Probab. Theory Relat. Fields (2016) 166:1175–1194
DOI 10.1007/s00440-016-0738-9

The lower tail of random quadratic forms with applications to ordinary least squares

Roberto Imbuzeiro Oliveira[1]

Received: 10 December 2013 / Revised: 22 August 2016 / Published online: 21 September 2016

Abstract Finite sample properties of random covariance-type matrices have been the subject of much research. In this paper we focus on the "lower tail" of such a matrix, and prove that it is sub-Gaussian under a simple fourth moment assumption on the one-dimensional marginals of the random vectors. A similar result holds for more general sums of random positive semidefinite matrices, and our (relatively simple) proof uses a variant of the so-called PAC-Bayesian method for bounding empirical processes. Using this bound, we obtain a nearly optimal finite-sample result for the ordinary least squares estimator under random design.

Keywords Random covariance matrices · Linear regression

Mathematics Subject Classification 60F99 · 94A15 · 62J05

1 Introduction

Let $X_1, \ldots, X_n$ be i.i.d. random (column) vectors in $\mathbb{R}^p$ with finite second moments. This paper contributes to the problem of obtaining finite-sample concentration bounds for the random covariance-type operator

Supported by a *Bolsa de Produtividade em Pesquisa* from CNPq, Brazil. This article was produced as part of the activities of FAPESP Center for Neuromathematics (Grant # 2013/ 07699-0, S. Paulo Research Foundation).

✉ Roberto Imbuzeiro Oliveira
rimfo@impa.br

1 IMPA: Estrada Dona Castorina, 110, Rio de Janeiro, RJ 22460-320, Brazil

$$\widehat{\Sigma}_n := \frac{1}{n}\sum_{i=1}^{n} X_i X_i^T \tag{1}$$

with mean $\Sigma := \mathbb{E}\left[X_1 X_1^T\right]$. This problem has received a great deal of attention recently, and has important applications to the estimation of covariance matrices [15,22], to the analysis of methods for least squares problems [10] and to compressed sensing and high dimensional, small sample size statistics [2,18,21].

The most basic problem is computing how many samples are needed to bring $\widehat{\Sigma}_n$ close to Σ. In general one needs at least $n \geq p$ to bring $\widehat{\Sigma}_n$ close to Σ, so that the ranks of the two matrices can match. A basic problem is to find conditions under which $n \geq C(\varepsilon)\, p$ samples are enough for guaranteeing

$$\Pr(\forall v \in \mathbb{R}^p,\ (1-\varepsilon) v^T \Sigma v \leq v^T \widehat{\Sigma}_n\, v \leq (1+\varepsilon)\, v^T \Sigma\, v) \approx 1, \tag{2}$$

where $C(\varepsilon)$ depends only on $\varepsilon > 0$ and on moment assumptions on the X_i's.

A well known bound by Rudelson [17,20] implies $C(\varepsilon)\, p \log p$ samples are necessary and sufficient if the vectors $\Sigma^{-1/2} X_i/\sqrt{p}$ have uniformly bounded norms. Removing the $\log p$ factor is relatively easy for sub-Gaussian vectors X_i, but even the seemingly nice case of log-concave random vectors (which have sub-exponential moments) had to wait for the breakthrough papers by Adamczak et al. [1,3]. A series of results [9,12,15,22] have proven similar results under finite-moment conditions on the one-dimensional marginals plus a (necessary) a high probability bound on $\max_{i \leq n} |X_i|_2$.

1.1 The sub-Gaussian lower tail

In this paper we focus on concentration properties of the lower tail of $\widehat{\Sigma}_n$. As it turns out, information about the lower tail is sufficient for many applications, including the analysis of regression-type problems (see Theorem 1.2 below for an example). Moreover, the asymmetry between upper and lower tails is interesting from a purely mathematical perspective.

Our main result is the following theorem.

Theorem 1.1 (Proven in Sect. 4) Let $X_1, \ldots, X_n$ be i.i.d. copies of a random vector $X \in \mathbb{R}^p$ with finite fourth moments. Define $\Sigma := \mathbb{E}\left[X X^T\right]$ and assume

$$\forall v \in \mathbb{R}^p \ : \ \sqrt{\mathbb{E}\left[(v^T X)^4\right]} \leq \mathsf{h}\, v^T \Sigma\, v \tag{3}$$

for some $\mathsf{h} \in (1, +\infty)$. Set

$$\widehat{\Sigma}_n := \frac{1}{n}\sum_{i=1}^{n} X_i X_i^T \text{ as in (1).}$$

Then, if the number n of samples satisfies

$$n \geq 81\mathsf{h}^2\,(p + 2\ln(2/\delta))/\varepsilon^2,$$

we have the bound

$$\Pr(\forall v \in \mathbb{R}^p \;:\; v^T\,\widehat{\Sigma}_n\,v \geq (1-\varepsilon)\,v^T\,\Sigma\,v) \geq 1-\delta. \tag{4}$$

Notice that the sample size n in Theorem 1.1 depends on p/ε^2, which is optimal if X_1 has i.i.d. entries due to the Bai-Yin theorem [5]. Moreover, the dependence of n on $\ln(1/\delta)/\varepsilon^2$ shows that the sample size depends on the confidence level δ in a sub-Gaussian fashion. More precisely, Theorem 1.1 implies that

$$V := 1 - \inf_{v^T\,\Sigma\,v=1} v^T\,\widehat{\Sigma}_n\,v$$

has a Gaussian-like right tail

$$\Pr\left(V \geq C\,\mathsf{h}\sqrt{\frac{p}{n}} + r\right) \leq e^{-r^2 n/C\mathsf{h}^2}\;(r \geq 0),$$

with $C > 0$ universal. We observe that Theorem 1.1 has a more general version involving general sums of i.i.d. positive semidefinite random matrices; see Theorem 4.1 below for details.

The main assumption in Theorem 1.1 is (3). This is a finite-moment assumption, and, from a theoretical perspective, it seems remarkable that one can obtain sub-Gaussian concentration from it. From the perspective of applications, there are reasonably natural settings where (3) is a sensible assumption.

1. Assume first $X = (X[1], X[2], \ldots, X[p])^T$ has diagonal Σ and satisfies a near unbiasedness assumption: for all $(i_1, i_2, i_3, i_4) \in \{1, \ldots, p\}^4$,

   $$i_4 \notin \{i_1, i_2, i_3\} \Rightarrow \mathbb{E}\,[X[i_1]\,X[i_2]\,X[i_3]\,X[i_4]] = 0.$$

   This is true if $X[1], X[2], \ldots, X[p]$ are mean-zero four-wise independent random variables, or if $X[1], X[2], \ldots, X[p]$ is unconditional (i.e. its law is preserved when each coordinate is multiplied by a sign). From this assumption we may obtain (3) with

   $$\mathsf{h} := 6\,\max\left\{\frac{\sqrt{\mathbb{E}\left[X[i]^4\right]}}{\mathbb{E}\left[X[i]^2\right]} \;:\; i = 1, 2, \ldots, p,\; \mathbb{E}\left[X^2[i]\right] > 0\right\}.$$

2. Assume now that some X satisfying (3) is replaced by $AX + \mu$ for some linear map $A \in \mathbb{R}^{p \times p'}$ and some $\mu \in \mathbb{R}^{p'}$. The new vector still satisfies $\mathsf{h} < +\infty$, although h may change by a universal constant factor. Note that the matrix A may be singular and/or that one may have $p' > p$, in which case $AX + \mu$ will have

highly correlated components. This is allowed if X "comes from a vector with uncorrelated entries".

3. The property $\mathsf{h} < +\infty$ is also preserved when X is multiplied by an independent scalar ξ, as long as $\mathbb{E}\left[\xi^4\right]/\mathbb{E}\left[\xi^2\right]^2$ is bounded by an absolute constant. As noted in [22], this is strictly weaker than what is needed for two-sided concentration as in (2).
4. An assumption *not* covered by our theorem is that of *bounded designs*: $\Sigma^{-1/2}X_1/\sqrt{p}$ a.s. bounded. This is verified when the coordinates X are a orthonormal functions such as the Fourier basis over [0, 1] (in this case $\Sigma = I_{p\times p}$). We note that this bounded design case is optimally covered by Rudelson's aforementioned bound [17,20].

One further attraction of Theorem 1.1 is its proof method, which is based on a PAC-Bayesian argument. The main feature of this method is that it provides a way to control empirical processes via entropic inequalities, as opposed to usual chaining methods. Further details about this method are given in Sect. 3 below. Although our application of this method is indebted to previous work by Audibert/Catoni [4] and Langford/Shawe-Taylor [13], we believe that this technique has much greater potential than what has been explored so far in the literature.

Remark 1 (Recent developments in lower tails) Many developments on variants of Theorem 1.1 have appeared since the first version of this paper. Almost simultaneously with us, Koltchinskii and Mendelson [11] obtained analogues of Theorem 1.1 under the assumption of $q > 4$ moments on the one dimensional marginals of X_1. They also obtained results under our assumption (3), albeit with suboptimal dependence on p and ε. Later, Yaskov [24,25] obtained bounds under the assumption of uniform bounds for the weak L^q norms of one dimensional marginals, where $q \geq 2$ is arbitrary. For each value of $q > 2$, he obtains the optimal exponent $\alpha_q > 0$ so that $n = \Theta(p/\varepsilon^{\alpha_q})$ samples are necessary and sufficient for (4) (with $\delta = e^{-p}$). His theorem is thus stronger Theorem 4.1 except possibly for the dependence of n on δ. However, the first part of [8, Theorem 3.1] by van de Geer and Muro achieves similar bounds as Yaskov, with the same dependence on δ as our own Theorem 1.1. There has also been some related progress in checking lower- and upper-tail properties that are relevant in the $p \gg n$ setting [9].

1.2 Application to ordinary least squares with random design

Theorem 1.1 will be illustrated with an application to random design linear regression when $n \gg p \gg 1$. In this setting one is given *data* in the form of n i.i.d. copies $(X_i, Y_i)_{i=1}^n$ of a random pair $(X, Y) \in \mathbb{R}^p \times \mathbb{R}$, where X is a vector of covariates and Y is a response variable. The goal is to find a vector $\widehat{\beta}_n$ that depends solely on the data so that the square loss

$$\ell(\beta) := \mathbb{E}\left[(Y - X^T\beta)^2\right]$$

is as small as possible. This setting of random design should be contrasted with the technically simpler case of fixed design, where the X_i are non-random. Fixed design results are not informative about out-of-sample prediction, which is important in many routine applications of OLS e.g. in Statistical Learning and in Linear Aggregation.

We show below that the usual ordinary least squares (OLS) estimator

$$\widehat{\beta}_n \in \mathrm{argmin}_{\beta\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n}(Y_i - \beta^T X_i)^2,$$

achieves error rates $\approx \sigma^2\,(p + \ln(1/\delta))/n$ in the random design setting, where σ^2 measures the magnitude of "errors". The formal theorem (modulo some definitions in Sect. 5.1) is as follows.

Theorem 1.2 (Proven in Sect. 5.2) Define $(X, Y), (X_1, Y_1), \dots, (X_n, Y_n)$ as above. Let $\beta_{\min}$ denote a minimizer of ℓ and let $\eta := Y - \beta_{\min}^T X$. Define $\Sigma := \mathbb{E}\left[X X^T\right]$, and let $\Sigma^{-1/2}$ be the Moore-Penrose pseudoinverse of $\Sigma^{1/2}$. Also set $Z := \eta\,\Sigma^{-1/2} X$. Let $\mathsf{h}, \sigma^2, \mathsf{h}_* > 0$ and $q > 2$ and assume that, for all $v \in \mathbb{R}^p$,

$$\sqrt{\mathbb{E}\left[(v^T X)^4\right]} \le \mathsf{h}\, v^T \Sigma\, v; \tag{5}$$

$$\mathbb{E}\left[(v^T Z)^2\right] \le \sigma^2\, |v|_2^2; \text{ and} \tag{6}$$

$$\sqrt[q]{\mathbb{E}\left[|Z|_2^q\right]} \le \mathsf{h}_*\, \sigma\, \sqrt{p}. \tag{7}$$

Then for any $\varepsilon \in (0, 1/2)$, there exists $C > 0$ depending only on h_*, ε and q such that, when $\delta \in (C/n^{q/2-1}, 1)$ and

$$n \ge C\mathsf{h}^2\,(p + 2\ln(4/\delta)),$$

then

$$\Pr\left(\ell(\widehat{\beta}_n) - \inf_{\beta\in\mathbb{R}^p} \ell(\beta) \le \frac{(1+\varepsilon)\,\sigma^2}{n}\,\left(\sqrt{p} + C\sqrt{\ln(4/\delta)}\right)^2\right) \ge 1 - \delta.$$

So for $n \gg p \gg 1$ the excess loss of OLS is bounded by $(1 + o\,(1))\,\sigma^2 p/n$, with high probability. This can be shown to be tight; cf. the end of Sect. 5.1 for details. An important point is that Theorem 1.2 makes minimal assumptions on the data, and works in a completely model-free, non-parametric, heteroskedastic setting. Our moment assumptions are reasonable e.g. when those of Theorem 1.1 are reasonable and $Z = \eta\,\Sigma^{-1/2} X$ is not too far from isotropic. For instance, if the "noise" term η is independent from X, this property follows from suitable moment assumptions on the noise and on the one-dimensional marginals of X. Even when there is no independence, one only needs higher moment assumptions on X and η (thanks to Hölder's inequality).

Theorem 1.2 extends recent papers Hsu et al. [10] and Audibert and Catoni [4]. Hsu et al. prove a variant of Theorem 1.2 where they assume an independent noise model with sub-Gaussian properties, as well as bounds on $\Sigma^{-1/2}X_i/\sqrt{p}$. Their bound does have the advantage of working up to much smaller values of δ. Audibert and Catoni obtain bounds for $\delta \geq 1/n$, albeit with worse constants and only by assuming that $(v^T X_1)^2 \leq B\, v^T \Sigma\, v$ almost surely for some $B > 0$. To the best of our knowledge, no excess loss bounds of optimal order were known under finite moment assumptions. We do note, however, that Theorem 1.2 is a simple consequence of our main result, Theorem 1.1, and a Fuk-Nagaev bound by Einmahl and Li [7, Theorem 4].

1.3 Organization

The remainder of the paper is organized as follows. Section 2 reviews some preliminaries and defines our notation. Section 3 discusses our PAC-Bayesian proof method, and Sect. 4 contains the proof of the sub-Gaussian lower tail (cf. Theorem 1.1). Some facts about OLS and our proof of Theorem 1.2 are presented in Sect. 5. The final Section contains some further remarks and open problems.

2 Notation and preliminaries

The coordinates of a vector $v \in \mathbb{R}^p$ are denoted by $v[1], v[2], \ldots, v[p]$. We denote the space of matrices with p rows, p' columns and real entries by $\mathbb{R}^{p\times p'}$. A is symmetric if it equals its own transpose A^T. Given $A \in \mathbb{R}^{p\times p}$, we let $\mathrm{tr}(A)$ denote the trace of A and $\lambda_{\max}(A)$ denote its largest eigenvalue. Also, $\mathrm{diag}(A)$ is the diagonal matrix whose diagonal entries match those of A. The $p \times p$ identity matrix is denoted by $I_{p\times p}$. We identify $\mathbb{R}^p$ with the space of column vectors $\mathbb{R}^{p\times 1}$, so that the standard Euclidean inner product of $v, w \in \mathbb{R}^p$ is $v^T w$. The Euclidean norm is denoted by $|v|_2 := \sqrt{v^T v}$.

We say that $A \in \mathbb{R}^{p\times p}$ is positive semidefinite, and write $A \succeq 0$, if it is symmetric and $v^T A v \geq 0$ for all $v \in \mathbb{R}^p$. In this case one can easily show that

$$v^T A v = 0 \Leftrightarrow v^T A = 0 \Leftrightarrow Av = 0. \tag{8}$$

The $2 \to 2$ operator norm of $A \in \mathbb{R}^{p\times p'}$ is

$$|A|_{2\to 2} := \max_{v\in\mathbb{R}^{p'}\,:\,|v|_2=1} |Av|_2.$$

For symmetric $A \in \mathbb{R}^{p\times p}$ this is the largest absolute value of its eigenvalues. Moreover, if A is positive semidefinite $|A|_{2\to 2} = \lambda_{\max}(A)$ is the largerst eigenvalue, and (when A is invertible)

$$|A^{-1}|_{2\to 2} = \frac{1}{\min_{v\in\mathbb{R}^p\,:\,|v|_2=1} v^T A v}. \tag{9}$$

Finally, we write $A \succeq B$ if $A - B \succeq 0$.

Throughout the paper we use big-oh and little-oh notation informally, mostly as shorthand. For instance, $a = O(b)$ means that a is at most of the same order of magnitude as b, whereas $a = o(b)$ or $a \ll b$ means a is much smaller than b.

3 The PAC-Bayesian method

In this section we give an overview of the PAC-Bayesian method as applied to our problem. The actual proof of Theorem 1.1 is presented in Sect. 4 below.

At first sight it may seem odd that we can obtain strong concentration as in Theorem 1.1 from finite moment assumptions. The key point here is that, for any $v \in \mathbb{R}^p$, the expression

$$v^T \widehat{\Sigma}_n v = \frac{1}{n} \sum_{i=1}^{n} (X_i^T v)^2$$

is a sum of random variables which are independent, identically distributed and *non negative*. Such sums are well known to have sub-Gaussian lower tails under weak assumptions; this follows e.g. Lemma A.1 below.

This fact may be used to show concentration of $v^T \widehat{\Sigma}_n v$ for any fixed $v \in \mathbb{R}^p$. It is less obvious how to turn this into a uniform bound. The standard techniques for this, such as chaining, involve looking at a discretized subsets of $\mathbb{R}^p$ and moving from this finite set to the whole space. In our case this second step is problematic, because it requires *upper* bounds on $v^T \widehat{\Sigma}_n v$, and we know that our assumptions are not strong enough to obtain this.

What we use instead is the so-called PAC-Bayesian method [6] for controlling empirical processes. At a very high level, this method replaces chaining and union bounds with arguments based on the relative entropy. What this means in our case is that a "smoothed-out" version of the process $v^T \widehat{\Sigma}_n v$ ($v \in \mathbb{R}^p$), where v is averaged over a Gaussian measure, automatically enjoys very strong concentration properties. This implies that the original process is also well behaved as long as the effect of the smoothing can be shown to be negligible. Many of our ideas come from Audibert and Catoni [4], who in turn credit Langford and Shawe-Taylor [13] for the idea of Gaussian smoothing.

To make our ideas more definite, we present a technical result that encapsulates the main ideas in our PAC-Bayesian approach. This requires some conditions.

Assumption 1 $\{Z_\theta : \theta \in \mathbb{R}^p\}$ is a family of random variables defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We assume that the map

$$\theta \mapsto Z_\theta(\omega) \in \mathbb{R}$$

is continuous for each $\omega \in \Omega$. Given $v \in \mathbb{R}^p$ and an invertible, positive semi-definite $C \in \mathbb{R}^{p\times p}$, we let $\Gamma_{v,C}$ denote the Gaussian probability measure over $\mathbb{R}^p$ with mean v and covariance matrix C. We will also assume that for all $\omega \in \Omega$ the integrals

$$(\Gamma_{v,C}\, Z_\theta)\,(\omega) := \int_{\mathbb{R}^p} Z_\theta(\omega)\, \Gamma_{v,C}(d\theta)$$

are well defined and depend continuously on v. We will use the notation $\Gamma_{v,C}\, f_\theta$ to denote the integral of f_θ (which may also depend on other parameters) over the variable θ with the measure $\Gamma_{v,C}$.

Proposition 3.1 (PAC-Bayesian Proposition) Assume the above setup, and also that C is invertible and $\mathbb{E}\left[e^{Z_\theta}\right] \leq 1$ for all $\theta \in \mathbb{R}^d$. Then for any $t \geq 0$,

$$\Pr\left(\forall v \in \mathbb{R}^p \;:\; \Gamma_{v,C} Z_\theta \leq t + \frac{|C^{-1/2}v|_2^2}{2}\right) \geq 1 - e^{-t}.$$

In the next subsection we will apply this to prove Theorem 4.1. Here is a brief overview: we will perform a change of coordinates under which $\Sigma = I_{p\times p}$. We will then define Z_θ as

$$Z_\theta = \xi\,|\theta|_2^2 - \xi\theta^T\,\widehat{\Sigma}_n\,\theta + \text{(other terms)}$$

where $\xi > 0$ will be chosen in terms of t and the "other terms" will ensure that $\mathbb{E}\left[e^{Z_\theta}\right] \leq 1$. Taking $C = \gamma\, I_{p\times p}$ will result in

$$\Gamma_{v,C} Z_\theta = \xi\,|v|_2^2 - \xi\, v^T\,\widehat{\Sigma}_n\, v + \xi\, S_v + \text{(other terms)}$$

where

$$S_v := \gamma\; p - \gamma\; \mathrm{tr}(\widehat{\Sigma}_n)$$

is a new term introduced by the"smoothing operator" $\Gamma_{v,\gamma C}$. The choice $\gamma = 1/p$ will ensure that this term is small, and the "other terms" will also turn out to be manageable. The actual proof will be slightly complicated by the fact that we need to truncate the operator $\widehat{\Sigma}_n$ to ensure that S_v is highly concentrated.

Proof As a preliminary step, we note that under our assumptions the map:

$$\omega \in \Omega \mapsto \sup_{v\in\mathbb{R}^p}\left\{\Gamma_{v,C} Z_\theta(\omega) - \frac{|C^{-1/2}v|_2^2}{2}\right\} \in \mathbb{R}\cup\{+\infty\}$$

is measurable, since (by continuity) we may take the supremum over $v \in \mathbb{Q}^p$, which is a countable set. In particular, the event in the statement of the proposition is indeed a measurable set.

To continue, recall the definition of Kullback Leiber divergence (or relative entropy) for probability measures over a measurable space $(\Theta, \mathcal{G})$:

$$K(\mu_1|\mu_0) := \begin{cases} \int_\Theta \ln\left(\frac{d\mu_1}{d\mu_0}\right)\, d\mu_1, & \text{if } \mu_1 \ll \mu_0; \\ +\infty, & \text{otherwise.} \end{cases} \tag{10}$$

A variational principle [14, eqn. (5.13)] implies that for any measurable function $h : \Theta \to \mathbb{R}$:

$$\int h\, d\mu_1 \leq \ln\left(\int e^h\, d\mu_0\right) + K(\mu_1|\mu_0). \tag{11}$$

We apply this when $\Theta = \mathbb{R}^d$ with $\mathcal{G}$ equal to the Borel σ-field $\mathcal{B}(\mathbb{R}^d)$, $\mu_1 = \Gamma_{v,C}$, $\mu_0 = \Gamma_{0,C}$ and $h = Z_\theta$. In this case it is well-known that the relative entropy of the two measures is $|C^{-1/2}v|_2^2/2$ [19, Appendix A.5]. This implies:

$$\sup_{v\in\mathbb{R}^p}\left(\Gamma_{v,C} Z_\theta - \frac{|C^{-1/2}v|_2^2}{2}\right) \leq \ln\left(\Gamma_{0,C}\, e^{Z_\theta}\right).$$

To finish, we prove that:

$$\Pr(\Gamma_{0,C}\, e^{Z_\theta} \geq e^t) \leq e^{-t}.$$

But this follows from Markov's inequality and Fubini's Theorem:

$$\Pr\left(\Gamma_{0,C}\, e^{Z_\theta} \geq e^t\right) \leq e^{-t}\, \mathbb{E}\left[\Gamma_{0,C}\, e^{Z_\theta}\right] = e^{-t}\, \Gamma_{0,C}\mathbb{E}\left[e^{Z_\theta}\right] \leq e^{-t},$$

because $\mathbb{E}\left[e^{Z_\theta}\right] \leq 1$ for any fixed θ. □

4 The sub-Gaussian lower tail

The goal of this section is to discuss and prove the following slight generalization of Theorem 1.1.

Theorem 4.1 *Assume $A_1, \ldots, A_n \in \mathbb{R}^{p\times p}$ are i.i.d. random self-adjoint, positive semidefinite matrices whose coordinates have bounded second moments. Define $\Sigma := \mathbb{E}[A_1]$ (this is an entrywise expectation) and*

$$\widehat{\Sigma}_n := \frac{1}{n}\sum_{i=1}^n A_i.$$

Assume $\mathsf{h} \in (1, +\infty)$ satisfies $\sqrt{\mathbb{E}\left[(v^T A_1\, v)^2\right]}^{1/2} \leq \mathsf{h}\, v^T\, \Sigma v$ for all $v \in \mathbb{R}^p$. Then for any $\delta \in (0, 1)$:

$$\Pr\left(\forall v \in \mathbb{R}^p\, :\, v^T\, \widehat{\Sigma}_n\, v \geq \left(1 - 9\,\mathsf{h}\sqrt{\frac{p + 2\ln(2/\delta)}{n}}\right) v^T\, \Sigma\, v\right) \geq 1 - \delta.$$

Theorem 1.1 is recovered when we set $A_i = X_i X_i^T$, check that the moment assumption on $v^T A_1 v$ translates into (3), and note that

$$n \geq 81\,\mathsf{h}^2 \left(\frac{p + 2\ln(2/\delta)}{\varepsilon^2}\right) \Rightarrow 9\,\mathsf{h}\sqrt{\frac{p + 2\ln(2/\delta)}{n}} \leq \varepsilon.$$

Theorem 4.1 is proved in several steps over the next subsections.

4.1 Preliminaries: normalization and truncation

We first note that we may assume that Σ is invertible. Indeed, if that is not the case, we can restrict ourselves to the range of Σ, which is isometric to $\mathbb{R}^{p'}$ for some $p' \leq p$, noting that $A_i v = 0$ and $v^T A_i = 0$ almost surely for any v that is orthogonal to the range (this follows from $\mathbb{E}\left[v^T A_1 v\right] = 0$ for v orthogonal to the range, combined with (8) above).

Granted invertibility, we may define:

$$B_i := \Sigma^{-1/2} A_i \Sigma^{-1/2} \ (1 \leq i \leq n) \tag{12}$$

and note that $B_1, \ldots, B_n$ are i.i.d. positive semidefinite with $\mathbb{E}[B_1] = I_{p\times p}$. Moreover,

$$\forall v \in \mathbb{R}^p \ : \ \sqrt{\mathbb{E}\left[(v^T B_1 v)^2\right]} = \sqrt{\mathbb{E}\left[((\Sigma^{-1/2}v)^T A_1 (\Sigma^{-1/2}v))^2\right]} \leq \mathsf{h}\,|v|_2^2. \tag{13}$$

Define

$$t := \ln(2/\delta) \quad \text{and} \quad \varepsilon := 9\,\mathsf{h}\sqrt{\frac{p+2t}{n}}. \tag{14}$$

Our goal is to show that the following holds with probability $\geq 1 - 2e^{-t}$:

$$\forall w \in \mathbb{R}^p \ : \ w^T \widehat{\Sigma}_n v \geq (1-\varepsilon)\, w^T \Sigma\, w.$$

Notice that, by homonegeity, it suffices to consider vectors of the form $w = \Sigma^{-1/2} v$ with $|v|_2 = 1$. Thus our goal may be restated as follows.

$$\textbf{Goal: } \Pr\left(\forall v \in \mathbb{R}^p \ : \ |v|_2 = 1 \Rightarrow \frac{1}{n}\sum_{i=1}^{n} v^T B_i\, v \geq 1 - \varepsilon\right) \geq 1 - \delta. \tag{15}$$

We will make yet another change to our goal. Fix some $R > 0$ and define (with hindsight) truncated operators

$$B_i^R := \left(1 \wedge \frac{R}{\mathrm{tr}(B_i)}\right) B_i, \tag{16}$$

with the convention that this is simply 0 if $\mathrm{tr}(B_i) = 0$. We collect some estimates for later use.

Lemma 4.1 *We have for all $v \in \mathbb{R}^p$ with unit norm*

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} v^T B_i^R v &\le \frac{1}{n}\sum_{i=1}^{n} v^T B_i v;\\
\mathbb{E}\left[(\mathrm{tr}(B_i^R))^2\right] &\le \mathbb{E}\left[(\mathrm{tr}(B_i))^2\right] \le (\mathsf{h}\, p)^2; \text{ and}\\
\mathbb{E}\left[v^T B_i^R v\right] &\ge \left(1 - \frac{\mathsf{h}^2\, p}{R}\right).
\end{aligned}$$

Proof The first assertion follows from the fact that the are positive semidefinite, so $v^T\, B_i v \ge 0$ and $v^T\, B_i^R v = \alpha_i\, v^T\, B_i b$ for each i, with $\alpha_i \in [0, 1]$ a scalar. This same reasoning implies $B_i^R \preceq B_i$, a fact that we will use below.

To prove the second assertion, we let $e_1, \ldots, e_p$ denote the canonical basis of $\mathbb{R}^p$, and apply Minkowski's inequality:

$$\begin{aligned}
\mathbb{E}\left[(\mathrm{tr}(B_i^R)^2)\right] &= \mathbb{E}\left[\left(\sum_{j=1}^{p} e_j^T\, B_i^R\, e_j\right)^2\right]\\
(\text{use } 0 \preceq B_i^R \preceq B_i) &\le \mathbb{E}\left[\left(\sum_{j=1}^{p} e_j^T\, B_i\, e_j\right)^2\right]\\
(\text{Minkowski}) &\le \left(\sum_{j=1}^{p}\sqrt{\mathbb{E}\left[(e_j^T\, B_i\, e_j)^2\right]}\right)^2\\
(\text{eqn. (13)}) &\le (\mathsf{h}\, p)^2
\end{aligned}$$

To prove the third assertion, we fix some $v \in \mathbb{R}^p$ with $|v|_2 = 1$. We use again that $v^T\, B_i\, v \ge 0$ to deduce

$$1 - \mathbb{E}\left[v^T\, B_i^R\, v\right] = \mathbb{E}\left[v^T\, (B_i - B_i^R)\, v\right] \le \mathbb{E}\left[(v^T\, B_i\, v)\, \chi_{\{\mathrm{tr}(B_i) > R\}}\right].$$

We may bound the RHS via Cauchy Schwarz, noting that $\mathbb{E}\left[(v^T\, B_i\, v)^2\right] \le \mathsf{h}^2$ by (13) and $\Pr \mathrm{tr}(B_i) > R \le \mathbb{E}\left[\mathrm{tr}(B_i)^2\right]/R^2 = (\mathsf{h}\, p/R)^2$. This gives:

$$1 - \mathbb{E}\left[v^T\, B_i^R\, v\right] \le \sqrt{\mathbb{E}\left[(v^T\, B_i\, v)^2\right] \Pr \mathrm{tr}(B_i) > R} \le \frac{\mathsf{h}^2\, p}{R}.$$

□

4.2 Applying the PAC-Bayesian method

We continue to use our definitions of $B_1, \ldots, B_n$ and $B_1^R, \ldots, B_n^R$, with the goal of proving (15). The parameters t and ε are as in (14). We also fix $\xi > 0$. We intend apply Proposition 3.1 with $C = I_{p\times p}/p$ and

$$Z_\theta := \xi\,\mathbb{E}\left[\theta^T B_1^R \theta\right] - \frac{\xi^2}{2n^2}\,\mathbb{E}\left[(\theta^T B_1^R \theta)^2\right] - \xi \sum_{i=1}^n \frac{\theta^T B_i^R \theta}{n}.$$

Let us check that the assumptions of the theorem are satisfied. First note that Z_θ is a quadratic form in θ, and is therefore a.s. continuous as a function of θ. The same argument combined with the square integrability of the normal distribution shows that $\Gamma_{v,C}\, Z_\theta$ is continuous in $v \in \mathbb{R}^p$. The inequality $\mathbb{E}\left[e^{Z_\theta}\right] \le 1$ follows from independence, which implies

$$\mathbb{E}\left[e^{Z_\theta}\right] = \prod_{i=1}^n \mathbb{E}\left[e^{\frac{\xi\,\mathbb{E}\left[\theta^T B_1^R \theta\right]}{n} - \frac{\xi \theta^T B_i^R \theta}{n} - \frac{\xi^2}{2n^2}\,\mathbb{E}\left[(\theta^T B_i^R \theta)^2\right]}\right],$$

plus the fact that, for any non-negative, square-integrable random variable W and any $\xi > 0$:

$$\mathbb{E}\left[e^{\xi\,\mathbb{E}[W] - \xi\, W - \frac{\xi^2}{2}\,\mathbb{E}[W^2]}\right] \le 1$$

(this is shown in Lemma A.1 in the Appendix). Therefore all assumptions of Proposition 3.1 are satisfied, and we may deduce from that result that, with probability $\ge 1 - e^{-t}$, for all $v \in \mathbb{R}^p$:

$$\xi\,\Gamma_{v,C}\mathbb{E}\left[\theta^T B_1^R \theta\right] - \frac{\xi^2}{2n}\,\mathbb{E}\left[(\theta^T B_1^R \theta)^2\right] - \xi \sum_{i=1}^n \Gamma_{v,C} \frac{\theta^T B_i^R \theta}{n} \le \frac{p|v|_2^2 + 2t}{2}.$$

This is the same as saying that, with probability $\ge 1 - e^{-t}$, the following inequality holds for all $v \in \mathbb{R}^p$ with $|v|_2 = 1$:

$$\sum_{i=1}^n \Gamma_{v,C} \frac{\theta^T B_i^R \theta}{n} \ge \Gamma_{v,C}\mathbb{E}\left[\theta^T B_1^R \theta\right] - \left(\frac{\xi}{2n}\,\Gamma_{v,C}\mathbb{E}\left[(\theta^T B_1^R \theta)^2\right] + \frac{p + 2t}{2\xi}\right). \tag{17}$$

4.3 Dealing with the terms

The next step in the proof is to control all the terms involving $\Gamma_{v,C}$ that appear in (17). For $v \in \mathbb{R}^p$ with $|v|_2 = 1$, explicit calculations reveal

$$\frac{1}{n}\sum_{i=1}^{n}\Gamma_{v,C}\,\theta^T B_i^R\theta = \frac{1}{n}\sum_{i=1}^{n} v^T B_i^R v + \sum_{i=1}^{n}\frac{\mathrm{tr}(B_i^R)}{pn}$$

$$\text{(use Lemma 4.1)} \leq \frac{1}{n}\sum_{i=1}^{n} v^T B_i v + \sum_{i=1}^{n}\frac{\mathrm{tr}(B_i^R)}{pn}; \tag{18}$$

$$\Gamma_{v,C}\,\mathbb{E}\left[\theta^T B_1^R\theta\right] = \mathbb{E}\left[v^T B_1^R v\right] + \frac{\mathbb{E}\left[\mathrm{tr}(B_1^R)\right]}{p}$$

$$\text{(use Lemma 4.1)} \geq 1 - \frac{\mathsf{h}^2\, p}{R} + \frac{\mathbb{E}\left[\mathrm{tr}(B_1^R)\right]}{p}. \tag{19}$$

We also need estimates for $\Gamma_{v,C}\mathbb{E}\left[(\theta^T B_1^R\theta)^2\right]$. Standard calculations with the normal distribution show that:

$$\Gamma_{v,C}(\theta^T B_1^R\theta)^2 = \Gamma_{0,C}\,(v^T B_1^R v + \theta^T B_1^R\theta + 2\theta^T B_1^R v)^2. \tag{20}$$

That is, instead of averaging θ over $\Gamma_{v,C}$, we may replace θ by $v+\theta$ and then average over $\Gamma_{0,C}$.

We now consider the RHS of (20). The first two terms inside the brackets in the RHS are non-negative. By Cauchy Schwarz and the AM/GM inequality, the third term satisfies

$$|2\theta^T B_1^R v| \leq 2\sqrt{(\theta^T B_1^R\,\theta)\,(v^T B_1^R\, v)} \leq (v^T B_1^R v + \theta^T B_1^R\theta).$$

We deduce that

$$0 \leq v^T B_1^R v + \theta^T B_1^R\theta + 2\theta^T B_1^R v \leq 2\,(v^T B_1^R v + \theta^T B_1^R\theta),$$

and plugging this into (20) gives

$$\Gamma_{0,C}\,(v^T B_1^R v + \theta^T B_1^R\theta + 2\theta^T B_1^R v)^2 \leq 4\Gamma_{0,C}\,[(v^T B_1^R v + \theta^T B_1^R\theta)^2] \tag{21}$$

$$\text{(use } (a+b)^2 \leq 2a^2 + 2b^2) \leq 8\,(v^T B_1^R v)^2 \tag{22}$$

$$+\,8\Gamma_{0,C}(\theta^T B_1^R\theta)^2. \tag{23}$$

We now compute the term in (23) as follows. First of all, since $C = I_{p\times p}/p$,

$$\text{Law of } \theta^T B_1^R\,\theta \text{ under } \Gamma_{0,C} \;= \text{Law of } \frac{1}{p}\sum_{i=1}^{p} N_i^2\lambda_i,$$

where the $\lambda_1,\ldots,\lambda_p \geq 0$ are the eigenvalues of B_1^R and the $N_1,\ldots,N_p$ are independent standard Gaussian random variables .We note that $\mathbb{E}\left[N_i^2 N_j^2\right] \leq \mathbb{E}\left[N_i^4\right] = 3$ for all $1 \leq i, j \leq p$ and that the eigenvalues of B_1^R are all real and nonnegative (since $B_1^R \succeq 0$), therefore

$$\Gamma_{0,C}(\theta^T B_1^R \theta)^2 \leq \frac{3\mathrm{tr}(B_i^R)^2}{p^2}.$$

We combine this with (21), (22), and (23), then apply Lemma 4.1 and recall $|v|_2 = 1$ to obtain:

$$\Gamma_{v,C}\mathbb{E}\left[(\theta^T B_i^R \theta)^2\right] \leq 8\mathsf{h}^2 + \frac{24}{p^2}\mathbb{E}\left[\mathrm{tr}(B_i^R)^2\right] \leq 32\mathsf{h}^2. \tag{24}$$

We plug this last estimate into (17) together with (19) and (18). This results in the following inequality, which holds with probability $\geq 1 - e^{-t}$ simultaneously for all $v \in \mathbb{R}^d$ with $|v|_2 = 1$:

$$\frac{1}{n}\sum_{i=1}^n v^T B_i^R v \geq 1 - \left\{\frac{\mathsf{h}^2 p}{R} + \frac{16\mathsf{h}^2}{n}\xi + \frac{p+2t}{2\xi} + \left(\sum_{i=1}^n \frac{\mathrm{tr}(B_i^R) - \mathbb{E}\left[\mathrm{tr}(B_i^R)\right]}{pn}\right)\right\}.$$

This holds for any choice of ξ. Optimizing over this parameter shows that, with probability $\geq 1 - e^{-t}$, we have the following inequality simultaneously for all $v \in \mathbb{R}^p$ with $|v|_2 = 1$.

$$\begin{aligned}\sum_{i=1}^n \frac{v^T B_i v}{n} \geq 1 &- \left\{\frac{\mathsf{h}^2 p}{R} + 4\sqrt{2}\,\mathsf{h}\sqrt{\frac{(p+2t)}{n}}\right\} \\ &- \left(\sum_{i=1}^n \frac{\mathrm{tr}(B_i^R) - \mathbb{E}\left[\mathrm{tr}(B_i^R)\right]}{pn}\right).\end{aligned} \tag{25}$$

4.4 The final step: control of the trace

We now take care of the term involving the traces on the RHS. This is precisely the moment when the truncation of B_i is useful, as it allows for the use of Bernstein's concentration inequality [23, Sect. 2.6]. This inequality states that, for independent random variables $Z_1, \ldots, Z_n$ with $\mathbb{E}[Z_i] = 0$, $\sum_{i=1}^n \mathbb{E}\left[Z_i^2\right] = \sigma^2$ and $|Z_i| \leq M$ for each $1 \leq i \leq n$ ($M > 0$ a constant), then

$$\Pr\left(\sum_{i=1}^n Z_i \geq \sigma\sqrt{2t} + \frac{2Mt}{3}\right) \leq e^{-t}.$$

The term involving traces in (25) is a sum of i.i.d. mean-zero random variables that (because of the truncation) lie between $-R/pn$ and R/pn. Moreover, the variance of each term is at most $\mathbb{E}\left[\mathrm{tr}(B_i^R)^2\right]/p^2n^2 \leq \mathsf{h}^2/n^2$ by Lemma 4.1. We deduce:

$$\Pr\left(\sum_{i=1}^n \frac{\mathrm{tr}(B_i^R) - \mathbb{E}\left[\mathrm{tr}(B_i^R)\right]}{pn} \leq \mathsf{h}\sqrt{\frac{2t}{n}} + \frac{2Rt}{3pn}\right) \geq 1 - e^{-t}.$$

Combining this with (25) implies that, for any $t \geq 0$, the following inequality holds with probability $\geq 1 - 2e^{-t}$, simultaneously for all $v \in \mathbb{R}^p$ with $|v|_2 = 1$:

$$\sum_{i=1}^{n} \frac{v^T B_i v}{n} \geq 1 - \left\{ \frac{\mathsf{h}^2 p}{R} + 4\sqrt{2}\,\mathsf{h}\sqrt{\frac{(p+2t)}{n}} + \mathsf{h}\sqrt{\frac{2t}{n}} + \frac{2R\,t}{3pn} \right\}. \tag{26}$$

This holds for any $R > 0$. Optimization over R gives

$$\inf_{R>0} \frac{\mathsf{h}^2 p}{R} + \frac{2R\,t}{3pn} = 2\mathsf{h}\sqrt{\frac{2t}{3n}} \leq \sqrt{2}\,\mathsf{h}\sqrt{\frac{2t}{n}},$$

so, with the right choice of R,

$$\left\{ \frac{\mathsf{h}^2 p}{R} + 4\sqrt{2}\,\mathsf{h}\sqrt{\frac{(p+2t)}{n}} + \mathsf{h}\sqrt{\frac{2\,t}{n}} + \frac{2R\,t}{3pn} \right\} \leq (5\sqrt{2}+1)\,\mathsf{h}\sqrt{\frac{(p+2t)}{n}} \leq \varepsilon,$$

according to the definition of ε in (14). We obtain

$$\Pr\left(\forall v \in \mathbb{R}^p \,:\, |v|_2 = 1 \Rightarrow 1 - \sum_{i=1}^{n} \frac{v^T B_i v}{n} \leq \varepsilon \right) \geq 1 - 2e^{-t}.$$

Inequality (15) follows. As noted in Sect. 4.1, (15) implies Theorem 4.1 and finishes the proof.

5 Applications in random-design linear regression

The main goal of this section is to prove Theorem 1.2.

5.1 Preliminaries

We begin by recalling the general facts about this problem. We assume $(X, Y) \in \mathbb{R}^p \times \mathbb{R}$ is a random pair, with $X \in \mathbb{R}^p$ a vector of covariates and $Y \in \mathbb{R}$ a response variable. We assume $\mathbb{E}\left[|X|_2^2\right] < +\infty$ and $\mathbb{E}\left[Y^2\right] < +\infty$. As in the introduction, we define the square loss function:

$$\ell(\beta) := \mathbb{E}\left[(Y - X^T\beta)^2\right] \quad (\beta \in \mathbb{R}^p). \tag{27}$$

It is not hard to show that ℓ has at least one minimizer $\beta_{\min} \in \mathbb{R}^p$, defined so that $\beta_{\min}^T X$ equals the L^2 projection of Y onto the linear space generated by the coordinates of X. In fact, this property uniquely defines the random variable $\beta_{\min}^T X$, if not necessarily the vector $\beta_{\min}$. It also implies

$$\eta := Y - \beta_{\min}^T X \text{ satisfies } \mathbb{E}\left[\eta\, X\right] = 0. \tag{28}$$

In fact, $\beta_{\min}$ is a minimizer of ℓ if and only if (28) holds. Another calculation shows that

$$\forall \beta \in \mathbb{R}^p \; : \; \ell(\beta) - \ell(\beta_{\min}) = |\Sigma^{1/2}(\beta - \beta_{\min})|_2^2, \tag{29}$$

where $\Sigma := \mathbb{E}\left[XX^T\right]$. In particular, $\beta_{\min}$ is the unique minimizer of ℓ if and only if Σ is non-singular.

Our main interest is in the OLS estimator, which satisfies

$$\widehat{\beta}_n \in \operatorname{argmin}_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} (Y_i - \beta^T X_i)^2.$$

If $\widehat{\Sigma}_n := n^{-1} \sum_{i=1}^n X_i X_i^T$ is invertible, $\widehat{\beta}_n$ is uniquely defined by the formula:

$$\widehat{\beta}_n := \widehat{\Sigma}_n^{-1} \frac{1}{n} \sum_{i=1}^{n} Y_i X_i. \tag{30}$$

If $\widehat{\Sigma}_n$ is not invertible, we may still *define* $\widehat{\beta}_n$ by (30) if we let $\widehat{\Sigma}_n^{-1}$ denote the Moore-Penrose pseudoinverse of $\widehat{\Sigma}_n$. This definition will be used implicitly below.

An interesting test case for our result is that of a linear model with Gaussian noise and Gaussian design, where we assume that X is mean-zero Gaussian with covariance matrix Σ and η is mean zero Gaussian with variance σ^2 and independent of X. Using the notation of Theorem 1.2, we see that $\mathbb{E}\left[\eta^2\right] = \sigma^2$ is the variance of the noise, and h_* does not depend on n or p. An explicit calculation (which we omit) implies that, for $n \gg p \gg 1$,

$$|\Sigma^{1/2}(\widehat{\beta}_n - \beta_{\min})|_2^2 \geq (1 - o(1)) \frac{\sigma^2 p}{n} \text{ with probability } 1 - o(1).$$

Theorem 1.2 guarantees that OLS achieves this error rate under much weaker assumptions on the distribution of (X, Y).

5.2 Proof of Theorem 1.2

Proof We will assume for convenience that Σ is invertible; the general case requires minor modifications. For each i, define

$$\eta_i := Y_i - X_i^T \beta_{\min} \text{ and} \tag{31}$$

$$Z_i := \eta_i \, \Sigma^{-1/2} X_i = (Y_i - X_i^T \beta_{\min}) \, \Sigma^{-1/2} X_i, \tag{32}$$

We note for later use that the Z_i are independent copies of the vector Z in the statement of Theorem 1.2.

The assumptions on X of Theorem 1.2 imply those of Theorem 1.1 with ε replaced by $\varepsilon/10$ and δ replaced by $\delta/2$ (at least when C is a large enough constant). This implies that the event

$$\mathsf{Lower} := \left\{\forall v \in \mathbb{R}^p \,:\, v^T \widehat{\Sigma}_n\, v \geq (1-\varepsilon/10)\; v^T \Sigma v\right\} \tag{33}$$

satisfies $\Pr(\mathsf{Lower}) \geq 1-\delta/2$ whenever the condition on n in Theorem 1.2 is satisfied. When Lower holds, $\widehat{\Sigma}_n$ is also invertible, so

$$\beta_{\min} = \widehat{\Sigma}_n^{-1}\, \widehat{\Sigma}_n\, \beta_{\min} = \widehat{\Sigma}_n^{-1}\, \frac{1}{n}\sum_{i=1}^n (\beta_{\min}^T X_i)\, X_i.$$

Comparing this with the definition of $\widehat{\beta}_n$ in (30), we see that:

$$\widehat{\beta}_n - \beta_{\min} = \widehat{\Sigma}_n^{-1}\, \frac{1}{n}\sum_{i=1}^n (Y_i - \beta_{\min}^T X_i)\, X_i = \widehat{\Sigma}_n^{-1}\, \Sigma^{1/2}\left(\frac{1}{n}\sum_{i=1}^n Z_i\right).$$

Therefore, when Lower holds, the excess loss $\ell(\widehat{\beta}_n) - \ell(\beta_{\min})$ satisfies

$$|\Sigma^{1/2}(\widehat{\beta}_n - \beta_{\min})|_2^2 = \left|(\Sigma^{1/2}\,\widehat{\Sigma}_n^{-1}\, \Sigma^{1/2})\left(\frac{1}{n}\sum_{i=1}^n Z_i\right)\right|_2^2 \leq \frac{\left|\frac{1}{n}\sum_{i=1}^n Z_i\right|_2^2}{(1-\varepsilon/10)^2}, \tag{34}$$

since Lower and (9) imply that the $2 \to 2$ operator norm of $\Sigma^{1/2}\,\widehat{\Sigma}_n^{-1}\, \Sigma^{1/2}$ is at most $1/(1-\varepsilon/10)$.

What we have discussed so far shows that (34) holds with probability $\geq 1-\delta/2$. We now show that

$$\Pr\left(\left|\frac{1}{n}\sum_{i=1}^n Z_i\right|_2^2 \geq \frac{(1+\varepsilon/5)^2\,\sigma^2}{n}\,(\sqrt{p} + C\ln(4/\delta))\right) \leq \frac{\delta}{3} \tag{35}$$

noting that this finishes the proof with some room to spare regarding the dependency on ε. To do this, we use the Fuk-Nagaev-type inequality by Einmahl and Li in [7, Theorem 4]. In the Euclidean setting, that result implies that if $U_1, \ldots, U_n$ are i.i.d. mean zero random vectors with $\Lambda_n := n\lambda_{\max}(\mathbb{E}\left[U_1 U_1^T\right])$, then for any $q > 2$, $\alpha, \phi \in (0,1)$ one can find $D > 0$ such that, for any $t > 0$,

$$\Pr\left(\left|\sum_{i=1}^n U_i\right|_2 \geq (1+\phi)\,\mathbb{E}\left[\left|\sum_{i=1}^n U_i\right|_2\right] + t\right) \leq e^{-\frac{t^2}{(2+\alpha)\,\Lambda_n}} + D\,n\,\frac{\mathbb{E}\left[|U_1|_2^q\right]}{t^q}.$$

To obtain (35), we apply the previous display with $U_i = Z_i/n$, $\phi = \varepsilon/10$, $\alpha = 1$, and

$$t := \sigma\sqrt{\frac{(\varepsilon\, p/100) \vee 3\ln(4/\delta)}{n}}.$$

We observe that

$$\mathbb{E}\left[\left|\sum_{i=1}^{n} U_i\right|_2\right] \leq \sqrt{\frac{1}{n}\,\mathbb{E}\left[|Z_1|_2^2\right]} \leq \sqrt{\sigma^2\,\frac{p}{n}},$$

$\Lambda_n \leq \sigma^2/n$ and $\mathbb{E}\left[|U_1|_2^q/t^q\right] \leq (100\mathsf{h}_*^2)^{q/2}/(\varepsilon\, n)^{q/2}$ under our assumptions. We deduce

$$\Pr\left|\sum_{i=1}^{n}\frac{Z_i}{n}\right|_2 \geq \left(1+\frac{\varepsilon}{5}\right)\frac{\sigma}{\sqrt{n}}\left(\sqrt{p}+\sqrt{3\ln(4/\delta)}\right) \leq \frac{\delta}{4} + D\,\frac{\mathsf{h}_*^q}{\varepsilon^{q/2}\,n^{q/2-1}}.$$

This clearly implies (35) after suitably adjusting the constants, at least in the desired range $\delta \geq C/n^{q/2-1}$. □

6 Final remarks

- The PAC-Bayesian method used in this paper seems an efficient alternative to chaining and other typical empirical processes methods. As such, it would be interesting to find other applications of it. One interesting question is if some variant of the method can be used to prove two-sided concentration of $\widehat{\Sigma}_n$.
- Consider the setting of Theorem 4.1. Let $\mathbb{R}^p_s$ denote the set of all $v \in \mathbb{R}^p$ that are s-sparse, i.e. have at most s nonzero coordinates. $\mathbb{R}^p_s$ is a union of $\binom{p}{s} \leq (ep/s)^s$ s-dimensional spaces, so if

$$n \geq 100\mathsf{h}^2\,\frac{s + s\ln(ep/s) + 2\ln(2/\delta)}{\varepsilon^2}$$

  one may apply Theorem 4.1 to these subspaces and deduce that $v^T\,\widehat{\Sigma}_n\,v \geq (1-\varepsilon)\,v^T\,\Sigma\,v$ for all $v \in \mathbb{R}^p_s$, with probability $\geq 1-\delta$. In a companion paper [16] we show that this result is relevant to prove that $\widehat{\Sigma}_n$ satisfies restricted eigenvalue properties when $p \gg n$. This result is relevant to the Compressed Sensing and High Dimensional Statistics.

Acknowledgments We thank the anonymous referees for their careful reading of the original manuscript, their many corrections and suggestions on the exposition. We thank them in particular for suggesting the use of the Fuk-Nagaev inequality from [7] in the analysis of OLS.

A Appendix: a moment generating function bound for non-negative random variables

Lemma A.1 *Let W be a nonnegative random variable with finite second moment. Then* $\forall\xi > 0$, $\mathbb{E}\left[e^{-\xi W}\right] \leq e^{-\xi\,\mathbb{E}[W] + \frac{\xi^2}{2}\,\mathbb{E}[W^2]}$.

Proof This follows from the fact that

$$\forall x \geq 0 \, : \, e^{-x} \leq 1 - x + \frac{x^2}{2}$$

applied to $x = \xi \, W$. Taking expectations of the resulting inequality gives

$$\mathbb{E}\left[e^{-\xi \, W}\right] \leq 1 - \xi \, \mathbb{E}\,[W] + \frac{\xi^2}{2} \, \mathbb{E}\left[W^2\right].$$

The result follows once we apply "$1 + y \leq e^y$", valid for all $y \in \mathbb{R}$, to $y := \xi \, \mathbb{E}\,[W] - \xi^2 \mathbb{E}\left[W^2\right]/2$. □

References

1. Adamczak, R., Litvak, A.E., Pajor, A., Tomczak-Jaegermann, N.: Quantitative estimates of the convergence of the empirical covariance matrix in log-concave ensembles. J. Am. Math. Soc. **23**, 535–561 (2010). doi:10.1090/S0894-0347-09-00650-X
2. Adamczak, R., Litvak, A.E., Pajor, A., Tomczak-Jaegermann, N.: Restricted isometry property of matrices with independent columns and neighborly polytopes by random sampling. Constr. Approx. **34**(1), 61–88 (2011). doi:10.1007/s00365-010-9117-4
3. Adamczak, R., Litvak, A.E., Pajor, A., Tomczak-Jaegermann, N.: Sharp bounds on the rate of convergence of the empirical covariance matrix. Comptes Rendus Mathematique 349(34):195– 200 (2011). doi:10.1016/j.crma.2010.12.014. http://www.sciencedirect.com/science/article/pii/S1631073X10003936
4. Audibert, J.Y., Catoni, O.: Robust linear least squares regression. Ann. Stat. **39**(5), 2766–2794 (2011). doi:10.1214/11-AOS918SUPP
5. Bai, Z.D., Yin, Y.Q.: Limit of the smallest eigenvalue of a large dimensional sample covariance matrix. Ann. Probab. **21**(3), 1275–1294 (1993)
6. Catoni, O.: PAC-Bayesian supervised classification (The thermodynamics of statistical learning). Institute of Mathematical Statistics (2007)
7. Einmahl, U., Li, D.: Characterization of lil behavior in banach space. Trans. Am. Math. Soc. **360**, 6677–6693 (2008)
8. van de Geer, S., Muro, A.: On higher order isotropy conditions and lower bounds for sparse quadratic forms. Electron. J. Stat. **8**(2), 3031–3061 (2014). doi:10.1214/15-EJS983
9. Guédon, O., Litvak, A.E., Pajor, A., Tomczak-Jaegermann, N.: On the interval of fluctuation of the singular values of random matrices. arXiv:1509.02322
10. Hsu, D., Kakade, S.M., Zhang, T.: Random design analysis of ridge regression. J. Mach. Learn. Res. Proc. Track **23**, 9.1–9.24 (2012)
11. Koltchinskii, V., Mendelson, S.: Bounding the smallest singular value of a random matrix without concentration. International Mathematics Research Notices (2015). doi:10.1093/imrn/rnv096. http://imrn.oxfordjournals.org/content/early/2015/03/31/imrn.rnv096.abstract
12. Tikhomirov, K.: Sample covariance matrices of heavy-tailed distributions. arXiv:1606.03557
13. Langford, J., Shawe-Taylor, J.: Pac-Bayes & margins. In: Becker, S., Thrun, S., Obermayer, K., (eds.) NIPS, pp. 423–430. MIT Press (2002)
14. Ledoux, M.: The Concentration of Measure Phenomenon. American Mathematical Society (2001)
15. Mendelson, S., Paouris, G.: On the singular values of random matrices. J. Eur. Math. Soc. **16**(4), 823–834 (2014)
16. Oliveira, R.I.: A simple method for lower bounding sparse quadratic forms. In preparation
17. Oliveira, R.I.: Sums of random Hermitian matrices and an inequality by Rudelson. Electron. Commun. Probab. **15**, 203–212 (2010)
18. Raskutti, G., Wainwright, M.J., Yu, B.: Restricted eigenvalue properties for correlated gaussian designs. J. Mach. Learn. Res. **11**, 2241–2259 (2010)

19. Rasmussen, C.E., Williams, C.K.I.: Gaussian processes for machine learning (adaptive computation and machine learning). The MIT Press, Cambridge (2005)
20. Rudelson, M.: Random vectors in the isotropic position. J. Funct. Anal. **164**(1), 60–72 (1999)
21. Rudelson, M., Zhou, S.: Reconstruction from anisotropic random measurements. IEEE Trans. Inf. Theory **59**(6), 3434–3447 (2013). doi:10.1109/TIT.2013.2243201
22. Srivastava, N., Vershynin, R.: Covariance estimation for distributions with 2+epsilon moments. Ann. Probab. **41**(5), 3081–3111 (2013)
23. Stéphane Boucheron Gábor Lugosi, P.M.: Concentration inequalities: a nonasymptotic theory of independence. Oxford University Press, Oxford (2013)
24. Yaskov, P.: Lower bounds on the smallest eigenvalue of a sample covariance matrix. Electron. Commun. Probab. **19**, no. 83, 1–10 (2014). doi:10.1214/ECP.v19-3807. http://ecp.ejpecp.org/article/view/3807
25. Yaskov, P.: Sharp lower bounds on the least singular value of a random matrix without the fourth moment condition. Electron. Commun. Probab. **20**, no. 44, 1–9 (2015). doi:10.1214/ECP.v20-4089. http://ecp.ejpecp.org/article/view/4089