - \frac{m^2}{2} - \frac{\sigma^2}{2}\right)\left[1 - F\left(\frac{-1+m}{\sigma}\right) - F\left(-\frac{m}{\sigma}\right)\right] + \frac{1}{2}F\left(\frac{-1+m}{\sigma}\right)$$
$$+ \frac{\sigma(1-m/2)}{\sqrt{2\pi}}\exp\left(-\frac{m^2}{2\sigma^2}\right) - \frac{\sigma(1-m)}{2\sqrt{2\pi}}\exp\left(-\frac{(1-m)^2}{2\sigma^2}\right),$$

using the expression

$$\psi(t_+) = \left(t - t^2/2\right)\left[\mathbb{1}(t \le 1) - \mathbb{1}(t \le 0)\right] + \frac{1}{2}\left[1 - \mathbb{1}(t \le 1)\right], \qquad t \in \mathbb{R}.$$

5.1. *Estimation of the mean of a random vector*   Consider a discrete set Λ of values of λ and a probability measure μ on Λ, to be chosen more precisely later on. Let β be some positive parameter that we will also choose later and put as previously $\rho_\theta = \mathcal{N}(\theta, \beta^{-1}I_d)$. Define for any $\theta \in \mathbb{S}_d$

$$\mathcal{E}_+(\theta) = \sup_{\lambda \in \Lambda} \frac{1}{n\lambda} \sum_{i=1}^{n} \int \psi\left(\lambda\langle\theta', X_i\rangle_+\right) d\rho_\theta(\theta') - \frac{\beta + 2\log(\delta^{-1}\mu(\lambda)^{-1})}{2\lambda n},$$

$$\mathcal{E}_-(\theta) = \sup_{\lambda \in \Lambda} \frac{1}{n\lambda} \sum_{i=1}^{n} \int \psi\left(\lambda\langle\theta', X_i\rangle_-\right) d\rho_\theta(\theta') - \frac{\beta + 2\log(\delta^{-1}\mu(\lambda)^{-1})}{2\lambda n},$$

and $\mathcal{E}(\theta) = \mathcal{E}_+(\theta) - \mathcal{E}_-(\theta)$.

Thoughtful readers may wonder why we introduce λ in this way and do not use instead $\rho_{\lambda\theta}$, to get a uniform result in $\lambda\theta$ in one shot, without introducing the discrete set Λ. It is because this option would produce the entropy factor $\dfrac{\lambda\beta}{2n}$ instead of $\dfrac{\beta}{2n\lambda}$, requiring a value of β depending on unknown moments of the distribution of X.

According to the PAC-Bayesian inequality of Proposition 2.1 on page 2, with probability at least $1 - 2\delta$,

$$\int \mathbb{E}(\langle\theta', X\rangle_+) d\rho_\theta(\theta')$$

$$- \inf_{\lambda \in \Lambda}\left\{\lambda \int \mathbb{E}(\langle\theta', X\rangle_+^2) d\rho_\theta(\theta') + \frac{\beta + 2\log(\delta^{-1}\mu(\lambda)^{-1})}{\lambda n}\right\}$$

16

$$\leq \mathcal{E}_+(\theta) \leq \int \mathbb{E}(\langle \theta', X \rangle_+) \, \mathrm{d}\rho_\theta(\theta').$$

More precisely, to obtain the above inequalities, we have used a union bound with respect to $\lambda \in \Lambda$, starting from the fact that, when we replace the infimum in λ in the previous equation with a fixed value of $\lambda \in \Lambda$, it holds with probability at least $1 - 2\mu(\lambda)\delta$. Since

$$\int f(-\theta') \, \mathrm{d}\rho_\theta(\theta') = \int f(\theta') \, \mathrm{d}\rho_{-\theta}(\theta'),$$

this implies also that

$$\int \mathbb{E}(\langle \theta', X \rangle_-) \, \mathrm{d}\rho_\theta(\theta')$$

$$- \inf_{\lambda \in \Lambda} \left\{ \lambda \int \mathbb{E}(\langle \theta', X \rangle_-^2) \, \mathrm{d}\rho_\theta(\theta') + \frac{\beta + 2\log(\delta^{-1}\mu(\lambda)^{-1})}{\lambda n} \right\}$$

$$\leq \mathcal{E}_-(\theta) \leq \int \mathbb{E}(\langle \theta', X \rangle_-) \, \mathrm{d}\rho_\theta(\theta').$$

Therefore, with probability at least $1 - 2\delta$,

$$- B_-(\theta) = - \inf_{\lambda \in \Lambda} \left\{ \lambda \int \mathbb{E}(\langle \theta', X \rangle_+^2) \, \mathrm{d}\rho_\theta(\theta') + \frac{\beta + 2\log(\delta^{-1}\mu(\lambda)^{-1})}{\lambda n} \right\}.$$

$$\leq \mathcal{E}(\theta) - \langle \theta, \mathbb{E}(X) \rangle \leq B_+(\theta) = \inf_{\lambda \in \Lambda} \left\{ \lambda \int \mathbb{E}(\langle \theta', X \rangle_-^2) \, \mathrm{d}\rho_\theta(\theta') \right.$$

$$\left. + \frac{\beta + 2\log(\delta^{-1}\mu(\lambda)^{-1})}{\lambda n} \right\}.$$

This defines for $\langle \theta, \mathbb{E}(X) \rangle$ a confidence interval of length no greater than

$$B(\theta) = \inf_{\lambda \in \Lambda} \left\{ \lambda \int \mathbb{E}(\langle \theta', X \rangle^2) \, \mathrm{d}\rho_\theta(\theta') + \frac{2\beta + 4\log(\delta^{-1}\mu(\lambda)^{-1})}{\lambda n} \right\}.$$

Unfortunately, neither $B_+(\theta)$, $B_-(\theta)$ nor $B(\theta)$ are observable. But, nevertheless, we can build an estimator $\widehat{m}$ such that

$$\sup_{\theta \in \mathbb{S}_d} \{ \langle \theta, \widehat{m} \rangle - \mathcal{E}(\theta) \} = \inf_{m \in \mathbb{R}^d} \sup_{\theta \in \mathbb{S}_d} \{ \langle \theta, m \rangle - \mathcal{E}(\theta) \}.$$

It satisfies with probability at least $1 - 2\delta$,

$$\|\widehat{m} - \mathbb{E}(X)\| = \sup_{\theta \in \mathbb{S}_d} \langle \theta, \widehat{m} - \mathbb{E}(X) \rangle$$

17

$$\leq \sup_{\theta \in \mathbb{S}_d} \left\{ \langle \theta, \widehat{m} \rangle - \mathcal{E}(\theta) \right\} + \sup_{\theta \in \mathbb{S}_d} \left\{ \mathcal{E}(\theta) - \langle \theta, \mathbb{E}(X) \rangle \right\} \leq 2 \sup_{\theta \in \mathbb{S}_d} \left\{ \mathcal{E}(\theta) - \langle \theta, \mathbb{E}(X) \rangle \right\}$$

$$\leq 2 \sup_{\theta} B_+(\theta) \leq \sup_{\theta \in \mathbb{S}_d} \inf_{\lambda \in \Lambda} 2\lambda \int \mathbb{E}(\langle \theta', X \rangle^2) \, \mathrm{d}\rho_\theta(\theta') + \frac{2\beta + 4\log(\delta^{-1}\mu(\lambda)^{-1})}{\lambda n}$$

$$= \sup_{\theta \in \mathbb{S}_d} \inf_{\lambda \in \Lambda} 2\lambda \left(\mathbb{E}(\langle \theta, X \rangle^2) + \frac{\mathbb{E}(\|X\|^2)}{\beta} \right) + \frac{2\beta + 4\log(\delta^{-1}\mu(\lambda)^{-1})}{\lambda n}.$$

LEMMA 5.2. *Let us choose $\beta = 2\log(\delta^{-1})$ and put $v = \sup_{\theta \in \mathbb{S}_d} \mathbb{E}(\langle \theta, X \rangle^2)$ and $T = \mathbb{E}(\|X\|^2)$. With probability at least $1 - 2\delta$,*

$$\|\widehat{m} - \mathbb{E}(X)\| \leq \inf_{\lambda \in \Lambda} B(\lambda),$$

where $B(\lambda) = \left\{ 2\lambda \left(v + \frac{T}{2\log(\delta^{-1})} \right) + \frac{8\log(\delta^{-1}) + 4\log(\mu(\lambda)^{-1})}{\lambda n} \right\}.$

To turn this lemma into an explicit bound, we need now to choose Λ and $\mu \in \mathcal{M}_+^1(\Lambda)$. Consider for some real parameters $\sigma > 0$ and $\alpha > 1$,

$$\Lambda = \left\{ \frac{\alpha^k}{\sigma \sqrt{n}} \; : \; k \in \mathbb{Z} \right\}.$$

For any $\lambda_k = \dfrac{\alpha^k}{\sigma \sqrt{n}} \in \Lambda$, put

$$\mu(\lambda) = \begin{cases} \dfrac{1}{2(|k| + 1)(|k| + 2)}, & k \neq 0, \\ 1/2, & k = 0, \end{cases}$$

and remark that

$$\mu(\lambda_k) \geq \frac{1}{2(|k| + 2)^2}, \qquad k \in \mathbb{Z}.$$

Put also

$$\lambda_* = \sqrt{\frac{4\log(\delta^{-1})}{n\left(v + \dfrac{T}{2\log(\delta^{-1})} \right)}}.$$

The bound $B(\lambda)$ appearing in the previous lemma can be written as

$$B(\lambda) = 4\sqrt{\frac{2}{n} \left(2v \log(\delta^{-1}) + T \right)} \left[\cosh\left(\log\left(\frac{\lambda}{\lambda_*} \right) \right) + \frac{\log(\mu(\lambda)^{-1})}{4\log(\delta^{-1})} \frac{\lambda_*}{\lambda} \right].$$

Since $\log(\lambda_k) = k\log(\alpha) - \log(\sigma) - \log(n)/2$, there exists $k_* \in \mathbb{Z}$ such that

$$|\log(\lambda_{k_*}/\lambda_*)| \leq \log(\alpha)/2,$$

18

so that

$$|k_*| \leq |\log(\sigma \lambda_* \sqrt{n})| / \log(\alpha) + 1/2.$$

Therefore

$$\inf_{\lambda \in \Lambda} B(\lambda) \leq B(\lambda_{k_*}) \leq 4C \sqrt{\frac{2}{n}\left(2v \log(\delta^{-1}) + T\right)}$$

where the constant C is equal to

$$C = \cosh\left(\frac{\log(\alpha)}{2}\right) + \frac{\sqrt{\alpha}}{2\log(\delta^{-1})} \log\left[\frac{1}{\sqrt{2}\log(\alpha)}\left|\log\left(\frac{2v\log(\delta^{-1}) + T}{8\sigma^2 \log(\delta^{-1})^2}\right)\right| + \frac{5}{\sqrt{2}}\right].$$

We see that the constant σ^2 can be interpreted as our best guess of the ratio

$$\frac{2v\log(\delta^{-1}) + T}{8\log(\delta^{-1})^2}.$$

However, this guess may be very loose without harming the constant C too much. Indeed, to give an example, if we choose $\alpha = e$ and we assume that we made an error of magnitude 10^6 on the choice of σ^2, compared to the optimal guess, we get

$$C \leq \cosh(1/2) + \frac{\exp(1/2)}{2\log(\delta^{-1})} \log\left[\frac{1}{\sqrt{2}}\log(10^6) + \frac{5}{\sqrt{2}}\right] \leq 1.13 + \frac{2.2}{\log(\delta^{-1})},$$

so that if we work at the confidence level corresponding to $\delta = 1/100$, we obtain that $C \leq 1.6$. In brief, the message is that C is typically between one and two.

5.2. *Adaptive estimation of the mean of a random matrix*   We consider here the same framework as in Section 3 on page 3. Let $M \in \mathbb{R}^{p \times q}$ be a random matrix and $(M_1, \ldots, M_n)$ be a sample made of n independent copies of M.

Using the asymmetric influence function φ defined by equation (3) on page 15, given $\xi \in \mathbb{S}_p, \theta \in \mathbb{S}_q$, we define the estimators

$$\mathcal{E}_+(\xi, \theta) = \sup_{\lambda \in \Lambda} \left\{ \frac{1}{\lambda n} \sum_{i=1}^{n} \int \psi\left[\lambda \langle \xi', M_i \theta' \rangle_+\right] d\nu_\xi(\xi') \, d\rho_\theta(\theta') \right.$$

$$\left. - \frac{\beta + \gamma + 2\log(\delta^{-1}\mu(\lambda)^{-1})}{2\lambda n} \right\},$$

$$\mathcal{E}_-(\xi, \theta) = -\mathcal{E}_+(-\xi, \theta) = \sup_{\lambda \in \Lambda} \left\{ \frac{1}{\lambda n} \sum_{i=1}^{n} \int \psi\left[\lambda \langle \xi', M_i \theta' \rangle_-\right] d\nu_\xi(\xi') \, d\rho_\theta(\theta') \right.$$

$$\left. - \frac{\beta + \gamma + 2\log(\delta^{-1}\mu(\lambda)^{-1})}{2\lambda n} \right\},$$

and $\mathcal{E}(\xi, \theta) = \mathcal{E}_+(\xi, \theta) - \mathcal{E}_-(\xi, \theta)$.

19

LEMMA 5.3.   *With probability at least $1 - 2\delta$, for any $\xi \in \mathbb{S}_p, \theta \in \mathbb{S}_q$,*

$$- \inf_{\lambda \in \Lambda} \left\{ \lambda \int \mathbb{E}(\langle \xi', M\theta' \rangle_+^2) \, \mathrm{d}\nu_\xi(\xi') \mathrm{d}\rho_\theta(\theta') + \frac{\beta + \gamma + 2\log(\delta^{-1}\mu(\lambda)^{-1})}{n\lambda} \right\}$$

$$\leq \mathcal{E}_+(\xi, \theta) - \int \mathbb{E}(\langle \xi', M\theta' \rangle_+) \, \mathrm{d}\nu_\xi(\xi') \mathrm{d}\rho_\theta(\theta') \leq 0,$$

so that

$$- B_+(-\xi, \theta) \leq \mathcal{E}(\xi, \theta) - \langle \xi, \mathbb{E}(M)\, \theta \rangle \leq B_+(\xi, \theta)$$

$$= \inf_{\lambda \in \Lambda} \left\{ \lambda \int \mathbb{E}(\langle \xi', M\theta' \rangle_-^2) \, \mathrm{d}\nu_\xi(\xi') \mathrm{d}\rho_\theta(\theta') + \frac{\beta + \gamma + 2\log(\delta^{-1}\mu(\lambda)^{-1})}{\lambda n} \right\}$$

$$\leq \inf_{\lambda \in \Lambda} \left\{ \lambda \left[\mathbb{E}(\langle \xi, M\theta \rangle^2) + \frac{\mathbb{E}(\|M\theta\|^2)}{\beta} + \frac{\mathbb{E}(\|M^\top \xi\|^2)}{\gamma} + \frac{\mathbb{E}(\|M\|_{\mathrm{HS}}^2)}{\beta\gamma} \right] \right.$$

$$\left. + \frac{\beta + \gamma + 2\log(\delta^{-1}\mu(\lambda)^{-1})}{\lambda n} \right\}.$$

Choose $\beta = \gamma = 2\chi \log(\delta^{-1})$, with $\chi > 0$. Let

$$\Lambda = \left\{ \lambda_k = \frac{\alpha^k}{\sigma \sqrt{n}} \; : \; k \in \mathbb{Z} \right\}$$

and

$$\mu(\lambda_k)^{-1} \leq 2(|k| + 2)^2,$$

as in the previous section. Put

$$v = \mathbb{E}(\langle \xi, M\theta \rangle^2) \; \leq \; \sup_{\xi \in \mathbb{S}_p, \theta \in \mathbb{S}_q} \mathbb{E}(\langle \xi, M\theta \rangle^2) = v_*,$$

$$t = \mathbb{E}(\|M\theta\|^2) \; \leq \; \sup_{\theta \in \mathbb{S}_q} \mathbb{E}(\|M\theta\|^2) = t_*,$$

$$u = \mathbb{E}(\|M^\top \xi\|^2) \; \leq \; \sup_{\xi \in \mathbb{S}_p} \mathbb{E}(\|M^\top \xi\|^2) = u_*,$$

$$T = \mathbb{E}(\|M\|_{\mathrm{HS}}^2),$$

$$\ell = \log(\delta^{-1}),$$

$$\lambda_*^2 = \frac{2(1 + \chi)\ell^2}{n \left(\ell v + \dfrac{t + u}{\chi} + \dfrac{T}{\ell \chi^2} \right)}.$$

Remark that, in a similar way to the case of a vector treated in the previous section,

$$B_+(\xi,\theta) = \inf_{\lambda \in \Lambda} \lambda \left(v + \frac{t+u}{\chi \ell} + \frac{T}{\chi^2 \ell^2} \right) + \frac{2(\chi+1)\ell + 2\log(\mu(\lambda)^{-1})}{\lambda n}$$

$$\leq \inf_{\lambda \in \Lambda} 2 \sqrt{\frac{2(1+\chi)}{n} \left(\ell v + \frac{t+u}{\chi} + \frac{T}{\ell \chi^2} \right)} \left\{ \cosh\left[\log\left(\frac{\lambda}{\lambda_*} \right) \right] + \frac{\lambda_* \log(\mu(\lambda)^{-1})}{2\lambda(1+\chi)\ell} \right\}.$$

Replacing λ_* by its value, choosing $\lambda = \lambda_{k_*}$ such that $|\log(\lambda/\lambda_*)| \leq \log(\alpha)/2$ and remarking that

$$|k_*| \leq \frac{|\log(\sqrt{n}\sigma\lambda_*)|}{\log(\alpha)} + \frac{1}{2},$$

we obtain

PROPOSITION 5.4. *With probability at least $1 - 2\delta$, for any $\xi \in \mathbb{S}_p$, any $\theta \in \mathbb{S}_q$,*

$$|\mathcal{E}(\xi,\theta) - \langle \xi, \mathbb{E}(M)\theta \rangle|$$

$$\leq B(\xi,\theta) = 2C \sqrt{\frac{2(1+\chi)}{n} \left(v \log(\delta^{-1}) + \frac{t+u}{\chi} + \frac{T}{\chi^2 \log(\delta^{-1})} \right)},$$

where, using the abbreviation $\ell = \log(\delta^{-1})$,

$$C = \cosh\left(\frac{\log(\alpha)}{2} \right)$$

$$+ \frac{\sqrt{\alpha}}{(1+\chi)\ell} \log \left\{ \frac{1}{\sqrt{2}\log(\alpha)} \left| \log\left(\frac{\ell v + \frac{t+u}{\chi} + \frac{T}{\ell \chi^2}}{2\sigma^2(1+\chi)\ell^2} \right) \right| + \frac{5}{\sqrt{2}} \right\}.$$

Let us now consider an estimator $\widehat{m}$ such that

$$\sup_{\xi \in \mathbb{S}_p, \theta \in \mathbb{S}_q} |\mathcal{E}(\xi,\theta) - \langle \xi, \widehat{m}\, \theta \rangle| \leq \inf_{m \in \mathbb{R}^{p \times q}} \sup_{\xi \in \mathbb{S}_p, \theta \in \mathbb{S}_q} |\mathcal{E}(\xi,\theta) - \langle \xi, m\, \theta \rangle|.$$

With probability at least $1 - 2\delta$,

$$\|\widehat{m} - \mathbb{E}(M)\|_\infty \leq 2 \sup_{\xi \in \mathbb{S}_p, \theta \in \mathbb{S}_q} B(\xi,\theta).$$

Remark that we can bound $\sup_{\xi \in \mathbb{S}_p, \theta \in \mathbb{S}_q} B(\xi,\theta)$ by the explicit expression for $B(\xi,\theta)$ where v, t and u are replaced by their upper bounds v_*, t_*, and u_* with respect to $\xi \in \mathbb{S}_p$ and $\theta \in \mathbb{S}_q$.

Remark also that we can weaken the influence of T by choosing $\chi > 1$, but that we can reach the optimal bound for $\|\widehat{m} - \mathbb{E}(M)\|_\infty$ only if we know an upper bound

for the ratio T/v_*. Indeed, if we know T/v_* (or an upper bound of the same order of magnitude, up to a constant), we can choose

$$\chi = \max\left\{\frac{1}{\log(\delta^{-1})}\sqrt{\frac{T}{v_*}}, 1\right\}.$$

In this case, with probability at least $1 - 2\delta$,

$$\|\widehat{m} - \mathbb{E}(M)\|_\infty \leq \frac{8C}{\sqrt{n}}\sqrt{v_*\log(\delta^{-1}) + t_* + u_* + \sqrt{v_*T}}.$$

Most likely we do not know

$$\sqrt{\frac{T}{v_*}} = \sqrt{\frac{\mathbb{E}(\|M\|^2_{\mathrm{HS}})}{\sup_{\xi\in\mathbb{S}_p, \theta\in\mathbb{S}_q}\mathbb{E}(\langle\xi, M\theta\rangle^2)}},$$

but we can still choose χ greater than one, to lower the influence of $T = \mathbb{E}(\|M\|^2_{\mathrm{HS}})$ in the bound.

6. Adaptive Gram matrix estimate   We devote a section to the adaptive estimation of a Gram matrix, since it is an important subject for applications to principal component analysis and to least squares regression. We recall that, given a random vector $X \in \mathbb{R}^d$, the Gram matrix of X is defined as

$$G = \mathbb{E}(XX^\top) \in \mathbb{R}^{d\times d}.$$

The general approach of the previous section uses an estimator that cannot be computed explicitly without recourse to a Monte Carlo sampling algorithm. In the special case of the Gram matrix, we will produce an estimator that does not suffer from this drawback.

Consequences of what is proved in this section regarding robust principal component analysis can easily be drawn from the method exposed in Giulini (2017b). We refer to this paper for further details. Consequences regarding least squares regression are discussed at the end of this paper.

In this section, we will use the asymmetric influence function defined by equation (3) on page 15. The explicit computation of our estimator however will use the modified auxiliary function

$$\varphi_2(m, \sigma) = \mathbb{E}\left[\psi\left((m + \sigma W)^2\right)\right], \qquad m \in \mathbb{R}, \sigma \in \mathbb{R}_+,$$

where $W \sim \mathcal{N}(0, 1)$ is a standard Gaussian random variable.
Observe that it is possible to explicitly compute the function φ_2 in terms of the Gaussian distribution function $F(a) = \mathbb{P}(W \leq a)$.

LEMMA 6.1. *For any $m \in \mathbb{R}$ and $\sigma \in \mathbb{R}_+$,*

$$\varphi_2(m,\sigma) = m^2 + \sigma^2 - \frac{1}{2}\left(m^4 + 6m^2\sigma^2 + 3\sigma^4\right) + r_2(m,\sigma),$$

where

$$r_2(m,\sigma) = \frac{1}{2}\left[(m^2-1)^2 + (6m^2-2)\sigma^2 + 3\sigma^4\right]\left[F\left(\frac{-1-m}{\sigma}\right) + F\left(\frac{-1+m}{\sigma}\right)\right]$$

$$+ \frac{\sigma}{2\sqrt{2\pi}}\left[\sigma^2(3-5m) - (1+m)(1-m)^2\right]\exp\left(-\frac{(1+m)^2}{2\sigma^2}\right)$$

$$+ \frac{\sigma}{2\sqrt{2\pi}}\left[\sigma^2(3+5m) - (1-m)(1+m)^2\right]\exp\left(-\frac{(1-m)^2}{2\sigma^2}\right).$$

PROOF. The proof is based on the expression

$$\psi(t) = t - t^2/2 + \mathbb{1}(t \geq 1)(1-t)^2/2, \qquad t \in \mathbb{R}_+,$$

and on the identities

$$\mathbb{E}\left[\mathbb{1}(W \leq a)\right] = F(a) = 1 - \mathbb{E}\left[\mathbb{1}(W \geq a)\right],$$

$$\mathbb{E}\left[W\mathbb{1}(W \leq a)\right] = -\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{a^2}{2}\right) = -\mathbb{E}\left[W\mathbb{1}(W \geq a)\right],$$

$$\mathbb{E}\left[W^2\mathbb{1}(W \leq a)\right] = -\frac{a}{\sqrt{2\pi}}\exp\left(-\frac{a^2}{2}\right) + F(a) = 1 - \mathbb{E}\left[W^2\mathbb{1}(W \geq a)\right],$$

$$\mathbb{E}\left[W^3\mathbb{1}(W \leq a)\right] = -\frac{a^2+2}{\sqrt{2\pi}}\exp\left(-\frac{a^2}{2}\right) = -\mathbb{E}\left[W^3\mathbb{1}(W \geq a)\right],$$

$$\mathbb{E}\left[W^4\mathbb{1}(W \leq a)\right] = -\frac{a^3+3a}{\sqrt{2\pi}}\exp\left(-\frac{a^2}{2}\right) + 3F(a) = 3 - \mathbb{E}\left[W^4\mathbb{1}(W \geq a)\right].$$

Let us put $G(t) = \dfrac{1}{\sqrt{2\pi}}\exp\left(-\dfrac{t^2}{2}\right)$.

$$\mathbb{E}\left\{\psi\left[(m+\sigma W)^2\right]\right\} = \mathbb{E}\left[(m+\sigma W)^2 - \frac{1}{2}(m+\sigma W)^4\right] + r_2(m,\sigma),$$

where

$$2r_2(m,\sigma) = \mathbb{E}\Bigg\{\left[(m+\sigma W)^4 - 2(m+\sigma W)^2 + 1\right]$$

$$\times \left[\mathbb{1}\left(W \leq \frac{-1-m}{\sigma}\right) + \mathbb{1}\left(W \geq \frac{1-m}{\sigma}\right)\right]\Bigg\}$$

23

$$= \mathbb{E}\Big\{ \big[(m^2 - 1)^2 + 4m(m^2 - 1)\sigma W + (6m^2 - 2)\sigma^2 W^2 + 4m\sigma^3 W^3 + \sigma^4 W^4 \big]$$

$$\times \Big[\mathbb{1}\Big(W \le \frac{-1 - m}{\sigma} \Big) + \mathbb{1}\Big(W \ge \frac{1 - m}{\sigma} \Big) \Big] \Big\}$$

$$= (m^2 - 1)^2 \Big[F\Big(\frac{-1 - m}{\sigma}\Big) + F\Big(\frac{-1 + m}{\sigma}\Big) \Big]$$

$$+ 4m(m^2 - 1)\sigma \Big[-G\Big(\frac{-1 - m}{\sigma}\Big) + G\Big(\frac{1 - m}{\sigma}\Big) \Big]$$

$$+ (6m^2 - 2)\sigma^2 \Big[\frac{1 + m}{\sigma} G\Big(\frac{-1 - m}{\sigma}\Big) + \frac{1 - m}{\sigma} G\Big(\frac{1 - m}{\sigma}\Big)$$

$$+ F\Big(\frac{-1 - m}{\sigma}\Big) + F\Big(\frac{-1 + m}{\sigma}\Big) \Big]$$

$$+ 4m\sigma^3 \Big\{ -\Big[\Big(\frac{1 + m}{\sigma}\Big)^2 + 2 \Big] G\Big(\frac{-1 - m}{\sigma}\Big) + \Big[\Big(\frac{1 - m}{\sigma}\Big)^2 + 2 \Big] G\Big(\frac{-1 + m}{\sigma}\Big) \Big\}$$

$$+ \sigma^4 \Big\{ \Big[\Big(\frac{1 + m}{\sigma}\Big)^3 + \frac{3(1 + m)}{\sigma} \Big] G\Big(\frac{-1 - m}{\sigma}\Big) + \Big[\Big(\frac{1 - m}{\sigma}\Big)^3 + \frac{3(1 - m)}{\sigma} \Big] G\Big(\frac{1 - m}{\sigma}\Big)$$

$$+ 3 \Big[F\Big(\frac{-1 - m}{\sigma}\Big) + F\Big(\frac{-1 + m}{\sigma}\Big) \Big] \Big\},$$

so that

$$r_2(m, \sigma) = \frac{1}{2} \Big[(m^2 - 1)^2 + (6m^2 - 2)\sigma^2 + 3\sigma^4 \Big] \Big[F\Big(\frac{-1 - m}{\sigma}\Big) + F\Big(\frac{-1 + m}{\sigma}\Big) \Big]$$

$$+ \frac{1}{2}\sigma \Big[\sigma^2(3 - 5m) - (1 + m)(1 - m)^2 \Big] G\Big(\frac{-1 - m}{\sigma}\Big)$$

$$+ \frac{1}{2}\sigma \Big[\sigma^2(3 + 5m) - (1 - m)(1 + m)^2 \Big] G\Big(\frac{1 - m}{\sigma}\Big).$$

$\square$

Observe now that, when θ' is distributed according to $\rho_\theta = \mathcal{N}(\theta, \beta^{-1} I_d)$, the real valued random variable $\langle \theta', x \rangle$ is Gaussian with mean $\langle \theta, x \rangle$ and standard deviation $\|x\| / \sqrt{\beta}$. Thus we can state the following.

LEMMA 6.2. *For any $\theta, x \in \mathbb{R}^d$,*

$$\int \psi(\langle \theta', x \rangle^2) \, d\rho_\theta(\theta') = \varphi_2\Big(\langle \theta, x \rangle, \frac{\|x\|}{\sqrt{\beta}} \Big).$$

Introduce

$$A_{\lambda, \beta}(\theta, x) = \varphi_2\Big(\lambda^{1/2}\langle \theta, x \rangle, \frac{\|x\|}{\sqrt{\beta}} \Big) - \log\Big(1 + \frac{\|x\|^2}{\beta} \Big),$$

24

where $\lambda \in \mathbb{R}_+$ is a constant modifying the norm of θ. Next proposition provides some upper and lower bounds.

PROPOSITION 6.3. *With probability at least $1 - \delta$, for any $\theta \in \mathbb{R}^d$, any $\lambda \in \mathbb{R}_+$,*

$$\frac{1}{n\lambda} \sum_{i=1}^{n} A_{\lambda,\beta}(\theta, X_i) - \frac{\beta \|\theta\|^2}{2n} - \frac{\log(\delta^{-1})}{n\lambda} \leq \mathbb{E}\left(\langle \theta, X \rangle^2\right) + \frac{\mathbb{E}(\|X\|^4)}{\lambda \beta^2}.$$

Moreover with probability at least $1 - \delta$, for any $\theta \in \mathbb{R}^d$, any $\lambda \in \mathbb{R}_+$,

$$\frac{1}{n\lambda} \sum_{i=1}^{n} A_{\lambda,\beta}(\theta, X_i) + \frac{\beta \|\theta\|^2}{2n} + \frac{\log(\delta^{-1})}{n\lambda}$$

$$\geq \mathbb{E}\left(\langle \theta, X \rangle^2\right) - \lambda \mathbb{E}\left(\langle \theta, X \rangle^4\right) - \frac{6\mathbb{E}\left(\|X\|^2 \langle \theta, X \rangle^2\right)}{\beta} - \frac{3\mathbb{E}(\|X\|^4)}{\lambda \beta^2}.$$

PROOF. According to Proposition 2.1 on page 2, with probability at least $1 - \delta$, for any $\theta \in \mathbb{R}^d$ and any $\lambda \in \mathbb{R}_+$,

$$\frac{1}{n\lambda} \sum_{i=1}^{n} \left[\int \psi(\langle \theta', X_i \rangle^2) \, \mathrm{d}\rho_{\lambda^{1/2}\theta}(\theta') - \log\left(1 + \frac{\|X_i\|^2}{\beta}\right) \right] - \frac{\beta \|\theta\|^2}{2n} + \frac{\log(\delta^{-1})}{n\lambda}$$

$$\leq \frac{1}{\lambda} \int \log\left\{ \mathbb{E}\left[\exp\left(\psi(\langle \theta', X \rangle^2) - \log\left(1 + \frac{\|X\|^2}{\beta}\right) \right) \right] \right\} \mathrm{d}\rho_{\lambda^{1/2}\theta}(\theta').$$

According to Lemma 5.1 on page 15,

$$\psi(t) - \log(1 + u) \leq \log\left(\frac{1 + t}{1 + u}\right)$$

$$= \log\left(1 - u + \frac{t + u^2}{1 + u}\right) \leq \log(1 + t - u + u^2), \qquad t, u \in \mathbb{R}_+.$$

Thus the right-hand side of the previous inequality is not greater than

$$\frac{1}{\lambda} \int \mathbb{E}(\langle \theta', X \rangle^2) \, \mathrm{d}\rho_{\lambda^{1/2}\theta}(\theta') - \frac{\mathbb{E}(\|X\|^2)}{\lambda \beta} + \frac{\mathbb{E}(\|X\|^4)}{\lambda \beta^2} = \mathbb{E}(\langle \theta, X \rangle^2) + \frac{\mathbb{E}(\|X\|^4)}{\lambda \beta^2}.$$

In the same time, due to Lemma 6.2, its left-hand side is equal to

$$\frac{1}{n\lambda} \sum_{i=1}^{n} A_{\lambda,\beta}(\theta, X_i) - \frac{\beta \|\theta\|^2}{2n} - \frac{\log(\delta^{-1})}{n\lambda}.$$

25

This achieves the proof for the upper bound. Let us now come to the lower bound. As a consequence of Lemma 5.1 on page 15, for any $t \in [0, 1]$ and any $y \in \mathbb{R}_+$,

$$
\begin{aligned}
-\psi(t) + \log(1 + y) &\leq \log(1 - t + t^2) + \log(1 + y) \\
&= \log\left(1 - t + t^2 + (1 - t + t^2)y\right) \\
&\leq \log(1 - t + t^2 + y).
\end{aligned}
$$

When $t \in [1, \infty[$ the same inequality is also obviously true:

$$
-\psi(t) + \log(1 + y) \leq \log(1 + y) \leq \log(1 - t + t^2 + y).
$$

As a consequence, for any $x \in \mathbb{R}^d$,

$$
\begin{aligned}
\int -\psi(\langle \theta', x \rangle^2) \, \mathrm{d}\rho_\theta(\theta') &+ \log\left(1 + \frac{\|x\|^2}{\beta}\right) \\
&\leq \int \log\left(1 - \langle \theta', x \rangle^2 + \langle \theta', x \rangle^4 + \frac{\|x\|^2}{\beta}\right) \, \mathrm{d}\rho_\theta(\theta')
\end{aligned}
$$

Thus, according to the PAC-Bayesian inequality stated in Proposition 2.1 on page 2, with probability al least $1 - \delta$, for any $\theta \in \mathbb{R}^d$ and any $\lambda \in \mathbb{R}_+$,

$$
\begin{aligned}
\frac{1}{n\lambda} \sum_{i=1}^{n} &\left[\int -\psi(\langle \theta', X_i \rangle^2) \, \mathrm{d}\rho_{\lambda^{1/2}\theta}(\theta') + \log\left(1 + \frac{\|X_i\|^2}{\beta}\right) \right] - \frac{\beta\|\theta\|^2}{2n} - \frac{\log(\delta^{-1})}{n\lambda} \\
&\leq \frac{1}{\lambda} \int \log\left\{ \mathbb{E}\left[\exp\left(-\psi(\langle \theta', X \rangle^2) + \log\left(1 + \frac{\|X\|^2}{\beta}\right) \right) \right] \right\} \mathrm{d}\rho_{\lambda^{1/2}\theta}(\theta') \\
&\leq \frac{1}{\lambda} \int \mathbb{E}\left(-\langle \theta', X \rangle^2 + \langle \theta', X \rangle^4 + \frac{\|X\|^2}{\beta} \right) \mathrm{d}\rho_{\lambda^{1/2}\theta}(\theta').
\end{aligned}
$$

To conclude the proof, it is enough to use the explicit expression of the moments of a Gaussian random variable, remembering that, when θ' is distributed according to $\rho_{\lambda^{1/2}\theta}$, the distribution of $\langle \theta', X \rangle$ is equal to $\mathcal{N}\left(\lambda^{1/2}\langle \theta, X \rangle, \|X\|^2/\beta\right)$. $\qquad\square$

The next proposition defines an estimator of the quadratic form $\mathbb{E}(\langle \theta, X \rangle^2)$. Note that, since we introduced a parameter λ that takes care of the norm of θ, we will assume in the following without loss of generality that $\theta \in \mathbb{S}_d$, the unit sphere of $\mathbb{R}^d$.

PROPOSITION 6.4. *Let us assume that*

$$
\mathbb{E}(\|X\|^4) \leq T < \infty,
$$

for a known constant T. For any $\theta \in \mathbb{S}_d$, consider the estimator of $\mathbb{E}(\langle \theta, X \rangle^2)$ defined as

$$\mathcal{E}(\theta) = \sup_{\lambda \in \mathbb{R}^d} \left\{ \frac{1}{n\lambda} \sum_{i=1}^{n} A_{\lambda,\beta}(\theta, X_i) - \frac{\beta}{2n} - \frac{\log(\delta^{-1})}{n\lambda} - \frac{T}{\lambda\beta^2} \right\}.$$

With probability at least $1 - \delta$, for any $\theta \in \mathbb{S}_d$,

$$\mathcal{E}(\theta) \leq \mathbb{E}\left(\langle \theta, X \rangle^2 \right).$$

Moreover, with probability at least $1 - \delta$, for any $\theta \in \mathbb{S}_d$,

$$\mathbb{E}\left(\langle \theta, X \rangle^2 \right) \leq \mathcal{E}(\theta) + 2\sqrt{2\mathbb{E}\left(\langle \theta, X \rangle^4 \right)\left(\frac{2T}{\beta^2} + \frac{\log(\delta^{-1})}{n} \right)}$$

$$+ \frac{6\mathbb{E}\left(\|X\|^2 \langle \theta, X \rangle^2 \right)}{\beta} + \frac{\beta}{n}.$$

REMARK 6.1. Introducing $\alpha = \dfrac{2Tn}{\beta^2}$, we can also express the previous bound as

$$\mathbb{E}(\langle \theta, X \rangle^2) \leq \mathcal{E}(\theta) + 2\sqrt{\frac{2}{n}\mathbb{E}(\langle \theta, X \rangle^4)\left[\alpha + \log(\delta^{-1}) \right]}$$

$$+ 3\sqrt{\frac{2\alpha}{Tn}}\mathbb{E}\left(\langle \theta, X \rangle^2 \|X\|^2 \right) + \sqrt{\frac{2T}{\alpha n}}$$

$$\leq \mathcal{E}(\theta) + \left[2\sqrt{2\mathbb{E}(\langle \theta, X \rangle^4)} + 3\sqrt{\frac{2}{T}}\mathbb{E}\left(\langle \theta, X \rangle^2 \|X\|^2 \right) \right]\sqrt{\frac{\alpha}{n}}$$

$$+ \sqrt{\frac{2T}{n\alpha}} + 2\sqrt{\frac{2}{n}\mathbb{E}(\langle \theta, X \rangle^4)\log(\delta^{-1})}$$

$$\leq \mathcal{E}(\theta) + 5\sqrt{\frac{2\alpha}{n}\mathbb{E}(\langle \theta, X \rangle^4)} + \sqrt{\frac{2T}{n\alpha}} + 2\sqrt{\frac{2}{n}\mathbb{E}(\langle \theta, X \rangle^4)\log(\delta^{-1})},$$

where the last inequality is a consequence of the Cauchy-Schwarz inequality

$$\mathbb{E}\left(\langle \theta, X \rangle^2 \|X\|^2 \right) \leq \sqrt{\mathbb{E}\left(\langle \theta, X \rangle^4 \right)\mathbb{E}(\|X\|^4)} \leq \sqrt{T\,\mathbb{E}\left(\langle \theta, X \rangle^4 \right)}.$$

PROOF. Proposition 6.4 follows from Proposition 6.3 and the definition of the estimator $\mathcal{E}$. To get the second inequality, observe that the value of λ minimizing

$$\lambda\mathbb{E}\left(\langle \theta, X \rangle^4 \right) + \frac{6\mathbb{E}\left(\|X\|^2 \langle \theta, X \rangle^2 \right)}{\beta} + \frac{4T}{\lambda\beta^2} + \frac{\beta}{n} + \frac{2\log(\delta^{-1})}{n\lambda}$$

27

is given by

$$\lambda = \sqrt{2\mathbb{E}(\langle\theta,X\rangle^4)^{-1}\left(\frac{2T}{\beta^2} + \frac{\log(\delta^{-1})}{n}\right)}.$$

$\square$

In the following proposition, we make the estimator adaptive in α as well as in λ and we introduce our estimator $\widehat{G}$ of the Gram matrix G.

PROPOSITION 6.5. *Let us assume that $\mathbb{E}(\|X\|^4) \le T < \infty$, where T is a known constant. Consider the estimator*

$$\widetilde{\mathcal{E}}(\theta) = \sup_{\lambda\in\mathbb{R}_+} \sup_{k\in\mathbb{N}} \frac{1}{n\lambda} \sum_{i=1}^{n}\left[\varphi_2\left(\sqrt{\lambda}\langle\theta,X_i\rangle,\left(\frac{\exp(k)}{10Tn}\right)^{1/4}\|X_i\|\right)\right.$$
$$\left. - \log\left(1 + \sqrt{\frac{\exp(k)}{10Tn}}\|X_i\|^2\right)\right] - \frac{1}{n}\sqrt{\frac{5T}{2\exp(k)}} - \frac{\exp(k)}{10n\lambda}$$
$$- \frac{\log[(k+1)(k+2)/\delta]}{n\lambda}, \qquad \theta\in\mathbb{S}_d.$$

With probability at least $1-\delta$, for any $\theta\in\mathbb{S}_d$,

$$\widetilde{\mathcal{E}}(\theta) \le \mathbb{E}(\langle\theta,X\rangle^2).$$

With probability at least $1-\delta$, for any $\theta\in\mathbb{S}_d$,

$$\mathbb{E}(\langle\theta,X\rangle^2) \le \widetilde{\mathcal{E}}(\theta) + B(\theta),$$

where

$$B(\theta) = 2\sqrt{\frac{\mathbb{E}(\langle\theta,X\rangle^4)}{n}}\left\{3.3\left(\frac{T}{\mathbb{E}(\langle\theta,X\rangle^4)}\right)^{1/4}\right.$$
$$\left. + \sqrt{4\log\left(\frac{1}{2}\log\left(\frac{T}{\mathbb{E}(\langle\theta,X\rangle^4)}\right) + \frac{5}{2}\right) + 2\log(\delta^{-1})}\right\}.$$

Consider an estimator $\widehat{G}\in\mathbb{R}^{d\times d}$ such that

$$0 \le \inf_{\theta\in\mathbb{S}_d} \langle\theta,\widehat{G}\theta\rangle - \widetilde{\mathcal{E}}(\theta)$$

and

28

$$\sup_{\theta \in \mathbb{S}_d} \langle \theta, \widehat{G}\, \theta \rangle - \widetilde{\mathcal{E}}(\theta) = \inf \Big\{ \sup_{\theta \in \mathbb{S}_d} \langle \theta, M\, \theta \rangle - \widetilde{\mathcal{E}}(\theta) \; : \; M \in \mathbb{R}^{d \times d},$$

$$M = M^\top, \; 0 \leq \inf_{\theta \in \mathbb{S}_d} \langle \theta, M\, \theta \rangle - \widetilde{\mathcal{E}}(\theta) \Big\}.$$

With probability at least $1 - 2\delta$,

$$\| \widehat{G} - G \|_\infty \leq \sup_{\theta \in \mathbb{S}_d} B(\theta).$$

REMARK 6.2. It is interesting to rephrase this result in terms of the directional kurtosis

$$\kappa(\theta) = \begin{cases} \dfrac{\mathbb{E}(\langle \theta, X \rangle^4)}{\mathbb{E}(\langle \theta, X \rangle^2)^2}, & \mathbb{E}(\langle \theta, X \rangle^2) > 0, \\ 1, & \text{otherwise.} \end{cases}$$

We obtain with probability at least $1 - 2\delta$,

$$1 - 2\sqrt{\frac{\kappa(\theta)}{n}} \left\{ 3.3 \left(\frac{T}{\kappa(\theta)\mathbb{E}(\langle \theta, X \rangle^2)^2} \right)^{1/4} \right.$$
$$\left. + \sqrt{4\log\left(\frac{1}{2}\log\left(\frac{T}{\kappa(\theta)\mathbb{E}(\langle \theta, X \rangle^2)^2} \right) + \frac{5}{2} \right) + 2\log(\delta^{-1})} \right\} \leq \frac{\widetilde{\mathcal{E}}(\theta)}{\mathbb{E}(\langle \theta, X \rangle^2)} \leq 1,$$

with the appropriate convention that $r/0 = +\infty$ when $r > 0$ and $0/0 = 1$. This inequality shows under which circumstances it is possible to estimate the order of magnitude of $\mathbb{E}(\langle \theta, X \rangle^2)$ and consequently the eigenvalues of the Gram matrix G. Indeed, introducing $\kappa_* = \sup_{\theta \in \mathbb{S}_d} \kappa(\theta)$, we deduce with probability at least $1 - 2\delta$ a bound of the form

$$1 - \sqrt{\frac{f\big(\kappa_*, \mathbb{E}(\langle \theta, X \rangle^2)\big)}{n}} \leq \frac{\widetilde{\mathcal{E}}(\theta)}{\mathbb{E}(\langle \theta, X \rangle^2)} \leq 1,$$

where the function $\sigma \mapsto F(\kappa_*, \sigma) = \sigma\left(1 - \sqrt{f(\kappa_*, \sigma)/n} \right)$ is non-decreasing. Let us write $G = \mathbb{E}(XX^\top)$ as

$$G = \sum_{i=1}^d \sigma_i e_i e_i^\top,$$

where $(e_1, \dots, e_d)$ is an orthonormal basis of eigenvectors and where $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_d$ are the eigenvalues of G counted with their multiplicities and sorted in

decreasing order. Introducing $\mathcal{L}_i$, the set of all linear subspaces of $\mathbb{R}^d$ of dimension i, it is well known that

$$\sigma_i = \sup\{\inf\{\langle\theta, G\theta\rangle, \ \theta \in L \cap \mathbb{S}_d\}, \ L \in \mathcal{L}_i\}.$$

A proof can for instance be found in (Kato, 1982, page 62). Based on this formula, we can introduce the estimator

$$\widehat{\sigma}_i = \sup\{\inf\{\widetilde{\mathcal{E}}(\theta), \ \theta \in L \cap \mathbb{S}_d\}, \ L \in \mathcal{L}_i\}.$$

It is such that

$$\begin{aligned}
F(\kappa_*, \sigma_i) &= F\Big(\kappa_*, \sup\{\inf\{\langle\theta, G\theta\rangle, \ \theta \in L \cap \mathbb{S}_d\}, \ L \in \mathcal{L}_i\}\Big) \\
&= \sup\{\inf\{F(\kappa_*, \langle\theta, G\theta\rangle), \ \theta \in L \cap \mathbb{S}_d\}, \ L \in \mathcal{L}_i\} \\
&\leq \widehat{\sigma}_i \leq \sup\{\inf\{\langle\theta, G\theta\rangle, \ \theta \in L \cap \mathbb{S}_d\}, \ L \in \mathcal{L}_i\} = \sigma_i,
\end{aligned}$$

proving that with probability at least $1 - 2\delta$,

$$1 - \sqrt{\frac{f(\kappa_*, \sigma_i)}{n}} \leq \frac{\widehat{\sigma}_i}{\sigma_i} \leq 1, \qquad 1 \leq i \leq d.$$

PROOF OF PROPOSITION 6.5 ON PAGE 28. The optimal value of α in the last bound given in Remark 6.1 on page 27 is given by

$$\alpha_* = \frac{1}{5}\sqrt{\frac{T}{\mathbb{E}(\langle\theta, X\rangle^4)}} \geq \frac{1}{5}\sqrt{\frac{\mathbb{E}(\|X\|^4)}{\mathbb{E}(\langle\theta, X\rangle^4)}} \geq \frac{1}{5}.$$

According to the simplified inequality stated at the end of Remark 6.1, with probability at least $1 - \delta$, for any $\theta \in \mathbb{S}_d$,

$$\begin{aligned}
\mathbb{E}(\langle\theta, X\rangle^2) &\leq \mathcal{E}(\theta) + 2\sqrt{\frac{10}{n}}\big[T\mathbb{E}(\langle\theta, X\rangle^4)\big]^{1/4}\frac{(\sqrt{\alpha/\alpha_*} + \sqrt{\alpha_*/\alpha})}{2} \\
&\qquad\qquad\qquad\qquad + 2\sqrt{\frac{2}{n}\mathbb{E}(\langle\theta, X\rangle^4)\log(\delta^{-1})} \\
&= \mathcal{E}(\theta) + 2\sqrt{\frac{10}{n}}\big[T\mathbb{E}(\langle\theta, X\rangle^4)\big]^{1/4}\cosh\Big(\frac{1}{2}\log(\alpha/\alpha_*)\Big) \\
&\qquad\qquad\qquad\qquad + 2\sqrt{\frac{2}{n}\mathbb{E}(\langle\theta, X\rangle^4)\log(\delta^{-1})}.
\end{aligned}$$

We will take a weighted union bound on all values of α belonging to $\{\exp(k)/5 \ : \ k \in \mathbb{N}\}$. To perform this, we have to modify accordingly the definition of the estimator and consider the estimator $\widetilde{\mathcal{E}}$ defined in the proposition. In this

30

change of definition, we have replaced β with $\sqrt{\dfrac{10\,Tn}{\exp(k)}}$ and δ with $\dfrac{\delta}{(k+1)(k+2)}$, and we have taken the supremum in $k \in \mathbb{N}$ as well as in $\lambda \in \mathbb{R}_+$. As

$$\sum_{k \in \mathbb{N}} \frac{\delta}{(k+1)(k+2)} = \delta$$

we get from Proposition 6.3 on page 25 that with probability at least $1 - \delta$, for any $\theta \in \mathbb{S}_d$, $\widetilde{\mathcal{E}}(\theta) \leq \mathbb{E}(\langle \theta, X \rangle^2)$. Recalling that $\alpha = \dfrac{2Tn}{\beta^2} = \dfrac{\exp(k)}{5}$, we get with probability at least $1 - \delta$, for any $\theta \in \mathbb{S}_d$,

$$\mathbb{E}(\langle \theta, X \rangle^2) \leq \widetilde{\mathcal{E}}(\theta)$$
$$+ \inf_{k \in \mathbb{N}} 2\sqrt{\frac{10}{n}} \left[T\mathbb{E}(\langle \theta, X \rangle^4) \right]^{1/4} \cosh\left(\frac{1}{2} \log\left(\frac{\exp(k)}{5\alpha_*} \right) \right)$$
$$+ 2\sqrt{\frac{2}{n}\mathbb{E}(\langle \theta, X \rangle^4) \log\left[(k+2)^2/\delta \right]}.$$

(We can take the infimum in k because the inequality holds with probability $1 - \delta$ for any value of $k \in \mathbb{N}$). We can now choose k to be the closest integer to $\log(5\alpha_*)$ (that is known to be a non-negative quantity). It is such that $\left| \log\left(\dfrac{\exp(k)}{5\alpha_*} \right) \right| \leq \dfrac{1}{2}$ and therefore

$$k + 2 \leq \log(5\alpha_*) + \frac{5}{2} = \frac{1}{2}\log\left(\frac{T}{\mathbb{E}(\langle \theta, X \rangle^4)} \right) + \frac{5}{2}.$$

Remarking that $\sqrt{10}\cosh(1/4) \leq 3.3$ ends the proof. $\qquad\square$

7. Linear least squares regression Consider a couple of random variables $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$ whose distribution is assumed to be unknown. Let

$$(X_1, Y_1), \ldots, (X_n, Y_n)$$

be an observed sample made of n independent copies of (X, Y). In this section, we consider the question of estimating

$$\inf_\theta \mathbb{E}\left[(\langle \theta, X \rangle - Y)^2 \right].$$

Introduce the Gram matrix

$$G = \mathbb{E}(XX^\top) \in \mathbb{R}^{d \times d},$$

31

the vector

$$V = \mathbb{E}(YX) \in \mathbb{R}^d,$$

and the risk function

$$R(\theta) = \langle \theta, G\theta \rangle - 2\langle \theta, V \rangle.$$

Remark that

$$\mathbb{E}\left[(\langle \theta, X \rangle - Y)^2\right] = \mathbb{E}(Y^2) + R(\theta), \qquad \theta \in \mathbb{R}^d,$$

so that minimizing the quadratic loss is equivalent to minimizing R.

We have seen in the previous sections various methods to estimate G and V. As a straightforward consequence, we state a first result, concerning the minimization over a bounded domain.

PROPOSITION 7.1. *Assume that $\widehat{G} \in \mathbb{R}^{d \times d}$ and $\widehat{V} \in \mathbb{R}^d$ are such that*

(4) $$\|\widehat{G} - G\|_\infty \le \epsilon, \text{ and } \|\widehat{V} - V\| \le \eta.$$

Assume also that $\widehat{G}$ is a symmetric positive semi-definite matrix. Let Θ be a closed bounded set in $\mathbb{R}^d$ and let $B = \sup_{\theta \in \Theta} \|\theta\|$. Consider the estimated risk

$$\widehat{R}(\theta) = \langle \theta, \widehat{G}\,\theta \rangle - 2\langle \theta, \widehat{V} \rangle, \qquad \theta \in \mathbb{R}^d,$$

and an estimator $\widehat{\theta} \in \arg\min_{\Theta} \widehat{R}$. It is such that

$$R(\widehat{\theta}) - \inf_{\Theta} R \le 2B(\epsilon B + 2\eta).$$

PROOF. Remark that

$$R(\widehat{\theta}) \le \widehat{R}(\widehat{\theta}) + B^2\epsilon + 2B\eta = \inf_{\theta \in \Theta} \widehat{R}(\theta) + B^2\epsilon + 2B\eta$$

$$\le \inf_{\theta \in \Theta} R(\theta) + 2B^2\epsilon + 4B\eta.$$

$\square$

COROLLARY 7.2. *Assume that we know constants v, T, v', T' such that*

$$\sup_{\theta \in \mathbb{S}_d} \mathbb{E}(\langle \theta, X \rangle^4) \le v < \infty,$$

$$\mathbb{E}(\|X\|^4) \le T < \infty,$$

$$\sup_{\theta \in \mathbb{S}_d} \mathbb{E}(Y^2 \langle \theta, X \rangle^2) \le v' < \infty,$$

32

$$\mathbb{E}(Y^2\|X\|^2) \le T' < \infty,$$

Using Propositions 3.3 on page 6 and 4.2 on page 11, we can define estimators $\widehat{G}$ and $\widehat{V}$ such that with probability at least $1 - 2\delta$,

$$\|\widehat{G} - G\|_\infty \le \epsilon = 2\sqrt{\frac{2v}{n}\left(2\log(\delta^{-1} + 12\sqrt{T/v}\right)}$$

and $\quad \|\widehat{V} - V\| \le \eta = 2\left(\sqrt{T'/n} + \sqrt{2v'\log(\delta^{-1})/n}\right).$

Consequently, the estimator $\widehat{\theta}$ of the previous proposition based on $\widehat{G}$ and $\widehat{V}$ is such that with probability at least $1 - 2\delta$,

$$R(\widehat{\theta}) - \inf_\Theta R \le \mathcal{O}\left(\sqrt{\frac{\log(\delta^{-1})}{n}}\right),$$

where the constant hiding behind the notation $\mathcal{O}$ depends only on v, T, v', T' and $\sup_{\theta\in\Theta}\|\theta\|$.

REMARK 7.1. We get only a slow speed of order $n^{-1/2}$ and not n^{-1}, but we think it is the price to pay to have a dimension-free bound under such hypotheses.

In the following, we will release the constraint that θ belongs to a bounded domain. We will also propose conditions under which a fast rate of order $\mathcal{O}(\log(\delta^{-1})/n)$ is possible. We will be interested first in defining some non-asymptotic confidence region for $\theta_* \in \arg\min_{\theta\in\mathbb{R}^d} R(\theta)$. We will broaden our analysis to the estimation of the ridge regression $\theta_\lambda \in \arg\min_{\theta\in\mathbb{R}^d}(R(\theta) + \lambda\|\theta\|^2)$, since this extension is quite natural in this context. Indeed, the ridge regression problem consists in minimizing R on a ball centered at the origin, and ridge regressors, as we will see, will anyhow play a role in the definition of a robust estimator.

PROPOSITION 7.3. *Make the same assumptions as at the beginning of Proposition 7.1 on the preceding page and consider some parameter $\lambda \in \mathbb{R}_+$. Introduce the ridge regression loss function*

$$R_\lambda(\theta) = R(\theta) + \lambda\|\theta\|^2 = \langle\theta, (G + \lambda I)\theta\rangle - 2\langle\theta, V\rangle$$

and its empirical counterpart

$$\widehat{R}_\lambda(\theta) = \widehat{R}(\theta) + \lambda\|\theta\|^2 = \langle\theta, (\widehat{G} + \lambda I)\theta\rangle - 2\langle\theta, \widehat{V}\rangle.$$

Let $\theta_\lambda \in \arg\min_{\theta\in\mathbb{R}^d} R_\lambda$ and $\widehat{\theta}_\lambda \in \arg\min_{\theta\in\mathbb{R}^d} \widehat{R}_\lambda(\theta)$. Define the confidence region

$$\widehat{\Theta}_\lambda = \left\{\theta \in \mathbb{R}^d \; : \; \|(\widehat{G} + \lambda)(\theta - \widehat{\theta}_\lambda)\| \le \|\theta\|\epsilon + \eta\right\}.$$

33

On the event defined by equation (4) *on page 32,*

$$\theta_\lambda \in \widehat{\Theta}_\lambda.$$

Moreover, for any estimator $\widehat{\theta} \notin \widehat{\Theta}_\lambda$, *the improved pick*

$$\widetilde{\theta} \in \arg\min_{\theta \in \mathbb{R}^d} \left\{ \widehat{R}_\lambda(\theta) - \widehat{R}_\lambda(\widehat{\theta}) + \epsilon\|\theta - \widehat{\theta}\|^2 + 2\|\theta - \widehat{\theta}\|\left(\epsilon\|\widehat{\theta}\| + \eta\right) \right\}$$

is such that

$$R_\lambda(\widetilde{\theta}) < R_\lambda(\widehat{\theta}),$$

and more precisely such that

$$R_\lambda(\widetilde{\theta}) - R_\lambda(\widehat{\theta}) \leq \widehat{R}_\lambda(\widetilde{\theta}) - \widehat{R}_\lambda(\widehat{\theta}) + \|\widetilde{\theta} - \widehat{\theta}\|\left(\epsilon\|\widetilde{\theta} + \widehat{\theta}\| + 2\eta\right) < 0.$$

PROOF. Note that for any $\theta, \xi \in \mathbb{R}^d$,

$$
\begin{aligned}
R_\lambda(\xi) - R_\lambda(\theta) &= \langle \theta - \xi, (G + \lambda I)(\theta + \xi)\rangle - 2\langle \theta - \xi, V\rangle \\
&\leq \widehat{R}_\lambda(\xi) - \widehat{R}_\lambda(\theta) + \|\xi - \theta\|(\epsilon\|\xi + \theta\| + 2\eta) \\
&\leq \widehat{R}_\lambda(\xi) - \widehat{R}_\lambda(\theta) + \epsilon\|\xi - \theta\|^2 + 2\|\xi - \theta\|(\epsilon\|\theta\| + \eta) \stackrel{\text{def}}{=} \gamma_\lambda(\theta, \xi).
\end{aligned}
$$

As $\xi \mapsto \gamma_\lambda(\theta, \xi)$ is strictly convex, $\inf_{\xi \in \mathbb{R}^d} \gamma_\lambda(\theta, \xi) = 0 = \gamma_\lambda(\theta, \theta)$ if and only if its subdifferential satisfies

$$0 \in \frac{\partial}{\partial\xi}_{|\xi=\theta} \gamma_\lambda(\theta, \xi) = 2(\widehat{G} + \lambda I)\theta - 2\widehat{V} + 2\mathbb{B}_d(\epsilon\|\theta\| + \eta),$$

where $\mathbb{B}_d$ is the unit ball of $\mathbb{R}^d$. Remarking that $\widehat{V} = (\widehat{G} + \lambda I)\widehat{\theta}_\lambda$, we see that this is equivalent to

$$\|(\widehat{G} + \lambda I)(\theta - \widehat{\theta}_\lambda)\| \leq \epsilon\|\theta\| + \eta.$$

To complete the proof, it is enough to remark that, due to its definition,

$$0 \geq \inf_{\xi \in \mathbb{R}^d} \gamma_\lambda(\theta_\lambda, \xi) \geq \inf_{\xi \in \mathbb{R}^d} R_\lambda(\xi) - R_\lambda(\theta_\lambda) = 0,$$

so that $\theta_\lambda \in \widehat{\Theta}_\lambda$. $\qquad\square$

Note that $\widetilde{\theta}$ is the solution of a strictly convex minimization problem. It is characterized by the equation

$$(\widehat{G} + \lambda I)\theta - \widehat{V} + \epsilon(\theta - \widehat{\theta}) + \frac{\theta - \widehat{\theta}}{\|\theta - \widehat{\theta}\|}(\epsilon\|\widehat{\theta}\| + \eta) = 0.$$

In view of the shape of the confidence region, it is natural to consider the estimator

$$\widetilde{\theta}_\lambda \in \arg\min_{\theta \in \widehat{\Theta}_\lambda} \|\theta\|.$$

34

PROPOSITION 7.4. *Let $\xi \in \widehat{\Theta}_\lambda$ be any parameter value within the above defined confidence region. Under the event defined by equation (4) on page 32, it is such that*

$$\|(G + \lambda I)(\xi - \theta_\lambda)\| \le 2(\epsilon\|\xi\| + \eta).$$

In particular, since $\theta_\lambda \in \widehat{\Theta}_\lambda$, we see from the definition of $\widetilde{\theta}_\lambda$ that $\|\widetilde{\theta}_\lambda\| \le \|\theta_\lambda\|$ and therefore that

$$\|(G + \lambda I)(\widetilde{\theta}_\lambda - \theta_\lambda)\|^2 \le 4(\epsilon\|\theta_\lambda\| + \eta)^2.$$

Thus, when $\epsilon = \mathcal{O}\big(\sqrt{\log(\delta^{-1})/n}\big)$ and $\eta = \mathcal{O}\big(\sqrt{\log(\delta^{-1})/n}\big)$, we get a convergence speed of order $\mathcal{O}(\log(\delta^{-1})/n)$, but for a modified definition of the loss function. Using a basis $(e_i, 1 \le i \le d)$ of eigenvectors of G, with corresponding eigenvalues $\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_d \ge 0$, we see more precisely that for any $\theta \in \mathbb{R}^d$,

$$R_\lambda(\theta) - R_\lambda(\theta_\lambda) = \sum_{i=1}^{d} (\sigma_i + \lambda)\langle\theta - \theta_\lambda, e_i\rangle^2,$$

whereas

$$\|(G + \lambda I)(\theta - \theta_\lambda)\|^2 = \sum_{i=1}^{d}(\sigma_i + \lambda)^2\langle\theta - \theta_\lambda, e_i\rangle^2 = \frac{1}{4}\big\|\nabla R_\lambda(\theta)\big\|^2.$$

The relation between the two risks is that

$$(\sigma_d + \lambda)\big[R_\lambda(\theta) - R_\lambda(\theta_\lambda)\big] \le \|(G + \lambda I)(\theta - \theta_\lambda)\|^2$$
$$\le (\sigma_1 + \lambda)\big[R_\lambda(\theta) - R_\lambda(\theta_\lambda)\big].$$

Consequently

$$R_\lambda(\widetilde{\theta}_\lambda) - R_\lambda(\theta_\lambda) \le \frac{4}{\sigma_d + \lambda}\big(\epsilon\|\widetilde{\theta}_\lambda\| + \eta\big)^2 \le \frac{4}{\sigma_d + \lambda}\big(\epsilon\|\theta_\lambda\| + \eta\big)^2.$$

PROOF. For any $\xi \in \widehat{\Theta}_\lambda$,

$$\|(G + \lambda I)(\xi - \theta_\lambda)\| = \|(G + \lambda I)\xi - V\| \le \|(\widehat{G} + \lambda I)\xi - \widehat{V}\| + \epsilon\|\xi\| + \eta \le 2(\epsilon\|\xi\| + \eta),$$

from which the other statements made in the proposition are straightforward consequences. □

From this proposition, we conclude that we have a dimension-free bound for $\|(G + \lambda I)(\widetilde{\theta}_\lambda - \theta_\lambda)\|^2$, whereas the bound we obtain for $R_\lambda(\widetilde{\theta}_\lambda) - R_\lambda(\theta_\lambda)$ depends on the dimension through $\sigma_d + \lambda$, so that it is dimension-free only for large enough values of λ.

For small values of λ depending on n, we can obtain a dimension-free slow rate in the following way. Remark that, since $\sigma_i \leq \dfrac{(\sigma_i + \lambda)^2}{4\lambda}$,

$$R_0(\widetilde{\theta}_\lambda) - R_0(\theta_0) = \sum_{i=1}^d \sigma_i \langle \widetilde{\theta}_\lambda - \theta_0, e_i \rangle^2 \leq \sum_{i=1}^d \frac{(\sigma_i + \lambda)^2}{4\lambda} \langle \widetilde{\theta}_\lambda - \theta_0, e_i \rangle^2$$

$$= \frac{1}{4\lambda} \|(G + \lambda I)(\widetilde{\theta}_\lambda - \theta_0)\|^2.$$

Since $V = G\theta_0 = (G + \lambda I)\theta_\lambda$,

$$\|(G + \lambda I)(\widetilde{\theta}_\lambda - \theta_0)\| = \|(G + \lambda I)(\widetilde{\theta}_\lambda - \theta_\lambda) - \lambda \theta_0\|$$

$$\leq \|(G + \lambda I)(\widetilde{\theta}_\lambda - \theta_\lambda)\| + \lambda \|\theta_0\| \leq 2(\epsilon \|\theta_\lambda\| + \eta) + \lambda \|\theta_0\|.$$

Moreover, $\|\theta_\lambda\| \leq \|\theta_0\|$, indeed,

$$R_\lambda(\theta_\lambda) = R_0(\theta_\lambda) + \lambda \|\theta_\lambda\|^2 \leq R_0(\theta_0) + \lambda \|\theta_0\|^2 \leq R_0(\theta_\lambda) + \lambda \|\theta_0\|^2.$$

Therefore,
$$\|(G + \lambda I)(\widetilde{\theta}_\lambda - \theta_0)\| \leq 2\big[(\epsilon + \lambda/2)\|\theta_0\| + \eta\big]$$

and coming back to R_0,

$$R_0(\widetilde{\theta}_\lambda) - R_0(\theta_0) \leq \frac{1}{\lambda}\big[(\epsilon + \lambda/2)\|\theta_0\| + \eta\big]^2.$$

Choose $\lambda = 2(\epsilon + \eta)$ to obtain

$$R_0(\widetilde{\theta}_{2(\epsilon+\eta)}) - R_0(\theta_0) \leq \big[\|\theta_0\| + 1/2\big]\big[(2\epsilon + \eta)\|\theta_0\| + \eta\big].$$

This is a dimension-free bound for $R_0(\widetilde{\theta}_\lambda) - R_0(\theta_0)$, but it is of order $\mathcal{O}\big(\sqrt{\log(\delta^{-1})/n}\big)$ instead of $\mathcal{O}(\log(\delta^{-1})/n)$. Notice that it is adaptive in $\|\theta_0\|$, though.

To get faster dimension-free rates for $R_0(\theta)$, we need to introduce some restrictions.

First of all, let us notice that the previous results hold uniformly in any linear subspace of $\mathbb{R}^d$.

PROPOSITION 7.5. *Let us make the same assumptions as in Proposition 7.1 on page 32. For any linear subspace L of $\mathbb{R}^d$, define*

$$\theta_{L,\lambda} \in \arg\min_{\xi \in L} R_\lambda(\xi),$$

$$\widehat{\theta}_{L,\lambda} \in \arg\min_{\xi \in L} \widehat{R}_\lambda(\xi).$$

Let

$$\pi_L \theta = \arg\min_{\xi \in L} \|\xi - \theta\|$$

be the orthogonal projection on L and let

$$\widehat{\Theta}_{L,\lambda} = \left\{ \xi \in L : \|\pi_L(\widehat{G} + \lambda I)(\xi - \widehat{\theta}_{L,\lambda})\| \leq \epsilon\|\xi\| + \eta \right\}$$

and $\quad \widetilde{\theta}_{L,\lambda} \in \arg\min_{\xi \in \widehat{\Theta}_{L,\lambda}} \|\xi\|.$

Finally introduce the least eigenvalue of $\pi_L G \pi_L$

$$\sigma_L = \inf\left\{ \|G\xi\| : \xi \in L, \|\xi\| = 1 \right\}.$$

Whenever equation (4) on page 32 is satisfied, for any linear subspace L of $\mathbb{R}^d$ *and any parameter* $\lambda \in \mathbb{R}_+$,

$$\|\pi_L(G + \lambda I)(\widetilde{\theta}_{L,\lambda} - \theta_{L,\lambda})\|^2 \leq 4(\epsilon\|\widetilde{\theta}_{L,\lambda}\| + \eta)^2$$
$$\leq 4(\epsilon\|\theta_{L,\lambda}\| + \eta)^2,$$
$$\text{and } R_\lambda(\widetilde{\theta}_{L,\lambda}) - R_\lambda(\theta_{L,\lambda}) \leq \frac{4}{\sigma_L + \lambda}(\epsilon\|\widetilde{\theta}_{L,\lambda}\| + \eta)^2$$
$$\leq \frac{4}{\sigma_L + \lambda}(\epsilon\|\theta_{L,\lambda}\| + \eta)^2.$$

Remark that we can estimate σ_L by

$$\widehat{\sigma}_L = \inf\left\{ \|\widehat{G}\xi\| : \xi \in L, \|\xi\| = 1 \right\}.$$

It is such that, for any linear subspace L,

$$\widehat{\sigma}_L - \epsilon \leq \sigma_L \leq \widehat{\sigma}_L + \epsilon.$$

Obtaining a fast convergence rate for the minimization of $R_\lambda(\theta)$ when λ is small or null and σ_d is small is possible in a sparse recovery framework.

PROPOSITION 7.6. *Consider a family* $\mathcal{L}$ *of linear subspaces of* $\mathbb{R}^d$. *Assume that* $\theta_\lambda \in L_* \in \mathcal{L}$ *and that* $\|\theta_\lambda\| \leq A$, *a known constant.*
Consider the confidence region

$$\widehat{\Theta}_\lambda = \left\{ \xi \in \mathbb{R}^d : \|(\widehat{G} + \lambda I)(\xi - \widehat{\theta}_\lambda)\| \leq \epsilon\|\xi\| + \eta, \|\xi\| \leq A \right\}.$$

Define the model selector

$$\widehat{\mathcal{L}} = \Big\{ L \in \mathcal{L} \ : \ \widehat{\Theta}_\lambda \cap L \neq \varnothing \Big\},$$

$$\widehat{L} \in \arg\max\Big\{ \widehat{\sigma}_L \ : \ L \in \widehat{\mathcal{L}} \Big\},$$

and the estimator

$$\widetilde{\theta} \in \arg\min\Big\{ \|\xi\| \ : \ \xi \in \widehat{\Theta}_\lambda \cap \widehat{L} \Big\}.$$

Define

$$\sigma_* = \inf\Big\{ \sigma_{L+\mathbb{R}\theta_\lambda} \ : \ L \in \mathcal{L}, \ \sigma_L \geq \sigma_{L_*} - 2\epsilon \Big\}$$

Under the event described by equation (4) *on page 32,*

$$(\sigma_* + \lambda)\|\widetilde{\theta} - \theta_\lambda\| \leq \|(G + \lambda I)(\widetilde{\theta} - \theta_\lambda)\| \leq 2(\epsilon\|\widetilde{\theta}\| + \eta) \leq 2(\epsilon A + \eta),$$

and

$$R_\lambda(\widetilde{\theta}) - R_\lambda(\theta_\lambda) \leq \frac{4}{\lambda + \sigma_*}(\epsilon\|\widetilde{\theta}\| + \eta)^2 \leq \frac{4}{\lambda + \sigma_*}(\epsilon A + \eta)^2.$$

PROOF. Since $\widetilde{\theta} \in \widehat{\Theta}_\lambda$,

$$\|(G + \lambda I)(\widetilde{\theta} - \theta_\lambda)\| \leq 2(\epsilon\|\widetilde{\theta}\| + \eta) \leq 2(\epsilon A + \eta).$$

On the other hand,

$$\|(G + \lambda I)(\widetilde{\theta} - \theta_\lambda)\| \geq \|\pi_{\widehat{L}+\mathbb{R}\theta_\lambda}(G + \lambda I)(\widetilde{\theta} - \theta_\lambda)\| \geq (\sigma_{\widehat{L}+\mathbb{R}\theta_\lambda} + \lambda)\|\widetilde{\theta} - \theta_\lambda\|.$$

Moreover, $L_* \in \widehat{\mathcal{L}}$, since $\theta_\lambda \in \widehat{\Theta}_\lambda \cap L_* \neq \varnothing$. Thus

$$\sigma_{\widehat{L}} \geq \widehat{\sigma}_{\widehat{L}} - \epsilon \geq \widehat{\sigma}_{L_*} - \epsilon \geq \sigma_{L_*} - 2\epsilon,$$

so that $\sigma_{\widehat{L}+\mathbb{R}\theta_\lambda} \geq \sigma_*$, according to the definition of σ_*, implying that

$$(\sigma_* + \lambda)\|\widetilde{\theta} - \theta_\lambda\| \leq \|(G + \lambda I)(\widetilde{\theta} - \theta_\lambda)\| \leq 2(\epsilon\|\widetilde{\theta}\| + \eta),$$

and consequently that

$$R_\lambda(\widetilde{\theta}) - R_\lambda(\theta_\lambda) \leq \|\widetilde{\theta} - \theta_\lambda\| \, \|(G + \lambda I)(\widetilde{\theta} - \theta_\lambda)\| \leq \frac{4}{\sigma_* + \lambda}(\epsilon\|\widetilde{\theta}\| + \eta)^2.$$

□

Remark that the constant σ_* is defined in terms of restricted eigenvalues of the Gram matrix, a concept that has been used by other authors, for example in Bickel, Ritov and Tsybakov (2009), to set the conditions of sparse recovery.

In the case of nested models, we can replace the constant σ_* with a simpler one, as in the following proposition.

PROPOSITION 7.7. *Consider a nested family of linear subspaces of $\mathbb{R}^d$*

$$\mathcal{L} = \{L_1 \subset L_2 \subset \cdots \subset L_K\}.$$

Assume that $\theta_\lambda \in L_* \in \mathcal{L}$, where L_* is unknown, and that $\|\theta_\lambda\| \leq A$, where A is known. Consider the confidence region

$$\widehat{\Theta}_\lambda = \left\{ \xi \in \mathbb{R}^d \; : \; \|(\widehat{G} + \lambda I)(\xi - \widehat{\theta}_\lambda)\| \leq \epsilon\|\xi\| + \eta, \; \|\xi\| \leq A \right\}.$$

Define the model selector

$$\widehat{k} = \arg\min\left\{ j \; : \; \widehat{\Theta}_\lambda \cap L_j \neq \varnothing \right\},$$
$$\widehat{L} = L_{\widehat{k}},$$

and the estimator

$$\widetilde{\theta} \in \arg\min \left\{ \|\xi\| \; : \; \xi \in \widehat{\Theta}_\lambda \cap \widehat{L} \right\}.$$

Under the event described by equation (4) on page 32,

$$(\sigma_{L_*} + \lambda)\|\widetilde{\theta} - \theta_\lambda\| \leq \|(G + \lambda I)(\widetilde{\theta} - \theta_\lambda)\| \leq 2(\epsilon\|\widetilde{\theta}\| + \eta) \leq 2(\epsilon A + \eta),$$

and

$$R_\lambda(\widetilde{\theta}) - R_\lambda(\theta_\lambda) \leq \frac{4}{\lambda + \sigma_{L_*}}(\epsilon\|\widetilde{\theta}\| + \eta)^2 \leq \frac{4}{\lambda + \sigma_{L_*}}(\epsilon A + \eta)^2.$$

PROOF. As in the previous proposition,

$$\widetilde{\theta} \in \widehat{\Theta}_\lambda,$$

so that $\|(G + \lambda I)(\widetilde{\theta} - \theta_\lambda)\| \leq 2(\epsilon\|\widetilde{\theta}\| + \eta)$. Moreover $L_* \cap \widehat{\Theta}_\lambda \neq \varnothing$, so that $\widehat{L} \subset L_*$, implying that

$$(\sigma_* + \lambda)\|\widetilde{\theta} - \theta_\lambda\| \leq \|\pi_{L_*}(G + \lambda I)(\widetilde{\theta} - \theta_\lambda)\| \leq \|(G + \lambda I)(\widetilde{\theta} - \theta_\lambda)\|$$

and that $R_\lambda(\widetilde{\theta}) - R_\lambda(\theta_\lambda) \leq \dfrac{4}{\sigma_* + \lambda}(\epsilon\|\widetilde{\theta}\| + \eta)^2.$ $\qquad\qquad\square$

References

BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics* **37** 1705–1732.

CATONI, O. (2004). *Statistical Learning Theory and Stochastic Optimization, Lectures on Probability Theory and Statistics, École d'Été de Probabilités de Saint-Flour XXXI – 2001. Lecture Notes in Mathematics* **1851**. Springer pages 1–269.

CATONI, O. (2012). Challenging the empirical mean and empirical variance: a deviation study. *Ann. Inst. Henri Poincaré* **48** 1148-1185.

CATONI, O. (2016). PAC-Bayesian bounds for the Gram matrix and least squares regression with a random design. *preprint on ArXiv*.

CATONI, O. and GIULINI, I. (2017). Dimension free PAC-Bayesian bounds for the estimation of the mean of a random vector. In *NIPS–2017, to appear*.

GIULINI, I. (2017a). Robust dimension-free Gram operator estimates. *Bernoulli, to appear*.

GIULINI, I. (2017b). Robust PCA and pairs of projections in a Hilbert space. *Electron. J. Statist.* **11** 3903–3926.

JOLY, E., LUGOSI, G. and OLIVEIRA, R. I. (2017). On the estimation of the mean of a random vector. *Electronic Journal of Statistics* **11** 440–451.

KATO, T. (1982). *A Short Introduction to Perturbation Theory for Linear Operators*. Springer-Verlag, New York.

LUGOSI, G. and MENDELSON, S. (2017). Sub-Gaussian estimators of the mean of a random vector. *Annals of Statistics, to appear*.

MINSKER, S. (2015). Geometric Median and Robust Estimation in Banach Spaces. *Bernoulli* **4** 2308–2335.

MINSKER, S. (2016). Sub-Gaussian estimators of the mean of a random matrix with heavy-tailed entries. *Annals of Statistics, to appear*.