Robust dimension-free Gram operator estimates

ILARIA GIULINI

Laboratoire de Probabilités et Modèles Aléatoires, Université Paris Diderot, 75013, Paris, France.
E-mail: giulini@math.univ-paris-diderot.fr

In this paper, we investigate the question of estimating the Gram operator by a robust estimator from an i.i.d. sample in a separable Hilbert space and we present uniform bounds that hold under weak moment assumptions. The approach consists in first obtaining non-asymptotic dimension-free bounds in finite-dimensional spaces using some PAC-Bayesian inequalities related to Gaussian perturbations of the parameter and then in generalizing the results in a separable Hilbert space. We show both from a theoretical point of view and with the help of some simulations that such a robust estimator improves the behavior of the classical empirical one in the case of heavy tail data distributions.

Keywords: dimension-free bounds; Gram operator; PAC-Bayesian learning; robust estimation

Contents

1. Introduction

Many algorithms, such as spectral clustering, kernel principal component analysis or more generally kernel-based methods, are based on estimating eigenvalues and eigenvectors of integral

operators defined by a kernel function, from a given random sample. To set the context from a statistical point of view, let $\mu \in \mathcal{M}^1_+(\mathcal{X})$ be an unknown probability distribution on a compact space $\mathcal{X}$ and let k be a kernel on $\mathcal{X}$. The goal is to estimate the integral operator

$$L_k f(x) = \int k(x, z) f(z) \, d\mu(z)$$

from an i.i.d. random sample drawn according to μ.

A first study on the relationship between the spectral properties of a kernel matrix and the corresponding integral operator can be found in [4] for the case of a symmetric square integrable kernel k. They prove that the ordered spectrum of the kernel matrix $K_{ij} = \frac{1}{n} k(X_i, X_j)$ converges to the ordered spectrum of the kernel integral operator L_k. Connections between this empirical matrix and its continuous counterpart have been subject of much research, for example in the framework of kernel-PCA (see [9,10,14]) and spectral clustering (see [13]). In [7], the authors study the connection between the spectral properties of the empirical kernel matrix K_{ij} and those of the corresponding integral operator L_k by introducing two extension operators on the (same) reproducing kernel Hilbert space defined by k, that have the same spectrum (and related eigenfunctions) as K and L_k respectively. In such a way they overcome the difficulty of dealing with objects (K and L_k) operating in different spaces.

The integral operator L_k is related to the Gram operator

$$\mathcal{G}v = \int \langle v, \phi(z) \rangle_{\mathcal{K}} \phi(z) \, d\mu(z), \qquad v \in \mathcal{K},$$

where $\mathcal{K}$ denotes the reproducing kernel Hilbert space defined by the kernel k and ϕ the corresponding feature map.

The main objective of this paper is to estimate Gram operators on (infinite-dimensional) Hilbert spaces. Some bounds on the deviation of the empirical Gram operator from the true Gram operator in separable Hilbert spaces can be found in [5] in the case of Gaussian random vectors.

Let us introduce some notation. We denote by $\mathcal{H}$ a separable Hilbert space and by $P \in \mathcal{M}^1_+(\mathcal{H})$ a (possibly unknown) probability distribution on $\mathcal{H}$. Remark that with the above notation $P = \mu \circ \phi^{-1}$. Our goal is to estimate the Gram operator $\mathcal{G} : \mathcal{H} \to \mathcal{H}$ defined as

$$\mathcal{G}v = \int \langle v, z \rangle_{\mathcal{H}} z \, dP(z)$$

from an i.i.d. sample drawn according to P. Our approach consists in first considering the finite-dimensional setting where X is a random vector in $\mathbb{R}^d$ and then in generalizing the results to the infinite-dimensional case of separable Hilbert spaces. To be able to go from finite to infinite dimension, we will establish dimension-free inequalities. To be more precise, we consider the related problem of estimating the quadratic form

$$N(\theta) = \langle \mathcal{G}\theta, \theta \rangle_{\mathcal{H}}, \qquad \theta \in \mathcal{H}$$

which rewrites explicitly as

$$N(\theta) = \int \langle \theta, z \rangle_{\mathcal{H}}^2 \, \mathrm{d}\mathrm{P}(z).$$

In the finite-dimensional setting, we construct an estimator of the quadratic form $N(\theta)$ and we provide non-asymptotic dimension-free bounds for the approximation error that hold under weak moment assumptions. Note that similar techniques have also been used in [2] in the finite-dimensional setting. However, the results presented here are not comparable to those in [2] as the complexity terms appearing in the bounds are not the same. As a result, in [2] the bounds on the approximation error depend explicitly on the dimension of the ambient space, and grow to $+\infty$ with it. Observe that in the finite-dimensional case the quadratic form $N(\theta)$ can be seen as the quadratic form associated with the Gram matrix

$$G = \int xx^\top \, \mathrm{d}\mathrm{P}(x).$$

Observe also that the study of the Gram matrix is of interest in the case of a non-centered criterion and that it coincides, in the case of centered data (i.e., $\mathbb{E}[X] = 0$), with the study of the covariance matrix

$$\Sigma = \mathbb{E}\big[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^\top\big].$$

Many theoretical results have been published on the estimation of covariance matrices, for example, [8,11,12]. These results follow from the study of random matrix theory and use as an estimator of G the matrix obtained by replacing the unknown probability distribution P with the sample distribution $\frac{1}{n} \sum_{i=1}^n \delta_{X_i}$. In [8], the non-commutative Khintchine inequality is used to obtain bounds on the sample covariance matrix of a bounded random vector. Non-asymptotic results are obtained in [12] as a consequence of the analysis of random matrices with independent rows, while in [11] the author uses an extension of the Bernstein inequality to matrices. However, such an empirical estimator becomes less efficient when the data have a long tail distribution. In [6], the author presents a different estimator based on the geometrical median which is more robust than the classical empirical one. Another way to construct a robust estimator is to use rank-based coefficients (as the Kendall's tau) but this requires strong hypotheses on the distribution.

   We first present a way to construct a robust estimator of the Gram matrix G in finite dimension and then we extend the results to the infinite-dimensional case. Note that the bounds we propose are not formulated in terms of matrix norm, and provide instead bounds for $|\theta^\top (\widehat{G} - G)\theta|$ depending on the direction θ, whereas a bound on the operator norm, for example, would provide a single bound for $\sup_{\theta \in \mathbb{S}_1} |\theta^\top (\widehat{G} - G)\theta|$.

   The paper is organized as follows. Section 2 deals with the finite-dimensional case. We provide a new robust estimator of the Gram matrix G and we use a PAC-Bayesian approach to obtain non-asymptotic dimension-independent bounds of its approximation error. In Section 3, we extend the results to the infinite-dimensional case, taking advantage of the fact that they are independent of the dimension of the ambient space. In Section 4, we propose some empirical results to show the performance of our estimator. In Section 5, we compare from a theoretical point of view the behavior of our robust estimator to the one of the classical empirical estimator. Finally in

Section 6, we extend the results to estimate the expectation of a symmetric random matrix and we consider the problem of estimating the covariance matrix in the case when the expectation is unknown.

2. The finite-dimensional setting

Let $P \in \mathcal{M}_+^1(\mathbb{R}^d)$ be an unknown probability distribution on $\mathbb{R}^d$ and let $X \in \mathbb{R}^d$ be a random vector of law P. We denote by $\mathbb{E}$ the expectation with respect to P. Our goal is to estimate the quadratic form

$$N(\theta) = \mathbb{E}[\langle \theta, X \rangle^2], \qquad \theta \in \mathbb{R}^d$$

(that computes the energy in the direction θ) from an i.i.d. sample $X_1, \ldots, X_n \in \mathbb{R}^d$ drawn according to P. Observe that $N(\theta)$ can be seen as the quadratic form associated to the Gram matrix

$$G = \mathbb{E}[XX^\top].$$

Indeed to recover the Gram matrix G from the above quadratic form it is sufficient to use the polarization identity

$$
\begin{aligned}
G_{ij} &= e_i^\top G e_j \\
&= \frac{1}{4}\big[(e_i + e_j)^\top G(e_i + e_j) - (e_i - e_j)^\top G(e_i - e_j)\big] \\
&= \frac{1}{4}\big[N(e_i + e_j) - N(e_i - e_j)\big],
\end{aligned}
$$

where $\{e_i\}_{i=1}^d$ is the canonical basis of $\mathbb{R}^d$.

A classical empirical estimator of the quadratic form $N(\theta)$ is

$$\bar{N}(\theta) = \frac{1}{n} \sum_{i=1}^n \langle \theta, X_i \rangle^2$$

obtained by replacing the unknown probability distribution P with the sample distribution. However, as shown in [1], if the distribution of $\langle \theta, X \rangle^2$ has a heavy tail for some values of θ, the quality of the approximation provided by the classical empirical estimator can be improved, using some M-estimator with a suitable smooth influence function and a scale parameter depending on the sample size. Thus in order to construct a robust estimator for N, we consider, for any $\theta \in \mathbb{R}^d$ and any $\lambda > 0$,

$$r_\lambda(\theta) = \frac{1}{n} \sum_{i=1}^n \psi(\langle \theta, X_i \rangle^2 - \lambda), \tag{2.1}$$

where the function $\psi : \mathbb{R} \to \mathbb{R}$, defined as

$$\psi(t) = \begin{cases} \log(2) & \text{if } t \geq 1, \\ -\log\left(1 - t + \dfrac{t^2}{2}\right) & \text{if } 0 \leq t \leq 1, \\ -\psi(-t) & \text{if } t \leq 0, \end{cases} \tag{2.2}$$

is symmetric non-decreasing, bounded, and satisfies

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

Introduce

$$\widehat{\alpha}_\lambda(\theta) = \sup\{\alpha \in \mathbb{R}_+ \mid r_\lambda(\alpha\theta) \leq 0\}. \tag{2.3}$$

In order to simplify notation, in the following we omit the dependence on λ and we write $\widehat{\alpha}$ instead of $\widehat{\alpha}_\lambda$.

Observe that, since the function $\alpha \mapsto r_\lambda(\alpha\theta)$ is continuous, $r_\lambda(\widehat{\alpha}(\theta)\theta) = 0$ as soon as $\widehat{\alpha}(\theta) < +\infty$. Moreover, since the function ψ is close to the identity in a neighborhood of the origin,

$$0 = r_\lambda\big(\widehat{\alpha}(\theta)\theta\big) \simeq \widehat{\alpha}(\theta)^2 \bar{N}(\theta) - \lambda$$

and therefore it is natural to consider as an estimator of $N(\theta)$ a quantity related to $\lambda/\widehat{\alpha}(\theta)^2$, for a suitable value of λ. We consider the family of (robust) estimators

$$\widetilde{N}_\lambda(\theta) = \frac{\lambda}{\widehat{\alpha}(\theta)^2} \tag{2.4}$$

and we observe that, since $\widehat{\alpha}(\theta)$ is homogeneous of degree -1 in θ,

$$\widetilde{N}_\lambda(\theta) = \|\theta\|^2 \widetilde{N}_\lambda\big(\theta/\|\theta\|\big).$$

In the following, we will use a PAC-Bayesian approach linked to Gaussian perturbations of the parameter θ to first construct a confidence region for $N(\theta)$ and then define and study a robust estimator by choosing a suitable value $\widehat{\lambda}$ for the parameter λ.

Given $\theta \in \mathbb{R}^d$, we consider the family of Gaussian perturbations $\pi_\theta \sim \mathcal{N}(\theta, \beta^{-1}\mathrm{I})$ of mean θ and covariance matrix $\beta^{-1}\mathrm{I}$ where $\beta > 0$ is a free parameter.

Let $\Lambda \subset (\mathbb{R}_+ \setminus \{0\})^2$ be a finite set of possible values of the couple of parameters (λ, β) and $|\Lambda|$ its cardinality. Let us introduce

$$s_4 = \mathbb{E}\big[\|X\|^4\big]^{1/4} \quad \text{and} \quad \kappa = \sup_{\substack{\theta \in \mathbb{R}^d \\ \mathbb{E}[\langle\theta, X\rangle^2] > 0}} \frac{\mathbb{E}[\langle\theta, X\rangle^4]}{\mathbb{E}[\langle\theta, X\rangle^2]^2} \tag{2.5}$$

assuming that these two quantities are finite. [1] We will prove later on that

$$s_4^2 \leq \kappa^{1/2} \mathbb{E}\big(\|X\|^2\big) = \kappa^{1/2} \mathbf{Tr}(G).$$

Note that in the case where the probability distribution P is Gaussian, $\kappa = 3$.

For any $(\lambda, \beta) \in \Lambda$ and $\varepsilon > 0$, we put

$$
\begin{aligned}
\xi &= \frac{\kappa \lambda}{2}, \\
\mu &= \lambda(\kappa - 1) + \frac{(2+c)\kappa^{1/2} s_4^2}{\beta}, \\
\gamma &= \frac{\lambda}{2}(\kappa - 1) + \frac{(2+c)\kappa^{1/2} s_4^2}{\beta} + \frac{(2+3c)s_4^4}{2\beta^2 \lambda} + \frac{\log(|\Lambda|/\varepsilon)}{n\lambda}, \\
\delta &= \frac{\beta}{2n\lambda},
\end{aligned}
\tag{2.6}
$$

where

$$c = \frac{15}{8\log(2)(\sqrt{2}-1)} \exp\left(\frac{1+2\sqrt{2}}{2}\right) \leq 44.3. \tag{2.7}$$

Proposition 2.1. *With probability at least $1 - 2\varepsilon$, for any $\theta \in \mathbb{R}^d$, any $(\lambda, \beta) \in \Lambda$,*

$$\Phi_{\theta,-}\left(\frac{\lambda}{\widehat{\alpha}(\theta)^2}\right) \leq N(\theta) \leq \Phi_{\theta,+}^{-1}\left(\frac{\lambda}{\widehat{\alpha}(\theta)^2}\right),$$

where $\Phi_{\theta,-}$ and $\Phi_{\theta,+}$ are non-decreasing functions defined as

$$\Phi_{\theta,-}(t) = t\left(1 - \frac{\gamma + \delta\lambda\|\theta\|^2/t}{1 + \mu - \gamma - \delta\lambda\|\theta\|^2/t}\right)\mathbb{1}\big[\xi - \mu + 2\gamma + 2\delta\lambda\|\theta\|^2/t < 1\big],$$

$$\Phi_{\theta,+}(t) = t\left(1 + \frac{\gamma + \delta\lambda\|\theta\|^2/t}{1 - \mu - \gamma - 2\delta\lambda\|\theta\|^2/t}\right)^{-1}\mathbb{1}\big[\xi + \mu + \gamma + 2\delta\lambda\|\theta\|^2/t < 1\big],$$

and where $\Phi_{\theta,+}^{-1}(u) = \sup\{t \in \mathbb{R}_+ : \Phi_{\theta,+}(t) \leq u\}$.

For the proof, we refer to Section 7.1.

Observe that since those functions depend on θ only through $\|\theta\|$, if θ is such that $\|\theta\| = 1$ it is natural to omit the dependence of θ and write Φ_- and Φ_+. In the following we will omit the dependence of θ of the functions defined in Proposition 2.1, so that we write Φ_- and Φ_+ instead of $\Phi_{\theta,-}$ and $\Phi_{\theta,+}$.

---

[1] As it will be explained later, it is sufficient to know upper bounds for these quantities since the following results still hold true replacing s_4 and κ by upper bounds.

Proposition 2.2. *Let $\sigma \in \mathbb{R}_+$ be any energy level. We consider the set*

$$\Gamma = \left\{ (\lambda, \beta, t) \in \Lambda \times \mathbb{R}_+ \mid \xi + \mu + \gamma + 2\frac{\delta\lambda}{\max\{t, \sigma\}} < 1 \right\}$$

and the bound

$$B_{\lambda,\beta}(t) = \begin{cases} \dfrac{\gamma + \lambda\delta/\max\{t, \sigma\}}{1 - \mu - \gamma - 2\lambda\delta/\max\{t, \sigma\}}, & (\lambda, \beta, t) \in \Gamma, \\ +\infty & \textit{otherwise.} \end{cases} \quad (2.8)$$

With probability at least $1 - 2\varepsilon$, for any $\theta \in \mathbb{R}^d$, any $(\lambda, \beta) \in \Lambda$,

$$\left| \frac{\max\{N(\theta), \sigma\|\theta\|^2\}}{\max\{\widetilde{N}_\lambda(\theta), \sigma\|\theta\|^2\}} - 1 \right| \leq B_{\lambda,\beta}\left[\|\theta\|^{-2} \widetilde{N}_\lambda(\theta) \right].$$

Proof. We observe that, for any $z, t, \sigma \in \mathbb{R}_+$, if $\Phi_+(z) \leq t$ then

$$\Phi_+\left(\max\{z, \sigma\}\right) \leq \max\{t, \sigma\}$$

since it is clear from the definition of Φ_+ that $\Phi_+(\sigma) \leq \sigma$. Similarly if $\Phi_-(z) \leq t$, then

$$\Phi_-\left(\max\{z, \sigma\}\right) \leq \max\{t, \sigma\}.$$

Thus, according to the definition of $B_{\lambda,\beta}$ in equation (2.8), we get

$$\Phi_+\left(\max\{z, \sigma\}\right) = \max\{z, \sigma\}\left(1 + B_{\lambda,\beta}(z)\right)^{-1}, \quad (2.9)$$

$$\Phi_-\left(\max\{z, \sigma\}\right) \geq \max\{z, \sigma\}\left(1 - B_{\lambda,\beta}(z)\right). \quad (2.10)$$

Therefore

$$\begin{aligned} \Phi_+^{-1}\left(\max\{z, \sigma\}\right) &= \sup\{t : \Phi_+(t) \leq \max\{z, \sigma\}\} \\ &= \sup\{t \geq z : \Phi_+(t) \leq \max\{z, \sigma\}\} \\ &= \sup\{t \geq z : \left(1 + B_{\lambda,\beta}(t)\right)^{-1}\max\{z, \sigma\}\} \\ &\leq \sup\{t \geq z : \left(1 + B_{\lambda,\beta}(z)\right)^{-1}t \leq \max\{z, \sigma\}\} \\ &= \max\{z, \sigma\}\left(1 + B_{\lambda,\beta}(z)\right), \end{aligned}$$

where we have have used the fact that $B_{\lambda,\beta}$ is non-decreasing. In view of these inequalities, the present proposition is a consequence of the previous one. □

From now on we fix the finite set Λ of all possible values of the couple (λ, β) as

$$\Lambda = \left\{ (\lambda_j, \beta_j) \mid 0 \leq j < K \right\}, \quad (2.11)$$

where $K = 1 + \lceil a^{-1}\log(\frac{n}{72(2+c)\kappa^{1/2}}) \rceil$, with $a > 0$ and c defined in equation (2.7), and

$$\lambda_j = \sqrt{\frac{2}{n(\kappa-1)}\left(\frac{(2+3c)}{4(2+c)\kappa^{1/2}\exp(-ja)} + \log(K/\varepsilon)\right)},$$

$$\beta_j = \sqrt{2(2+c)\kappa^{1/2}s_4^4 n \exp\left[-(j-1/2)a\right]}.$$

We put

$$(\widehat{\lambda}, \widehat{\beta}) = \arg\min_{(\lambda,\beta)\in\Lambda} B_{\lambda,\beta}\left[\|\theta\|^{-2}\widetilde{N}_\lambda(\theta)\right]$$

and we define our robust estimator as

$$\widehat{N}(\theta) = \widetilde{N}_{\widehat{\lambda}}(\theta). \tag{2.12}$$

Proposition 2.3. *Let us fix a threshold* $\sigma \le s_4^2$ *and set the value of the parameter* a *used to define* Λ *to* $1/2$. *Introduce*

$$\zeta_*(t) = \sqrt{2.032(\kappa-1)\left(\frac{0.73\,\mathbf{Tr}(G)}{t} + \log(K) + \log(\varepsilon^{-1})\right)} + \sqrt{\frac{98.5\kappa\,\mathbf{Tr}(G)}{t}}, \qquad t \in \mathbb{R}_+,$$

where $\mathbf{Tr}(G) = \mathbb{E}[\|X\|^2]$ *denotes the trace of the Gram matrix, and*

$$B_*(t) = \begin{cases} \dfrac{n^{-1/2}\zeta_*(\max\{t,\sigma\})}{1 - 4n^{-1/2}\zeta_*(\max\{t,\sigma\})}, & \left[6 + (\kappa-1)^{-1}\right]\zeta_*\left(\max\{t,\sigma\}\right) \le \sqrt{n}, \\ +\infty & \text{otherwise.} \end{cases}$$

With probability at least $1 - 2\varepsilon$, *for any* $\theta \in \mathbb{R}^d$,

$$\left|\frac{\max\{N(\theta), \sigma\|\theta\|^2\}}{\max\{\widehat{N}(\theta), \sigma\|\theta\|^2\}} - 1\right| \le B_*\left[\|\theta\|^{-2}N(\theta)\right].$$

For the proof, we refer to Section 7.2. Remark that equation (7.20) of the proof provides a bound for any choice of the parameter $a > 0$ and that we report here only the numerical value of the bound when $a = 1/2$ for the sake of simplicity.

Assuming any reasonable bound on the sample size we can bound the logarithmic factor $\log\log(n)$ hidden in $\log(K)$ with a relatively small constant. In particular, if we choose $n \le 10^{20}$, we get $\log(K) \le 4.35$.

In order to provide a more friendly version of the above bound, we introduce here the $\mathcal{O}$ notation, where $A = \mathcal{O}(B)$ means that there exists a numerical constant τ such that $A \le \tau B$. Thus, the result stated in Proposition 2.3 becomes, with probability at least $1 - 2\varepsilon$, for any $\theta \in \mathbb{R}^d$,

$$\left|\frac{\max\{N(\theta), \sigma\|\theta\|^2\}}{\max\{\widehat{N}(\theta), \sigma\|\theta\|^2\}} - 1\right| \le \mathcal{O}\left(\sqrt{\frac{\kappa}{n}\left(\frac{\mathbf{Tr}(G)}{\max\{\|\theta\|^{-2}N(\theta), \sigma\}} + \log\left(\log(n)/\varepsilon\right)\right)}\right).$$

Observe that the bound B_*, unlike the bound provided in [2], does not depend explicitly on the dimension d of the ambient space. More specifically, the dimension has been replaced by the effective dimension coming from the entropy term of the PAC-Bayesian bound

$$\mathbf{Tr}(G)/\max\{\|\theta\|^{-2}N(\theta), \sigma\}.$$

To get an intuition of why this entropy term replaces the dimension, it is sufficient to consider the case where the energy N is uniformly distributed, so that $N(\theta) = N_* \geq \sigma$ for any $\theta \in \mathbb{R}^d$ with $\|\theta\| = 1$. In this case indeed

$$\frac{\mathbf{Tr}(G)}{\max\{\|\theta\|^{-2}N(\theta), \sigma\}} = \frac{\sum_{i=1}^d N(v_i)}{N_*} = d,$$

where v_i denotes an orthonormal basis of eigenvectors of G. In summary, when the Gram matrix G to be estimated is proportional to the identity matrix, then our bound coincides (up to some moderate increase in the constants) with the bound proved in [2], but when the eigenvalues of G are decreasing, then our bound balances the complexity in a different way and is looser when θ is in the span of eigenvectors with low eigenvalues, but tighter when θ is in the span of eigenvectors with high eigenvalues.

Let us also remark that the variance $\mathbf{Var}(\langle\theta, X\rangle^2)$ of $\langle\theta, X\rangle^2$ is related to κ by the relation

$$\kappa = 1 + \sup\left\{\frac{\mathbf{Var}(\langle\theta, X\rangle^2)}{\mathbb{E}(\langle\theta, X\rangle^2)^2} : \theta \in \mathbb{R}^d, \mathbb{E}(\langle\theta, X\rangle^2) > 0\right\}.$$

Moreover, we do not need to know the exact values of κ and $\mathbf{Tr}(G) = \mathbb{E}(\|X\|^2)$ to compute the estimator and evaluate the bound, it is sufficient to know upper bounds instead. Indeed, if we use those upper bounds to define our estimator, the above result is still true with κ and $\mathbf{Tr}(G)$ replaced by their upper bounds.

We also observe that in order to have a meaningful (finite) bound we can choose the threshold σ such that

$$8\zeta_*(\sigma) \leq \sqrt{n} \tag{2.13}$$

so that $B_*(t) < +\infty$ for any $t \in \mathbb{R}_+$, assuming that $\kappa \geq 3/2$. More precisely, using the inequality $(\sqrt{a} + \sqrt{b})^2 \leq 2(a + b)$, we see that equation (2.13) holds when

$$\sigma = \frac{100\kappa\,\mathbf{Tr}(G)}{n/128 - 4.35 - \log(\varepsilon^{-1})}.$$

With this choice the threshold σ decays to zero at speed $1/n$ as the sample size grows to infinity.

Remark that the estimator $\hat{N}$ is not necessarily a quadratic form. We conclude this section by introducing and studying a quadratic estimator of N, that is an estimator of the form $\theta^\top Q\theta$, where Q is an estimate of the Gram matrix G.

We observe that Proposition 2.1 provides a confidence region for $N(\theta)$. Define

$$B_-(\theta) = \max_{(\lambda,\beta)\in\Lambda} \Phi_-\big(\widetilde{N}_\lambda(\theta)\big) \quad \text{and} \quad B_+(\theta) = \min_{(\lambda,\beta)\in\Lambda} \Phi_+^{-1}\big(\widetilde{N}_\lambda(\theta)\big),$$

where we recall that $\widetilde{N}_\lambda(\theta) = \frac{\lambda}{\widehat{\alpha}(\theta)^2}$ and Λ is defined in equation (2.11). According to Proposition 2.1, with probability at least $1 - 2\varepsilon$, for any $\theta \in \mathbb{R}^d$,

$$B_-(\theta) \leq N(\theta) \leq B_+(\theta). \tag{2.14}$$

From a theoretical point of view we can consider as an estimator of N any quadratic form belonging to the confidence interval $[B_-(\theta), B_+(\theta)]$ for any θ. Such a quadratic form exists with probability at least $1 - 2\varepsilon$ according to equation (2.14). However, from an algorithmic point of view, we would like to impose these constraints only for a finite number of directions θ. In particular, in the following we are going to study the properties of a symmetric matrix Q that satisfies $\mathbf{Tr}(Q^2) \leq \mathbf{Tr}(G^2)$ and

$$B_-(\theta) \leq \theta^\top Q \theta \leq B_+(\theta), \qquad \theta \in \Theta_\delta,$$

where Θ_δ is any finite δ-net of the unit sphere $\mathbb{S}_d = \{\theta \in \mathbb{R}^d, \|\theta\| = 1\}$, meaning that

$$\sup_{\theta \in \mathbb{S}_d} \min_{\xi \in \Theta_\delta} \|\theta - \xi\| \leq \delta.$$

The matrix Q can be computed using a convex optimization algorithm as described in Section 1.2.4 of [3].

Note that a more straightforward choice would have been to set

$$Q_{i,j} = \frac{1}{4}\big[\widehat{N}(e_i + e_j) - \widehat{N}(e_i - e_j)\big], \tag{2.15}$$

where $\{e_i\}_{i=1}^d$ denotes the canonical basis of $\mathbb{R}^d$. Unfortunately this simple choice is not adequate to control the approximation error independently of the dimension d in all directions. To get a dimension-free bound we need an estimator that behaves well in a far larger set of directions Θ_δ than the d^2 directions $e_i \pm e_j$.

From now on let $\sigma \in]0, s_4^2]$ be a threshold such that $8\zeta_*(\sigma) \leq \sqrt{n}$. The next proposition provides the analogous for the quadratic form $\theta^\top Q \theta$ of the dimension-free bound presented in Proposition 2.3 for $\widehat{N}(\theta)$.

Proposition 2.4. *With the same notation as in Proposition 2.3, with probability at least $1 - 2\varepsilon$, for any $\theta \in \mathbb{S}_d$,*

$$\big|\max\{\theta^\top Q \theta, \sigma\} - \max\{N(\theta), \sigma\}\big|$$
$$\leq 2\max\{N(\theta), \sigma\} B_*(N(\theta)) + 5\delta\sqrt{\mathbf{Tr}(G^2)},$$
$$\big|\max\{\theta^\top Q \theta, \sigma\} - \max\{N(\theta), \sigma\}\big|$$
$$\leq 2\max\{\theta^\top Q \theta, \sigma\} B_*(\min\{\theta^\top Q \theta, s_4^2\}) + 5\delta\sqrt{\mathbf{Tr}(G^2)}.$$

Remark that the parameter δ of the net Θ_δ governs the computation cost of Q. Thus, we can in theory (that is if we have an arbitrarily fast computer at our disposal), take δ as close to zero as we want.

Proof. Since for any $\theta \in \mathbb{S}_d$ there is $\xi \in \Theta_\delta$ such that $\|\theta - \xi\| \leq \delta$, we have

$$
\begin{aligned}
\left| \theta^\top Q \theta - \xi^\top Q \xi \right| &= (\theta + \xi)^\top Q (\theta - \xi) \\
&\leq \|\theta + \xi\| \|Q\|_\infty \|\theta - \xi\| \leq 2\delta \sqrt{\mathbf{Tr}(Q^2)} \leq 2\delta \sqrt{\mathbf{Tr}(G^2)}.
\end{aligned}
\tag{2.16}
$$

Let us put $\eta = 2\delta \sqrt{\mathbf{Tr}(G^2)}$. We observe that, with probability at least $1 - 2\varepsilon$,

$$
\begin{aligned}
\Phi_- \circ \Phi_+ \left(\theta^\top Q \theta - \eta \right) &\leq N(\theta) + \eta, \\
\Phi_- \circ \Phi_+ \left(N(\theta) - \eta \right) &\leq \theta^\top Q \theta + \eta,
\end{aligned}
\tag{2.17}
$$

where Φ_+ and Φ_- are defined in Proposition 2.1 and depend on θ only through $\|\theta\|$. Indeed, in the event of probability at least $1 - 2\varepsilon$ described in equation (2.14),

$$
\theta^\top Q \theta \leq \Phi_+^{-1}\left(\widetilde{N}_\lambda(\xi) \right) + \eta \leq \Phi_+^{-1} \circ \Phi_-^{-1}\left(N(\xi) \right) + \eta \leq \Phi_+^{-1} \circ \Phi_-^{-1}\left(N(\theta) + \eta \right) + \eta,
$$

since equation (2.16) also holds for N, and in the same way we get

$$
\theta^\top Q \theta \geq \Phi_-\left(\widetilde{N}_\lambda(\xi) \right) - \eta \geq \Phi_- \circ \Phi_+\left(N(\xi) \right) - \eta \geq \Phi_- \circ \Phi_+\left(N(\theta) - \eta \right) - \eta
$$

which proves equation (2.17). We conclude the proof using Corollary 7.1 on page 3912 in Section 7.4 $\qquad \square$

Note that the estimated matrix Q of the previous proposition is not necessarily positive semidefinite. We can remedy that shortcoming by considering instead its positive part Q_+ (obtained by taking the positive part of its eigenvalues in the framework of functional calculus on symmetric matrices).

Based on the fact that $\theta^\top Q \theta \geq B_-(\theta) \geq 0$ on Θ_δ and on equation (2.16), stating that for any $\theta \in \mathbb{S}_d$,

$$
\left| \theta^\top Q \theta - \xi^\top Q \xi \right| \leq 2\delta \sqrt{\mathbf{Tr}(G^2)},
$$

where $\xi \in \Theta_\delta$ is such that $\|\theta - \xi\| \leq \delta$, we can see that we do not loose much when replacing Q by Q_+.

Proposition 2.5. *With probability at least $1 - 2\varepsilon$, for any $\theta \in \mathbb{S}_d$,*

$$
\left| \max\left\{ \theta^\top Q_+ \theta, \sigma \right\} - \max\left\{ N(\theta), \sigma \right\} \right|
$$

$$
\leq 2 \max\left\{ N(\theta), \sigma \right\} B_*\left(N(\theta) \right) + 7\delta \sqrt{\mathbf{Tr}(G^2)},
$$

$$
\left| \max\left\{ \theta^\top Q_+ \theta, \sigma \right\} - \max\left\{ N(\theta), \sigma \right\} \right|
$$

$$\leq 2 \max\{\theta^\top Q_+\theta, \sigma\} B_*\big(\min\{\theta^\top Q_+\theta, s_4^2\}\big) + 7\delta\sqrt{\mathbf{Tr}(G^2)},$$

where B_* is defined in Proposition 2.3.

Proof. Let us put as before $\eta = 2\delta\sqrt{\mathbf{Tr}(G^2)}$. For any $\theta \in \mathbb{S}_d$, there is $\xi \in \Theta_\delta$ such that $\|\theta - \xi\| \leq \delta$, so that, according to equation (2.16),

$$\theta^\top Q\theta \geq \xi^\top Q\xi - \eta \geq -\eta.$$

Then we deduce that

$$\|Q_-\|_\infty = \sup_{\theta \in \mathbb{S}_d} \theta^\top Q_-\theta = -\inf_{\theta \in \mathbb{S}_d} \theta^\top Q\theta \leq \eta.$$

Therefore, for any $\theta \in \mathbb{S}_d$,

$$\big|\max\{\theta^\top Q\theta, \sigma\} - \max\{\theta^\top Q_+\theta, \sigma\}\big| \leq \big|\theta^\top Q\theta - \theta^\top Q_+\theta\big| = \theta^\top Q_-\theta \leq \eta.$$

Combining the above equation with Proposition 2.4 we conclude the proof. $\qquad\square$

Since for any $a, b \in \mathbb{R}_+$, $a - b - \sigma \leq \max\{a, \sigma\} - \max\{b, \sigma\}$, we obtain as a consequence:

Corollary 2.1. *With probability at least $1 - 2\varepsilon$, for any $\theta \in \mathbb{S}_d$,*

$$\big|\theta^\top Q_+\theta - N(\theta)\big| \leq 2\max\{N(\theta), \sigma\} B_*\big(N(\theta)\big) + 7\delta\sqrt{\mathbf{Tr}(G^2)} + \sigma,$$

$$\big|\theta^\top Q_+\theta - N(\theta)\big| \leq 2\max\{\theta^\top Q_+\theta, \sigma\} B_*\big(\min\{\theta^\top Q_+\theta, s_4^2\}\big) + 7\delta\sqrt{\mathbf{Tr}(G^2)} + \sigma.$$

To conclude, we mention that it is possible to get similar dimension-free bounds under light tail hypotheses for the classical empirical estimator

$$\bar{G} = \frac{1}{n}\sum_{i=1}^{n} X_i X_i^\top.$$

For more details, we refer to Section 5.

3. The infinite-dimensional setting

In this section, we extend the results obtained in the previous section to the infinite-dimensional setting.

Let $\mathcal{H}$ be a separable Hilbert space and let $\mathrm{P} \in \mathcal{M}_+^1(\mathcal{H})$ be an unknown probability distribution on $\mathcal{H}$. We consider the Gram operator $\mathcal{G} : \mathcal{H} \to \mathcal{H}$ defined by

$$\mathcal{G}\theta = \int \langle \theta, v\rangle_\mathcal{H}\, v\, \mathrm{dP}(v).$$

and we assume $\mathbf{Tr}(\mathcal{G}) = \mathbb{E}(\|X\|_{\mathcal{H}}^2) < +\infty$, where $X \in \mathcal{H}$ denotes a random vector with law P. In analogy to the previous section we denote by N the quadratic form associated with the Gram operator $\mathcal{G}$ so that

$$N(\theta) = \langle \mathcal{G}\theta, \theta \rangle_{\mathcal{H}} = \int \langle \theta, v \rangle_{\mathcal{H}}^2 \, \mathrm{dP}(v), \qquad \theta \in \mathcal{H}.$$

We consider $(\mathcal{H}_k)_k$ an increasing sequence of subspaces of $\mathcal{H}$ of finite dimensions such that $\bigcup_k \mathcal{H}_k = \mathcal{H}$. For instance, if $\{e_i, i \in \mathbb{N}^*\}$ is an orthonormal basis of $\mathcal{H}$, we can take $\mathcal{H}_k = \mathbf{span}\{e_1, \ldots, e_k\}$. To give a more concrete example, in the case when $\mathcal{H} = \mathbb{L}^2([0, 1])$, we can take for e_i the Fourier basis. In this example, the orthogonal projector Π_k on $\mathcal{H}_k$ is given by

$$\Pi_k v = \sum_{i=1}^{k} \langle v, e_i \rangle_{\mathcal{H}} e_i, \qquad v \in \mathcal{H}.$$

We denote by N_k the quadratic form in $\mathcal{H}_k$ associated with the probability distribution of $\Pi_k X$, so that

$$N_k(\theta) = \mathbb{E}\big(\langle \theta, \Pi_k X \rangle^2\big) = N(\theta), \qquad \theta \in \mathcal{H}_k.$$

Remark that for any $\theta \in \mathcal{H}$,

$$N_k(\Pi_k \theta) = N(\Pi_k \theta).$$

In the following we consider an i.i.d. sample of size n in $\mathcal{H}$ drawn according to P. According to Proposition 2.1, the event

$$\mathcal{A}_k = \left\{ \forall \theta \in \mathcal{H}_k, \forall (\lambda, \beta) \in \Lambda, \ \Phi_{\theta,-}\left(\frac{\lambda}{\widehat{\alpha}(\theta)^2}\right) \leq N(\theta) \leq \Phi_{\theta,+}^{-1}\left(\frac{\lambda}{\widehat{\alpha}(\theta)^2}\right) \right\}$$

is such that $\mathrm{P}^{\otimes n}(\mathcal{A}_k) \geq 1 - 2\varepsilon$. Since $\mathcal{A}_{k+1} \subset \mathcal{A}_k$, by the continuity of measure,

$$\mathrm{P}^{\otimes n}\left(\bigcap_{k \in \mathbb{N}} \mathcal{A}_k\right) \geq 1 - 2\varepsilon.$$

This means that with probability at least $1 - 2\varepsilon$, for any $\theta \in \bigcup_k \mathcal{H}_k$ and any $(\lambda, \beta) \in \Lambda$,

$$\Phi_{\theta,-}\left(\frac{\lambda}{\widehat{\alpha}(\theta)^2}\right) \leq N(\theta) \leq \Phi_{\theta,+}^{-1}\left(\frac{\lambda}{\widehat{\alpha}(\theta)^2}\right).$$

Consequently, since $N(\theta) = \lim_{k \to +\infty} N(\Pi_k(\theta))$, for any $\theta \in \mathcal{H}$, the following result holds.

Proposition 3.1. *With probability at least $1 - 2\varepsilon$, for any $\theta \in \mathcal{H}$,*

$$B_-(\theta) \leq N(\theta) \leq B_+(\theta),$$

where

$$B_-(\theta) = \lim_{k \to +\infty} \sup_{(\lambda,\beta) \in \Lambda} \Phi_{\Pi_k\theta,-}\left(\frac{\lambda}{\widehat{\alpha}(\Pi_k\theta)^2}\right),$$

$$B_+(\theta) = \lim_{k \to +\infty} \inf_{(\lambda,\beta) \in \Lambda} \Phi^{-1}_{\Pi_k\theta,+}\left(\frac{\lambda}{\widehat{\alpha}(\Pi_k\theta)^2}\right).$$

If we do not want to go to the limit, we can use the explicit bound

$$\left|N(\theta) - N(\Pi_k(\theta))\right| = \left|\langle \theta + \Pi_k(\theta), \mathcal{G}(\theta - \Pi_k(\theta))\rangle_{\mathcal{H}}\right|$$

$$\leq 2\|\theta\|_{\mathcal{H}}\|\mathcal{G}\|_\infty\|\theta - \Pi_k(\theta)\|_{\mathcal{H}}$$

$$\leq 2\|\theta\|_{\mathcal{H}}\,\mathbf{Tr}(\mathcal{G})\|\theta - \Pi_k(\theta)\|_{\mathcal{H}}$$

$$= 2\|\theta\|_{\mathcal{H}}\mathbb{E}\left(\|X\|^2_{\mathcal{H}}\right)\|\theta - \Pi_k(\theta)\|_{\mathcal{H}}.$$

This bound depends on $\|\theta - \Pi_k\theta\|_{\mathcal{H}}$. We will see in the following another bound that goes uniformly to zero for any $\theta \in \mathbb{S}_{\mathcal{H}}$ when k tends to infinity. In the same way, proceeding as already done in the previous section we state the analogous of Proposition 2.3.

Let

$$\kappa \geq \sup_{\substack{\theta \in \mathcal{H} \\ \mathbb{E}(\langle\theta,X\rangle^2_{\mathcal{H}})>0}} \frac{\mathbb{E}(\langle\theta,X\rangle^4_{\mathcal{H}})}{\mathbb{E}(\langle\theta,X\rangle^2_{\mathcal{H}})^2} \quad \text{and} \quad s_4 \geq \mathbb{E}\left(\|X\|^4_{\mathcal{H}}\right)^{1/4}$$

be known constants and put

$$K = 1 + \left\lceil 2\log\left(\frac{n}{72(2+c)\kappa^{1/2}}\right)\right\rceil,$$

where

$$c = \frac{15}{8\log(2)(\sqrt{2}-1)}\exp\left(\frac{1+2\sqrt{2}}{2}\right)$$

as in equation (2.7). Define

$$\zeta_*(t) = \sqrt{2.032(\kappa - 1)\left(\frac{0.73\,\mathbf{Tr}(\mathcal{G})}{t} + \log(K) + \log(\varepsilon^{-1})\right)} + \sqrt{\frac{98.5\kappa\,\mathbf{Tr}(\mathcal{G})}{t}}$$

and consider, according to equation (2.12), the estimators

$$\widehat{N}_k(\theta) = \widetilde{N}_{\widehat{\lambda}}(\theta), \qquad \theta \in \mathcal{H}_k.$$

For any $\theta \in \mathcal{H}$, define $\widehat{N}_{\mathcal{H}}(\theta)$ by choosing any limit point of $\widehat{N}_k(\Pi_k\theta)$, such as for example, $\liminf_{k \to \infty} \widehat{N}_k(\Pi_k\theta)$.

Proposition 3.2. *Define the bound*

$$B_*(t) = \frac{n^{-1/2}\zeta_*(\max\{t,\sigma\})}{1 - 4n^{-1/2}\zeta_*(\max\{t,\sigma\})},$$

where $\sigma \in {]0, s_4^2]}$ *is some energy level such that*

$$[6 + (\kappa - 1)^{-1}]\zeta_*(\sigma) \le \sqrt{n}.$$

With probability at least $1 - 2\varepsilon$, *for any* $\theta \in \mathcal{H}$,

$$\left| \frac{\max\{N(\theta), \sigma\|\theta\|_{\mathcal{H}}^2\}}{\max\{\widehat{N}_{\mathcal{H}}(\theta), \sigma\|\theta\|_{\mathcal{H}}^2\}} - 1 \right| \le B_*[\|\theta\|_{\mathcal{H}}^{-2}N(\theta)].$$

Proof. This is a consequence of the fact that $\lim_{k\to+\infty} N(\Pi_k(\theta)) = N(\theta)$ and of the continuity of B_*. □

As already discussed at the end of Proposition 2.3, any reasonable bound on the sample size n allows bounding the logarithmic factor $\log(K)$ by a relatively small constant. In particular, assuming that $n \le 10^{20}$, we get $\log(K) \le 4.35$.

In the following we construct an estimator of the Gram operator $\mathcal{G}$. Let $X_1, \ldots, X_n \in \mathcal{H}$ be an i.i.d. sample drawn according to P. Define $V = \mathbf{span}\{X_1, \ldots, X_n\}$ and

$$V_k = \mathbf{span}\{\Pi_k X_1, \ldots, \Pi_k X_n\} = \Pi_k(V).$$

Let Θ_δ be a δ-net of $\mathbb{S}_{\mathcal{H}} \cap V_k$ (where $\mathbb{S}_{\mathcal{H}}$ denotes the unit sphere in $\mathcal{H}$) and remark that Θ_δ is finite because $\dim(V_k) \le n < +\infty$. We compute a linear operator $\widehat{\mathcal{G}}_k : V_k \to V_k$ with minimal Hilbert Schmidt norm – so that $\mathbf{Tr}(\widehat{\mathcal{G}}_k^2) \le \mathbf{Tr}(\mathcal{G}^2)$ – satisfying

$$\max_{(\lambda,\beta)\in\Lambda} \Phi_-\big(\widetilde{N}_\lambda(\theta)\big) \le \langle\widehat{\mathcal{G}}_k\theta, \theta\rangle_{\mathcal{H}} \le \min_{(\lambda,\beta)\in\Lambda} \Phi_+^{-1}\big(\widetilde{N}_\lambda(\theta)\big), \qquad \theta \in \Theta_\delta.$$

Observe that $\widehat{\mathcal{G}}_k$ plays the same role as the symmetric matrix Q in the finite-dimensional setting. We consider as an estimator of $\mathcal{G}$ the operator

$$\mathcal{Q} = \widehat{\mathcal{G}}_k \circ \Pi_{V_k}, \tag{3.1}$$

where Π_{V_k} is the orthogonal projector on V_k. Let us decompose $\mathcal{Q}$ in its positive and negative parts and write $\mathcal{Q} = \mathcal{Q}_+ - \mathcal{Q}_-$.

Proposition 3.3. *For any threshold* $\sigma \in \mathbb{R}_+$ *such that* $\sigma \le s_4^2$ *and* $8\zeta_*(\sigma) \le \sqrt{n}$, *with probability at least* $1 - 2\varepsilon$, *for any* $\theta \in \mathbb{S}_{\mathcal{H}}$ *and for any* k,

$$\big|\max\{\langle\theta, \mathcal{Q}_+\theta\rangle_{\mathcal{H}}, \sigma\} - \max\{\langle\Pi_k\theta, \mathcal{G}\Pi_k\theta\rangle_{\mathcal{H}}, \sigma\}\big|$$

$$\le 2\max\{\langle\Pi_k\theta, \mathcal{G}\Pi_k\theta\rangle_{\mathcal{H}}, \sigma\}B_*\big(\langle\Pi_k\theta, \mathcal{G}\Pi_k\theta\rangle_{\mathcal{H}}\big) + 7\delta\sqrt{\mathbf{Tr}(\mathcal{G}^2)}$$

$$\left|\max\big\{\langle\theta,\mathcal{Q}_+\theta\rangle_\mathcal{H},\sigma\big\} - \max\big\{\langle\Pi_k\theta,\mathcal{G}\Pi_k\theta\rangle_\mathcal{H},\sigma\big\}\right|$$
$$\leq 2\max\big\{\langle\theta,\mathcal{Q}_+\theta\rangle_\mathcal{H},\sigma\big\}B_*\big(\min\{\langle\theta,\mathcal{Q}_+\theta\rangle_\mathcal{H},s_4^2\}\big) + 7\delta\sqrt{\mathbf{Tr}(\mathcal{G}^2)}.$$

For the proof, we refer to Section 7.3.

Let us consider $\{p_i\}_{i=1}^{+\infty}$ an orthonormal basis of eigenvectors of $\mathcal{G}$ such that the corresponding sequence of eigenvalues $\{\lambda_i, i = 1,\ldots,+\infty\}$ is non-increasing.

Proposition 3.4. *Consider some threshold* $\sigma \in \mathbb{R}_+$ *such that* $\sigma \leq s_4^2$ *and* $8\zeta_*(\sigma) \leq \sqrt{n}$. *With probability at least* $1 - 2\varepsilon$, *for any* $\theta \in \mathbb{S}_\mathcal{H}$ *and for any* k,

$$\left|\max\big\{\langle\theta,\mathcal{Q}_+\theta\rangle_\mathcal{H},\sigma\big\} - \max\big\{\langle\theta,\mathcal{G}\theta\rangle_\mathcal{H},\sigma\big\}\right|$$
$$\leq 2\max\big\{\langle\theta,\mathcal{G}\theta\rangle_\mathcal{H},\sigma\big\}B_*\big(\langle\theta,\mathcal{G}\theta\rangle_\mathcal{H}\big) + 7\delta\sqrt{\mathbf{Tr}(\mathcal{G}^2)} + 3\nu_k,$$
$$\left|\max\big\{\langle\theta,\mathcal{Q}_+\theta\rangle_\mathcal{H},\sigma\big\} - \max\big\{\langle\theta,\mathcal{G}\theta\rangle_\mathcal{H},\sigma\big\}\right|$$
$$\leq 2\max\big\{\langle\theta,\mathcal{Q}_+\theta\rangle_\mathcal{H},\sigma\big\}B_*\big(\min\{\langle\theta,\mathcal{Q}_+\theta\rangle_\mathcal{H},s_4^2\}\big) + 7\delta\sqrt{\mathbf{Tr}(\mathcal{G}^2)} + 2\nu_k,$$

where

$$\nu_k = \inf_{m=1,\ldots,+\infty}\left(\sum_{i=1}^{m-1}\lambda_i\|p_i - \Pi_k p_i\|_\mathcal{H} + \lambda_m/2\right)$$

$$\leq \inf_{m=1,\ldots,+\infty}\left(\sum_{i=1}^{m-1}\lambda_i\|p_i - \Pi_k p_i\|_\mathcal{H} + \mathbf{Tr}(\mathcal{G})/(2m)\right) \xrightarrow[k\to+\infty]{} 0.$$

Proof. It is enough to observe that

$$\left|\max\big\{\langle\theta,\mathcal{Q}_+\theta\rangle_\mathcal{H},\sigma\big\} - \max\big\{\langle\theta,\mathcal{G}\theta\rangle_\mathcal{H},\sigma\big\}\right|$$
$$\leq \left|\max\big\{\langle\theta,\mathcal{Q}_+\theta\rangle_\mathcal{H},\sigma\big\} - \max\big\{\langle\Pi_k\theta,\mathcal{G}\Pi_k\theta\rangle_\mathcal{H},\sigma\big\}\right| + \left|\langle\theta,\mathcal{G}\theta\rangle_\mathcal{H} - \langle\Pi_k\theta,\mathcal{G}\Pi_k\theta\rangle_\mathcal{H}\right|$$

and, for any $\theta \in \mathbb{S}_\mathcal{H}$, we have

$$\left|\langle\theta,\mathcal{G}\theta\rangle_\mathcal{H} - \langle\Pi_k\theta,\mathcal{G}\Pi_k\theta\rangle_\mathcal{H}\right| = \left|\sum_{i=1}^{+\infty}\big(\langle\Pi_k\theta,p_i\rangle_\mathcal{H}^2 - \langle\theta,p_i\rangle_\mathcal{H}^2\big)\lambda_i\right|$$
$$= \left|\sum_{i=1}^{+\infty}\big(\langle\theta,\Pi_k p_i\rangle_\mathcal{H}^2 - \langle\theta,p_i\rangle_\mathcal{H}^2\big)\lambda_i\right|$$
$$= \left|\sum_{i=1}^{m-1}\langle\theta,\Pi_k p_i + p_i\rangle_\mathcal{H}\langle\theta,p_i - \Pi_k p_i\rangle_\mathcal{H}\lambda_i\right|$$

$$+ \left| \sum_{i=m}^{+\infty} \left(\langle \theta, \Pi_k p_i \rangle_{\mathcal{H}}^2 - \langle \theta, p_i \rangle_{\mathcal{H}}^2 \right) \lambda_i \right|$$

$$\leq \sum_{i=1}^{m-1} 2\lambda_i \| p_i - \Pi_k p_i \|_{\mathcal{H}}$$

$$+ \max \left\{ \sum_{i=m}^{+\infty} \lambda_i \langle \Pi_k \theta, p_i \rangle_{\mathcal{H}}^2, \sum_{i=m}^{+\infty} \lambda_i \langle \theta, p_i \rangle_{\mathcal{H}}^2 \right\}$$

$$\leq \inf_{m=1,\dots,+\infty} \left(\sum_{i=1}^{m-1} 2\lambda_i \| p_i - \Pi_k p_i \|_{\mathcal{H}} + \lambda_m \right).$$

Indeed, $\sum_{i=m}^{+\infty} \langle \Pi_k \theta, p_i \rangle_{\mathcal{H}}^2 \leq \| \Pi_k \theta \|_{\mathcal{H}}^2 \leq \| \theta \|^2 \leq 1$, so that

$$\sum_{i=m}^{+\infty} \lambda_i \langle \Pi_k \theta, p_i \rangle_{\mathcal{H}}^2 \leq \left(\sup_{i \geq m} \lambda_i \right) \left(\sum_{i=m}^{+\infty} \langle \Pi_k \theta, p_i \rangle_{\mathcal{H}}^2 \right) \leq \lambda_m,$$

and in the same way $\sum_{i=m}^{+\infty} \lambda_i \langle \theta, p_i \rangle_{\mathcal{H}}^2 \leq \lambda_m$.   □

Remark that we can use this result to bound $|\langle \theta, (\mathcal{G} - \mathcal{Q}_+)\theta \rangle_{\mathcal{H}}|$, using the inequality

$$\left| \langle \theta, (\mathcal{G} - \mathcal{Q}_+)\theta \rangle_{\mathcal{H}} \right| \leq \left| \max\{ \langle \theta, \mathcal{Q}_+\theta \rangle_{\mathcal{H}}, \sigma \} - \max\{ \langle \theta, \mathcal{G}\theta \rangle_{\mathcal{H}}, \sigma \} \right| + \sigma.$$

4. Empirical results

We present some empirical results that show the performance of our robust estimator. We use here a simplified construction that does not follow exactly the definition of the estimator Q, but exhibits the same kind of behaviour. We have simplified the construction in two ways. First, we do less intensive computations by using more directions than in equation (2.15) on page 3873 but less than specified in the δ-net Θ_δ required by the theory. More precisely, we estimate repeatedly using equation (2.15) on page 3873 in an eigen-basis of the previous iterate of the estimation. Second, we do not use the theoretical value of λ, that is necessarily pessimistic for the sake of mathematical correctness. We use instead the optimal constant for estimating $\mathbb{E}(\langle \theta, X \rangle^2)$ in a single direction, as given in [1].

Let $X_1, \dots, X_n \in \mathbb{R}^d$ be a sample drawn according to the probability distribution P and let $\lambda > 0$. Let $p \in \mathbb{R}^n$ and define $S(p, \lambda)$ as the solution of

$$\sum_{i=1}^{n} \psi \left[\lambda \left(S(p, \lambda)^{-1} p_i^2 - 1 \right) \right] = 0.$$

In practice we compute $S(p, \lambda)$ using the Newton algorithm. We observe that, when $p_i = \langle \theta, X_i \rangle$ and λ is suitably chosen, $S(p, \lambda)$ is an approximation of the estimator $\widehat{N}(\theta)$ of the quadratic form $N(\theta)$ and more precisely, in this case, $S(p, \lambda)$ is exactly $\widetilde{N}_\lambda(\theta)$ defined in equation (2.4).

Define $S(p)$ as the solution obtained when the parameter λ is set to

$$\lambda = m\sqrt{\frac{1}{v}\left[\frac{2}{n}\log(\varepsilon^{-1})\left(1 - \frac{2}{n}\log(\varepsilon^{-1})\right)\right]},$$

where $m = \frac{1}{n}\sum_{i=1}^{n} p_i^2$, $v = \frac{1}{n-1}\sum_{i=1}^{n}(p_i^2 - m)^2$ and $\varepsilon = 0.1$.

Let $\bar{\lambda}_1 \geq \cdots \geq \bar{\lambda}_d \geq 0$ be the eigenvalues of the empirical Gram matrix $\bar{G}$, that will be our starting point, and let $u_1, \ldots, u_d$ be a corresponding orthonormal basis of eigenvectors. We decompose the empirical Gram matrix as

$$\bar{G} = UDU^\top,$$

where U is the orthogonal matrix whose columns are the eigenvectors of $\bar{G}$ and D is the diagonal matrix $D = \text{diag}(\bar{\lambda}_1, \ldots, \bar{\lambda}_d)$. Based on the polarization formula,

$$u_i^\top G u_j = \frac{1}{4}\left[N(u_i + u_j) - N(u_i - u_j)\right], \qquad i, j = 1, \ldots, d,$$

we revise iteratively our estimation of G by estimating $N(u_i + \sigma u_j)$, with $\sigma \in \{+1, -1\}$, by

$$S\big(\langle u_i + \sigma u_j, X_\ell\rangle, 1 \leq \ell \leq n\big).$$

More precisely, for any $n \times d$ matrix W, we define $C(W)$ as the $d \times d$ matrix with entries

$$C(W)_{i,j} = \frac{1}{4}\big[S\big((W_{\ell,i} + W_{\ell,j}) \mid 1 \leq \ell \leq n\big) - S\big((W_{\ell,i} - W_{\ell,j}) \mid 1 \leq \ell \leq n\big)\big].$$

Let Y be the matrix whose ℓth row is the vector X_ℓ, so that

$$(YU)_{\ell,k} = \langle u_k, X_\ell\rangle, \qquad 1 \leq \ell \leq n, 1 \leq k \leq d.$$

We update the Gram matrix estimate to

$$Q_0 = UC(YU)U^\top.$$

Then we iterate the update scheme, decomposing Q_0 as

$$Q_0 = O_0 D_0 O_0^\top,$$

where $O_0 O_0^\top = O_0^\top O_0 = I$ and D_0 is a diagonal matrix and computing

$$Q_1 = O_0 C(Y O_0) O_0^\top.$$

The inductive update step is more generally the following. Assuming we have constructed Q_k, we decompose it as

$$Q_k = O_k D_k O_k^\top,$$

where $O_k O_k^\top = O_k^\top O_k = \mathrm{I}$ and D_k is a diagonal matrix and we define the new updated estimator of G as

$$Q_{k+1} = O_k C(Y O_k) O_k^\top.$$

We stop this iterative estimation scheme when $\|Q_k - Q_{k-1}\|_F$ falls under a given threshold. In the following numerical experiment we more simply performed four updates. We take as our robust estimator of G the last update Q_k.

We now present an example of the performance of this estimator, for some i.i.d. sample of size $n = 100$ in $\mathbb{R}^{10}$ drawn according to the Gaussian mixture distribution

$$\mathrm{P} = (1 - \alpha)\mathcal{N}(0, M_1) + \alpha\mathcal{N}(0, 16\mathbf{I}),$$

where $\alpha = 0.05$ and

$$M_1 = \begin{bmatrix} \begin{matrix} 2 & 1 \\ 1 & 1 \end{matrix} & & \text{\Large 0} \\ & \begin{matrix} 0.01 & & \\ & 0 & \\ & & \ddots \end{matrix} & \\ \text{\Large 0} & & 0.01 \end{bmatrix}.$$

The Gram matrix of P is equal to

$$G = (1 - \alpha)M_1 + 16\alpha\mathbf{I} = \begin{bmatrix} \begin{matrix} 2.7 & 0.95 \\ 0.95 & 1.75 \end{matrix} & & \text{\Large 0} \\ & \begin{matrix} 0.8095 & & \\ & 0 & \\ & & \ddots \end{matrix} & \\ \text{\Large 0} & & 0.8095 \end{bmatrix}.$$

This example illustrates a favorable situation where the performance of the robust estimator is particularly striking when compared to the empirical Gram matrix. As it can be seen on the expression of the sample distribution as well as on the configuration plots below, this is a situation of intermittent high variance: the sample is a mixture of a rare high variance signal and a frequent low variance more structured signal.

We tested the algorithm on 500 different samples, of size $n = 100$ each, drawn according to the Gaussian mixture distribution defined above. Random sample configurations are presented in Figure 1.

Figure 2 shows that the robust estimator Q significantly improves the error in terms of square of the Frobenius norm when compared to the empirical estimator $\bar{G}$. The red solid line represents the empirical quantile function of the errors of the robust estimator, whereas the blue dotted line represents the quantiles of $\|\bar{G} - G\|_F^2$.

This quantile function is obtained by sorting the 500 empirical errors in increasing order.

We also represented in Figure 2 the boxplots of the distributions of $\|\bar{G} - G\|_F^2$ and $\|Q - G\|_F^2$ on 500 statistical experiments. (The boxplots show the first, second and third quartiles, with whiskers extending to the most extreme data point within 1.5 of interquartile range. Further away extreme data points are printed individually).
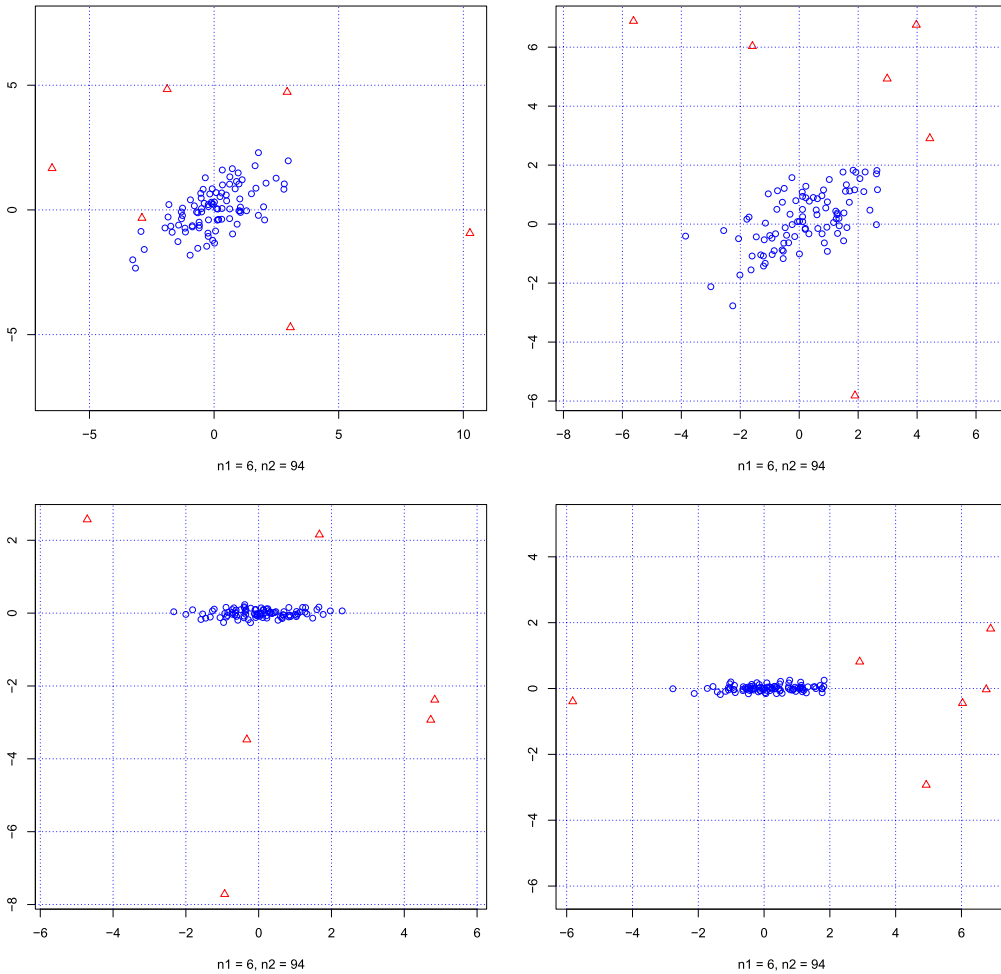
Figure 1. Two data samples projected onto the two first coordinates (above) and the second and third coordinates (below). Blue circles are drawn from the most frequent distribution and red triangles from the less frequent one.

The mean of the square distances $\|Q - G\|_F^2$ on 500 trials is 5.6 ± 0.4, where the indicated mean estimator and confidence interval is the non-asymptotic confidence interval given by Proposition 2.4 of [1] at confidence level 0.99. In the case of the empirical estimator, the mean is 15.5 ± 2. The empirical standard deviations accross 500 trials of $\|Q - G\|_F^2$ and $\|\bar{G} - G\|_F^2$ respectively, are close to 2 and 10. So we see that in this case the robust estimator reliably decreases the expected error by a factor larger than 2 and also produces errors with a much smaller standard deviation from sample to sample.
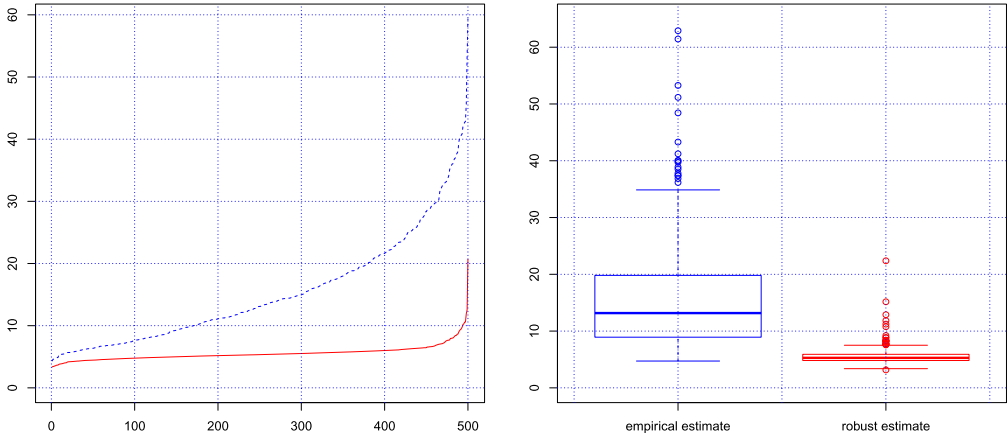
Figure 2. The solid red line represents the empirical quantile function of the square distances $\|Q - G\|_F^2$ in 500 statistical experiments, the dotted blue line represents the empirical quantile function of the square distances $\|\bar{G} - G\|_F^2$. The corresponding boxplots are also displayed on the right figure.

In Section 5, we show from a theoretical point of view that the two estimators Q and $\bar{G}$ behave in a similar way in light tail situations (meaning that in this case their predictions are the same).

5. The classical empirical estimator

The main goal of Section 2 is to estimate the Gram matrix $G = \mathbb{E}(XX^\top)$, where $X \in \mathbb{R}^d$ is a random vector of unknown law $P \in \mathcal{M}_+^1(\mathbb{R}^d)$, from an i.i.d. sample $X_1, \ldots, X_n \in \mathbb{R}^d$ drawn according to P. We have constructed a robust estimator of the Gram matrix and in Section 4 we have shown empirically its performance in the case of a Gaussian mixture distribution. In this section, we show from a theoretical point of view that the classical empirical estimator

$$\bar{G} = \frac{1}{n} \sum_{i=1}^{n} X_i X_i^\top$$

behaves similarly to our robust estimator in light tail situations, while it may perform worse otherwise.

As already done in Section 2, we consider the quadratic form

$$N(\theta) = \theta^\top G \theta = \mathbb{E}\big(\langle \theta, X \rangle^2\big)$$

and we denote by

$$\bar{N}(\theta) = \theta^\top \bar{G} \theta = \frac{1}{n} \sum_{i=1}^{n} \langle \theta, X_i \rangle^2$$

the quadratic form associated to the empirical Gram matrix $\bar{G}$. According to the notation introduced in Section 2, let $a > 0$ and let

$$K = 1 + \left\lceil a^{-1} \log\left(\frac{n}{72(2+c)\kappa^{1/2}} \right) \right\rceil,$$

where $\kappa = \sup\limits_{\substack{\theta \in \mathbb{R}^d \\ \mathbb{E}(\langle \theta, X \rangle^2) > 0}} \frac{\mathbb{E}(\langle \theta, X \rangle^4)}{\mathbb{E}(\langle \theta, X \rangle^2)^2}$ and $c = \frac{15}{8 \log(2)(\sqrt{2}-1)} \exp(\frac{1+2\sqrt{2}}{2})$. Let us put

$$R = \max_{i=1,\dots,n} \|X_i\| \tag{5.1}$$

and let us introduce

$$\tau_*(t) = \frac{\lambda_*(t)^2 \exp(a/2) R^4}{3 \max\{t, \sigma\}^2}, \qquad t \in \mathbb{R}_+,$$

where λ_* is defined in equation (7.16) as

$$\lambda_*(t) = \sqrt{\frac{2}{n(\kappa-1)} \left(\frac{(2+3c)\mathbb{E}(\|X\|^4)^{1/2}}{4(2+c)\kappa^{1/2} \max\{t, \sigma\}} + \log(K/\varepsilon) \right)}.$$

At the end of the section, we mention some assumptions under which it is possible to give a non-random bound for R.

The following proposition, compared with the result obtained for the robust estimator $\widehat{N}(\theta)$, presented in Proposition 2.3, shows that the different behavior of the two estimators $\widehat{N}$ and $\bar{N}$ can appear only for heavy tail data distributions.

Proposition 5.1. *Consider any threshold $\sigma \in \mathbb{R}_+$ such that $\sigma \leq \mathbb{E}(\|X\|^4)^{1/2}$. With probability at least $1 - 2\varepsilon$, for any $\theta \in \mathbb{S}_d$,*

$$\left| \frac{\max\{\bar{N}(\theta), \sigma\}}{\max\{N(\theta), \sigma\}} - 1 \right| \leq B_*\big(N(\theta)\big) + \frac{\tau_*(N(\theta))}{[1 - \tau_*(N(\theta))]_+ [1 - B_*(N(\theta))]_+},$$

where B_* is defined in Proposition 2.3.

For the proof we refer to Section 7.5.

Observe that, also in this case, the bound does not depend explicitly on the dimension d of the ambient space and thus the result can be extended to any infinite-dimensional Hilbert space.

We continue this section by stating assumptions under which it is possible to give a non-random bound for R, defined in equation (5.1).

Assume that, for some exponent $p \geq 1$ and some positive constants α and η,

$$\mathbb{E}\left[\exp\left(\frac{\alpha}{2} \left(\frac{\|X\|^{2/p}}{\mathbf{Tr}(G)^{1/p}} - 1 - \eta^{2/p} \right) \right) \right] \leq 1.$$

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

$$R \leq \mathbf{Tr}(G)^{1/2}\big(1 + \eta^{2/p} + 2\alpha^{-1}\log(n/\varepsilon)\big)^{p/2}, \tag{5.2}$$

where we recall that $\mathbf{Tr}(G) = \mathbb{E}[\|X\|^2]$.

To give a point of comparison, in the centered Gaussian case where $X \sim \mathcal{N}(0, G)$ is a Gaussian vector, we have, for any $\alpha \in]0, \lambda_1^{-1}\mathbf{Tr}(G)[$,

$$\mathbb{E}\left[\exp\left[\frac{\alpha}{2}\left(\frac{\|X\|^2}{\mathbf{Tr}(G)} + \frac{1}{\alpha}\sum_{i=1}^d\log\left(1 - \frac{\alpha\lambda_i}{\mathbf{Tr}(G)}\right)\right)\right]\right] = 1,$$

where $\lambda_1 \geq \cdots \geq \lambda_d$ are the eigenvalues of G. Therefore, with probability at least $1 - \varepsilon$,

$$R \leq \mathbf{Tr}(G)^{1/2}\left(-\frac{1}{\alpha}\sum_{i=1}^d\log\left(1 - \frac{\alpha\lambda_i}{\mathbf{Tr}(G)}\right) + \frac{2\log(n/\varepsilon)}{\alpha}\right)^{1/2}.$$

We can then consider the optimal value of α in the right-hand side of the previous equation, to establish that with probability at least $1 - \varepsilon$,

$$R \leq \inf_{\alpha \in]0,\mathbf{Tr}(G)/\lambda_1[} \mathbf{Tr}(G)^{1/2}\left(-\frac{1}{\alpha}\sum_{i=1}^d\log\left(1 - \frac{\alpha\lambda_i}{\mathbf{Tr}(G)}\right) + \frac{2\log(n/\varepsilon)}{\alpha}\right)^{1/2}$$

$$\leq \inf_{\alpha \in]0,\mathbf{Tr}(G)/\lambda_1[} \mathbf{Tr}(G)^{1/2}\left(\sum_{i=1}^d\frac{\lambda_i}{\mathbf{Tr}(G) - \alpha\lambda_i} + \frac{2\log(n/\varepsilon)}{\alpha}\right)^{1/2}$$

$$\leq \inf_{\alpha \in]0,\mathbf{Tr}(G)/\lambda_1[}\left(\frac{\mathbf{Tr}(G)^2}{\mathbf{Tr}(G) - \alpha\lambda_1} + \frac{2\,\mathbf{Tr}(G)\log(n/\varepsilon)}{\alpha}\right)^{1/2}$$

$$\leq \inf_{\alpha \in]0,\mathbf{Tr}(G)/\lambda_1[}\left(\mathbf{Tr}(G) + \frac{\alpha\lambda_1\,\mathbf{Tr}(G)}{\mathbf{Tr}(G) - \alpha\lambda_1}\right.$$

$$\left. + \frac{2\log(n/\varepsilon)\lambda_1(\mathbf{Tr}(G) - \alpha\lambda_1)}{\alpha\lambda_1} + 2\lambda_1\log(n/\varepsilon)\right)^{1/2}$$

$$= \big(\mathbf{Tr}(G) + 2\sqrt{2\log(n/\varepsilon)\lambda_1\,\mathbf{Tr}(G)} + 2\lambda_1\log(n/\varepsilon)\big)^{1/2} = \sqrt{\mathbf{Tr}(G)} + \sqrt{2\lambda_1\log(n/\varepsilon)}.$$

In order to replace hypothesis (5.2) by some polynomial assumptions it is convenient to replace R with

$$\widetilde{R} = \left(\frac{1}{n}\sum_{i=1}^n\|X_i\|^6\right)^{1/6}.$$

Indeed, by the Bienaymé Chebyshev inequality, we get that, with probability at least $1 - \varepsilon$,

$$\widetilde{R} \leq \left(\mathbb{E}\big[\|X\|^6\big] + \left(\frac{\mathbb{E}[\|X\|^{12}]}{n\varepsilon} \right)^{1/2} \right)^{1/6} \leq \left(1 + (n\varepsilon)^{-1/2} \right)^{1/6} \mathbb{E}\big[\|X\|^{12}\big]^{1/12}$$

and hence, with probability at least $1 - n^{-1}$,

$$\widetilde{R} \leq 2^{1/6} \mathbb{E}\big[\|X\|^{12}\big]^{1/12}.$$

We can prove an analogue of Proposition 5.1, where $\widetilde{R}$ plays the role of R.

Proposition 5.2. *Let $0 < \sigma \leq \mathbb{E}(\|X\|^4)^{1/2}$ and let us put*

$$\zeta_*(t) = \frac{\lambda_*(t)^2 \exp(a/2) \widetilde{R}^6}{3 \max\{t, \sigma\}^3}, \qquad t \in \mathbb{R}_+.$$

With probability at least $1 - 2\varepsilon$, for any $\theta \in \mathbb{S}_d$,

$$\left| \frac{\max\{\bar{N}(\theta), \sigma\}}{\max\{N(\theta), \sigma\}} - 1 \right|$$

$$\leq B_*\big(N(\theta)\big) + \frac{\zeta_*(N(\theta))}{[1 - B_*(N(\theta))]_+}$$

$$\leq \mathcal{O}\left(\sqrt{\frac{\kappa}{n} \left(\frac{\mathbf{Tr}(G)}{\max\{N(\theta), \sigma\}} + \log\big(\log(n)/\varepsilon\big) \right)} \right)$$

$$+ \mathcal{O}\left(\frac{\mathbb{E}(\|X\|^{12})^{1/2}}{n\kappa (1 + (n\varepsilon)^{-1/2})(\max\{N(\theta), \sigma\})^3} \left(\frac{\mathbf{Tr}(G)}{\max\{N(\theta), \sigma\}} + \log\big(\log(n)/\varepsilon\big) \right) \right).$$

For the proof we refer to Section 7.6.

6. Generalization

In this section, we come back to the finite-dimensional framework and we consider the problem of estimating the expectation of a symmetric random matrix. We will use these results to estimate the covariance matrix in the case of unknown expectation.

6.1. Symmetric random matrix

Let $A \in M_d(\mathbb{R})$ be a symmetric random matrix of size d. As already observed for the Gram matrix, the expectation of A can be recovered via the polarization identity from the quadratic form

$$N_A(\theta) = \theta^\top \mathbb{E}(A)\theta, \qquad \theta \in \mathbb{R}^d,$$

where the expectation is taken with respect to the unknown probability distribution of A on the space of symmetric matrices of size d. Observe that, if we decompose A in its positive and negative parts

$$A = A_+ - A_-,$$

where A_+ and A_- are defined by keeping respectively the positive and negative parts of the eigenvalues of A, in the framework of functional calculus on symmetric matrices, the quadratic form N_A rewrites as

$$N_A(\theta) = \mathbb{E}[\theta^\top A_+ \theta] - \mathbb{E}[\theta^\top A_- \theta] = N_{A_+}(\theta) - N_{A_-}(\theta).$$

Thus in the following, we will consider the case of a symmetric positive semi-definite random matrix of size d.

From now on, let $A \in M_d(\mathbb{R})$ be a symmetric positive semi-definite random matrix of size d and let P be a probability distribution on the space of symmetric positive semi-definite random matrices of size d. Our goal is to estimate

$$N(\theta) = \mathbb{E}[\theta^\top A \theta], \qquad \theta \in \mathbb{R}^d,$$

from an i.i.d. sample $A_1, \ldots, A_n \in M_d(\mathbb{R})$ of symmetric positive semi-definite matrices drawn according to P. We observe that the quadratic form $N(\theta)$ rewrites as

$$N(\theta) = \mathbb{E}\big[\big\|A^{1/2}\theta\big\|^2\big],$$

where $A^{1/2}$ denotes the square root of A.

The construction of the (robust) estimator $\widehat{N}(\theta)$ follows the one already done in the case of the Gram matrix with the necessary adjustments. For any $\lambda > 0$ and for any $\theta \in \mathbb{R}^d$, we consider the empirical criterion

$$r_\lambda(\theta) = \frac{1}{n} \sum_{i=1}^n \psi\big(\big\|A_i^{1/2}\theta\big\|^2 - \lambda\big),$$

where the influence function ψ is defined as in equation (2.2), and we perturb the parameter θ with the Gaussian perturbation $\pi_\theta \sim \mathcal{N}(\theta, \beta^{-1}\mathrm{I})$ of mean θ and covariance matrix $\beta^{-1}\mathrm{I}$, where $\beta > 0$ is a free real parameter. We consider the family of estimators

$$\widetilde{N}(\theta) = \frac{\lambda}{\widehat{\alpha}(\theta)^2},$$

where $\widehat{\alpha}(\theta) = \sup\{\alpha \in \mathbb{R}_+ \mid r_\lambda(\alpha\theta) \leq 0\}$. Let us put

$$s_4 = \mathbb{E}\big[\|A\|_\infty^2\big]^{1/4} \quad \text{and} \quad \kappa = \sup_{\substack{\theta \in \mathbb{R}^d \\ \mathbb{E}[\|A^{1/2}\theta\|^2] > 0}} \frac{\mathbb{E}[\|A^{1/2}\theta\|^4]}{\mathbb{E}[\|A^{1/2}\theta\|^2]^2}, \tag{6.1}$$

where $\|A\|_\infty$ is the operator norm, that is in this context of symmetric positive semi-definite matrices equal to the largest eigenvalue of A. The finite set $\Lambda \subset (\mathbb{R}_+ \setminus \{0\})^2$ of possible values of the couple of parameters (λ, β) is defined as

$$\Lambda = \big\{ (\lambda_j, \beta_j) \mid 0 \le j < K \big\},$$

where

$$K = 1 + \left\lceil a^{-1} \log\left(\frac{n}{72(2+c)\kappa^{1/2}} \right) \right\rceil, \tag{6.2}$$

for some real positive parameter a to be chosen later on, and

$$\lambda_j = \sqrt{ \frac{2}{(\kappa-1)n} \left(\frac{(2+3c)\mathbb{E}(\mathbf{Tr}(A^2))}{4(2+c)\kappa^{1/2}\mathbb{E}(\|A\|_\infty^2)} \exp(ja) + \log(K/\varepsilon) \right) },$$

$$\beta_j = \sqrt{ 2(2+c)\kappa^{1/2}\mathbb{E}(\|A\|_\infty^2) \exp\big[-(j-1/2)a \big] }.$$

Note that in the case of the Gram matrix, the picture is simplified by the fact that $\|XX^\top\|_\infty^2 = \mathbf{Tr}[(XX^\top)^2]$, whereas here we have to take into account the fact that the operator norm and the Frobenius norm of A are different when the rank of A is larger than one.

We recall that c is defined in equation (2.7) as

$$c = \frac{15}{8 \log(2)(\sqrt{2}-1)} \exp\left(\frac{1+2\sqrt{2}}{2} \right).$$

According to equation (2.8), let

$$B_{\lambda,\beta}(t) = \begin{cases} \dfrac{\gamma + \lambda\delta/\max\{t,\sigma\}}{1 - \mu - \gamma - 2\lambda\delta/\max\{t,\sigma\}}, & (\lambda, \beta, t) \in \Gamma, \\ +\infty, & \text{otherwise}, \end{cases}$$

and put $(\widehat{\lambda}, \widehat{\beta}) = \arg\min_{(\lambda,\beta)\in\Lambda} B_{\lambda,\beta}[\|\theta\|^{-2}\widetilde{N}_\lambda(\theta)]$. Define the estimator $\widehat{N}$ as

$$\widehat{N}(\theta) = \widetilde{N}_{\widehat{\lambda}}(\theta). \tag{6.3}$$

Proposition 6.1. *Let $\sigma \in]0, s_4^2]$ be some energy level. With probability at least $1 - 2\varepsilon$, for any $\theta \in \mathbb{R}^d$,*

$$\left| \frac{\max\{N(\theta), \sigma\|\theta\|^2\}}{\max\{\widehat{N}(\theta), \sigma\|\theta\|^2\}} - 1 \right| \le B_*\big[\|\theta\|^{-2}N(\theta) \big],$$

where B_* is defined as

$$B_*(t) = \begin{cases} \dfrac{n^{-1/2}\zeta_*(\max\{t,\sigma\})}{1 - 4n^{-1/2}\zeta_*(\max\{t,\sigma\})}, & \big[6 + (\kappa-1)^{-1}\big]\zeta_*(\max\{t,\sigma\}) \le \sqrt{n}, \\ +\infty, & \text{otherwise}, \end{cases}$$

and

$$\zeta_*(t) = \sqrt{2.032(\kappa - 1)\left(\frac{0.73\mathbb{E}[\mathbf{Tr}(A^2)]}{\kappa^{1/2}\mathbb{E}[\|A\|_\infty^2]^{1/2}t} + \log(K) + \log(\varepsilon^{-1})\right)}$$

$$+ \sqrt{\frac{98.5\kappa^{1/2}\mathbb{E}[\|A\|_\infty^2]^{1/2}}{t}}.$$

As already discuss at the end of Proposition 2.3, if $a = 1/2$ and $n \leq 10^{20}$, we can bound the logarithmic factor $\log(K)$ with the (small) constant 4.35.

For the proof, we refer to Section 7.7.

Remark that to obtain the above result, we have used the fact that

$$\mathbb{E}[\|A\theta\|^2] \leq \mathbb{E}[\|A\|_\infty^2]^{1/2}\kappa^{1/2}N(\theta).$$

However, if we use any upper bound of the form

$$\mathbb{E}[\|A\theta\|^2] \leq f(\mathbb{E}[A])N(\theta)$$

Proposition 6.1 holds replacing $\mathbb{E}[\|A\|_\infty^2]^{1/2}\kappa^{1/2}$ with $f(\mathbb{E}[A])$ in the definition of ζ_*. Similarly, we can replace $\mathbb{E}[\mathbf{Tr}(A^2)]$ by an upper bound.

We observe in particular that

$$\frac{\mathbb{E}[\|A\|_\infty^2]^{1/2}}{\kappa^{1/2}} \leq \frac{\mathbb{E}[\mathbf{Tr}(A^2)]}{\kappa^{1/2}\mathbb{E}[\|A\|_\infty^2]^{1/2}} \leq \mathbb{E}[\mathbf{Tr}(A)] = \mathbf{Tr}[\mathbb{E}(A)].$$

Indeed, to see the first inequality it is sufficient to observe that $\|A\|_\infty^2 \leq \mathbf{Tr}(A^2)$. Moreover, we have that

$$\mathbb{E}[\mathbf{Tr}(A^2)] \leq \mathbb{E}[\|A\|_\infty \mathbf{Tr}(A)] \leq \mathbb{E}[\|A\|_\infty^2]^{1/2}\mathbb{E}[\mathbf{Tr}(A)^2]^{1/2},$$

and, denoting by $\{e_i\}_{i=1}^d$ an orthonormal basis of $\mathbb{R}^d$,

$$\mathbb{E}[\mathbf{Tr}(A)^2] = \sum_{\substack{1 \leq i \leq d, \\ 1 \leq j \leq d}} \mathbb{E}[\|A^{1/2}e_i\|^2\|A^{1/2}e_j\|^2]$$

$$\leq \sum_{\substack{1 \leq i \leq d, \\ 1 \leq j \leq d}} \mathbb{E}[\|A^{1/2}e_i\|^4]^{1/2}\mathbb{E}[\|A^{1/2}e_j\|^4]^{1/2}$$

$$\leq \kappa \sum_{\substack{1 \leq i \leq d, \\ 1 \leq j \leq d}} \mathbb{E}[\|A^{1/2}e_i\|^2]\mathbb{E}[\|A^{1/2}e_j\|^2] = \kappa\mathbb{E}[\mathbf{Tr}(A)]^2.$$

This implies that we can bound ζ_* in Proposition 6.1 by

$$\zeta_*(t) = \sqrt{2.032(\kappa - 1)\left(\frac{0.73\mathbb{E}[\mathbf{Tr}(A)]}{t} + \log(K) + \log(\varepsilon^{-1})\right)} + \sqrt{\frac{98.5\kappa\mathbb{E}[\mathbf{Tr}(A)]}{t}}. \quad (6.4)$$

We conclude this section observing that, since the entropy terms are dominated by $\mathbb{E}[\mathbf{Tr}(A)]$, the result can be generalized to the case where A is a random symmetric positive semi-definite operator in a infinite-dimensional Hilbert space with the only additional assumption that $\mathbb{E}[\mathbf{Tr}(A)] < +\infty$.

6.2. Covariance matrix

Let $X \in \mathbb{R}^d$ be a random vector distributed according to the unknown probability measure $\mathrm{P} \in \mathcal{M}_+^1(\mathbb{R}^d)$. The covariance matrix of X is defined as

$$\Sigma = \mathbb{E}\big[\big(X - \mathbb{E}[X]\big)\big(X - \mathbb{E}[X]\big)^\top\big]$$

and our goal is to estimate, uniformly in θ, the quadratic form

$$N(\theta) = \theta^\top \Sigma \theta = \mathbb{E}\big[\langle \theta, X - \mathbb{E}[X]\rangle^2\big], \qquad \theta \in \mathbb{R}^d,$$

from an i.i.d. sample $X_1, \ldots, X_n \in \mathbb{R}^d$ drawn according to P. We cannot use the results we have proved for the Gram matrix, since the quadratic form N depends on the unknown quantity $\mathbb{E}[X]$. However we can find a workaround, using the results of the previous section about symmetric random matrices. Indeed, we do not need to estimate $\mathbb{E}[X]$ in order to estimate N but it is sufficient to observe that the quadratic form N can be written as

$$N(\theta) = \frac{1}{2}\mathbb{E}\big[\langle \theta, X - X'\rangle^2\big],$$

where X' is an independent copy of X. More generally, given $q \in \mathbb{N}$, we may consider q independent copies $X^{(1)}, \ldots, X^{(q)}$ of X and the random matrix

$$A = \frac{1}{q(q-1)}\sum_{1 \leq j < k \leq q}\big(X^{(j)} - X^{(k)}\big)\big(X^{(j)} - X^{(k)}\big)^\top$$

so that

$$N(\theta) = \frac{1}{q(q-1)}\mathbb{E}\bigg[\sum_{1 \leq j < k \leq q}\langle \theta, X^{(j)} - X^{(k)}\rangle^2\bigg] = \mathbb{E}\big[\theta^\top A\theta\big].$$

We will discuss later how to choose q. In the following we use a robust block estimate which consists in dividing the sample $X_1, \ldots, X_n$ in blocks of size q and then in considering the original

sample as a "new" sample of $\lfloor n/q \rfloor$ symmetric matrices $A_1, \ldots, A_{\lfloor n/q \rfloor}$ (of independent copies of A) defined as

$$A_i = \frac{1}{q(q-1)} \sum_{(i-1)q < j < k \leq iq} (X_i - X_j)(X_i - X_j)^\top$$

that thus correspond to the empirical covariance estimates on each block. We can use the results of the previous section to define a robust estimator of $N(\theta)$.

Let us introduce

$$\kappa' = \sup_{\substack{\theta \in \mathbb{R}^d, \\ \mathbb{E}(\|A^{1/2}\theta\|^2) > 0}} \frac{\mathbb{E}[\|A^{1/2}\theta\|^4]}{\mathbb{E}[\|A^{1/2}\theta\|^2]^2},$$

and

$$\kappa = \sup_{\substack{\theta \in \mathbb{R}^d \\ \mathbb{E}[\langle \theta, X - \mathbb{E}(X) \rangle^2] > 0}} \frac{\mathbb{E}[\langle \theta, X - \mathbb{E}(X) \rangle^4]}{\mathbb{E}[\langle \theta, X - \mathbb{E}(X) \rangle^2]^2}.$$

Lemma 6.1. *The two kurtosis coefficients introduced above are related by the relation*

$$\kappa' \leq 1 + \tau_q(\kappa)/q,$$

where $\tau_q(\kappa) = \kappa - 1 + \frac{2}{q-1}$.

For the proof we refer to Section 7.8.

Let $\widehat{N}(\theta)$ be the estimator defined in equation (6.3) and remark that

$$\mathbb{E}\big(\mathbf{Tr}(A)\big) = \mathbf{Tr}\big(\mathbb{E}(A)\big) = \mathbf{Tr}(\Sigma) = \mathbb{E}\big(\|X - \mathbb{E}(X)\|^2\big).$$

Proposition 6.2. *For any energy level $\sigma \in]0, \mathbf{Tr}(\Sigma)]$, with probability at least $1 - 2\varepsilon$, for any $\theta \in \mathbb{R}^d$,*

$$\left| \frac{\max\{N(\theta), \sigma \|\theta\|^2\}}{\max\{\widehat{N}(\theta), \sigma \|\theta\|^2\}} - 1 \right| \leq B_*\big(\|\theta\|^{-2} N(\theta)\big),$$

where

$$B_*(t) = \begin{cases} \dfrac{(q\lfloor n/q \rfloor)^{1/2} \zeta_q(\max\{t, \sigma\})}{1 - 4(q\lfloor n/q \rfloor)^{1/2} \zeta_q(\max\{t, \sigma\})}, & \text{if } \big(6 + q/\tau_q(\kappa)\big)\zeta_q\big(\max\{t, \sigma\}\big) \leq \big(q\lfloor n/q \rfloor\big)^{1/2}, \\ +\infty, & \text{otherwise} \end{cases}$$

and

$$\zeta_q(t) = \sqrt{2.032\tau_q(\kappa)\left(\frac{0.73\,\mathbf{Tr}(\Sigma)}{t} + \log(K) + \log\big(\varepsilon^{-1}\big)\right)} + \sqrt{\frac{98.5(q + \tau_q(\kappa))\,\mathbf{Tr}(\Sigma)}{t}}$$

with

$$K = 1 + \left\lceil \frac{1}{2} \log\left(\frac{\lfloor n/q \rfloor}{72(2+c)(1+\tau_q(\kappa)/q)^{1/2}} \right) \right\rceil.$$

Proof. The result follows from Proposition 6.1, using the definition of ζ_* given in equation (6.4) on page 3891, where we replace κ by κ' and n by $\lfloor n/q \rfloor$. We conclude the proof according to Lemma 6.1. $\qquad\square$

Here, we have used the upper bound for the entropy factor defined in terms of $\mathbb{E}[\mathbf{Tr}(A)] = \mathbf{Tr}(\Sigma)$, as mentioned in the remarks following Proposition 6.1. We can improve somehow the constants by evaluating more carefully $\mathbb{E}[\|A\theta\|^2]$ and $\mathbb{E}[\mathbf{Tr}(A^2)]$ as shown in the next lemma proved in Section 7.9.

Lemma 6.2. *It holds true that*

$$\mathbb{E}\big[\|A\theta\|^2\big] \leq \left(1 - \frac{q-2}{q(q-1)}\right)\|\Sigma\|_\infty N(\theta) + \frac{1}{q}\left(\kappa + \frac{1}{q-1}\right)\mathbf{Tr}(\Sigma)N(\theta) \qquad (6.5)$$

and

$$\mathbb{E}\big[\mathbf{Tr}(A^2)\big] \leq \left(1 - \frac{q-2}{q(q-1)}\right)\mathbf{Tr}(\Sigma^2) + \frac{1}{q}\left(\kappa + \frac{1}{q-1}\right)\mathbf{Tr}(\Sigma)^2. \qquad (6.6)$$

Using these tighter bounds, we can improve ζ_q to

$$\zeta_q(t) = \left[2.032\tau_q(\kappa)\left(\frac{0.73[(1 - \frac{q-2}{q(q-1)})\mathbf{Tr}(\Sigma^2) + \frac{1}{q}(\kappa + \frac{1}{q-1})\mathbf{Tr}(\Sigma)^2]}{[(1 - \frac{q-2}{q(q-1)})\|\Sigma\|_\infty + \frac{1}{q}(\kappa + \frac{1}{q-1})\mathbf{Tr}(\Sigma)]t}\right.\right.$$
$$\left.\left. + \log(K) + \log(\varepsilon^{-1})\right)\right]^{1/2}$$
$$+ \sqrt{\frac{98.5[q(1 - \frac{q-2}{q(q-1)})\|\Sigma\|_\infty + (\kappa + \frac{1}{q-1})\mathbf{Tr}(\Sigma)]}{t}}.$$

Therefore, in the case when

$$q\|\Sigma\|_\infty \leq \mathbf{Tr}(\Sigma),$$

we have

$$\mathbb{E}\big[\|A\theta\|^2\big] \leq \frac{1}{q}\left(\kappa + 1 + \frac{2}{q(q-1)}\right)\mathbf{Tr}(\Sigma)N(\theta)$$

and hence, recalling that $\mathbf{Tr}(\Sigma^2) \leq \mathbf{Tr}(\Sigma)^2$, we can take

$$\zeta_q(t) = \sqrt{2.032\tau_q(\kappa)\left(\frac{0.73\,\mathbf{Tr}(\Sigma)}{t} + \log(K) + \log(\varepsilon^{-1})\right)} + \sqrt{\frac{98.5(\kappa + 1 + \frac{2}{q(q-1)})\mathbf{Tr}(\Sigma)}{t}}.$$

If we compare the above result with the bound obtained in Proposition 2.3 on page 3871 for the Gram matrix estimator, we see that the first appearance of κ in the definition of ζ_q has been replaced with

$$\tau_q(\kappa) + 1 = \kappa + \frac{2}{q-1},$$

and that the second appearance of κ has been replaced with

$$\kappa + 1 + \frac{2}{q(q-1)}.$$

Thus, when $\|\Sigma\|_\infty \leq \mathbf{Tr}(\Sigma)/2$, and this is not a very strong hypothesis, we can take at least $q = 2$, and obtain an improved bound for the estimation of Σ. This bound is not much larger than for the estimation of the centered Gram matrix, that we could have performed if we had known $\mathbb{E}(X)$, since the difference is just a matter of replacing κ with $\kappa + 2$.

7. Proofs

In this section, we give the proofs of the results presented in the previous sections. More precisely, Section 7.1 deals with Proposition 2.1 (on page 3869), Section 7.2 with Proposition 2.3 (on page 3871) and Section 7.3 with Proposition 3.3 (on page 3878), some preliminary lemmas being postponed to Section 7.4.

7.1. Proof of Proposition 2.1

The proof of Proposition 2.1 requires a sequence of preliminary results.

Our approach relies on perturbing the parameter θ with the Gaussian perturbation $\pi_\theta \sim \mathcal{N}(\theta, \beta^{-1}I)$, where $\beta > 0$ is a free parameter.

Lemma 7.1. *We have*

$$\int \langle \theta', x \rangle^2 \, d\pi_\theta(\theta') = \langle \theta, x \rangle^2 + \frac{\|x\|^2}{\beta}.$$

Proof. Let $W \in \mathbb{R}^d$ be a random variable drawn according to $\pi_\theta \sim \mathcal{N}(\theta, \beta^{-1}I)$. It follows that $\langle W, x \rangle$ is a one-dimensional Gaussian random variable with mean $\langle \theta, x \rangle$ and variance

$$x^\top(\beta^{-1}I)x = \frac{\|x\|^2}{\beta}.$$

Consequently

$$\int \langle \theta', x \rangle^2 \, d\pi_\theta(\theta') = \mathbb{E}\big[\langle W, x \rangle\big]^2 + \mathbf{Var}\big[\langle W, x \rangle\big] = \langle \theta, x \rangle^2 + \frac{\|x\|^2}{\beta}. \qquad \square$$

Accordingly, we get

$$r_\lambda(\theta) = \frac{1}{n} \sum_{i=1}^n \psi \left[\int \left(\langle \theta', x \rangle^2 - \frac{\|x\|^2}{\beta} - \lambda \right) d\pi_\theta(\theta') \right].$$

In order to pull the expectation with respect to π_θ out of the influence function ψ, with a minimal loss of accuracy, we introduce the function

$$\chi(z) = \begin{cases} \psi(z), & z \leq z_1, \\ \psi(z_1) + p_1(z - z_1) - (z - z_1)^2/8, & z_1 \leq z \leq z_1 + 4p_1, \\ \psi(z_1) + 2p_1^2, & z \geq z_1 + 4p_1, \end{cases} \tag{7.1}$$

where $z_1 \in [0, 1]$ is such that $\psi''(z_1) = -1/4$ and p_1 is defined by the condition $p_1 = \psi'(z_1)$. Using the explicit expression of the first and second derivative of ψ, we get

$$z_1 = 1 - \sqrt{4\sqrt{2} - 5},$$

$$p_1 = \psi'(z_1) = \frac{\sqrt{4\sqrt{2} - 5}}{2(\sqrt{2} - 1)}$$

and

$$\sup \chi = \psi(z_1) + 2p_1^2 = -\log\left[2(\sqrt{2} - 1)\right] + \frac{1 + 2\sqrt{2}}{2}.$$

Lemma 7.2. *For any $z \in \mathbb{R}$,*

$$\psi(z) \leq \chi(z) \leq \log\left(1 + z + z^2/2\right). \tag{7.2}$$

Proof. We first prove that $\psi(z) \leq \chi(z)$. The inequality is trivial for $z \leq z_1$, since $\chi(z) = \psi(z)$. For $z \in [z_1, z_1 + 4p_1]$, performing a Taylor expansion at z_1, we obtain that

$$\psi(z) = \psi(z_1) + p_1(z - z_1) - \frac{1}{8}(z - z_1)^2 + \int_{z_1}^z \frac{\psi'''(u)}{2}(z - u)^2 \, du$$

$$\leq \psi(z_1) + p_1(z - z_1) - \frac{1}{8}(z - z_1)^2 = \chi(z),$$

since $\psi'''(u) \leq 0$ for $u \in [0, 1[$. Finally we observe that, for any $z \geq z_1 + 4p_1$,

$$\chi(z) = \psi(z_1) + 2p_1^2 > \log(2) \geq \psi(z).$$

Let us now show that $\chi(z) \leq \log(1 + z + z^2/2)$. For $z \leq z_1$, we have already seen that the inequality is satisfied since $\chi(z) = \psi(z)$. Moreover, we observe that the function

$$f(z) = \log\left(1 + z + z^2/2\right)$$

is such that $f(z_1) \geq \chi(z_1)$ and also $f'(z_1) \geq \chi'(z_1)$. Performing a Taylor expansion at z_1, we get

$$f(z) = f(z_1) + f'(z_1)(z - z_1) + \int_{z_1}^{z} f''(u)(z - u)^2 \, du$$

$$\geq \chi(z_1) + \chi'(z_1)(z - z_1) + \inf f'' \frac{(z - z_1)^2}{2}.$$

Since for any $t \in [z_1, z_1 + 4p_1]$,

$$\inf f'' = f''(\sqrt{3} - 1) = -1/4 = \chi''(t),$$

we deduce that

$$f(z) \geq \chi(z_1) + p_1(z - z_1) - \frac{1}{8}(z - z_1)^2 = \chi(z).$$

In particular, $f(z_1 + 4p_1) \geq \chi(z_1 + 4p_1)$. Recalling that f is an increasing function while χ is constant on the interval $[z_1 + 4p_1, +\infty[$, we conclude the proof. $\qquad\square$

Next, lemma allows us to pull the expectation with respect to π_θ out of the function χ.

Lemma 7.3. *Let Θ be a measurable space. For any $\rho \in \mathcal{M}_+^1(\Theta)$ and any $h \in L_\rho^1(\Theta)$,*

$$\chi\left(\int h \, d\rho\right) \leq \int \chi(h) \, d\rho + \frac{1}{8}\mathbf{Var}(h \, d\rho), \qquad (7.3)$$

where by definition

$$\mathbf{Var}(h \, d\rho) = \int \left(h(\theta) - \int h \, d\rho\right)^2 d\rho(\theta) \in \mathbb{R} \cup \{+\infty\}.$$

Moreover,

$$\psi\left(\int h \, d\rho\right) \leq \int \chi(h) \, d\rho + \min\left\{\log(4), \frac{1}{8}\mathbf{Var}(h \, d\rho)\right\}.$$

Proof. To prove equation (7.3) we observe that performing a Taylor expansion of the function χ at $z = \int h \, d\rho$

$$\chi[h(\theta)] \geq \chi(z) + (h(\theta) - z)\chi'(z) + \inf \chi'' \frac{(h(\theta) - z)^2}{2},$$

so that, recalling that $\inf \chi'' = -1/4$, we get

$$\int \chi[h(\theta)] \, d\rho(\theta) \geq \chi\left(\int h \, d\rho\right) - \frac{1}{8}\int \left(h(\theta) - \int h(\theta) \, d\rho\right)^2 d\rho(\theta)$$

$$= \chi \left(\int h \, d\rho \right) - \frac{1}{8} \mathbf{Var}(h \, d\rho).$$

Combining equation (7.3) with the fact that $\psi(z) \leq \chi(z)$, for any $z \in \mathbb{R}$, we obtain that

$$\psi \left(\int h \, d\rho \right) \leq \int \chi(h) \, d\rho + \frac{1}{8} \mathbf{Var}(h \, d\rho).$$

We conclude the proof by remarking that

$$\psi \left(\int h \, d\rho \right) - \int \chi(h) \, d\rho \leq \sup \psi - \inf \chi \leq \log(4). \qquad \square$$

Applying this result to our problem, we obtain

$$\psi \left(\langle \theta, x \rangle^2 - \lambda \right) \leq \int \chi \left(\langle \theta', x \rangle^2 - \frac{\|x\|^2}{\beta} - \lambda \right) d\pi_\theta (\theta')$$

$$+ \min \left\{ \log(4), \frac{1}{8} \mathbf{Var} \left[\langle \theta', x \rangle^2 \, d\pi_\theta (\theta') \right] \right\},$$

where, putting $m = \langle \theta, x \rangle$, $\sigma = \frac{\|x\|}{\sqrt{\beta}}$ and denoting by $W \sim \mathcal{N}(0, \sigma^2)$ a centered Gaussian random variable,

$$\mathbf{Var} \left[\langle \theta', x \rangle^2 \, d\pi_\theta (\theta') \right] = \mathbf{Var} \left[(m + W)^2 \right]$$

$$= 4m^2 \sigma^2 + 2\sigma^4 = \frac{4 \langle \theta, x \rangle^2 \|x\|^2}{\beta} + \frac{2\|x\|^4}{\beta^2}. \tag{7.4}$$

Let us remark that, for any $a, b, c \in \mathbb{R}_+$ and $W \sim \mathcal{N}(0, \sigma^2)$,

$$\min \{ a, bm^2 + c \} \leq \min \{ a, b(m + W)^2 + c \} + \min \{ a, b(m - W)^2 + c \}, \tag{7.5}$$

since $bm^2 + c \leq \max \{ b(m + W)^2 + c, b(m - W)^2 + c \}$. Therefore, taking the expectation with respect to W of this inequality and remarking that W and $-W$ have the same probability distribution we get

$$\min \{ a, bm^2 + c \} \leq 2\mathbb{E} \left[\min \{ a, b(m + W)^2 + c \} \right].$$

Thus in our context we put $a = \log(4)$, $b = \|x\|^2/(2\beta)$ and $c = \|x\|^4/(4\beta^2)$ and we obtain

$$\psi \left(\langle \theta, x \rangle^2 - \lambda \right) \leq \int \chi \left(\langle \theta', x \rangle^2 - \frac{\|x\|^2}{\beta} - \lambda \right) d\pi_\theta (\theta')$$

$$+ \int \min \left\{ 4 \log(2), \frac{\langle \theta', x \rangle^2 \|x\|^2}{\beta} + \frac{\|x\|^4}{2\beta^2} \right\} d\pi_\theta (\theta').$$

Lemma 7.4. *For any positive constants b, y and any $z \in \mathbb{R}$,*

$$\chi(z) + \min\{b, y\} \leq \log\left(1 + z + \frac{z^2}{2} + y\exp(\sup \chi)\frac{(\exp(b) - 1)}{b}\right).$$

Proof. For any positive real constants a, b, y,

$$\log(a) + \min\{b, y\} = \log\left(a\exp(\min\{b, y\})\right)$$

$$\leq \log\left(a + a\min\{b, y\}\frac{(\exp(b) - 1)}{b}\right),$$

since the function $x \mapsto \frac{\exp(x) - 1}{x}$ is non-decreasing for $x \geq 0$. It follows that

$$\log(a) + \min\{b, y\} \leq \log\left[a + ya\left(\exp(b) - 1\right)/b\right].$$

Applying this inequality to $a = \exp[\chi(z)]$ and reminding that $\chi(z) \leq \log(1 + z + z^2/2)$, we conclude the proof. $\qquad\square$

As a consequence, choosing $b = 4\log(2)$, $z = \langle\theta', x\rangle^2 - \|x\|^2/\beta - \lambda$ and $y = \langle\theta', x\rangle^2\|x\|^2/\beta + \|x\|^4/2\beta^2$, we get

$$\psi\left(\langle\theta, x\rangle^2 - \lambda\right) \leq \int \log\left[1 + \langle\theta'x\rangle^2 - \frac{\|x\|^2}{\beta} - \lambda + \frac{1}{2}\left(\langle\theta', x\rangle^2 - \frac{\|x\|^2}{\beta} - \lambda\right)^2\right.$$

$$\left. + \frac{c\|x\|^2}{\beta}\left(\langle\theta', x\rangle^2 + \frac{\|x\|^2}{2\beta}\right)\right]d\pi_\theta(\theta'),$$

where

$$c = \frac{15}{8\log(2)(\sqrt{2} - 1)}\exp\left(\frac{1 + 2\sqrt{2}}{2}\right) \leq 44.3.$$

We observe that the above inequality allows to compare $\psi(\langle\theta, x\rangle^2 - \lambda)$ with the expectation with respect to the Gaussian perturbation π_θ. In terms of the empirical criterion r_λ, we have proved that

$$r_\lambda(\theta) \leq \frac{1}{n}\sum_{i=1}^{n}\int \log\left[1 + \langle\theta', X_i\rangle^2 - \frac{\|X_i\|^2}{\beta} - \lambda + \frac{1}{2}\left(\langle\theta', X_i\rangle^2 - \frac{\|X_i\|^2}{\beta} - \lambda\right)^2\right.$$

$$\left. + \frac{c\|X_i\|^2}{\beta}\left(\langle\theta', X_i\rangle^2 + \frac{\|X_i\|^2}{2\beta}\right)\right]d\pi_\theta(\theta').$$

We are now ready to use the following general purpose PAC-Bayesian inequality.

Proposition 7.1. *Let $v \in \mathcal{M}_+^1(\mathbb{R}^d)$ be a prior probability distribution on $\mathbb{R}^d$ and let $f : \mathbb{R}^d \times \mathbb{R}^d \to [a, +\infty]$ be a measurable function where $a \in \mathbb{R}$. With probability at least $1 - \varepsilon$, for any posterior distribution $\rho \in \mathcal{M}_+^1(\mathbb{R}^d)$,*

$$\int \frac{1}{n} \sum_{i=1}^{n} f(X_i, \theta') \, d\rho(\theta')$$

$$\leq \int \log\{\mathbb{E}[\exp(f(X, \theta'))]\} \, d\rho(\theta') + \frac{\mathcal{K}(\rho, v) + \log(\varepsilon^{-1})}{n},$$

where

$$\mathcal{K}(\rho, v) = \begin{cases} \int \log\left(\frac{d\rho}{dv}\right) d\rho, & \text{if } \rho \ll v, \\ +\infty, & \text{otherwise,} \end{cases}$$

is the Kullback divergence of ρ with respect to v. By convention, a non-measurable event is said to happen with probability at least $1 - \varepsilon$ when it includes a measurable event of probability non-smaller than $1 - \varepsilon$.

For the proof, we refer to page 1164 of [1].

In our context, we consider as prior distribution $v = \pi_0$ and we restrict the result to posterior distributions ρ belonging to the family of Gaussian perturbations

$$\left\{ \pi_\theta \sim \mathcal{N}(\theta, \beta^{-1}\mathrm{I}) \mid \theta \in \mathbb{R}^d \right\}$$

so that the Kullback divergence is given by

$$\mathcal{K}(\pi_\theta, \pi_0) = \frac{\beta \|\theta\|^2}{2}.$$

We observe that since the result holds for any choice of the posterior, it allows us to obtain uniform results in θ. More precisely, we apply the above PAC-Bayesian inequality to

$$f(X_i, \theta') = \log\left[1 + t(X_i, \theta') + \frac{1}{2} t(X_i, \theta')^2 + \frac{c\|X_i\|^2}{\beta} \left(\langle \theta', X_i \rangle + \frac{\|X_i\|^2}{2\beta} \right) \right],$$

where

$$t(x, \theta') = \langle \theta', x \rangle^2 - \frac{\|x\|^2}{\beta} - \lambda.$$

Using the fact that $\log(1 + t) \leq t$, we get that, with probability at least $1 - \varepsilon$, for any $\theta \in \mathbb{R}^d$,

$$r_\lambda(\theta) \leq \int \mathbb{E}\left[t(X, \theta') + \frac{1}{2} t(X, \theta')^2 + \frac{c\|X\|^2}{\beta} \left(\langle \theta', X \rangle^2 + \frac{\|X\|^2}{2\beta} \right) \right] d\pi_\theta(\theta')$$

$$+ \frac{\beta \|\theta\|^2}{2n} + \frac{\log(\varepsilon^{-1})}{n}$$

$$= \mathbb{E}\left[\langle \theta, X \rangle^2 - \lambda + \frac{1}{2}\left((\langle \theta, X \rangle^2 - \lambda)^2 + \frac{4\langle \theta, X \rangle^2 \|X\|^2}{\beta} + \frac{2\|X\|^4}{\beta^2} \right) \right.$$
$$\left. + \frac{c\|X\|^2}{\beta}\left(\langle \theta, X \rangle^2 + \frac{3\|X\|^2}{2\beta} \right) \right] + \frac{\beta\|\theta\|^2}{2n} + \frac{\log(\varepsilon^{-1})}{n}.$$

To obtain the last line, we have used Lemma 7.1 on page 3894 and equation (7.4) on page 3897.

Let us recall the definition of s_4 and κ introduced in equation (2.5). We have defined

$$s_4 = \mathbb{E}\left[\|X\|^4 \right]^{1/4} \quad \text{and} \quad \kappa = \sup_{\substack{\theta \in \mathbb{R}^d \\ \mathbb{E}[\langle \theta, X \rangle^2] > 0}} \frac{\mathbb{E}[\langle \theta, X \rangle^4]}{\mathbb{E}[\langle \theta, X \rangle^2]^2}.$$

Using the Cauchy–Schwarz inequality and

$$\mathbb{E}\left[\langle \theta, X \rangle^4 \right] \le \kappa N(\theta)^2 \qquad \text{we deduce that } \mathbb{E}\left[\langle \theta, X \rangle^2 \|X\|^2 \right] \le \kappa^{1/2} s_4^2 N(\theta),$$

and we get that, with probability at least $1 - \varepsilon$, for any $\theta \in \mathbb{R}^d$,

$$\begin{aligned}
r_\lambda(\theta) &\le \frac{\kappa}{2}\left[N(\theta) - \lambda \right]^2 + \left[1 + (\kappa - 1)\lambda + \frac{(2 + c)\kappa^{1/2} s_4^2}{\beta} \right]\left[N(\theta) - \lambda \right] \\
&\quad + \frac{(\kappa - 1)\lambda^2}{2} + \frac{(2 + c)\kappa^{1/2} s_4^2 \lambda}{\beta} + \frac{(2 + 3c) s_4^4}{2\beta^2} + \frac{\beta\|\theta\|^2}{2n} + \frac{\log(\varepsilon^{-1})}{n}.
\end{aligned} \tag{7.6}$$

According to the (compact) notation introduced in equation (2.6) on page 3869, the above inequality rewrites as

$$\frac{r_\lambda(\theta)}{\lambda} \le \xi\left(\frac{N(\theta)}{\lambda} - 1 \right)^2 + (1 + \mu)\left(\frac{N(\theta)}{\lambda} - 1 \right) + \gamma + \delta\|\theta\|^2. \tag{7.7}$$

Similarly, observing that

$$-r_\lambda(\theta) = \frac{1}{n}\sum_{i=1}^n \psi\left(\lambda - \langle \theta, X_i \rangle^2 \right),$$

we obtain a lower bound for the empirical criterion r_λ. Namely, with probability at least $1 - \varepsilon$, for any $\theta \in \mathbb{R}^d$, any $(\lambda, \beta) \in \Lambda$,

$$\frac{r_\lambda(\theta)}{\lambda} \ge -\xi\left(\frac{N(\theta)}{\lambda} - 1 \right)^2 + (1 - \mu)\left(\frac{N(\theta)}{\lambda} - 1 \right) - \gamma - \delta\|\theta\|^2. \tag{7.8}$$

We now combine the two bounds above to get the confidence region for $N(\theta)$ defined in Proposition 2.1 on page 3869. Assume that both equation (7.7) and equation (7.8) hold for any $\theta \in \mathbb{R}^d$, an event that happens with probability at least $1 - 2\varepsilon$.

Let us introduce

$$\tau(\theta) = \frac{\lambda\delta\|\theta\|^2}{N(\theta)}$$

and

$$p_\theta(z) = -\xi z^2 + \big[1 - \mu - \tau(\theta)\big]z - \gamma - \tau(\theta), \qquad z \in \mathbb{R}.$$

We observe that $\tau(\alpha\theta) = \tau(\theta)$, and consequently $p_{\alpha\theta}(z) = p_\theta(z)$ for any $\alpha \in \mathbb{R}_+$. We consider the case when $p_\theta(1) > 0$, meaning that

$$\xi + \mu + \gamma + 2\tau(\theta) < 1. \tag{7.9}$$

In this case, the second degree polynomial p_θ has two distinct real roots, z_{-1} and z_{+1}, where

$$z_\sigma = \frac{1 - \mu - \tau(\theta) + \sigma\sqrt{[1 - \mu - \tau(\theta)]^2 - 4\xi[\gamma + \tau(\theta)]}}{2\xi}, \qquad \sigma \in \{1, -1\}.$$

Since equation (7.9) can also be written as $-\xi > -[1 - \mu - \gamma - 2\tau(\theta)]$, we get

$$p_\theta\left(\frac{\gamma + \tau(\theta)}{1 - \mu - \gamma - 2\tau(\theta)}\right)$$
$$> \frac{-[\gamma + \tau(\theta)]^2}{1 - \mu - \gamma - 2\tau(\theta)} + \big[1 - \mu - \tau(\theta)\big]\frac{\gamma + \tau(\theta)}{1 - \mu - \gamma - 2\tau(\theta)} - \gamma - \tau(\theta) = 0,$$

which implies

$$z_{-1} < \frac{\gamma + \tau(\theta)}{1 - \mu - \gamma - 2\tau(\theta)} < z_{+1}.$$

Therefore, since according to equation (7.8), for any $\alpha \in [0, \widehat{\alpha}(\theta)]$,

$$p_\theta\left(\frac{\alpha^2 N(\theta)}{\lambda} - 1\right) \leq \frac{r_\lambda(\alpha\theta)}{\lambda} \leq 0,$$

it is true that

$$\left[-1, \frac{\widehat{\alpha}(\theta)^2 N(\theta)}{\lambda} - 1\right] \cap \,]z_{-1}, z_{+1}[= \varnothing.$$

Observing that $z_{-1} \geq 0 > -1$, it follows that $\widehat{\alpha}(\theta)^2 N(\theta)/\lambda - 1 \leq z_{-1}$. This proves that, for any $\theta \in \mathbb{R}^d$ satisfying equation (7.9),

$$N(\theta) \leq \frac{\lambda}{\widehat{\alpha}(\theta)^2}(1 + z_{-1}) \leq \frac{\lambda}{\widehat{\alpha}(\theta)^2}\left(1 + \frac{\gamma + \tau(\theta)}{1 - \mu - \gamma - 2\tau(\theta)}\right),$$

which rewrites as

$$\Phi_{\theta,+}\big[N(\theta)\big] \leq \frac{\lambda}{\widehat{\alpha}(\theta)^2}.$$

Moreover, this inequality is trivially true when condition (7.9) is not satisfied, because its left-hand side is equal to zero and its right-hand side is non-negative.

Proving the second part of the proposition requires a new argument and not a mere update of signs in the proof of the first part. Although it may seem at first sight that we are just aiming at a reverse inequality, the situation is more subtle than that.

Let us first remark that in the case when

$$\xi - \mu + \gamma + \delta \widehat{\alpha}(\theta)^2 \|\theta\|^2 < 1, \tag{7.10}$$

is not satisfied, the bound

$$\Phi_{\theta,-}\left(\frac{\lambda}{\widehat{\alpha}(\theta)^2}\right) \leq N(\theta) \tag{7.11}$$

is trivially satisfied because the left-hand size is equal to zero. In the case when equation (7.10) is true, it is also true that $\widehat{\alpha}(\theta) < +\infty$, so that $r_\lambda[\widehat{\alpha}(\theta)\theta] = 0$, and therefore, according to equation (7.7),

$$0 \leq q_{\widehat{\alpha}(\theta)\theta}\left(\frac{\widehat{\alpha}(\theta)^2 N(\theta)}{\lambda} - 1\right),$$

where $q_\theta(z) = \xi z^2 + (1 + \mu)z + \gamma + \delta\|\theta\|^2$.

Since condition (7.10) can also be written as $q_{\widehat{\alpha}(\theta)\theta}(-1) < 0$, it implies that the second order polynomial $q_{\widehat{\alpha}(\theta)\theta}$ has two roots and that $\frac{\widehat{\alpha}(\theta)^2 N(\theta)}{\lambda} - 1$ is on the right of its largest root, which is larger than -1. On the other hand, we observe that, under condition (7.10), putting $\widehat{\tau}(\theta) = \delta\widehat{\alpha}(\theta)^2\|\theta\|^2$, we get

$$q_{\widehat{\alpha}(\theta)\theta}\left(-\frac{\gamma + \widehat{\tau}(\theta)}{1 + \mu - \gamma - \widehat{\tau}(\theta)}\right) < \frac{(\gamma + \widehat{\tau}(\theta))^2}{1 + \mu - \gamma - \widehat{\tau}(\theta)} - \frac{(1 + \mu)[\gamma + \widehat{\tau}(\theta)]}{1 + \mu - \gamma + \widehat{\tau}(\theta)} + \gamma + \widehat{\tau}(\theta) = 0.$$

Therefore, when condition (7.10) is satisfied,

$$\frac{\widehat{\alpha}(\theta)^2 N(\theta)}{\lambda} - 1 \geq -\frac{\gamma + \widehat{\tau}(\theta)}{1 + \mu - \gamma - \widehat{\tau}(\theta)},$$

which rewrites as equation (7.11).

7.2. Proof of Proposition 2.3

We first observe that, according to Proposition 2.2 on page 3870, with probability at least $1 - 2\varepsilon$,

$$\left|\frac{\max\{N(\theta), \sigma\|\theta\|^2\}}{\max\{\widehat{N}(\theta), \sigma\|\theta\|^2\}} - 1\right| \leq B_{\widehat{\lambda},\widehat{\beta}}\left[\|\theta\|^{-2}\widehat{N}(\theta)\right] = \inf_{(\lambda,\beta)\in\Lambda} B_{\lambda,\beta}\left[\|\theta\|^{-2}\widetilde{N}_\lambda(\theta)\right]$$

since, by definition, $(\widehat{\lambda}, \widehat{\beta})$ are the values which minimize $B_{\lambda,\beta}[\|\theta\|^{-2}\widetilde{N}_\lambda(\theta)]$. According to equation (2.9) on page 3870, since $\Phi_+(\|\theta\|^{-2}N(\theta)) \leq \|\theta\|^{-2}\widehat{N}(\theta)$ and $B_{\lambda,\beta}$ is a decreasing

function, we get

$$\left| \frac{\max\{N(\theta), \sigma \|\theta\|^2\}}{\max\{\widehat{N}(\theta), \sigma \|\theta\|^2\}} - 1 \right| \le \inf_{(\lambda, \beta) \in \Lambda} B_{\lambda, \beta}\left(\frac{\|\theta\|^{-2} N(\theta)}{1 + B_{\lambda, \beta}[\|\theta\|^{-2} N(\theta)]} \right).$$

With the same notation as in Proposition 2.2, we introduce the subset Γ' of Γ defined as

$$\Gamma' = \big\{ (\lambda, \beta, t) \in \Lambda \times \mathbb{R}_+ \mid \xi + \mu + \gamma + 4\delta\lambda / \max\{t, \sigma\} < 1,$$
$$\mu + \gamma + 2\delta\lambda / \max\{t, \sigma\} \le 1/2,$$
$$\text{and } 2\gamma + \delta\lambda / \max\{t, \sigma\} \le 1/2 \big\}$$

and the function

$$\widetilde{B}_{\lambda, \beta}(t) = \begin{cases} \dfrac{\gamma + \lambda\delta / \max\{t, \sigma\}}{1 - \mu - 2\gamma - 4\lambda\delta / \max\{t, \sigma\}}, & (\lambda, \beta, t) \in \Gamma', \\ +\infty, & \text{otherwise.} \end{cases} \tag{7.12}$$

Lemma 7.5. *For any $(\lambda, \beta) \in \Lambda$ and any $t \in \mathbb{R}_+$,*

$$B_{\lambda, \beta}\left(\frac{t}{1 + B_{\lambda, \beta}(t)} \right) \le \widetilde{B}_{\lambda, \beta}(t),$$

$$\frac{B_{\lambda, \beta}(t)}{1 - B_{\lambda, \beta}(t)} \le \widetilde{B}_{\lambda, \beta}(t).$$

Proof. We first observe that when $(\lambda, \beta, t) \notin \Gamma'$ then $\widetilde{B}_{\lambda, \beta}(t) = +\infty$ and hence the two inequalities are trivial. We now assume $(\lambda, \beta, t) \in \Gamma'$ and we put $\tau = \lambda\delta / \max\{t, \sigma\}$. We prove the second inequality first. Since $\Gamma' \subset \Gamma$, we have

$$\frac{B_{\lambda, \beta}(t)}{1 - B_{\lambda, \beta}(t)} = \frac{\gamma + \tau}{1 - \mu - 2\gamma - 3\tau} \le \widetilde{B}_{\lambda, \beta}(t).$$

In order to prove the first inequality, we first check that $(\lambda, \beta, \frac{t}{1 + B_{\lambda, \beta}(t)}) \in \Gamma$. We start observing that, since

$$\max\{t / [1 + B_{\lambda, \beta}(t)], \sigma\} \ge \max\{t, \sigma\} / [1 + B_{\lambda, \beta}(t)],$$

then

$$\xi + \mu + \gamma + 2\delta\lambda / \max\left\{ \frac{t}{1 + B_{\lambda, \beta}(t)}, \sigma \right\} \le \xi + \mu + \gamma + 2[1 + B_{\lambda, \beta}(t)]\frac{\delta\lambda}{\max\{t, \sigma\}}$$

$$= \xi + \mu + \gamma + 2\tau + 2\tau B_{\lambda, \beta}(t).$$

Moreover, as $(\lambda, \beta, t) \in \Gamma'$, we get

$$B_{\lambda,\beta}(t) = \frac{\gamma + \tau}{1 - \mu - \gamma - 2\tau} \le 1,$$

so that

$$\xi + \mu + \gamma + 2\delta\lambda / \max\{t / [1 + B_{\lambda,\beta}(t)], \sigma\} \le \xi + \mu + \gamma + 4\tau < 1,$$

which proves that, indeed, $(\lambda, \beta, \frac{t}{1 + B_{\lambda,\beta}(t)}) \in \Gamma$. Therefore

$$\begin{aligned}
B_{\lambda,\beta}\left(\frac{t}{1 + B_{\lambda,\beta}(t)}\right) &\le \frac{\gamma + \tau(1 + B_{\lambda,\beta}(t))}{1 - \mu - \gamma - 2\tau[1 + \tau B_{\lambda,\beta}(t)]} \\
&= \frac{(\gamma + \tau)(1 + \tau / (1 - \mu - \gamma - 2\tau))}{1 - \mu - \gamma - 2\tau - 2\tau B_{\lambda,\beta}(t)},
\end{aligned}$$

where in the last line we have used the definition of $B_{\lambda,\beta}$. Observing that

$$1 - \mu - \gamma - 2\tau - 2\tau B_{\lambda,\beta}(t) = \frac{(1 - \mu - \gamma - 2\tau)^2 - 2\tau(\gamma + \tau)}{1 - \mu - \gamma - 2\tau},$$

we obtain

$$\begin{aligned}
B_{\lambda,\beta}\left(\frac{t}{1 + B_{\lambda,\beta}(t)}\right) &\le \frac{(\gamma + \tau)(1 - \mu - \gamma - \tau)}{(1 - \mu - \gamma - 2\tau)^2 - 2\tau(\gamma + \tau)} \\
&= \frac{(\gamma + \tau)(1 - \mu - \gamma - \tau)}{(1 - \mu - \gamma - \tau)^2 + \tau^2 - 2\tau(1 - \mu - \gamma - \tau) - 2\tau^2 - 2\gamma\tau} \\
&= \frac{\gamma + \tau}{1 - \mu - \gamma - \tau - 2\tau - (\tau^2 + 2\gamma\tau) / (1 - \mu - \gamma - \tau)}.
\end{aligned}$$

Considering that

$$(\tau^2 + 2\gamma\tau) / (1 - \mu - \gamma - \tau) \le \tau,$$

since when $(\lambda, \beta, t) \in \Gamma'$, it is true that $1 - \mu - \gamma - \tau \ge 1/2$ and $2\gamma + \tau \le 1/2$, we conclude that

$$B_{\lambda,\beta}\left(\frac{t}{1 + B_{\lambda,\beta}(t)}\right) \le \frac{\gamma + \tau}{1 - \mu - \gamma - 4\tau} = \widetilde{B}_{\lambda,\beta}(t). \qquad \square$$

Applying the above lemma to our problem we get that, with probability at least $1 - 2\varepsilon$, for any $\theta \in \mathbb{R}^d$,

$$\left| \frac{\max\{N(\theta), \sigma\|\theta\|^2\}}{\max\{\widehat{N}(\theta), \sigma\|\theta\|^2\}} - 1 \right| \le \inf_{(\lambda,\beta)\in\Lambda} \widetilde{B}_{\lambda,\beta}[\|\theta\|^{-2}N(\theta)]. \tag{7.13}$$

Let us recall the definition of the finite set Λ given in (2.11) on page 3870. Let $a > 0$ and

$$K = 1 + \left\lceil a^{-1} \log\left(\frac{n}{72(2+c)\kappa^{1/2}}\right) \right\rceil.$$

We define

$$\Lambda = \{(\lambda_j, \beta_j) \mid 0 \le j < K\},$$

where

$$\lambda_j = \sqrt{\frac{2}{n(\kappa-1)}\left(\frac{(2+3c)}{4(2+c)\kappa^{1/2}\exp(-ja)} + \log(K/\varepsilon)\right)}$$

and

$$\beta_j = \sqrt{2(2+c)\kappa^{1/2}s_4^4 n \exp\left[-(j-1/2)a\right]}.$$

We introduce the explicit bound

$$\zeta(t) = \sqrt{2(\kappa-1)\left(\frac{(2+3c)s_4^2}{4(2+c)\kappa^{1/2}t} + \log(K/\varepsilon)\right)}\cosh(a/4) + \sqrt{\frac{2(2+c)\kappa^{1/2}s_4^2}{t}}\cosh(a/2)$$

and

$$B_*(t) = \begin{cases} \dfrac{n^{-1/2}\zeta(\max\{t,\sigma\})}{1-4n^{-1/2}\zeta(\max\{t,\sigma\})}, & \left[6+(\kappa-1)^{-1}\right]\zeta\left(\max\{t,\sigma\}\right) \le \sqrt{n}, \\ +\infty, & \text{otherwise.} \end{cases}$$

Lemma 7.6. *For any $t \in \mathbb{R}_+$, we have*

$$\inf_{(\lambda,\beta)\in\Lambda} \widetilde{B}_{\lambda,\beta}(t) \le B_*\left(\min\{t, s_4^2\}\right). \tag{7.14}$$

Proof. We recall that the function $\widetilde{B}_{\lambda,\beta}$ is non-increasing so that

$$\widetilde{B}_{\lambda,\beta}(t) \le \widetilde{B}_{\lambda,\beta}\left(\min\{t, s_4^2\}\right).$$

Moreover, since

$$\max\{\min\{t, s_4^2\}, \sigma\} = \min\{\max\{t, \sigma\}, s_4^2\},$$

it is sufficient to prove the result for $\max\{t, \sigma\} \in [0, s_4^2]$.

As equation (7.14) is trivial when $B_*(t) = +\infty$, we may assume that $B_*(t) < +\infty$, so that $6\zeta(\max\{t, \sigma\}) \le \sqrt{n}$. In particular, by considering only the second term in the definition of ζ, we obtain that

$$\sqrt{\frac{2(2+c)\kappa^{1/2}s_4^2}{\max\{t,\sigma\}}} \le \sqrt{\frac{2(2+c)\kappa^{1/2}s_4^2}{\max\{t,\sigma\}}}\cosh(a/2) \le \frac{\sqrt{n}}{6},$$

which implies

$$\frac{\max\{t, \sigma\}}{s_4^2} \geq \frac{72(2+c)\kappa^{1/2}}{n} \geq \exp(-a(K-1)).$$

Therefore, since

$$\log\left(\frac{\max\{t, \sigma\}}{s_4^2}\right) \in \left[-a(K-1), 0\right],$$

there exists $\widehat{j} \in \{0, \ldots, K-1\}$ for which

$$\left|\log\left(\frac{\max\{t, \sigma\}}{s_4^2}\right) + \widehat{j}a\right| \leq a/2. \tag{7.15}$$

We recall that by equation (2.6) on page 3869,

$$\gamma + \delta\lambda/\max\{t, \sigma\} = \frac{\lambda}{2}(\kappa-1) + \frac{(2+c)\kappa^{1/2}s_4^2}{\beta} + \frac{(2+3c)s_4^4}{2\beta^2\lambda} + \frac{\log(K/\varepsilon)}{n\lambda} + \frac{\beta}{2n\max\{t, \sigma\}}$$

and we observe that (λ_*, β_*) defined as

$$\lambda_* = \sqrt{\frac{2}{n(\kappa-1)}\left(\frac{(2+3c)s_4^2}{4(2+c)k^{1/2}\max\{t, \sigma\}} + \log(K/\varepsilon)\right)}, \tag{7.16}$$

$$\beta_* = \sqrt{2(2+c)k^{1/2}s_4^2\max\{t, \sigma\}n} \tag{7.17}$$

are the desired values that optimize $\gamma + \delta\lambda/\max\{t, \sigma\}$. We also remark that, by equation (7.15),

$$\beta_{\widehat{j}}\exp(-a/2) \leq \beta_* \leq \beta_{\widehat{j}}, \tag{7.18}$$

$$\lambda_{\widehat{j}}\exp(-a/4) \leq \lambda_* \leq \lambda_{\widehat{j}}\exp(a/4). \tag{7.19}$$

Thus, evaluating $\gamma + \delta\lambda/\max\{t, \sigma\}$ in $(\lambda_{\widehat{j}}, \beta_{\widehat{j}}) \in \Lambda$, we obtain that

$$\gamma_{\widehat{j}} + \delta_{\widehat{j}}\lambda_{\widehat{j}}/\max\{t, \sigma\}$$

$$= \frac{\lambda_*(\kappa-1)}{2}\frac{\lambda_{\widehat{j}}}{\lambda_*}$$

$$+ \frac{(2+c)\kappa^{1/2}s_4^2}{\beta_*}\frac{\beta_*}{\beta_{\widehat{j}}} + \frac{(2+3c)s_4^4}{2\beta_{\widehat{j}}^2\lambda_*}\frac{\lambda_*}{\lambda_{\widehat{j}}} + \frac{\log(K/\varepsilon)}{n\lambda_*}\frac{\lambda_*}{\lambda_{\widehat{j}}} + \frac{\beta_*}{2n\max\{t, \sigma\}}\frac{\beta_{\widehat{j}}}{\beta_*}$$

$$\leq \frac{\lambda_*(\kappa-1)}{2}\frac{\lambda_{\widehat{j}}}{\lambda_*} + \frac{1}{n\lambda_*}\left[\frac{(2+3c)s_4^2}{4(2+c)k^{1/2}\max\{t, \sigma\}} + \log(K/\varepsilon)\right]\frac{\lambda_*}{\lambda_{\widehat{j}}}$$

$$+ \sqrt{\frac{(2+c)\kappa^{1/2}s_4^2}{2n\max\{t, \sigma\}}\left(\frac{\beta_*}{\beta_{\widehat{j}}} + \frac{\beta_{\widehat{j}}}{\beta_*}\right)}$$

$$\leq \sqrt{\frac{2(\kappa-1)}{n}\left[\frac{(2+3c)s_4^2}{4(2+c)k^{1/2}\max\{t,\sigma\}} + \log(K/\varepsilon)\right]\cosh\left[\log\left(\frac{\lambda_{\hat{\jmath}}}{\lambda_*}\right)\right]}$$

$$+ \sqrt{\frac{2(2+c)\kappa^{1/2}s_4^2}{n\max\{t,\sigma\}}}\cosh\left[\log\left(\frac{\beta_{\hat{\jmath}}}{\beta_*}\right)\right].$$

By equation (7.18), we get

$$\gamma_{\hat{\jmath}} + \delta_{\hat{\jmath}}\lambda_{\hat{\jmath}}/\max\{t,\sigma\}$$

$$\leq \sqrt{\frac{2(\kappa-1)}{n}\left[\frac{(2+3c)s_4^2}{4(2+c)k^{1/2}\max\{t,\sigma\}} + \log(K/\varepsilon)\right]\cosh\left(\frac{a}{4}\right)}$$

$$+ \sqrt{\frac{2(2+c)\kappa^{1/2}s_4^2}{n\max\{t,\sigma\}}}\cosh\left(\frac{a}{2}\right).$$

We also observe that

$$\mu_{\hat{\jmath}} + \gamma_{\hat{\jmath}} + 4\delta_{\hat{\jmath}}\lambda_{\hat{\jmath}}/\max\{t,\sigma\} \leq 4\left[\gamma_{\hat{\jmath}} + \delta_{\hat{\jmath}}\lambda_{\hat{\jmath}}/\max\{t,\sigma\}\right] \leq 4n^{-1/2}\zeta(t),$$

since by definition $\mu_{\hat{\jmath}} \leq 2\gamma_{\hat{\jmath}}$. In the same way, observing that

$$\xi_{\hat{\jmath}} = \frac{\kappa\lambda_{\hat{\jmath}}}{2} \leq \gamma_{\hat{\jmath}}\left(1 + \frac{1}{\kappa-1}\right)$$

we obtain

$$\xi_{\hat{\jmath}} + \mu_{\hat{\jmath}} + \gamma_{\hat{\jmath}} + 4\delta_{\hat{\jmath}}\lambda_{\hat{\jmath}}/\max\{t,\sigma\} < \left[4 + (\kappa-1)^{-1}\right]\gamma_{\hat{\jmath}} + 4\delta_{\hat{\jmath}}\lambda_{\hat{\jmath}}/\max\{t,\sigma\}$$

$$\leq \left[6 + (\kappa-1)^{-1}\right]n^{-1/2}\zeta\left(\max\{t,\sigma\}\right)$$

and similarly,

$$2\left[\mu_{\hat{\jmath}} + \gamma_{\hat{\jmath}} + 2\delta_{\hat{\jmath}}\lambda_{\hat{\jmath}}/\max\{t,\sigma\}\right] \leq 6n^{-1/2}\zeta\left(\max\{t,\sigma\}\right),$$

$$2\left[2\gamma_{\hat{\jmath}} + \delta_{\hat{\jmath}}\lambda_{\hat{\jmath}}/\max\{t,\sigma\}\right] \leq 4n^{-1/2}\zeta\left(\max\{t,\sigma\}\right).$$

This implies that, whenever $B_*(t) < +\infty$, then $(\lambda_{\hat{\jmath}}, \beta_{\hat{\jmath}}, t) \in \Gamma'$. We have then proved that

$$\inf_{(\lambda,\beta)\in\Lambda} \widetilde{B}_{\lambda,\beta}(t) \leq \widetilde{B}_{\lambda_{\hat{\jmath}},\beta_{\hat{\jmath}}}(t) \leq \frac{n^{-1/2}\zeta\left(\max\{t,\sigma\}\right)}{1 - 4n^{-1/2}\zeta\left(\max\{t,\sigma\}\right)} = B_*(t). \qquad \square$$

Applying the above lemma to equation (7.13) on page 3904 and observing that, for any $\theta \in \mathbb{R}^d$,

$$\|\theta\|^{-2}N(\theta) \leq \mathbb{E}\left[\|X\|^2\right] \leq s_4^2,$$

we obtain that, with probability at least $1 - 2\varepsilon$, for any $\theta \in \mathbb{R}^d$,

$$\left|\frac{\max\{N(\theta), \sigma\|\theta\|^2\}}{\max\{\widehat{N}(\theta), \sigma\|\theta\|^2\}} - 1\right| \le B_*\big[\|\theta\|^{-2}N(\theta)\big].$$

Since by the Cauchy–Schwarz inequality

$$s_4^2 \le \sqrt{\kappa}\,\mathbf{Tr}(G)$$

we get

$$\zeta(t) \le \sqrt{2(\kappa - 1)\bigg(\frac{(2 + 3c)\,\mathbf{Tr}(G)}{4(2 + c)t} + \log(K/\varepsilon)\bigg)}\cosh(a/4)$$

$$+ \sqrt{\frac{2(2 + c)\kappa\,\mathbf{Tr}(G)}{t}}\cosh(a/2). \tag{7.20}$$

Choosing $a = 1/2$ and computing explicitly numerical constants concludes the proof.

7.3. Proof of Proposition 3.3

We observe that it is sufficient to prove that with probability at least $1 - 2\varepsilon$

$$\Phi_- \circ \Phi_+\big(\langle\theta, \mathcal{Q}\theta\rangle_{\mathcal{H}} - \eta\big) \le N(\Pi_k\theta) + \eta,$$
$$\Phi_- \circ \Phi_+\big(N(\Pi_k\theta) - \eta\big) \le \langle\theta, \mathcal{Q}\theta\rangle_{\mathcal{H}} + \eta, \tag{7.21}$$

where $\eta = 2\delta\sqrt{\mathbf{Tr}(\mathcal{G}^2)}$ and $N(\Pi_k\theta) = \langle\Pi_k\theta, \mathcal{G}\Pi_k\theta\rangle_{\mathcal{H}}$. Indeed, if equation (7.21) is satisfied, according to the postponed Corollary 7.1 on page 3912,

$$\big|\max\big\{\langle\theta, \mathcal{Q}\theta\rangle_{\mathcal{H}}, \sigma\big\} - \max\big\{N(\Pi_k\theta), \sigma\big\}\big|$$
$$\le 2\max\big\{N(\Pi_k\theta), \sigma\big\}B_*\big(N(\Pi_k\theta)\big) + 5\eta/2,$$
$$\big|\max\big\{\langle\theta, \mathcal{Q}\theta\rangle_{\mathcal{H}}, \sigma\big\} - \max\big\{N(\Pi_k\theta), \sigma\big\}\big|$$
$$\le 2\max\big\{\langle\theta, \mathcal{Q}\theta\rangle_{\mathcal{H}}, \sigma\big\}B_*\big(\min\big\{\langle\theta, \mathcal{Q}\theta\rangle_{\mathcal{H}}, s_4^2\big\}\big) + 5\eta/2,$$

which is the analogous, in the infinite-dimensional setting, of Proposition 2.4 on page 3873. Thus, following the proof of Proposition 2.5 we obtain the desired bounds.

Let us now prove equation (7.21). Observe that, for any $\theta \in \mathbb{S}_{\mathcal{H}}$,

$$\langle\theta, \mathcal{Q}\theta\rangle_{\mathcal{H}} = \langle\Pi_{V_k}\theta, \mathcal{Q}\Pi_{V_k}\theta\rangle_{\mathcal{H}} \le \|\Pi_{V_k}\theta\|_{\mathcal{H}}^2\big(\langle\xi, \mathcal{Q}\xi\rangle_{\mathcal{H}} + \eta\big),$$

where $\xi \in \Theta_\delta$ is the closest point in Θ_δ to $\|\Pi_{V_k}\theta\|_{\mathcal{H}}^{-1}\Pi_{V_k}\theta$. Since $\xi \in \mathcal{H}_k$, with probability at least $1 - \varepsilon$, for any $(\lambda, \beta) \in \Lambda$,

$$\langle\xi, \mathcal{Q}\xi\rangle_{\mathcal{H}} \le \Phi_+^{-1}\big(\widetilde{N}_\lambda(\xi)\big) = \Phi_+^{-1}\big[\widetilde{N}_\lambda\big(\xi + \|\Pi_{V_k}\theta\|_{\mathcal{H}}^{-1}(\Pi_k - \Pi_{V_k})\theta\big)\big].$$

Let us now remark that for any $a \in [0, 1]$, we have $\Phi_+(at) \leq a\Phi_+(t)$, so that $a\Phi_+^{-1}(t) \leq \Phi_+^{-1}(at)$. Therefore

$$
\begin{aligned}
\langle \theta, \mathcal{Q}\theta \rangle_{\mathcal{H}} &\leq \|\Pi_{V_k}\theta\|_{\mathcal{H}}^2 \Phi_+^{-1}\big\{\widetilde{N}_\lambda\big[\|\Pi_{V_k}\theta\|_{\mathcal{H}}^{-1}\big(\|\Pi_{V_k}\theta\|_{\mathcal{H}}\xi + (\Pi_k - \Pi_{V_k})\theta\big)\big]\big\} + \eta \\
&\leq \|\Pi_{V_k}\theta\|_{\mathcal{H}}^2 \Phi_+^{-1} \circ \Phi_-^{-1}\big\{N\big[\|\Pi_{V_k}\theta\|_{\mathcal{H}}^{-1}\big(\|\Pi_{V_k}\theta\|_{\mathcal{H}}\xi + (\Pi_k - \Pi_{V_k})\theta\big)\big]\big\} + \eta \\
&\leq \Phi_+^{-1} \circ \Phi_-^{-1}\big\{N\big[\|\Pi_{V_k}\theta\|_{\mathcal{H}}\xi + (\Pi_k - \Pi_{V_k})\theta\big)\big]\big\} + \eta \\
&\leq \Phi_+^{-1} \circ \Phi_-^{-1}\big(N(\Pi_k\theta) + \eta\big) + \eta.
\end{aligned}
$$

Indeed,

$$
\big\|\big(\|\Pi_{V_k}\theta\|_{\mathcal{H}}\xi + (\Pi_k - \Pi_{V_k})\theta\big) - \Pi_k\theta\big\| \leq \delta,
$$

and this is a difference of two vectors belonging to the unit ball. In the same way

$$
\begin{aligned}
\langle \theta, \mathcal{Q}\theta \rangle_{\mathcal{H}} &\geq \|\Pi_{V_k}\theta\|_{\mathcal{H}}^2\big(\langle \xi, \mathcal{Q}\xi \rangle_{\mathcal{H}} - \eta\big) \\
&\geq \|\Pi_{V_k}\theta\|_{\mathcal{H}}^2 \Phi_-\big\{\widetilde{N}_\lambda\big[\|\Pi_{V_k}\theta\|_{\mathcal{H}}^{-1}\big(\|\Pi_{V_k}\theta\|_{\mathcal{H}}\xi + (\Pi_k - \Pi_{V_k})\theta\big)\big]\big\} - \eta \\
&\geq \|\Pi_{V_k}\theta\|_{\mathcal{H}}^2 \Phi_- \circ \Phi_+\big\{N\big[\|\Pi_k\theta\|_{\mathcal{H}}^{-1}\big(\|\Pi_{V_k}\theta\|_{\mathcal{H}}\xi + (\Pi_k - \Pi_{V_k})\theta\big)\big]\big\} - \eta \\
&\geq \Phi_- \circ \Phi_+\big(N(\Pi_k\theta) - \eta\big) - \eta
\end{aligned}
$$

which proves equation (7.21).

7.4. A technical result

In all this section, we use the same notation as in Section 2. Let $\sigma \in {]}0, s_4^2]$ be such that $8\zeta_*(\sigma) \leq \sqrt{n}$ where ζ_* is defined in Proposition 2.3 on page 3871.

Lemma 7.7. *The function*

$$
t \mapsto F(t) = \max\{t, \sigma\} B_*\big(\min\{t, s_4^2\}\big),
$$

where B_* is defined in Proposition 2.3, is non-decreasing for any $t \in \mathbb{R}_+$.

Proof. If $\sigma \geq s_4^2$, then $B_*(\min\{t, s_4^2\}) = B_*(\sigma)$, so that $F(t) = \max\{t, \sigma\} B_*(\sigma)$ is obviously non-decreasing. Otherwise, $\sigma \leq s_4^2$, so that

$$
\zeta\big(\max\{\min\{t, s_4^2\}, \sigma\}\big) = \zeta\big(\min\{\max\{t, \sigma\}, s_4^2\}\big).
$$

Therefore, the function F is of the form

$$
F(t) = c\frac{ug(u)}{(1 - g(u))},
$$

where $u = \max\{t, \sigma\}$,

$$g(u) = \sqrt{a_1/u + a_2} + \sqrt{a_3/u},$$

$g(\sigma) \leq 1/2$, and the constants $c, a_1, a_2,$ and a_3 are positive. Let $h(u) = \sqrt{a_1/u} + \sqrt{a_3/u}$ and observe that

$$g'(u) = -\frac{1}{2u}\left(\frac{a_1/u}{(a_1/u + a_2)^{1/2}} + \sqrt{a_3/u}\right) \geq -\frac{1}{2u}(\sqrt{a_1/u} + \sqrt{a_3/u}) = h'(u)$$

and that $g(u) \geq h(u)$. Therefore $h(u) \leq g(u) \leq 1/2$, for any $u \geq \sigma$, and

$$\frac{\partial}{\partial u} \log\left(\frac{ug(u)}{1 - g(u)}\right) = \frac{1}{u} + \frac{g'(u)}{g(u)(1 - g(u))}$$

$$\geq \frac{1}{u} + \frac{h'(u)}{h(u)(1 - h(u))}$$

$$= \frac{1}{u} - \frac{1}{2u(1 - h(u))} \geq 0,$$

showing that F is non-decreasing. □

Lemma 7.8. *For any $(a, b) \in \mathbb{R}^2$ such that, for any $(\lambda, \beta) \in \Lambda$,*

$$\Phi_- \circ \Phi_+(a - \eta) \leq b + \eta, \quad \text{and} \quad \Phi_- \circ \Phi_+(b - \eta) \leq a + \eta,$$

and any threshold $\sigma \in \mathbb{R}_+$ such that $8\zeta(\sigma) \leq \sqrt{n}$ and $\sigma \leq s_4^2$, we have

$$\left|\max\{a, \sigma\} - \max\{b, \sigma\}\right| \leq 2\max\{a + \eta, \sigma\}B_*\left(\min\{a + \eta, s_4^2\}\right) + 2\eta, \quad (7.22)$$

$$\left|\max\{a, \sigma\} - \max\{b, \sigma\}\right| \leq 2\max\{b + \eta, \sigma\}B_*\left(\min\{b + \eta, s_4^2\}\right) + 2\eta. \quad (7.23)$$

Proof. By symmetry of a and b, equation (7.23) is a consequence of equation (7.22).
　　Step 1. We will prove that

$$\max\{b - \eta, \sigma\} \leq \max\{a + \eta, \sigma\}\left(1 + 2\widetilde{B}_{\lambda,\beta}(a + \eta)\right), \quad (7.24)$$

where $\widetilde{B}_{\lambda,\beta}$ is defined in equation (7.12) on page 3903.
　　Case 1. Assume that

$$\max\{\Phi_+(b - \eta), \sigma\} \leq \max\{a + \eta, \sigma\},$$

and remark that, since Φ_+ is non-decreasing and $\Phi_+(\sigma) \leq \sigma$,

$$\max\{\Phi_+(b - \eta), \sigma\} \geq \max\{\Phi_+(b - \eta), \Phi_+(\sigma)\}$$

$$= \Phi_+\left(\max\{(b - \eta), \sigma\}\right) = \frac{\max\{b - \eta, \sigma\}}{1 + B_{\lambda,\beta}(b - \eta)},$$

according to equation (2.9) on page 3870, where $B_{\lambda,\beta}$ is defined in equation (2.8). Therefore in this case,

$$\max\{b - \eta, \sigma\} \leq \max\{a + \eta, \sigma\}\big(1 + B_{\lambda,\beta}(b - \eta)\big), \tag{7.25}$$

but when $\max\{b - \eta, \sigma\} > \max\{a + \eta, \sigma\}$,

$$B_{\lambda,\beta}(b - \eta) \leq B_{\lambda,\beta}(a + \eta)$$

because $B_{\lambda,\beta}(t)$ is a non-increasing function of $\max\{t, \sigma\}$, thus equation (7.25) implies that

$$\max\{b - \eta, \sigma\} \leq \max\{a + \eta, \sigma\}\big(1 + B_{\lambda,\beta}(a + \eta)\big).$$

Since $B_{\lambda,\beta} \leq \widetilde{B}_{\lambda,\beta}$, equation (7.24) holds true.

Case 2. Assume now that we are not in **Case 1**, implying that

$$\max\{b - \eta, \sigma\} \geq \max\big\{\Phi_+(b - \eta), \sigma\big\} > \max\{a + \eta, \sigma\}.$$

In this case

$$\begin{aligned}
\max\{a + \eta, \sigma\} &\geq \max\big\{\Phi_- \circ \Phi_+(b - \eta), \sigma\big\} \\
&\geq \max\big\{\Phi_- \circ \Phi_+(b - \eta), \Phi_-(\sigma)\big\} \\
&\geq \Phi_-\big(\max\big\{\Phi_+(b - \eta), \sigma\big\}\big) \\
&\geq \max\big\{\Phi_+(b - \eta), \sigma\big\}\big[1 - B_{\lambda,\beta}\big(\max\big\{\Phi_+(b - \eta), \sigma\big\}\big)\big]
\end{aligned}$$

according to equation (2.10) on page 3870. Moreover, continuing the above chain of inequalities,

$$\begin{aligned}
\max\{a + \eta, \sigma\} &\geq \max\big\{\Phi_+(b - \eta), \Phi_+(\sigma)\big\}\big[1 - B_{\lambda,\beta}\big(\max\{a + \eta, \sigma\}\big)\big] \\
&= \Phi_+\big(\max\{b - \eta, \sigma\}\big)\big[1 - B_{\lambda,\beta}(a + \eta)\big] \\
&\geq \max\{b - \eta, \sigma\}\frac{1 - B_{\lambda,\beta}(a + \eta)}{1 + B_{\lambda,\beta}(\max\{b - \eta, \sigma\})} \\
&\geq \max\{b - \eta, \sigma\}\frac{1 - B_{\lambda,\beta}(a + \eta)}{1 + B_{\lambda,\beta}(a + \eta)}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\max\{b - \eta, \sigma\} &\leq \max\{a + \eta, \sigma\}\frac{1 + B_{\lambda,\beta}(a + \eta)}{1 - B_{\lambda,\beta}(a + \eta)} \\
&= \max\{a + \eta, \sigma\}\bigg(1 + \frac{2B_{\lambda,\beta}(a + \eta)}{1 - B_{\lambda,\beta}(a + \eta)}\bigg) \\
&\leq \max\{a + \eta, \sigma\}\big(1 + 2\widetilde{B}_{\lambda,\beta}(a + \eta)\big)
\end{aligned}$$

according to Lemma 7.5 on page 3903. This concludes the proof of **Step 1**.

Step 2. Taking the infimum in $(\lambda, \beta) \in \Lambda$ in equation (7.24), according to equation (7.14) on page 3905, we obtain that

$$\max\{b - \eta, \sigma\} \leq \max\{a + \eta, \sigma\}\big(1 + 2B_*\big(\min\{a + \eta, s_4^2\}\big)\big).$$

We can then use the fact that $t \mapsto \max\{t, \sigma\}B_*(\min\{t, s_4^2\})$ is non-decreasing (proved in Lemma 7.7 on page 3909) to deduce that

$$\max\{b - \eta, \sigma\} \leq \max\{a + \eta, \sigma\} + 2\max\{b + \eta, \sigma\}B_*\big(\min\{b + \sigma, s_4^2\}\big),$$

since there is nothing to prove when already $\max\{b + \eta, \sigma\} \leq \max\{a + \eta, \sigma\}$. Remark that $\max\{a + \eta, \sigma\} \leq \max\{a + \eta, \sigma + \eta\} \leq \max\{a, \sigma\} + \eta$ and that in the same way $\max\{b - \eta, \sigma\} \geq \max\{b, \sigma\} - \eta$. This proves that

$$\max\{b, \sigma\} - \max\{a, \sigma\} \leq 2\max\{a + \eta, \sigma\}B_*\big(\min\{a + \eta, s_4^2\}\big) + 2\eta$$

and

$$\max\{b, \sigma\} - \max\{a, \sigma\} \leq 2\max\{b + \eta, \sigma\}B_*\big(\min\{b + \eta, s_4^2\}\big) + 2\eta.$$

By symmetry, we can then exchange a and b to prove the same bounds for $\max\{a, \sigma\} - \max\{b, \sigma\}$, and therefore also for the absolute value of this quantity, which ends the proof of the lemma. $\qquad\square$

As a consequence the following result holds.

Corollary 7.1. *For any $(a, b) \in \mathbb{R}^2$ such that, for any $(\lambda, \beta) \in \Lambda$,*

$$\Phi_- \circ \Phi_+(a - \eta) \leq b + \eta, \quad \text{and} \quad \Phi_- \circ \Phi_+(b - \eta) \leq a + \eta,$$

and any threshold $\sigma \in \mathbb{R}_+$ such that $8\zeta(\sigma) \leq \sqrt{n}$ and $\sigma \leq s_4^2$, we have

$$\big|\max\{a, \sigma\} - \max\{b, \sigma\}\big| \leq 2\max\{a, \sigma\}B_*\big(\min\{a, s_4^2\}\big) + 5\eta/2,$$

$$\big|\max\{a, \sigma\} - \max\{b, \sigma\}\big| \leq 2\max\{b, \sigma\}B_*\big(\min\{b, s_4^2\}\big) + 5\eta/2.$$

Proof. This is a consequence of the previous lemma, of the fact that $B_*(\min\{t, s_4^2\}) \leq 1/4$, and of the fact that $\max\{a + \eta, \sigma\} \leq \max\{a, \sigma\} + \eta$. $\qquad\square$

7.5. Proof of Proposition 5.1

Let $\Lambda \subset (\mathbb{R}_+ \setminus \{0\})^2$ be the finite set defined in equation (2.11) on page 3870. We use as a tool the family of estimators

$$\widetilde{N}_\lambda(\theta) = \frac{\lambda}{\widehat{\alpha}(\theta)^2}$$

introduced in equation (2.4), where $\widehat{\alpha}(\theta)$ is defined in equation (2.3). Let us put

$$\tau_\lambda(t) = \frac{\lambda^2 R^4}{3 \max\{t, \sigma\}^2}, \qquad t \in \mathbb{R}_+.$$

We divide the proof into 4 steps.

Step 1. The first step consists in linking the empirical estimator $\bar{N}$ with $\widetilde{N}_\lambda$.

We claim that, with probability at least $1 - \varepsilon$, for any $\theta \in \mathbb{S}_d$, any $(\lambda, \beta) \in \Lambda$, such that $\Phi_+(N(\theta)) > 0$,

$$\frac{\bar{N}(\theta)}{\max\{\widetilde{N}_\lambda(\theta), \sigma\}} \leq [1 - \tau_\lambda(\widetilde{N}_\lambda(\theta))]_+^{-1},$$

with the convention that $\frac{1}{0} = +\infty$. Moreover, with probability at least $1 - \varepsilon$, for any $\theta \in \mathbb{S}_d$, any $(\lambda, \beta) \in \Lambda$, such that $\Phi_+(N(\theta)) > 0$,

$$\frac{\bar{N}(\theta)}{\widetilde{N}_\lambda(\theta)} \geq 1 - \frac{\lambda^2}{3}.$$

We first observe that, according to the definition of $\widehat{\alpha}(\theta)$, for any threshold $\sigma \in \mathbb{R}_+$,

$$\frac{1}{n} \sum_{i=1}^n \psi\left[\lambda\left(\max\{\widetilde{N}_\lambda(\theta), \sigma\}^{-1} \langle \theta, X_i \rangle^2 - 1\right)\right] \leq r\left(\lambda^{1/2}\widetilde{N}_\lambda(\theta)^{-1/2}\theta\right) = r_\lambda\left(\widehat{\alpha}(\theta)\theta\right) \leq 0,$$

where we have used the fact that the function ψ, introduce in equation (2.2), is non-decreasing. Moreover

$$r_\lambda\left(\widehat{\alpha}(\theta)\theta\right) = 0$$

as soon as $\widehat{\alpha}(\theta) < +\infty$ and this holds true, according to Proposition 2.1, with probability at least $1 - \varepsilon$, for any $\theta \in \mathbb{S}_d$ and any $(\lambda, \beta) \in \Lambda$ such that $\Phi_+(N(\theta)) > 0$. Indeed, by Proposition 2.1, with probability at least $1 - \varepsilon$, for any $\theta \in \mathbb{S}_d$, any $(\lambda, \beta) \in \Lambda$,

$$\widetilde{N}_\lambda(\theta) \geq \Phi_+\left(N(\theta)\right).$$

Defining $g(z) = z - \psi(z)$, we get

$$\frac{\bar{N}(\theta)}{\max\{\widetilde{N}_\lambda(\theta), \sigma\}} - 1 = \frac{1}{n\lambda} \sum_{i=1}^n \lambda\left(\langle \theta, X_i \rangle^2 \max\{\widetilde{N}_\lambda(\theta), \sigma\}^{-1} - 1\right)$$

$$\leq \frac{1}{n\lambda} \sum_{i=1}^n g\left[\lambda\left(\langle \theta, X_i \rangle^2 \max\{\widetilde{N}_\lambda(\theta), \sigma\}^{-1} - 1\right)\right]. \tag{7.26}$$

In the same way, with probability at least $1 - \varepsilon$, for any $\theta \in \mathbb{S}_d$, any $(\lambda, \beta) \in \Lambda$ such that $\Phi_+(N(\theta)) > 0$, we obtain

$$1 - \frac{\bar{N}(\theta)}{\widetilde{N}_\lambda(\theta)} \leq \frac{1}{n\lambda} \sum_{i=1}^n g\left[\lambda\left(1 - \langle \theta, X_i \rangle^2 \widetilde{N}_\lambda(\theta)^{-1}\right)\right]. \tag{7.27}$$

We remark that the derivative of g is

$$g'(z) = 1 - \psi'(z) = \begin{cases} 1 & \text{if } z \notin [-1, 1], \\[2mm] \dfrac{\frac{z^2}{2}}{1 + z + \frac{z^2}{2}} & \text{if } z \in [-1, 0], \\[4mm] \dfrac{\frac{z^2}{2}}{1 - z + \frac{z^2}{2}} & \text{if } z \in [0, 1], \end{cases}$$

showing that $0 \le g'(z) \le z^2$, and therefore that g is a non-decreasing function satisfying

$$g(z) \le \frac{1}{3} z_+^3. \tag{7.28}$$

Applying equation (7.28) to equation (7.26) we obtain

$$\frac{\bar{N}(\theta)}{\max\{\widetilde{N}_\lambda(\theta), \sigma\}} - 1 \le \frac{\lambda^2}{3n} \sum_{i=1}^n \big(\langle \theta, X_i \rangle^2 \max\{\widetilde{N}_\lambda(\theta), \sigma\}^{-1} - 1\big)_+^3$$

$$\le \frac{\lambda^2}{3n \max\{\widetilde{N}_\lambda(\theta), \sigma\}^3} \sum_{i=1}^n \langle \theta, X_i \rangle^6,$$

where we have used the fact that $(z^2 - 1)_+ \le z^2$. Since, by the Cauchy–Schwarz inequality, $\langle \theta, X_i \rangle^2 \le \|\theta\|^2 R^2 = R^2$, we get

$$\frac{\bar{N}(\theta)}{\max\{\widetilde{N}_\lambda(\theta), \sigma\}} - 1 \le \frac{\lambda^2}{3n \max\{\widetilde{N}_\lambda(\theta), \sigma\}^3} R^4 \sum_{i=1}^n \langle \theta, X_i \rangle^2$$

$$= \frac{\lambda^2}{3} \times \frac{R^4}{\max\{\widetilde{N}_\lambda(\theta), \sigma\}^2} \times \frac{\bar{N}(\theta)}{\max\{\widetilde{N}_\lambda(\theta), \sigma\}},$$

which proves the first inequality. Similarly, since g in non-decreasing, we obtain that, with probability at least $1 - \varepsilon$, for any $\theta \in \mathbb{S}_d$, any $(\lambda, \beta) \in \Lambda$ such that $\Phi_+(N(\theta)) > 0$,

$$1 - \frac{\bar{N}(\theta)}{\widetilde{N}_\lambda(\theta)} \le \frac{1}{n\lambda} \sum_{i=1}^n g(\lambda) \le \frac{\lambda^2}{3},$$

where the last inequality follows from equation (7.28).

Step 2. This is an intermediate step. We claim that, with probability at least $1 - 2\varepsilon$, for any $\theta \in \mathbb{S}_d$, any $(\lambda, \beta) \in \Lambda$, any $\sigma > 0$,

$$\max\{\bar{N}(\theta), \sigma\} \le \Phi_-^{-1}\big(\max\{N(\theta), \sigma\}\big)\big[1 - \tau_\lambda\big(N(\theta)\big)\big]_+^{-1},$$

$$\max\{\bar{N}(\theta), \sigma\} \le \Phi_-^{-1}\big(\max\{N(\theta), \sigma\}\big)\big[1 - \tau_\lambda\big(\bar{N}(\theta)\big[1 - \tau_\lambda(\sigma)\big]_+\big)\big]_+^{-1},$$

$$\bar{N}(\theta) \geq \left(1 - \frac{\lambda^2}{3}\right)_+ \Phi_+\big(N(\theta)\big),$$

where Φ_+ and Φ_- are defined in Proposition 2.1.

We consider the threshold

$$\sigma' = \Phi_-^{-1}\big(\max\{N(\theta), \sigma\}\big) \geq \max\{N(\theta), \sigma\},$$

where we have used the fact that, by definition, $\Phi_-(t)^{-1} \geq t$, for any $t \in \mathbb{R}_+$. We assume that we are in the intersection of the two events of Proposition 2.1, which holds true with probability at least $1 - 2\varepsilon$, so that

$$\sigma' \geq \max\{N(\theta), \sigma, \widetilde{N}_\lambda(\theta)\}. \tag{7.29}$$

According to **Step 1**, choosing as a threshold $\max\{\sigma, \sigma'\}$, we get

$$\frac{\bar{N}(\theta)}{\max\{\widetilde{N}_\lambda(\theta), \sigma, \sigma'\}} \leq \big[1 - \tau_\lambda\big(\max\{\widetilde{N}_\lambda(\theta), \sigma'\}\big)\big]_+^{-1}$$

(where τ_λ is still defined with respect to σ), so that, according to equation (7.29),

$$\bar{N}(\theta) \leq \sigma'\big[1 - \tau_\lambda(\sigma')\big]_+^{-1}. \tag{7.30}$$

As a consequence, recalling the definition of σ', we have

$$\bar{N}(\theta) \leq \Phi_-^{-1}\big(\max\{N(\theta), \sigma\}\big)\big[1 - \tau_\lambda\big(N(\theta)\big)\big]_+^{-1}.$$

Thus, observing that

$$\sigma \leq \Phi_-^{-1}(\sigma) \leq \Phi_-^{-1}\big(\max\{N(\theta), \sigma\}\big)\big[1 - \tau_\lambda\big(N(\theta)\big)\big]_+^{-1},$$

we obtain the first inequality. To prove the second inequality, we use equation (7.30) once to see that

$$\sigma' \geq \bar{N}(\theta)\big[1 - \tau_\lambda(\sigma')\big]_+ \geq \bar{N}(\theta)\big[1 - \tau_\lambda(\sigma)\big]_+,$$

and we use it again to get

$$\bar{N}(\theta) \leq \Phi_-^{-1}\big(\max\{N(\theta), \sigma\}\big)\big[1 - \tau_\lambda(\sigma')\big]_+^{-1}$$
$$\leq \Phi_-^{-1}\big(\max\{N(\theta), \sigma\}\big)\big[1 - \tau_\lambda\big(\bar{N}(\theta)\big[1 - \tau_\lambda(\sigma)\big]_+\big)\big]_+^{-1}.$$

To complete the proof of the second inequality, it is enough to remark that

$$\sigma \leq \Phi_-^{-1}(\sigma) \leq \Phi_-^{-1}\big(\max\{N(\theta), \sigma\}\big)\big[1 - \tau_\lambda\big(\bar{N}(\theta)\big[1 - \tau_\lambda(\sigma)\big]_+\big)\big]_+^{-1}.$$

To prove the last inequality, it is sufficient to remark that $\widetilde{N}_\lambda(\theta) \geq \Phi_+(N(\theta))$ by Proposition 2.1 and hence, when $\Phi_+(N(\theta)) > 0$,

$$\bar{N}(\theta) \geq \left(1 - \frac{\lambda^2}{3}\right)_+ \Phi_+\big(N(\theta)\big).$$

On the other hand, when $\Phi_+(N(\theta)) = 0$, this inequality is also obviously satisfied.

Step 3. We now prove that, with probability at least $1 - 2\varepsilon$, for any $\theta \in \mathbb{S}_d$, any $(\lambda, \beta) \in \Lambda$, any $\sigma > 0$,

$$\frac{\max\{\bar{N}(\theta), \sigma\}}{\max\{N(\theta), \sigma\}} - 1 \leq \widetilde{B}_{\lambda,\beta}\big(N(\theta)\big) + \frac{\tau_\lambda(N(\theta))}{[1 - \tau_\lambda(N(\theta))]_+[1 - B_{\lambda,\beta}(N(\theta))]_+},$$

$$1 - \frac{\max\{\bar{N}(\theta), \sigma\}}{\max\{N(\theta), \sigma\}} \leq B_{\lambda,\beta}\big(N(\theta)\big) + \frac{\lambda^2}{3},$$

where $B_{\lambda,\beta}$ is defined in equation (2.8) on page 3870 and $\widetilde{B}_{\lambda,\beta}$ in equation (7.12) on page 3903.

We observe that, according to **Step 2**,

$$\max\big\{\bar{N}(\theta), \sigma\big\} \leq \Phi_-^{-1}\big(\max\{N(\theta), \sigma\}\big)[1 - \tau_\lambda\big(N(\theta)\big)]_+^{-1}$$

$$\leq \frac{\max\{N(\theta), \sigma\}}{[1 - \tau_\lambda(N(\theta))]_+[1 - B_{\lambda,\beta}(N(\theta))]_+},$$

which implies

$$\frac{\max\{\bar{N}(\theta), \sigma\}}{\max\{N(\theta), \sigma\}} - 1 \leq \frac{B_{\lambda,\beta}(N(\theta))}{[1 - B_{\lambda,\beta}(N(\theta))]_+} + \frac{\tau_\lambda(N(\theta))}{[1 - \tau_\lambda(N(\theta))]_+[1 - B_{\lambda,\beta}(N(\theta))]_+}.$$

Applying Lemma 7.5 on page 3903, we obtain the first inequality.

To prove the second inequality we observe that, using again **Step 2**,

$$\max\big\{\bar{N}(\theta), \sigma\big\} \geq \left(1 - \frac{\lambda^2}{3}\right)_+ \Phi_+\big(\max\{N(\theta), \sigma\}\big)$$

$$= \left(1 - \frac{\lambda^2}{3}\right)_+ \max\big\{N(\theta), \sigma\big\}[1 + B_{\lambda,\beta}\big(N(\theta)\big)]^{-1},$$

where we have used the fact that $\Phi_+(\max\{z, \sigma\}) = \max\{z, \sigma\}(1 + B_{\lambda,\beta}(z))^{-1}$ as shown in equation (2.9) on page 3870. Thus, we conclude that

$$1 - \frac{\max\{\bar{N}(\theta), \sigma\}}{\max\{N(\theta), \sigma\}} \leq \frac{B_{\lambda,\beta}(N(\theta)) + \lambda^2/3}{1 + B_{\lambda,\beta}(N(\theta))} \leq B_{\lambda,\beta}\big(N(\theta)\big) + \frac{\lambda^2}{3}.$$

Step 4. From **Step 3**, we deduce that

$$\left|\frac{\max\{\bar{N}(\theta), \sigma\}}{\max\{N(\theta), \sigma\}} - 1\right| \leq \widetilde{B}_{\lambda,\beta}\big(N(\theta)\big) + \frac{\tau_\lambda(N(\theta))}{[1 - \tau_\lambda(N(\theta))]_+[1 - B_{\lambda,\beta}(N(\theta))]_+}.$$

To conclude the proof it is sufficient to apply **Step 3** to $(\lambda_{\widehat{\jmath}}, \beta_{\widehat{\jmath}}) \in \Lambda$ defined in equation (2.11) on page 3870. Indeed, by equation (7.14) on page 3905, for any $t \in \mathbb{R}_+$,

$$B_{\lambda_{\widehat{\jmath}}, \beta_{\widehat{\jmath}}}(t) \leq B_*(t)$$

and, by equation (7.19) on page 3906, we have $\lambda_{\widehat{\jmath}} \leq \lambda_*(\theta) \exp(a/4)$.

7.6. Proof of Proposition 5.2

We observe that another way to take advantage of equation (7.26) on page 3913 is to write

$$\frac{\bar{N}(\theta)}{\max\{\widetilde{N}_\lambda(\theta), \sigma\}} - 1 \leq \frac{\lambda^2 \|\theta\|^6}{3 \max\{\widetilde{N}_\lambda(\theta), \sigma\}^3} \frac{1}{n} \sum_{i=1}^n \|X_i\|^6.$$

Thus, putting

$$\zeta_\lambda(t) = \frac{\lambda^2 \widetilde{R}^6}{3 \max\{t, \sigma\}^3}, \qquad t \in \mathbb{R}_+,$$

we get that, for any $\theta \in \mathbb{S}_d$,

$$\frac{\bar{N}(\theta)}{\max\{\widetilde{N}_\lambda(\theta), \sigma\}} \leq 1 + \zeta_\lambda\big(\widetilde{N}_\lambda(\theta)\big).$$

The same reasoning used to prove **Step 2** of Proposition 5.1 on page 3885 shows that, with probability at least $1 - \varepsilon$, for any $\theta \in \mathbb{S}_d$, any $(\lambda, \beta) \in \Lambda$, any $\sigma > 0$,

$$\max\big\{\bar{N}(\theta), \sigma\big\} \leq \Phi_-^{-1}\big(\max\{N(\theta), \sigma\}\big)\big[1 + \zeta_\lambda\big(N(\theta)\big)\big].$$

As a consequence, with probability at least $1 - 2\varepsilon$, for any $\theta \in \mathbb{S}_d$, any $(\lambda, \beta) \in \Lambda$,

$$\left| \frac{\max\{\bar{N}(\theta), \sigma\}}{\max\{N(\theta), \sigma\}} - 1 \right| \leq \widetilde{B}_{\lambda,\beta}\big(N(\theta)\big) + \frac{\zeta_\lambda(N(\theta))}{[1 - B_{\lambda,\beta}(N(\theta))]_+}.$$

7.7. Proof of Proposition 6.1

To prove Proposition 6.1, we use many results already proved in the case of the Gram matrix (with the necessary adjustments).

We first observe that, denoting by $W \in \mathbb{R}^d$ a gaussian random vector with mean $A^{1/2}\theta$ and covariance matrix $\beta^{-1}A$, we have

$$\mathbb{E}\big[\big\|A^{1/2}\theta'\big\|^2 \, d\pi_\theta\big(\theta'\big)\big] = \sum_{i=1}^d \mathbb{E}\big(\langle W, e_i\rangle^2\big)$$

$$= \big\|A^{1/2}\theta\big\|^2 + \frac{\mathbf{Tr}(A)}{\beta},$$

where $\{e_i\}_{i=1}^d$ is the canonical basis of $\mathbb{R}^d$ (and $\langle W, e_i \rangle$ is a one-dimensional Gaussian random variable with mean $\langle A^{1/2}\theta, e_i \rangle$ and variance $\beta^{-1}e_i^\top A e_i$). Therefore the empirical criterion rewrites as

$$r_\lambda(\theta) = \frac{1}{n} \sum_{i=1}^n \psi \left[\int \left(\|A_i^{1/2}\theta'\|^2 - \frac{\mathbf{Tr}(A_i)}{\beta} - \lambda \right) d\pi_\theta(\theta') \right].$$

We now use Lemma 7.3 on page 3896 to pull the expectation outside the influence function ψ. We decompose A into $A = UDU^\top$, where $UU^\top = I$ and $D = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$ and we observe that $U^\top \theta'$ has the same distribution as $U^\top \theta + W$, where $W \sim \mathcal{N}(0, \beta^{-1}I)$ so that

$$\mathbf{Var}\left[\|A^{1/2}\theta'\|^2 d\pi_\theta(\theta')\right] = \mathbf{Var}\left(\sum_{i=1}^d ((U^\top\theta)_i + W_i)^2 \lambda_i \right)$$

$$= \sum_{i=1}^d \lambda_i^2 \mathbf{Var}\left[((U^\top\theta)_i + W_i)^2 \right]$$

$$= \sum_{i=1}^d \left(\frac{2}{\beta^2} + \frac{4}{\beta}(U^\top\theta) \right)_i^2 \lambda_i^2$$

$$= \frac{2}{\beta^2} \mathbf{Tr}(A^2) + \frac{4}{\beta} \|A\theta\|^2.$$

As a consequence we get

$$r_\lambda(\theta) \le \frac{1}{n} \sum_{i=1}^n \left[\int \chi \left(\|A_i^{1/2}\theta'\|^2 - \frac{\mathbf{Tr}(A_i)}{\beta} - \lambda \right) d\pi_\theta(\theta') \right.$$

$$\left. + \min \left\{ \log(4), \frac{1}{2\beta} \|A_i\theta\|^2 + \frac{\mathbf{Tr}(A_i^2)}{4\beta^2} \right\} \right],$$

where the function χ is defined in equation (7.1) on page 3895. We then apply equation (7.5) on page 3897 with $m = \|A\theta\|$, $a = \log(4)$, $b = 1/(2\beta)$ and $c = \mathbf{Tr}(A^2)/(4\beta^2)$ to obtain

$$r_\lambda(\theta) \le \frac{1}{n} \sum_{i=1}^n \left[\int \chi \left(\|A^{1/2}\theta'\|^2 - \frac{\mathbf{Tr}(A)}{\beta} - \lambda \right) d\pi_\theta(\theta') \right.$$

$$\left. + \int \min \left\{ 4\log(2), \frac{1}{\beta} \|A\theta'\|^2 + \frac{\mathbf{Tr}(A^2)}{2\beta^2} \right\} d\pi_\theta(\theta') \right]$$

and we conclude, by Lemma 7.4 on page 3898, that

$$r_\lambda(\theta) \le \frac{1}{n} \sum_{i=1}^n \int \log\left[1 + \|A_i^{1/2}\theta'\|^2 - \frac{\mathbf{Tr}(A_i)}{\beta} - \lambda + \frac{1}{2}\left(\|A_i^{1/2}\theta'\|^2 - \frac{\mathbf{Tr}(A_i)}{\beta} - \lambda \right)^2 \right.$$

$$+ \frac{c}{\beta}\left(\|A_i\theta'\|^2 + \frac{\mathbf{Tr}(A_i^2)}{2\beta}\right)\bigg]\mathrm{d}\pi_\theta(\theta'),$$

where

$$c = \frac{15}{8\log(2)(\sqrt{2}-1)}\exp\left(\frac{1+2\sqrt{2}}{2}\right).$$

We then apply the PAC-Bayesian inequality (7.1) to

$$f(A_i,\theta') = \log\left[1 + t(A_i,\theta') + \frac{1}{2}t(A_i,\theta')^2 + \frac{c}{\beta}\left(\|A_i\theta'\|^2 + \frac{\mathbf{Tr}(A_i^2)}{2\beta}\right)\right],$$

where

$$t(A,\theta') = \|A^{1/2}\theta'\|^2 - \frac{\mathbf{Tr}(A)}{\beta} - \lambda$$

and we choose as posterior distributions the family of Gaussian perturbations π_θ. We obtain that, with probability at least $1 - \varepsilon$, for any $\theta \in \mathbb{R}^d$,

$$r_\lambda(\theta) \leq \int \mathbb{E}\left[t(A,\theta') + \frac{1}{2}t(A,\theta')^2 + \frac{c}{\beta}\left(\|A\theta'\|^2 + \frac{\mathbf{Tr}(A^2)}{2\beta}\right)\right]\mathrm{d}\pi_\theta(\theta')$$

$$+ \frac{\beta\|\theta\|^2}{2n} + \frac{\log(\varepsilon^{-1})}{n}$$

$$= \mathbb{E}\left[\|A^{1/2}\theta\|^2 - \lambda + \frac{1}{2}\left(\|A^{1/2}\theta\|^2 - \lambda\right)^2 + \frac{(c+2)\|A\theta\|^2}{\beta} + \frac{(2+3c)\mathbf{Tr}(A^2)}{2\beta^2}\right]$$

$$+ \frac{\beta\|\theta\|^2}{2n} + \frac{\log(\varepsilon^{-1})}{n}.$$

Using the Cauchy–Schwarz inequality, we remark that

$$\mathbb{E}\left[\|A\theta\|^2\right] \leq \mathbb{E}\left[\|A\|_\infty\|A^{1/2}\theta\|^2\right]$$

$$\leq \mathbb{E}\left[\|A\|_\infty^2\right]^{1/2}\mathbb{E}\left[\|A^{1/2}\theta\|^4\right]^{1/2}$$

$$\leq \mathbb{E}\left[\|A\|_\infty^2\right]^{1/2}\kappa^{1/2}N(\theta),$$

where κ is defined in equation (6.1) on page 3888. Thus

$$r_\lambda(\theta) \leq \frac{\kappa}{2}\left[N(\theta) - \lambda\right]^2 + \left[1 + (\kappa-1)\lambda + \frac{(2+c)\kappa^{1/2}\mathbb{E}[\|A\|_\infty^2]^{1/2}}{\beta}\right]\left[N(\theta) - \lambda\right]$$

$$+ \frac{(\kappa-1)\lambda^2}{2} + \frac{(2+c)\kappa^{1/2}\mathbb{E}[\|A\|_\infty^2]^{1/2}\lambda}{\beta}$$

$$+ \frac{(2+3c)\mathbb{E}[\mathbf{Tr}(A^2)]}{2\beta^2} + \frac{\beta\|\theta\|^2}{2n} + \frac{\log(\varepsilon^{-1})}{n}.$$

This is the analogous of equation (7.6) on page 3900. The end of the proof is similar to the case of the Gram matrix.

7.8. Proof of Lemma 6.1

Replacing X with $X - \mathbb{E}[X]$ we may assume during the proof that $\mathbb{E}[X] = 0$. It is true that

$$\mathbb{E}\big[\|A^{1/2}\theta\|^4\big] = \mathbb{E}\left[\left(\frac{1}{q(q-1)}\sum_{1 \leq j < k \leq q} \langle \theta, X^{(j)} - X^{(k)}\rangle^2\right)^2\right]$$

$$= \frac{1}{q^2(q-1)^2} \sum_{\substack{1 \leq j < k \leq q \\ 1 \leq s < t \leq q}} \mathbb{E}\big[\langle \theta, X^{(j)} - X^{(k)}\rangle^2 \langle \theta, X^{(s)} - X^{(t)}\rangle^2\big].$$

Recalling the definition of the covariance, we have

$$\mathbb{E}\big[\|A^{1/2}\theta\|^4\big] = \frac{1}{q^2(q-1)^2}\Bigg\{\sum_{\substack{1 \leq j < k \leq q \\ 1 \leq s < t \leq q}} \mathbb{E}\big[\langle \theta, X^{(j)} - X^{(k)}\rangle^2\big]\mathbb{E}\big[\langle \theta, X^{(s)} - X^{(t)}\rangle^2\big]$$

$$+ \sum_{1 \leq j < k \leq q} \mathbb{E}\big[\langle \theta, X^{(j)} - X^{(k)}\rangle^4\big] - \mathbb{E}\big[\langle \theta, X^{(j)} - X^{(k)}\rangle^2\big]^2$$

$$+ \sum_{\substack{1 \leq j < k \leq q \\ 1 \leq s < t \leq q \\ |\{j,k\} \cap \{s,t\}| = 1}} \big(\mathbb{E}\big[\langle \theta, X^{(j)} - X^{(k)}\rangle^2 \langle \theta, X^{(s)} - X^{(t)}\rangle^2\big]$$

$$- \mathbb{E}\big[\langle \theta, X^{(j)} - X^{(k)}\rangle^2\big]\mathbb{E}\big[\langle \theta, X^{(s)} - X^{(t)}\rangle^2\big]\big)\Bigg\}$$

$$= \frac{1}{4}\mathbb{E}\big[\langle \theta, X^{(2)} - X^{(1)}\rangle^2\big]^2$$

$$+ \frac{1}{2q(q-1)}\mathbb{E}\big[\langle \theta, X^{(2)} - X^{(1)}\rangle^4\big] - \mathbb{E}\big[\langle \theta, X^{(2)} - X^{(1)}\rangle^2\big]^2$$

$$+ \frac{q-2}{q(q-1)}\big(\mathbb{E}\big[\langle \theta, X^{(1)} - X^{(2)}\rangle^2 \langle \theta, X^{(1)} - X^{(3)}\rangle^2\big]$$

$$- \mathbb{E}\big[\langle \theta, X^{(1)} - X^{(2)}\rangle^2\big]^2\big).$$

Define $W_j = \langle \theta, X^{(j)}\rangle$ and observe that

$$\mathbb{E}\big[(W_1 - W_2)^2\big]^2 = 4N(\theta)^2,$$

$$\mathbb{E}\big[(W_1 - W_2)^4\big] = \mathbb{E}\big[W_1^4\big] + 6\mathbb{E}\big[W_1^2\big]\mathbb{E}\big[W_2^2\big] + \mathbb{E}\big[W_2^4\big]$$

$$= 2\mathbb{E}\big[W_1^4\big] + 6\mathbb{E}\big[W_1^2\big]^2 \le (2\kappa + 6)N(\theta)^2,$$

$$\mathbb{E}\big[(W_1 - W_2)^2(W_1 - W_3)^2\big] = \mathbb{E}\big[W_1^4\big] + 3\mathbb{E}\big[W_2^2\big]^2 \le (\kappa + 3)N(\theta)^2.$$

Therefore

$$\mathbb{E}\big[\|A^{1/2}\theta\|^4\big] \le \left(1 + \frac{(q-2)(\kappa-1)}{q(q-1)} + \frac{(\kappa+1)}{q(q-1)}\right)N(\theta)^2 = \left(1 + \frac{\tau_q(\kappa)}{q}\right)N(\theta)^2,$$

and hence $\kappa' \le 1 + \tau_q(\kappa)/q$, since $\mathbb{E}[\|A^{1/2}\theta\|^2] = N(\theta)$.

7.9. Proof of Lemma 6.2

Replacing X with $X - \mathbb{E}[X]$ we may assume that $\mathbb{E}[X] = 0$. Recall that

$$\mathbb{E}\big[\|X\|^4\big] \le \kappa \mathbb{E}\big[\|X\|^2\big]^2 = \kappa \, \mathbf{Tr}(\Sigma)^2$$

and $\mathbb{E}[\langle X^{(1)}, X^{(2)}\rangle^2] = \mathbf{Tr}(\Sigma^2)$. We observe that

$$\mathbb{E}\big[\|A\theta\|^2\big] = \mathbb{E}\left[\frac{1}{q^2(q-1)^2}\sum_{\substack{1 \le j < k \le q \\ 1 \le s < t \le q}} \langle\theta, X^{(j)} - X^{(k)}\rangle\langle X^{(j)} - X^{(k)}, X^{(s)} - X^{(t)}\rangle\langle X^{(s)} - X^{(t)}, \theta\rangle\right]$$

and

$$\mathbb{E}\big[\langle\theta, X^{(1)} - X^{(2)}\rangle\langle X^{(1)} - X^{(2)}, X^{(3)} - X^{(4)}\rangle\langle X^{(3)} - X^{(4)}, \theta\rangle\big]$$

$$= 4\mathbb{E}\big[\langle\theta, X^{(1)}\rangle\langle X^{(1)}, X^{(2)}\rangle\langle X^{(2)}, \theta\rangle\big]$$

$$= 4\|\Sigma\theta\|^2 \le 4\|\Sigma\|_\infty N(\theta),$$

$$\mathbb{E}\big[\langle\theta, X^{(1)} - X^{(2)}\rangle\langle X^{(1)} - X^{(2)}, X^{(1)} - X^{(3)}\rangle\langle X^{(1)} - X^{(3)}, \theta\rangle\big]$$

$$= \mathbb{E}\big[\langle\theta, X^{(1)}\rangle^2\|X^{(1)}\|^2\big] + 3\mathbb{E}\big[\langle\theta, X^{(1)}\rangle\langle X^{(1)}, X^{(2)}\rangle\langle X^{(2)}, \theta\rangle\big]$$

$$\le \kappa \, \mathbf{Tr}(\Sigma)N(\theta) + 3\|\Sigma\|_\infty N(\theta),$$

$$\mathbb{E}\big[\langle\theta, X^{(1)} - X^{(2)}\rangle\langle X^{(1)} - X^{(2)}, X^{(1)} - X^{(2)}\rangle\langle X^{(1)} - X^{(2)}, \theta\rangle\big]$$

$$= 2\mathbb{E}\big[\langle\theta, X^{(1)}\rangle^2\|X^{(1)}\|^2\big] + 2\mathbb{E}\big[\langle\theta, X^{(1)}\rangle^2\big]\mathbb{E}\big[\|X^{(1)}\|^2\big]$$

$$+ 4\mathbb{E}\big[\langle\theta, X^{(1)}\rangle\langle X^{(1)}, X^{(2)}\rangle\langle X^{(2)}, \theta\rangle\big]$$

$$\le 2(\kappa + 1)\mathbf{Tr}(\Sigma)N(\theta) + 4\|\Sigma\|_\infty N(\theta),$$

which proves the first inequality. In the same way,

$$\mathbb{E}\big[\mathbf{Tr}(A^2)\big] = \mathbb{E}\left[\frac{1}{q^2(q-1)^2}\sum_{\substack{1 \le j < k \le q \\ 1 \le s < t \le q}} \langle X^{(j)} - X^{(k)}, X^{(s)} - X^{(t)}\rangle^2\right],$$

and

$$\mathbb{E}\big[\langle X^{(1)} - X^{(2)}, X^{(3)} - X^{(4)}\rangle^2\big] = 4\mathbb{E}\big[\langle X^{(1)}, X^{(2)}\rangle^2\big]$$
$$= 4\,\mathbf{Tr}\big(\Sigma^2\big),$$
$$\mathbb{E}\big[\langle X^{(1)} - X^{(2)}, X^{(1)} - X^{(3)}\rangle^2\big] = \mathbb{E}\big[\|X^{(1)}\|^4\big] + 3\mathbb{E}\big[\langle X^{(1)}, X^{(2)}\rangle^2\big]$$
$$\le \kappa\,\mathbf{Tr}(\Sigma)^2 + 3\,\mathbf{Tr}\big(\Sigma^2\big),$$
$$\mathbb{E}\big[\langle X^{(1)} - X^{(2)}, X^{(1)} - X^{(2)}\rangle^2\big] = 2\mathbb{E}\big[\|X^{(1)}\|^4\big] + 2\mathbb{E}\big[\|X^{(1)}\|^2\big]^2 + 4\mathbb{E}\big[\langle X^{(1)}, X^{(2)}\rangle^2\big]$$
$$\le 2(\kappa + 1)\,\mathbf{Tr}(\Sigma)^2 + 4\,\mathbf{Tr}\big(\Sigma^2\big),$$

which concludes the proof.

Acknowledgements

References

[1] Catoni, O. (2012). Challenging the empirical mean and empirical variance: A deviation study. *Ann. Inst. Henri Poincaré Probab. Stat*. **48** 1148–1185. MR3052407

[2] Catoni, O. (2016). PAC-Bayesian bounds for the Gram matrix and least square regression with a random design. Preprint. Available at arXiv:1603.05229.

[3] Giulini, I. (2015). Generalization bounds for random samples in Hilbert spaces. Ph.D. thesis.

[4] Koltchinskii, V. and Giné, E. (2000). Random matrix approximation of spectra of integral operators. *Bernoulli* **6** 113–167. MR1781185

[5] Koltchinskii, V. and Lounici, K. (2016). Asymptotics and concentration bounds for bilinear forms of spectral projectors of sample covariance. *Ann. Inst. Henri Poincaré Probab. Stat*. **52** 1976–2013. MR3573302

[6] Minsker, S. (2015). Geometric median and robust estimation in Banach spaces. *Bernoulli* **21** 2308–2335. MR3378468

[7] Rosasco, L., Belkin, M. and De Vito, E. (2010). On learning with integral operators. *J. Mach. Learn. Res*. **11** 905–934. MR2600634

[8] Rudelson, M. (1999). Random vectors in the isotropic position. *J. Funct. Anal*. **164** 60–72. MR1694526

[9] Shawe-Taylor, J., Williams, C., Cristianini, N. and Kandola, J. (2002). On the eigenspectrum of the Gram matrix and its relationship to the operator eigenspectrum. In *Algorithmic Learning Theory*. *Lecture Notes in Computer Science* **2533** 23–40. Berlin: Springer. MR2071605

[10] Shawe-Taylor, J., Williams, C.K.I., Cristianini, N. and Kandola, J. (2005). On the eigenspectrum of the Gram matrix and the generalization error of kernel-PCA. *IEEE Trans. Inform. Theory* **51** 2510–2522. MR2246374

[11] Tropp, J.A. (2012). User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.* **12** 389–434. MR2946459

[12] Vershynin, R. (2012). Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing* 210–268. Cambridge: Cambridge Univ. Press. MR2963170

[13] von Luxburg, U., Belkin, M. and Bousquet, O. (2008). Consistency of spectral clustering. *Ann. Statist.* **36** 555–586. MR2396807

[14] Zwald, L., Bousquet, O. and Blanchard, G. (2004). Statistical properties of kernel principal component analysis. In *Learning Theory. Lecture Notes in Computer Science* **3120** 594–608. Berlin: Springer. MR2177937