Introduction to the non-asymptotic analysis of random matrices

Roman Vershynin[1]
University of Michigan
romanv@umich.edu

August 11, 2010; final revision November 23, 2011

Contents

This is a tutorial on some basic non-asymptotic methods and concepts in random matrix theory. The reader will learn several tools for the analysis of the extreme singular values of random matrices with independent rows or columns. Many of these methods sprung off from the development of geometric functional analysis since the 1970's. They have applications in several fields, most notably in theoretical computer science, statistics and signal processing. A few basic applications are covered in this text, particularly for the problem of estimating covariance matrices in statistics and for validating probabilistic constructions of measurement matrices in compressed sensing. These notes are written particularly for graduate students and beginning researchers in different areas, including functional analysts, probabilists, theoretical statisticians, electrical engineers, and theoretical computer scientists.

5.1   Introduction

Asymptotic and non-asymptotic regimes   Random matrix theory studies properties of $N \times n$ matrices A chosen from some distribution on the set of all matrices. As dimensions N and n grow to infinity, one observes that the spectrum of A tends to stabilize. This is manifested in several *limit laws*, which may be regarded as random matrix versions of the central limit theorem. Among them is Wigner's semicircle law for the eigenvalues of symmetric Gaussian matrices, the circular law for Gaussian matrices, the Marchenko-Pastur law for Wishart matrices $W = A^*A$ where A is a Gaussian matrix, the Bai-Yin and Tracy-Widom laws for the extreme eigenvalues of Wishart matrices W. The books [51, 5, 23, 6] offer thorough introduction to the classical problems of random matrix theory and its fascinating connections.

The asymptotic regime where the dimensions $N, n \to \infty$ is well suited for the purposes of statistical physics, e.g. when random matrices serve as finite-dimensional models of infinite-dimensional operators. But in some other areas including statistics, geometric functional analysis, and compressed sensing, the limiting regime may not be very useful [69]. Suppose, for example, that we ask about the largest singular value $s_{\max}(A)$ (i.e. the largest eigenvalue of $(A^*A)^{1/2}$); to be specific assume that A is an $n \times n$ matrix whose entries are independent standard normal random variables. The asymptotic random matrix theory answers this question as follows: the Bai-Yin law (see Theorem 5.31) states that

$$s_{\max}(A)/2\sqrt{n} \to 1 \quad \text{almost surely}$$

as the dimension $n \to \infty$. Moreover, the limiting distribution of $s_{\max}(A)$ is known to be the Tracy-Widom law (see [71, 27]). In contrast to this, a non-asymptotic answer to the same question is the following: in *every* dimension n, one has

$$s_{\max}(A) \leq C\sqrt{n} \quad \text{with probability at least } 1 - e^{-n},$$

here C is an absolute constant (see Theorems 5.32 and 5.39). The latter answer is less precise (because of an absolute constant C) but more quantitative because for fixed dimensions n it gives an exponential probability of success.[1] This is the kind of answer

---

[1]For this specific model (Gaussian matrices),Theorems 5.32 and 5.35 even give a sharp absolute constant $C \approx 2$ here. But the result mentioned here is much more general as we will see later; it only requires independence of rows or columns of A.

we will seek in this text – guarantees up to absolute constants in all dimensions, and with large probability.

Tall matrices are approximate isometries The following heuristic will be our guideline: *tall random matrices should act as approximate isometries.* So, an $N \times n$ random matrix A with $N \gg n$ should act almost like an isometric embedding of ℓ_2^n into ℓ_2^N:

$$(1 - \delta)K\|x\|_2 \leq \|Ax\|_2 \leq (1 + \delta)K\|x\|_2 \quad \text{for all } x \in \mathbb{R}^n$$

where K is an appropriate normalization factor and $\delta \ll 1$. Equivalently, this says that all the singular values of A are close to each other:

$$(1 - \delta)K \leq s_{\min}(A) \leq s_{\max}(A) \leq (1 + \delta)K,$$

where $s_{\min}(A)$ and $s_{\max}(A)$ denote the smallest and the largest singular values of A. Yet equivalently, this means that tall matrices are well conditioned: the *condition number* of A is $\kappa(A) = s_{\max}(A)/s_{\min}(A) \leq (1 + \delta)/(1 - \delta) \approx 1$.

In the asymptotic regime and for random matrices with independent entries, our heuristic is justified by Bai-Yin's law, which is Theorem 5.31 below. Loosely speaking, it states that as the dimensions N, n increase to infinity while the aspect ratio N/n is fixed, we have

$$\sqrt{N} - \sqrt{n} \approx s_{\min}(A) \leq s_{\max}(A) \approx \sqrt{N} + \sqrt{n}. \tag{5.1}$$

In these notes, we study $N \times n$ random matrices A with independent rows or independent columns, but not necessarily independent entries. We develop non-asymptotic versions of (5.1) for such matrices, which should hold for all dimensions N and n. The desired results should have the form

$$\sqrt{N} - C\sqrt{n} \leq s_{\min}(A) \leq s_{\max}(A) \leq \sqrt{N} + C\sqrt{n} \tag{5.2}$$

with large probability, e.g. $1 - e^{-N}$, where C is an absolute constant.[2] For tall matrices, where $N \gg n$, both sides of this inequality would be close to each other, which would guarantee that A is an approximate isometry.

Models and methods We shall study quite general models of random matrices – those with independent rows or independent columns that are sampled from high-dimensional distributions. We will place either strong moment assumptions on the distribution (sub-gaussian growth of moments), or no moment assumptions at all (except finite variance). This leads us to four types of main results:

1. Matrices with independent sub-gaussian rows: Theorem 5.39
2. Matrices with independent heavy-tailed rows: Theorem 5.41
3. Matrices with independent sub-gaussian columns: Theorem 5.58
4. Matrices with independent heavy-tailed columns: Theorem 5.62

---

[2]More accurately, we should expect $C = O(1)$ to depend on easily computable quantities of the distribution, such as its moments. This will be clear from the context.

These four models cover many natural classes of random matrices that occur in applications, including random matrices with independent entries (Gaussian and Bernoulli in particular) and random sub-matrices of orthogonal matrices (random Fourier matrices in particular).

The analysis of these four models is based on a variety of tools of probability theory and geometric functional analysis, most of which have not been covered in the texts on the "classical" random matrix theory. The reader will learn basics on sub-gaussian and sub-exponential random variables, isotropic random vectors, large deviation inequalities for sums of independent random variables, extensions of these inequalities to random matrices, and several basic methods of high dimensional probability such as symmetrization, decoupling, and covering (ε-net) arguments.

Applications   In these notes we shall emphasize two applications, one in statistics and one in compressed sensing. Our analysis of random matrices with independent rows immediately applies to a basic problem in statistics – *estimating covariance matrices* of high-dimensional distributions. If a random matrix A has i.i.d. rows A_i, then $A^*A = \sum_i A_i \otimes A_i$ is the *sample covariance matrix*. If A has independent columns A_j, then $A^*A = (\langle A_j, A_k \rangle)_{j,k}$ is the *Gram matrix*. Thus our analysis of the row-independent and column-independent models can be interpreted as a study of sample covariance matrices and Gram matrices of high dimensional distributions. We will see in Section 5.4.3 that for a general distribution in $\mathbb{R}^n$, its covariance matrix can be estimated from a sample of size $N = O(n \log n)$ drawn from the distribution. Moreover, for sub-gaussian distributions we have an even better bound $N = O(n)$. For low-dimensional distributions, much fewer samples are needed – if a distribution lies close to a subspace of dimension r in $\mathbb{R}^n$, then a sample of size $N = O(r \log n)$ is sufficient for covariance estimation.

In compressed sensing, the best known measurement matrices are random. A sufficient condition for a matrix to succeed for the purposes of compressed sensing is given by the *restricted isometry property*. Loosely speaking, this property demands that all sub-matrices of given size be well-conditioned. This fits well in the circle of problems of the non-asymptotic random matrix theory. Indeed, we will see in Section 5.6 that all basic models of random matrices are nice restricted isometries. These include Gaussian and Bernoulli matrices, more generally all matrices with sub-gaussian independent entries, and even more generally all matrices with sub-gaussian independent rows or columns. Also, the class of restricted isometries includes random Fourier matrices, more generally random sub-matrices of bounded orthogonal matrices, and even more generally matrices whose rows are independent samples from an isotropic distribution with uniformly bounded coordinates.

Related sources   This text is a tutorial rather than a survey, so we focus on explaining methods rather than results. This forces us to make some concessions in our choice of the subjects. *Concentration of measure* and its applications to random matrix theory are only briefly mentioned. For an introduction into concentration of measure suitable for a beginner, see [9] and [49, Chapter 14]; for a thorough exposition see [56, 43]; for connections with random matrices see [21, 44]. The monograph [45] also offers an introduction into concentration of measure and related probabilistic methods in analysis

and geometry, some of which we shall use in these notes.

We completely avoid the important (but more difficult) model of *symmetric random matrices* with independent entries on and above the diagonal. Starting from the work of Füredi and Komlos [29], the largest singular value (the spectral norm) of symmetric random matrices has been a subject of study in many works; see e.g. [50, 83, 58] and the references therein.

We also did not even attempt to discuss sharp small *deviation inequalities* (of Tracy-Widom type) for the extreme eigenvalues. Both these topics and much more are discussed in the surveys [21, 44, 69], which serve as bridges between asymptotic and non-asymptotic problems in random matrix theory.

Because of the absolute constant C in (5.2), our analysis of the smallest singular value (the *"hard edge"*) will only be useful for sufficiently tall matrices, where $N \geq C^2 n$. For square and almost square matrices, the hard edge problem will be only briefly mentioned in Section 5.3. The surveys [76, 69] discuss this problem at length, and they offer a glimpse of connections to other problems of random matrix theory and additive combinatorics.

Many of the results and methods presented in these notes are known in one form or another. Some of them are published while some others belong to the folklore of probability in Banach spaces, geometric functional analysis, and related areas. When available, historic references are given in Section 5.7.

5.2   Preliminaries

5.2.1   Matrices and their singular values

The main object of our study will be an $N \times n$ matrix A with real or complex entries. We shall state all results in the real case; the reader will be able to adjust them to the complex case as well. Usually but not always one should think of tall matrices A, those for which $N \geq n > 1$. By passing to the adjoint matrix A^*, many results can be carried over to "flat" matrices, those for which $N \leq n$.

It is often convenient to study A through the $n \times n$ symmetric positive-semidefinite matrix the matrix A^*A. The eigenvalues of $|A| := \sqrt{A^*A}$ are therefore non-negative real numbers. Arranged in a non-decreasing order, they are called the *singular values*[3] of A and denoted $s_1(A) \geq \cdots \geq s_n(A) \geq 0$. Many applications require estimates on the extreme singular values

$$s_{\max}(A) := s_1(A), \quad s_{\min}(A) := s_n(A).$$

---

[3]In the literature, singular values are also called *s-numbers*.

The smallest singular value is only of interest for tall matrices, since for $N < n$ one automatically has $s_{\min}(A) = 0$.

Equivalently, $s_{\max}(A)$ and $s_{\min}(A)$ are respectively the smallest number M and the largest number m such that

$$m\|x\|_2 \leq \|Ax\|_2 \leq M\|x\|_2 \quad \text{for all } x \in \mathbb{R}^n. \tag{5.3}$$

In order to interpret this definition geometrically, we look at A as a linear operator from $\mathbb{R}^n$ into $\mathbb{R}^N$. The Euclidean distance between any two points in $\mathbb{R}^n$ can increase by at most the factor $s_{\max}(A)$ and decrease by at most the factor $s_{\max}(A)$ under the action of A. Therefore, the extreme singular values control the distortion of the Euclidean geometry under the action of A. If $s_{\max}(A) \approx s_{\min}(A) \approx 1$ then A acts as an *approximate isometry*, or more accurately an approximate isometric embedding of ℓ_2^n into ℓ_2^N.

The extreme singular values can also be described in terms of the *spectral norm of A*, which is by definition

$$\|A\| = \|A\|_{\ell_2^n \to \ell_2^N} = \sup_{x \in \mathbb{R}^n \setminus \{0\}} \frac{\|Ax\|_2}{\|x\|_2} = \sup_{x \in S^{n-1}} \|Ax\|_2. \tag{5.4}$$

(5.3) gives a link between the extreme singular values and the spectral norm:

$$s_{\max}(A) = \|A\|, \quad s_{\min}(A) = 1/\|A^\dagger\|$$

where $A^\dagger$ denotes the pseudoinverse of A; if A is invertible then $A^\dagger = A^{-1}$.

5.2.2 Nets

Nets are convenient means to discretize compact sets. In our study we will mostly need to discretize the unit Euclidean sphere S^{n-1} in the definition of the spectral norm (5.4). Let us first recall a general definition of an ε-net.

Definition 5.1 (Nets, covering numbers). *Let (X, d) be a metric space and let $\varepsilon > 0$. A subset $\mathcal{N}_\varepsilon$ of X is called an ε-net of X if every point $x \in X$ can be approximated to within ε by some point $y \in \mathcal{N}_\varepsilon$, i.e. so that $d(x, y) \leq \varepsilon$. The minimal cardinality of an ε-net of X, if finite, is denoted $\mathcal{N}(X, \varepsilon)$ and is called the* covering number[4] *of X (at scale ε).*

From a characterization of compactness we remember that X is compact if and only if $\mathcal{N}(X, \varepsilon) < \infty$ for each $\varepsilon > 0$. A quantitative estimate on $\mathcal{N}(X, \varepsilon)$ would give us a *quantitative version of compactness* of X.[5] Let us therefore take a simple example of a metric space, the unit Euclidean sphere S^{n-1} equipped with the Euclidean metric[6] $d(x, y) = \|x - y\|_2$, and estimate its covering numbers.

---

[4]Equivalently, $\mathcal{N}(X, \varepsilon)$ is the minimal number of balls with radii ε and with centers in X needed to cover X.

[5]In statistical learning theory and geometric functional analysis, $\log \mathcal{N}(X, \varepsilon)$ is called *the metric entropy of X*. In some sense it measures the "complexity" of metric space X.

[6]A similar result holds for the geodesic metric on the sphere, since for small ε these two distances are equivalent.

Lemma 5.2 (Covering numbers of the sphere). *The unit Euclidean sphere S^{n-1} equipped with the Euclidean metric satisfies for every $\varepsilon > 0$ that*

$$\mathcal{N}(S^{n-1}, \varepsilon) \leq \left(1 + \frac{2}{\varepsilon}\right)^n.$$

Proof. This is a simple *volume argument.* Let us fix $\varepsilon > 0$ and choose $\mathcal{N}_\varepsilon$ to be a maximal ε-separated subset of S^{n-1}. In other words, $\mathcal{N}_\varepsilon$ is such that $d(x, y) \geq \varepsilon$ for all $x, y \in \mathcal{N}_\varepsilon$, $x \neq y$, and no subset of S^{n-1} containing $\mathcal{N}_\varepsilon$ has this property.[7]

The maximality property implies that $\mathcal{N}_\varepsilon$ is an ε-net of S^{n-1}. Indeed, otherwise there would exist $x \in S^{n-1}$ that is at least ε-far from all points in $\mathcal{N}_\varepsilon$. So $\mathcal{N}_\varepsilon \cup \{x\}$ would still be an ε-separated set, contradicting the minimality property.

Moreover, the separation property implies via the triangle inequality that the balls of radii $\varepsilon/2$ centered at the points in $\mathcal{N}_\varepsilon$ are disjoint. On the other hand, all such balls lie in $(1 + \varepsilon/2)B_2^n$ where B_2^n denotes the unit Euclidean ball centered at the origin. Comparing the volume gives $\text{vol}\left(\frac{\varepsilon}{2}B_2^n\right) \cdot |\mathcal{N}_\varepsilon| \leq \text{vol}\left((1 + \frac{\varepsilon}{2})B_2^n\right)$. Since $\text{vol}\left(rB_2^n\right) = r^n \text{vol}(B_2^n)$ for all $r \geq 0$, we conclude that $|\mathcal{N}_\varepsilon| \leq (1 + \frac{\varepsilon}{2})^n / (\frac{\varepsilon}{2})^n = (1 + \frac{2}{\varepsilon})^n$ as required. $\square$

Nets allow us to reduce the complexity of computations with linear operators. One such example is the computation of the spectral norm. To evaluate the spectral norm by definition (5.4) one needs to take the supremum over the whole sphere S^{n-1}. However, one can essentially replace the sphere by its ε-net:

Lemma 5.3 (Computing the spectral norm on a net). *Let A be an $N \times n$ matrix, and let $\mathcal{N}_\varepsilon$ be an ε-net of S^{n-1} for some $\varepsilon \in [0, 1)$. Then*

$$\max_{x \in \mathcal{N}_\varepsilon} \|Ax\|_2 \leq \|A\| \leq (1 - \varepsilon)^{-1} \max_{x \in \mathcal{N}_\varepsilon} \|Ax\|_2$$

Proof. The lower bound in the conclusion follows from the definition. To prove the upper bound let us fix $x \in S^{n-1}$ for which $\|A\| = \|Ax\|_2$, and choose $y \in \mathcal{N}_\varepsilon$ which approximates x as $\|x - y\|_2 \leq \varepsilon$. By the triangle inequality we have $\|Ax - Ay\|_2 \leq \|A\| \|x - y\|_2 \leq \varepsilon \|A\|$. It follows that

$$\|Ay\|_2 \geq \|Ax\|_2 - \|Ax - Ay\|_2 \geq \|A\| - \varepsilon \|A\| = (1 - \varepsilon) \|A\|.$$

Taking maximum over all $y \in \mathcal{N}_\varepsilon$ in this inequality, we complete the proof. $\square$

A similar result holds for symmetric $n \times n$ matrices A, whose spectral norm can be computed via the associated quadratic form: $\|A\| = \sup_{x \in S^{n-1}} |\langle Ax, x \rangle|$. Again, one can essentially replace the sphere by its ε-net:

Lemma 5.4 (Computing the spectral norm on a net). *Let A be a symmetric $n \times n$ matrix, and let $\mathcal{N}_\varepsilon$ be an ε-net of S^{n-1} for some $\varepsilon \in [0, 1)$. Then*

$$\|A\| = \sup_{x \in S^{n-1}} |\langle Ax, x \rangle| \leq (1 - 2\varepsilon)^{-1} \sup_{x \in \mathcal{N}_\varepsilon} |\langle Ax, x \rangle|.$$

---

[7]One can in fact construct $\mathcal{N}_\varepsilon$ inductively by first selecting an arbitrary point on the sphere, and at each next step selecting a point that is at distance at least ε from those already selected. By compactness, this algorithm will terminate after finitely many steps and it will yield a set $\mathcal{N}_\varepsilon$ as we required.

Proof. Let us choose $x \in S^{n-1}$ for which $\|A\| = |\langle Ax, x \rangle|$, and choose $y \in \mathcal{N}_\varepsilon$ which approximates x as $\|x - y\|_2 \leq \varepsilon$. By the triangle inequality we have

$$|\langle Ax, x \rangle - \langle Ay, y \rangle| = |\langle Ax, x - y \rangle + \langle A(x - y), y \rangle|$$
$$\leq \|A\|\|x\|_2\|x - y\|_2 + \|A\|\|x - y\|_2\|y\|_2 \leq 2\varepsilon\|A\|.$$

It follows that $|\langle Ay, y \rangle| \geq |\langle Ax, x \rangle| - 2\varepsilon\|A\| = (1 - 2\varepsilon)\|A\|$. Taking the maximum over all $y \in \mathcal{N}_\varepsilon$ in this inequality completes the proof. $\square$

5.2.3 Sub-gaussian random variables

In this section we introduce the class of sub-gaussian random variables,[8] those whose distributions are dominated by the distribution of a centered gaussian random variable. This is a convenient and quite wide class, which contains in particular the standard normal and all bounded random variables.

Let us briefly recall some of the well known properties of the *standard normal random variable X*. The distribution of X has density $\frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ and is denoted $N(0, 1)$. Estimating the integral of this density between t and ∞ one checks that the tail of a standard normal random variable X decays super-exponentially:

$$\mathbb{P}\{|X| > t\} = \frac{2}{\sqrt{2\pi}} \int_t^\infty e^{-x^2/2} \, dx \leq 2e^{-t^2/2}, \quad t \geq 1, \tag{5.5}$$

see e.g. [26, Theorem 1.4] for a more precise two-sided inequality. The absolute moments of X can be computed as

$$(\mathbb{E}|X|^p)^{1/p} = \sqrt{2}\Big[\frac{\Gamma((1 + p)/2)}{\Gamma(1/2)}\Big]^{1/p} = O(\sqrt{p}), \quad p \geq 1. \tag{5.6}$$

The moment generating function of X equals

$$\mathbb{E} \exp(tX) = e^{t^2/2}, \quad t \in \mathbb{R}. \tag{5.7}$$

Now let X be a general random variable. We observe that these three properties are equivalent – a super-exponential tail decay like in (5.5), the moment growth (5.6), and the growth of the moment generating function like in (5.7). We will then focus on the class of random variables that satisfy these properties, which we shall call sub-gaussian random variables.

Lemma 5.5 (Equivalence of sub-gaussian properties). *Let X be a random variable. Then the following properties are equivalent with parameters $K_i > 0$ differing from each other by at most an absolute constant factor.*[9]

---

[8]It would be more rigorous to say that we study *sub-gaussian probability distributions*. The same concerns some other properties of random variables and random vectors we study later in this text. However, it is convenient for us to focus on random variables and vectors because we will form random matrices out of them.

[9]The precise meaning of this equivalence is the following. There exists an absolute constant C such that property i implies property j with parameter $K_j \leq CK_i$ for any two properties $i, j = 1, 2, 3$.

1. *Tails:* $\mathbb{P}\{|X| > t\} \leq \exp(1 - t^2/K_1^2)$ *for all* $t \geq 0$;

2. *Moments:* $(\mathbb{E}|X|^p)^{1/p} \leq K_2\sqrt{p}$ *for all* $p \geq 1$;

3. *Super-exponential moment:* $\mathbb{E}\exp(X^2/K_3^2) \leq e$.

Moreover, if $\mathbb{E}X = 0$ *then properties 1–3 are also equivalent to the following one:*

4. *Moment generating function:* $\mathbb{E}\exp(tX) \leq \exp(t^2 K_4^2)$ *for all* $t \in \mathbb{R}$.

Proof. **1.** $\Rightarrow$ **2.** Assume property 1 holds. By homogeneity, rescaling X to X/K_1 we can assume that $K_1 = 1$. Recall that for every non-negative random variable Z, integration by parts yields the identity $\mathbb{E}Z = \int_0^\infty \mathbb{P}\{Z \geq u\}\,du$. We apply this for $Z = |X|^p$. After change of variables $u = t^p$, we obtain using property 1 that

$$\mathbb{E}|X|^p = \int_0^\infty \mathbb{P}\{|X| \geq t\}\, pt^{p-1}\,dt \leq \int_0^\infty e^{1-t^2} pt^{p-1}\,dt = \left(\frac{ep}{2}\right)\Gamma\left(\frac{p}{2}\right) \leq \left(\frac{ep}{2}\right)\left(\frac{p}{2}\right)^{p/2}.$$

Taking the p-th root yields property 2 with a suitable absolute constant K_2.

2. $\Rightarrow$ **3.** Assume property 2 holds. As before, by homogeneity we may assume that $K_2 = 1$. Let $c > 0$ be a sufficiently small absolute constant. Writing the Taylor series of the exponential function, we obtain

$$\mathbb{E}\exp(cX^2) = 1 + \sum_{p=1}^\infty \frac{c^p \mathbb{E}(X^{2p})}{p!} \leq 1 + \sum_{p=1}^\infty \frac{c^p(2p)^p}{p!} \leq 1 + \sum_{p=1}^\infty (2c/e)^p.$$

The first inequality follows from property 2; in the second one we use $p! \geq (p/e)^p$. For small c this gives $\mathbb{E}\exp(cX^2) \leq e$, which is property 3 with $K_3 = c^{-1/2}$.

3. $\Rightarrow$ **1.** Assume property 3 holds. As before we may assume that $K_3 = 1$. Exponentiating and using Markov's inequality[10] and then property 3, we have

$$\mathbb{P}\{|X| > t\} = \mathbb{P}\{e^{X^2} \geq e^{t^2}\} \leq e^{-t^2}\mathbb{E}e^{X^2} \leq e^{1-t^2}.$$

This proves property 1 with $K_1 = 1$.

2. $\Rightarrow$ **4.** Let us now assume that $\mathbb{E}X = 0$ and property 2 holds; as usual we can assume that $K_2 = 1$. We will prove that property 4 holds with an appropriately large absolute constant $C = K_4$. This will follow by estimating Taylor series for the exponential function

$$\mathbb{E}\exp(tX) = 1 + t\mathbb{E}X + \sum_{p=2}^\infty \frac{t^p \mathbb{E}X^p}{p!} \leq 1 + \sum_{p=2}^\infty \frac{t^p p^{p/2}}{p!} \leq 1 + \sum_{p=2}^\infty \left(\frac{e|t|}{\sqrt{p}}\right)^p. \qquad (5.8)$$

The first inequality here follows from $\mathbb{E}X = 0$ and property 2; the second one holds since $p! \geq (p/e)^p$. We compare this with Taylor's series for

$$\exp(C^2 t^2) = 1 + \sum_{k=1}^\infty \frac{(C|t|)^{2k}}{k!} \geq 1 + \sum_{k=1}^\infty \left(\frac{C|t|}{\sqrt{k}}\right)^{2k} = 1 + \sum_{p \in 2\mathbb{N}} \left(\frac{C|t|}{\sqrt{p/2}}\right)^p. \qquad (5.9)$$

---

[10]This simple argument is sometimes called exponential Markov's inequality.

The first inequality here holds because $p! \leq p^p$; the second one is obtained by substitution $p = 2k$. One can show that the series in (5.8) is bounded by the series in (5.9) with large absolute constant C. We conclude that $\mathbb{E} \exp(tX) \leq \exp(C^2 t^2)$, which proves property 4.

4. $\Rightarrow$ 1. Assume property 4 holds; we can also assume that $K_4 = 1$. Let $\lambda > 0$ be a parameter to be chosen later. By exponential Markov inequality, and using the bound on the moment generating function given in property 4, we obtain

$$\mathbb{P}\{X \geq t\} = \mathbb{P}\{e^{\lambda X} \geq e^{\lambda t}\} \leq e^{-\lambda t} \mathbb{E} e^{\lambda X} \leq e^{-\lambda t + \lambda^2}.$$

Optimizing in λ and thus choosing $\lambda = t/2$ we conclude that $\mathbb{P}\{X \geq t\} \leq e^{-t^2/4}$. Repeating this argument for $-X$, we also obtain $\mathbb{P}\{X \leq -t\} \leq e^{-t^2/4}$. Combining these two bounds we conclude that $\mathbb{P}\{|X| \geq t\} \leq 2e^{-t^2/4} \leq e^{1-t^2/4}$. Thus property 1 holds with $K_1 = 2$. The lemma is proved. $\qquad\square$

Remark 5.6.   1. The constants 1 and e in properties 1 and 3 respectively are chosen for convenience. Thus the value 1 can be replaced by any positive number and the value e can be replaced by any number greater than 1.

2. The assumption $\mathbb{E}X = 0$ is only needed to prove the necessity of property 4; the sufficiency holds without this assumption.

Definition 5.7 (Sub-gaussian random variables). *A random variable X that satisfies one of the equivalent properties 1 – 3 in Lemma 5.5 is called a* sub-gaussian random variable. *The* sub-gaussian norm *of X, denoted $\|X\|_{\psi_2}$, is defined to be the smallest K_2 in property 2. In other words,*[11]

$$\|X\|_{\psi_2} = \sup_{p \geq 1} p^{-1/2} (\mathbb{E}|X|^p)^{1/p}.$$

The class of sub-gaussian random variables on a given probability space is thus a normed space. By Lemma 5.5, every sub-gaussian random variable X satisfies:

$$\mathbb{P}\{|X| > t\} \leq \exp(1 - ct^2 / \|X\|_{\psi_2}^2) \quad \text{for all } t \geq 0; \tag{5.10}$$

$$(\mathbb{E}|X|^p)^{1/p} \leq \|X\|_{\psi_2} \sqrt{p} \quad \text{for all } p \geq 1; \tag{5.11}$$

$$\mathbb{E} \exp(cX^2 / \|X\|_{\psi_2}^2) \leq e;$$

$$\text{if } \mathbb{E}X = 0 \text{ then } \mathbb{E} \exp(tX) \leq \exp(Ct^2 \|X\|_{\psi_2}^2) \quad \text{for all } t \in \mathbb{R}, \tag{5.12}$$

where $C, c > 0$ are absolute constants. Moreover, up to absolute constant factors, $\|X\|_{\psi_2}$ is the smallest possible number in each of these inequalities.

Example 5.8. Classical examples of sub-gaussian random variables are Gaussian, Bernoulli and all bounded random variables.

1. **(Gaussian):** A standard normal random variable X is sub-gaussian with $\|X\|_{\psi_2} \leq C$ where C is an absolute constant. This follows from (5.6). More generally, if X is a centered normal random variable with variance σ^2, then X is sub-gaussian with $\|X\|_{\psi_2} \leq C\sigma$.

---

[11] The sub-gaussian norm is also called ψ_2 norm in the literature.

2. **(Bernoulli):** Consider a random variable X with distribution $\mathbb{P}\{X = -1\} = \mathbb{P}\{X = 1\} = 1/2$. We call X a *symmetric Bernoulli random variable*. Since $|X| = 1$, it follows that X is a sub-gaussian random variable with $\|X\|_{\psi_2} = 1$.

3. **(Bounded):** More generally, consider any bounded random variable X, thus $|X| \leq M$ almost surely for some M. Then X is a sub-gaussian random variable with $\|X\|_{\psi_2} \leq M$. We can write this more compactly as $\|X\|_{\psi_2} \leq \|X\|_\infty$.

A remarkable property of the normal distribution is *rotation invariance*. Given a finite number of independent centered normal random variables X_i, their sum $\sum_i X_i$ is also a centered normal random variable, obviously with $\mathrm{Var}(\sum_i X_i) = \sum_i \mathrm{Var}(X_i)$. Rotation invariance passes onto sub-gaussian random variables, although approximately:

Lemma 5.9 (Rotation invariance). *Consider a finite number of independent centered sub-gaussian random variables X_i. Then $\sum_i X_i$ is also a centered sub-gaussian random variable. Moreover,*

$$\big\| \sum_i X_i \big\|_{\psi_2}^2 \leq C \sum_i \|X_i\|_{\psi_2}^2$$

where C is an absolute constant.

Proof. The argument is based on estimating the moment generating function. Using independence and (5.12) we have for every $t \in \mathbb{R}$:

$$\mathbb{E} \exp\big(t \sum_i X_i\big) = \mathbb{E} \prod_i \exp(tX_i) = \prod_i \mathbb{E} \exp(tX_i) \leq \prod_i \exp(Ct^2 \|X_i\|_{\psi_2}^2)$$

$$= \exp(t^2 K^2) \quad \text{where } K^2 = C \sum_i \|X_i\|_{\psi_2}^2.$$

Using the equivalence of properties 2 and 4 in Lemma 5.5 we conclude that $\|\sum_i X_i\|_{\psi_2} \leq C_1 K$ where C_1 is an absolute constant. The proof is complete. $\square$

The rotation invariance immediately yields a *large deviation inequality* for sums of independent sub-gaussian random variables:

Proposition 5.10 (Hoeffding-type inequality). *Let $X_1, \ldots, X_N$ be independent centered sub-gaussian random variables, and let $K = \max_i \|X_i\|_{\psi_2}$. Then for every $a = (a_1, \ldots, a_N) \in \mathbb{R}^N$ and every $t \geq 0$, we have*

$$\mathbb{P}\Big\{ \Big| \sum_{i=1}^N a_i X_i \Big| \geq t \Big\} \leq e \cdot \exp\Big(-\frac{ct^2}{K^2 \|a\|_2^2} \Big)$$

where $c > 0$ is an absolute constant.

Proof. The rotation invariance (Lemma 5.9) implies the bound $\| \sum_i a_i X_i \|_{\psi_2}^2 \leq C \sum_i a_i^2 \|X_i\|_{\psi_2}^2 \leq CK^2 \|a\|_2^2$. Property (5.10) yields the required tail decay. $\square$

Remark 5.11. One can interpret these results (Lemma 5.9 and Proposition 5.10) as one-sided *non-asymptotic manifestations of the central limit theorem.* For example, consider the normalized sum of independent symmetric Bernoulli random variables $S_N = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \varepsilon_i$. Proposition 5.10 yields the tail bounds $\mathbb{P}\{|S_N| > t\} \le e \cdot e^{-ct^2}$ for any number of terms N. Up to the absolute constants e and c, these tails coincide with those of the standard normal random variable (5.5).

Using moment growth (5.11) instead of the tail decay (5.10), we immediately obtain from Lemma 5.9 a general form of the well known Khintchine inequality:

Corollary 5.12 (Khintchine inequality). *Let X_i be a finite number of independent subgaussian random variables with zero mean, unit variance, and $\|X_i\|_{\psi_2} \le K$. Then, for every sequence of coefficients a_i and every exponent $p \ge 2$ we have*

$$\Big(\sum_i a_i^2\Big)^{1/2} \le \big(\mathbb{E}\big|\sum_i a_i X_i\big|^p\big)^{1/p} \le CK\sqrt{p}\,\Big(\sum_i a_i^2\Big)^{1/2}$$

where C is an absolute constant.

Proof. The lower bound follows by independence and Hölder's inequality: indeed, $\big(\mathbb{E}\big|\sum_i a_i X_i\big|^p\big)^{1/p} \ge \big(\mathbb{E}\big|\sum_i a_i X_i\big|^2\big)^{1/2} = \big(\sum_i a_i^2\big)^{1/2}$. For the upper bound, we argue as in Proposition 5.10, but use property (5.11). $\qquad\square$

5.2.4 Sub-exponential random variables

Although the class of sub-gaussian random variables is natural and quite wide, it leaves out some useful random variables which have tails heavier than gaussian. One such example is a standard exponential random variable – a non-negative random variable with exponential tail decay

$$\mathbb{P}\{X \ge t\} = e^{-t}, \quad t \ge 0. \tag{5.13}$$

To cover such examples, we consider a class of *sub-exponential random variables*, those with at least an exponential tail decay. With appropriate modifications, the basic properties of sub-gaussian random variables hold for sub-exponentials. In particular, a version of Lemma 5.5 holds with a similar proof for sub-exponential properties, except for property 4 of the moment generating function. Thus for a random variable X the following properties are equivalent with parameters $K_i > 0$ differing from each other by at most an absolute constant factor:

$$\mathbb{P}\{|X| > t\} \le \exp(1 - t/K_1) \quad \text{for all } t \ge 0; \tag{5.14}$$

$$(\mathbb{E}|X|^p)^{1/p} \le K_2 p \quad \text{for all } p \ge 1; \tag{5.15}$$

$$\mathbb{E}\exp(X/K_3) \le e. \tag{5.16}$$

Definition 5.13 (Sub-exponential random variables). *A random variable X that satisfies one of the equivalent properties* (5.14) *–* (5.16) *is called a* sub-exponential random variable. *The* sub-exponential norm *of X, denoted $\|X\|_{\psi_1}$, is defined to be the smallest parameter K_2. In other words,*

$$\|X\|_{\psi_1} = \sup_{p \ge 1} p^{-1}(\mathbb{E}|X|^p)^{1/p}.$$

Lemma 5.14 (Sub-exponential is sub-gaussian squared). *A random variable X is sub-gaussian if and only if X^2 is sub-exponential. Moreover,*

$$\|X\|_{\psi_2}^2 \le \|X^2\|_{\psi_1} \le 2\|X\|_{\psi_2}^2.$$

Proof. This follows easily from the definition. $\qquad\square$

The moment generating function of a sub-exponential random variable has a similar upper bound as in the sub-gaussian case (property 4 in Lemma 5.5). The only real difference is that the bound only holds in a neighborhood of zero rather than on the whole real line. This is inevitable, as the moment generating function of an exponential random variable (5.13) does not exist for $t \ge 1$.

Lemma 5.15 (Mgf of sub-exponential random variables). *Let X be a centered sub-exponential random variable. Then, for t such that $|t| \le c/\|X\|_{\psi_1}$, one has*

$$\mathbb{E}\exp(tX) \le \exp(Ct^2\|X\|_{\psi_1}^2)$$

where $C, c > 0$ are absolute constants.

Proof. The argument is similar to the sub-gaussian case. We can assume that $\|X\|_{\psi_1} = 1$ by replacing X with $X/\|X\|_{\psi_1}$ and t with $t\|X\|_{\psi_1}$. Repeating the proof of the implication $2 \Rightarrow 4$ of Lemma 5.5 and using $\mathbb{E}|X|^p \le p^p$ this time, we obtain that $\mathbb{E}\exp(tX) \le 1 + \sum_{p=2}^{\infty}(e|t|)^p$. If $|t| \le 1/2e$ then the right hand side is bounded by $1 + 2e^2t^2 \le \exp(2e^2t^2)$. This completes the proof. $\qquad\square$

Sub-exponential random variables satisfy a *large deviation inequality* similar to the one for sub-gaussians (Proposition 5.10). The only significant difference is that *two tails* have to appear here – a gaussian tail responsible for the central limit theorem, and an exponential tail coming from the tails of each term.

Proposition 5.16 (Bernstein-type inequality). *Let $X_1, \ldots, X_N$ be independent centered sub-exponential random variables, and $K = \max_i \|X_i\|_{\psi_1}$. Then for every $a = (a_1, \ldots, a_N) \in \mathbb{R}^N$ and every $t \ge 0$, we have*

$$\mathbb{P}\Big\{\Big|\sum_{i=1}^{N} a_i X_i\Big| \ge t\Big\} \le 2\exp\Big[-c\min\Big(\frac{t^2}{K^2\|a\|_2^2}, \frac{t}{K\|a\|_\infty}\Big)\Big]$$

where $c > 0$ is an absolute constant.

Proof. Without loss of generality, we assume that $K = 1$ by replacing X_i with X_i/K and t with t/K. We use the exponential Markov inequality for the sum $S = \sum_i a_i X_i$ and with a parameter $\lambda > 0$:

$$\mathbb{P}\{S \ge t\} = \mathbb{P}\{e^{\lambda S} \ge e^{\lambda t}\} \le e^{-\lambda t}\mathbb{E}e^{\lambda S} = e^{-\lambda t}\prod_i \mathbb{E}\exp(\lambda a_i X_i).$$

If $|\lambda| \le c/\|a\|_\infty$ then $|\lambda a_i| \le c$ for all i, so Lemma 5.15 yields

$$\mathbb{P}\{S \ge t\} \le e^{-\lambda t}\prod_i \exp(C\lambda^2 a_i^2) = \exp(-\lambda t + C\lambda^2\|a\|_2^2).$$

Choosing $\lambda = \min(t/2C\|a\|_2^2, c/\|a\|_\infty)$, we obtain that

$$\mathbb{P}\{S \geq t\} \leq \exp\Big[-\min\Big(\frac{t^2}{4C\|a\|_2^2}, \frac{ct}{2\|a\|_\infty}\Big)\Big].$$

Repeating this argument for $-X_i$ instead of X_i, we obtain the same bound for $\mathbb{P}\{-S \geq t\}$. A combination of these two bounds completes the proof. $\qquad\square$

Corollary 5.17. *Let $X_1, \ldots, X_N$ be independent centered sub-exponential random variables, and let $K = \max_i \|X_i\|_{\psi_1}$. Then, for every $\varepsilon \geq 0$, we have*

$$\mathbb{P}\Big\{\Big|\sum_{i=1}^{N} X_i\Big| \geq \varepsilon N\Big\} \leq 2\exp\Big[-c\min\Big(\frac{\varepsilon^2}{K^2}, \frac{\varepsilon}{K}\Big)N\Big]$$

where $c > 0$ is an absolute constant.

Proof. This follows from Proposition 5.16 for $a_i = 1$ and $t = \varepsilon N$. $\qquad\square$

Remark 5.18 (Centering). The definitions of sub-gaussian and sub-exponential random variables X do not require them to be centered. In any case, one can always center X using the simple fact that if X is sub-gaussian (or sub-exponential), then so is $X - \mathbb{E}X$. Moreover,
$$\|X - \mathbb{E}X\|_{\psi_2} \leq 2\|X\|_{\psi_2}, \quad \|X - \mathbb{E}X\|_{\psi_1} \leq 2\|X\|_{\psi_1}.$$
This follows by triangle inequality $\|X - \mathbb{E}X\|_{\psi_2} \leq \|X\|_{\psi_2} + \|\mathbb{E}X\|_{\psi_2}$ along with $\|\mathbb{E}X\|_{\psi_2} = |\mathbb{E}X| \leq \mathbb{E}|X| \leq \|X\|_{\psi_2}$, and similarly for the sub-exponential norm.

5.2.5 Isotropic random vectors

Now we carry our work over to higher dimensions. We will thus be working with random vectors X in $\mathbb{R}^n$, or equivalently probability distributions in $\mathbb{R}^n$.

While the concept of the mean $\mu = \mathbb{E}Z$ of a random variable Z remains the same in higher dimensions, the second moment $\mathbb{E}Z^2$ is replaced by the $n \times n$ *second moment matrix* of a random vector X, defined as

$$\Sigma = \Sigma(X) = \mathbb{E}X \otimes X = \mathbb{E}XX^T$$

where $\otimes$ denotes the outer product of vectors in $\mathbb{R}^n$. Similarly, the concept of variance $\mathrm{Var}(Z) = \mathbb{E}(Z - \mu)^2 = \mathbb{E}Z^2 - \mu^2$ of a random variable is replaced in higher dimensions with the *covariance matrix* of a random vector X, defined as

$$\mathrm{Cov}(X) = \mathbb{E}(X - \mu) \otimes (X - \mu) = \mathbb{E}X \otimes X - \mu \otimes \mu$$

where $\mu = \mathbb{E}X$. By translation, many questions can be reduced to the case of centered random vectors, for which $\mu = 0$ and $\mathrm{Cov}(X) = \Sigma(X)$. We will also need a higher-dimensional version of unit variance:

Definition 5.19 (Isotropic random vectors). *A random vector X in $\mathbb{R}^n$ is called* isotropic *if $\Sigma(X) = I$. Equivalently, X is isotropic if*

$$\mathbb{E}\langle X, x\rangle^2 = \|x\|_2^2 \quad \text{for all } x \in \mathbb{R}^n. \tag{5.17}$$

Suppose $\Sigma(X)$ is an invertible matrix, which means that the distribution of X is not essentially supported on any proper subspace of $\mathbb{R}^n$. Then $\Sigma(X)^{-1/2}X$ is an isotropic random vector in $\mathbb{R}^n$. Thus every non-degenerate random vector can be made isotropic by an appropriate linear transformation.[12] This allows us to mostly focus on studying isotropic random vectors in the future.

Lemma 5.20. *Let X, Y be independent isotropic random vectors in $\mathbb{R}^n$. Then $\mathbb{E}\|X\|_2^2 = n$ and $\mathbb{E}\langle X, Y\rangle^2 = n$.*

Proof. The first part follows from $\mathbb{E}\|X\|_2^2 = \mathbb{E}\operatorname{tr}(X \otimes X) = \operatorname{tr}(\mathbb{E}X \otimes X) = \operatorname{tr}(I) = n$. The second part follows by conditioning on Y, using isotropy of X and using the first part for Y: this way we obtain $\mathbb{E}\langle X, Y\rangle^2 = \mathbb{E}\|Y\|_2^2 = n$. $\square$

Example 5.21. 1. **(Gaussian):** The (standard) *Gaussian random vector* X in $\mathbb{R}^n$ chosen according to the standard normal distribution $N(0, I)$ is isotropic. The coordinates of X are independent standard normal random variables.

2. **(Bernoulli):** A similar example of a discrete isotropic distribution is given by a *Bernoulli random vector* X in $\mathbb{R}^n$ whose coordinates are independent symmetric Bernoulli random variables.

3. **(Product distributions):** More generally, consider a random vector X in $\mathbb{R}^n$ whose coordinates are independent random variables with zero mean and unit variance. Then clearly X is an isotropic vector in $\mathbb{R}^n$.

4. **(Coordinate):** Consider a *coordinate random vector* X, which is uniformly distributed in the set $\{\sqrt{n}\,e_i\}_{i=1}^n$ where $\{e_i\}_{i=1}^n$ is the canonical basis of $\mathbb{R}^n$. Clearly X is an isotropic random vector in $\mathbb{R}^n$.[13]

5. **(Frame):** This is a more general version of the coordinate random vector. A *frame* is a set of vectors $\{u_i\}_{i=1}^M$ in $\mathbb{R}^n$ which obeys an approximate Parseval's identity, i.e. there exist numbers $A, B > 0$ called *frame bounds* such that

$$A\|x\|_2^2 \leq \sum_{i=1}^M \langle u_i, x\rangle^2 \leq B\|x\|_2^2 \quad \text{for all } x \in \mathbb{R}^n.$$

If $A = B$ the set is called a *tight frame*. Thus, tight frames are generalizations of orthogonal bases without linear independence. Given a tight frame $\{u_i\}_{i=1}^M$ with bounds $A = B = M$, the random vector X uniformly distributed in the set $\{u_i\}_{i=1}^M$ is clearly isotropic in $\mathbb{R}^n$.[14]

---

[12]This transformation (usually preceded by centering) is a higher-dimensional version of *standardizing* of random variables, which enforces zero mean and unit variance.

[13]The examples of Gaussian and coordinate random vectors are somewhat opposite – one is very continuous and the other is very discrete. They may be used as test cases in our study of random matrices.

[14]There is clearly a reverse implication, too, which shows that the class of tight frames can be identified with the class of discrete isotropic random vectors.

6. **(Spherical):** Consider a random vector X uniformly distributed on the unit Euclidean sphere in $\mathbb{R}^n$ with center at the origin and radius $\sqrt{n}$. Then X is isotropic. Indeed, by rotation invariance $\mathbb{E}\langle X, x\rangle^2$ is proportional to $\|x\|_2^2$; the correct normalization $\sqrt{n}$ is derived from Lemma 5.20.

7. **(Uniform on a convex set):** In convex geometry, a convex set K in $\mathbb{R}^n$ is called isotropic if a random vector X chosen uniformly from K according to the volume is isotropic. As we noted, every full dimensional convex set can be made into an isotropic one by an affine transformation. Isotropic convex sets look "well conditioned", which is advantageous in geometric algorithms (e.g. volume computations).

We generalize the concepts of sub-gaussian random variables to higher dimensions using one-dimensional marginals.

Definition 5.22 (Sub-gaussian random vectors). *We say that a random vector X in $\mathbb{R}^n$ is sub-gaussian if the one-dimensional marginals $\langle X, x\rangle$ are sub-gaussian random variables for all $x \in \mathbb{R}^n$. The* sub-gaussian norm *of X is defined as*

$$\|X\|_{\psi_2} = \sup_{x \in S^{n-1}} \|\langle X, x\rangle\|_{\psi_2}.$$

Remark 5.23 (Properties of high-dimensional distributions). The definitions of isotropic and sub-gaussian distributions suggest that more generally, natural properties of high-dimensional distributions may be defined via one-dimensional marginals. This is a natural way to generalize properties of random variables to random vectors. For example, we shall call a random vector sub-exponential if all of its one-dimensional marginals are sub-exponential random variables, etc.

One simple way to create sub-gaussian distributions in $\mathbb{R}^n$ is by taking a product of n sub-gaussian distributions on the line:

Lemma 5.24 (Product of sub-gaussian distributions). *Let $X_1, \ldots, X_n$ be independent centered sub-gaussian random variables. Then $X = (X_1, \ldots, X_n)$ is a centered sub-gaussian random vector in $\mathbb{R}^n$, and*

$$\|X\|_{\psi_2} \leq C \max_{i \leq n} \|X_i\|_{\psi_2}$$

where C is an absolute constant.

Proof. This is a direct consequence of the rotation invariance principle, Lemma 5.9. Indeed, for every $x = (x_1, \ldots, x_n) \in S^{n-1}$ we have

$$\|\langle X, x\rangle\|_{\psi_2} = \Big\| \sum_{i=1}^n x_i X_i \Big\|_{\psi_2} \leq C \sum_{i=1}^n x_i^2 \|X_i\|_{\psi_2}^2 \leq C \max_{i \leq n} \|X_i\|_{\psi_2}$$

where we used that $\sum_{i=1}^n x_i^2 = 1$. This completes the proof. $\qquad\square$

Example 5.25. Let us analyze the basic examples of random vectors introduced earlier in Example 5.21.

1. **(Gaussian, Bernoulli):** Gaussian and Bernoulli random vectors are sub-gaussian; their sub-gaussian norms are bounded by an absolute constant. These are particular cases of Lemma 5.24.

2. **(Spherical):** A spherical random vector is also sub-gaussian; its sub-gaussian norm is bounded by an absolute constant. Unfortunately, this does not follow from Lemma 5.24 because the coordinates of the spherical vector are not independent. Instead, by rotation invariance, the claim clearly follows from the following geometric fact. For every $\varepsilon \geq 0$, the spherical cap $\{x \in S^{n-1} : x_1 > \varepsilon\}$ makes up at most $\exp(-\varepsilon^2 n/2)$ proportion of the total area on the sphere.[15] This can be proved directly by integration, and also by elementary geometric considerations [9, Lemma 2.2].

3. **(Coordinate):** Although the coordinate random vector X is formally sub-gaussian as its support is finite, its sub-gaussian norm is too big: $\|X\|_{\psi_2} = \sqrt{n} \gg 1$. So we would not think of X as a sub-gaussian random vector.

4. **(Uniform on a convex set):** For many isotropic convex sets K (called ψ_2 bodies), a random vector X uniformly distributed in K is sub-gaussian with $\|X\|_{\psi_2} = O(1)$. For example, the cube $[-1, 1]^n$ is a ψ_2 body by Lemma 5.24, while the appropriately normalized cross-polytope $\{x \in \mathbb{R}^n : \|x\|_1 \leq M\}$ is not. Nevertheless, Borell's lemma (which is a consequence of Brunn-Minkowski inequality) implies a weaker property, that X is always *sub-exponential*, and $\|X\|_{\psi_1} = \sup_{x \in S^{n-1}} \|\langle X, x \rangle\|_{\psi_1}$ is bounded by absolute constant. See [33, Section 2.2.b$_3$] for a proof and discussion of these ideas.

5.2.6 Sums of independent random matrices

In this section, we mention without proof some results of classical probability theory in which scalars can be replaced by matrices. Such results are useful in particular for problems on random matrices, since we can view a random matrix as a generalization of a random variable. One such remarkable generalization is valid for Khintchine inequality, Corollary 5.12. The scalars a_i can be replaced by matrices, and the absolute value by the *Schatten norm*. Recall that for $1 \leq p \leq \infty$, the p-Schatten norm of an $n \times n$ matrix A is defined as the ℓ_p norm of the sequence of its singular values:

$$\|A\|_{C_p^n} = \|(s_i(A))_{i=1}^n\|_p = \Big(\sum_{i=1}^n s_i(A)^p \Big)^{1/p}.$$

For $p = \infty$, the Schatten norm equals the spectral norm $\|A\| = \max_{i \leq n} s_i(A)$. Using this one can quickly check that already for $p = \log n$ the Schatten and spectral norms are equivalent: $\|A\|_{C_p^n} \leq \|A\| \leq e\|A\|_{C_p^n}$.

---

[15]This fact about spherical caps may seem counter-intuitive. For example, for $\varepsilon = 0.1$ the cap looks similar to a hemisphere, but the proportion of its area goes to zero very fast as dimension n increases. This is a starting point of the study of the *concentration of measure phenomenon*, see [43].

Theorem 5.26 (Non-commutative Khintchine inequality, see [61] Section 9.8). *Let $A_1, \ldots, A_N$ be self-adjoint $n \times n$ matrices and $\varepsilon_1, \ldots, \varepsilon_N$ be independent symmetric Bernoulli random variables. Then, for every $2 \le p < \infty$, we have*

$$\left\| \Big(\sum_{i=1}^{N} A_i^2 \Big)^{1/2} \right\|_{C_p^n} \le \Big(\mathbb{E} \Big\| \sum_{i=1}^{N} \varepsilon_i A_i \Big\|_{C_p^n}^p \Big)^{1/p} \le C\sqrt{p} \left\| \Big(\sum_{i=1}^{N} A_i^2 \Big)^{1/2} \right\|_{C_p^n}$$

where C is an absolute constant.

Remark 5.27. 1. The scalar case of this result, for $n = 1$, recovers the classical Khintchine inequality, Corollary 5.12, for $X_i = \varepsilon_i$.

2. By the equivalence of Schatten and spectral norms for $p = \log n$, a version of non-commutative Khintchine inequality holds for the spectral norm:

$$\mathbb{E} \Big\| \sum_{i=1}^{N} \varepsilon_i A_i \Big\| \le C_1 \sqrt{\log n} \left\| \Big(\sum_{i=1}^{N} A_i^2 \Big)^{1/2} \right\| \tag{5.18}$$

where C_1 is an absolute constant. The logarithmic factor is unfortunately essential; it role will be clear when we discuss applications of this result to random matrices in the next sections.

Corollary 5.28 (Rudelson's inequality [65]). *Let $x_1, \ldots, x_N$ be vectors in $\mathbb{R}^n$ and $\varepsilon_1, \ldots, \varepsilon_N$ be independent symmetric Bernoulli random variables. Then*

$$\mathbb{E} \Big\| \sum_{i=1}^{N} \varepsilon_i x_i \otimes x_i \Big\| \le C\sqrt{\log \min(N, n)} \cdot \max_{i \le N} \|x_i\|_2 \cdot \Big\| \sum_{i=1}^{N} x_i \otimes x_i \Big\|^{1/2}$$

where C is an absolute constant.

Proof. One can assume that $n \le N$ by replacing $\mathbb{R}^n$ with the linear span of $\{x_1, \ldots, x_N\}$ if necessary. The claim then follows from (5.18), since

$$\left\| \Big(\sum_{i=1}^{N} (x_i \otimes x_i)^2 \Big)^{1/2} \right\| = \Big\| \sum_{i=1}^{N} \|x_i\|_2^2 \, x_i \otimes x_i \Big\|^{1/2} \le \max_{i \le N} \|x_i\|_2 \Big\| \sum_{i=1}^{N} x_i \otimes x_i \Big\|^{1/2}. \quad \square$$

Ahlswede and Winter [4] pioneered a different approach to matrix-valued inequalities in probability theory, which was based on trace inequalities like Golden-Thompson inequality. A development of this idea leads to remarkably sharp results. We quote one such inequality from [77]:

Theorem 5.29 (Non-commutative Bernstein-type inequality [77]). *Consider a finite sequence X_i of independent centered self-adjoint random $n \times n$ matrices. Assume we have for some numbers K and σ that*

$$\|X_i\| \le K \text{ almost surely,} \quad \Big\| \sum_i \mathbb{E}X_i^2 \Big\| \le \sigma^2.$$

Then, for every $t \geq 0$ we have

$$\mathbb{P}\Big\{\big\|\sum_i X_i\big\| \geq t\Big\} \leq 2n \cdot \exp\Big(\frac{-t^2/2}{\sigma^2 + Kt/3}\Big). \tag{5.19}$$

Remark 5.30. This is a direct matrix generalization of a classical Bernstein's inequality for bounded random variables. To compare it with our version of Bernstein's inequality for sub-exponentials, Proposition 5.16, note that the probability bound in (5.19) is equivalent to $2n \cdot \exp\big[-c\min\big(\frac{t^2}{\sigma^2}, \frac{t}{K}\big)\big]$ where $c > 0$ is an absolute constant. In both results we see a mixture of gaussian and exponential tails.

5.3 Random matrices with independent entries

We are ready to study the extreme singular values of random matrices. In this section, we consider the classical model of random matrices whose entries are independent and centered random variables. Later we will study the more difficult models where only the rows or the columns are independent.

The reader may keep in mind some classical examples of $N \times n$ random matrices with independent entries. The most classical example is the *Gaussian random matrix* A whose entries are independent standard normal random variables. In this case, the $n \times n$ symmetric matrix A^*A is called Wishart matrix; it is a higher-dimensional version of chi-square distributed random variables.

The simplest example of discrete random matrices is the *Bernoulli random matrix A* whose entries are independent symmetric Bernoulli random variables. In other words, Bernoulli random matrices are distributed uniformly in the set of all $N \times n$ matrices with ± 1 entries.

5.3.1 Limit laws and Gaussian matrices

Consider an $N \times n$ random matrix A whose entries are independent centered identically distributed random variables. By now, the *limiting behavior* of the extreme singular values of A, as the dimensions $N, n \to \infty$, is well understood:

Theorem 5.31 (Bai-Yin's law, see [8])**.** *Let $A = A_{N,n}$ be an $N \times n$ random matrix whose entries are independent copies of a random variable with zero mean, unit variance, and finite fourth moment. Suppose that the dimensions N and n grow to infinity while the aspect ratio n/N converges to a constant in $[0, 1]$. Then*

$$s_{\min}(A) = \sqrt{N} - \sqrt{n} + o(\sqrt{n}), \quad s_{\max}(A) = \sqrt{N} + \sqrt{n} + o(\sqrt{n}) \quad \text{almost surely.}$$

As we pointed out in the introduction, our program is to find non-asymptotic versions of Bai-Yin's law. There is precisely one model of random matrices, namely Gaussian, where an *exact* non-asymptotic result is known:

Theorem 5.32 (Gordon's theorem for Gaussian matrices)**.** *Let A be an $N \times n$ matrix whose entries are independent standard normal random variables. Then*

$$\sqrt{N} - \sqrt{n} \leq \mathbb{E}s_{\min}(A) \leq \mathbb{E}s_{\max}(A) \leq \sqrt{N} + \sqrt{n}.$$

The proof of the upper bound, which we borrowed from [21], is based on Slepian's comparison inequality for Gaussian processes.[16]

Lemma 5.33 (Slepian's inequality, see [45] Section 3.3). *Consider two Gaussian processes $(X_t)_{t \in T}$ and $(Y_t)_{t \in T}$ whose increments satisfy the inequality $\mathbb{E}|X_s - X_t|^2 \leq \mathbb{E}|Y_s - Y_t|^2$ for all $s, t \in T$. Then $\mathbb{E}\sup_{t \in T} X_t \leq \mathbb{E}\sup_{t \in T} Y_t$.*

Proof of Theorem 5.32. We recognize $s_{\max}(A) = \max_{u \in S^{n-1},\, v \in S^{N-1}} \langle Au, v \rangle$ to be the supremum of the Gaussian process $X_{u,v} = \langle Au, v \rangle$ indexed by the pairs of vectors $(u, v) \in S^{n-1} \times S^{N-1}$. We shall compare this process to the following one whose supremum is easier to estimate: $Y_{u,v} = \langle g, u \rangle + \langle h, v \rangle$ where $g \in \mathbb{R}^n$ and $h \in \mathbb{R}^N$ are independent standard Gaussian random vectors. The rotation invariance of Gaussian measures makes it easy to compare the increments of these processes. For every $(u, v), (u', v') \in S^{n-1} \times S^{N-1}$, one can check that

$$\mathbb{E}|X_{u,v} - X_{u',v'}|^2 = \sum_{i=1}^{n} \sum_{j=1}^{N} |u_i v_j - u'_i v'_j|^2 \leq \|u - u'\|_2^2 + \|v - v'\|_2^2 = \mathbb{E}|Y_{u,v} - Y_{u',v'}|^2.$$

Therefore Lemma 5.33 applies, and it yields the required bound

$$\mathbb{E}s_{\max}(A) = \mathbb{E}\max_{(u,v)} X_{u,v} \leq \mathbb{E}\max_{(u,v)} Y_{u,v} = \mathbb{E}\|g\|_2 + \mathbb{E}\|h\|_2 \leq \sqrt{N} + \sqrt{n}.$$

Similar ideas are used to estimate $\mathbb{E}s_{\min}(A) = \mathbb{E}\max_{v \in S^{N-1}} \min_{u \in S^{n-1}} \langle Au, v \rangle$, see [21]. One uses in this case Gordon's generalization of Slepian's inequality for minimax of Gaussian processes [35, 36, 37], see [45, Section 3.3]. □

While Theorem 5.32 is about the expectation of singular values, it also yields a large deviation inequality for them. It can be deduced formally by using the *concentration of measure* in the Gauss space.

Proposition 5.34 (Concentration in Gauss space, see [43]). *Let f be a real valued Lipschitz function on $\mathbb{R}^n$ with Lipschitz constant K, i.e. $|f(x) - f(y)| \leq K\|x - y\|_2$ for all $x, y \in \mathbb{R}^n$ (such functions are also called K-Lipschitz). Let X be the standard normal random vector in $\mathbb{R}^n$. Then for every $t \geq 0$ one has*

$$\mathbb{P}\{f(X) - \mathbb{E}f(X) > t\} \leq \exp(-t^2/2K^2).$$

Corollary 5.35 (Gaussian matrices, deviation; see [21]). *Let A be an $N \times n$ matrix whose entries are independent standard normal random variables. Then for every $t \geq 0$, with probability at least $1 - 2\exp(-t^2/2)$ one has*

$$\sqrt{N} - \sqrt{n} - t \leq s_{\min}(A) \leq s_{\max}(A) \leq \sqrt{N} + \sqrt{n} + t.$$

Proof. Note that $s_{\min}(A)$, $s_{\max}(A)$ are 1-Lipschitz functions of matrices A considered as vectors in $\mathbb{R}^{Nn}$. The conclusion now follows from the estimates on the expectation (Theorem 5.32) and Gaussian concentration (Proposition 5.34). □

---

[16]Recall that a Gaussian process $(X_t)_{t \in T}$ is a collection of centered normal random variables X_t on the same probability space, indexed by points t in an abstract set T.

Later in these notes, we find it more convenient to work with the $n \times n$ positive-definite symmetric matrix A^*A rather than with the original $N \times n$ matrix A. Observe that the normalized matrix $\bar{A} = \frac{1}{\sqrt{N}}A$ is an approximate isometry (which is our goal) if and only if $\bar{A}^*\bar{A}$ is an approximate identity:

Lemma 5.36 (Approximate isometries). *Consider a matrix B that satisfies*

$$\|B^*B - I\| \leq \max(\delta, \delta^2) \tag{5.20}$$

for some $\delta > 0$. Then

$$1 - \delta \leq s_{\min}(B) \leq s_{\max}(B) \leq 1 + \delta. \tag{5.21}$$

Conversely, if B satisfies (5.21) *for some $\delta > 0$ then $\|B^*B - I\| \leq 3\max(\delta, \delta^2)$.*

Proof. Inequality (5.20) holds if and only if $\left|\|Bx\|_2^2 - 1\right| \leq \max(\delta, \delta^2)$ for all $x \in S^{n-1}$. Similarly, (5.21) holds if and only if $\left|\|Bx\|_2 - 1\right| \leq \delta$ for all $x \in S^{n-1}$. The conclusion then follows from the elementary inequality

$$\max(|z - 1|, |z - 1|^2) \leq |z^2 - 1| \leq 3\max(|z - 1|, |z - 1|^2) \quad \text{for all } z \geq 0. \quad \square$$

Lemma 5.36 reduces our task of proving inequalities (5.2) to showing an equivalent (but often more convenient) bound

$$\left\|\frac{1}{N}A^*A - I\right\| \leq \max(\delta, \delta^2) \quad \text{where } \delta = O(\sqrt{n/N}).$$

5.3.2 General random matrices with independent entries

Now we pass to a more general model of random matrices whose entries are independent centered random variables with some general distribution (not necessarily normal). The largest singular value (the spectral norm) can be estimated by Latala's theorem for general random matrices with non-identically distributed entries:

Theorem 5.37 (Latala's theorem [42]). *Let A be a random matrix whose entries a_{ij} are independent centered random variables with finite fourth moment. Then*

$$\mathbb{E}s_{\max}(A) \leq C\Big[\max_i \big(\sum_j \mathbb{E}a_{ij}^2\big)^{1/2} + \max_j \big(\sum_i \mathbb{E}a_{ij}^2\big)^{1/2} + \big(\sum_{i,j} \mathbb{E}a_{ij}^4\big)^{1/4}\Big].$$

If the variance and the fourth moments of the entries are uniformly bounded, then Latala's result yields $s_{\max}(A) = O(\sqrt{N} + \sqrt{n})$. This is slightly weaker than our goal (5.2), which is $s_{\max}(A) = \sqrt{N} + O(\sqrt{n})$ but still satisfactory for most applications. Results of the latter type will appear later in the more general model of random matrices with independent rows or columns.

Similarly, our goal (5.2) for the smallest singular value is $s_{\min}(A) \geq \sqrt{N} - O(\sqrt{n})$. Since the singular values are non-negative anyway, such inequality would only be useful for sufficiently tall matrices, $N \gg n$. For almost square and square matrices, estimating the smallest singular value (known also as the *hard edge* of spectrum) is considerably more difficult. The progress on estimating the hard edge is summarized in [69]. If A has independent entries, then indeed $s_{\min}(A) \geq c(\sqrt{N} - \sqrt{n})$, and the following is an optimal probability bound:

Theorem 5.38 (Independent entries, hard edge [68])**.** *Let A be an $n \times n$ random matrix whose entries are independent identically distributed subgaussian random variables with zero mean and unit variance. Then for $\varepsilon \geq 0$,*

$$\mathbb{P}\big(s_{\min}(A) \leq \varepsilon(\sqrt{N} - \sqrt{n-1})\big) \leq (C\varepsilon)^{N-n+1} + c^N$$

where $C > 0$ and $c \in (0,1)$ depend only on the subgaussian norm of the entries.

This result gives an optimal bound for square matrices as well ($N = n$).

5.4   Random matrices with independent rows

In this section, we focus on a more general model of random matrices, where we only assume independence of the rows rather than all entries. Such matrices are naturally *generated by high-dimensional distributions*. Indeed, given an arbitrary probability distribution in $\mathbb{R}^n$, one takes a sample of N independent points and arranges them as the rows of an $N \times n$ matrix A. By studying spectral properties of A one should be able to learn something useful about the underlying distribution. For example, as we will see in Section 5.4.3, the extreme singular values of A would tell us whether the covariance matrix of the distribution can be estimated from a sample of size N.

The picture will vary slightly depending on whether the rows of A are sub-gaussian or have arbitrary distribution. For heavy-tailed distributions, an extra logarithmic factor has to appear in our desired inequality (5.2). The analysis of sub-gaussian and heavy-tailed matrices will be completely different.

There is an abundance of examples where the results of this section may be useful. They include all matrices with independent entries, whether sub-gaussian such as Gaussian and Bernoulli, or completely general distributions with mean zero and unit variance. In the latter case one is able to surpass the fourth moment assumption which is necessary in Bai-Yin's law, Theorem 5.31.

Other examples of interest come from non-product distributions, some of which we saw in Example 5.21. Sampling from discrete objects (matrices and frames) fits well in this framework, too. Given a deterministic matrix B, one puts a uniform distribution on the set of the rows of B and creates a random matrix A as before – by sampling some N random rows from B. Applications to sampling will be discussed in Section 5.4.4.

5.4.1   Sub-gaussian rows

The following result goes in the direction of our goal (5.2) for random matrices with independent sub-gaussian rows.

Theorem 5.39 (Sub-gaussian rows)**.** *Let A be an $N \times n$ matrix whose rows A_i are independent sub-gaussian isotropic random vectors in $\mathbb{R}^n$. Then for every $t \geq 0$, with probability at least $1 - 2\exp(-ct^2)$ one has*

$$\sqrt{N} - C\sqrt{n} - t \leq s_{\min}(A) \leq s_{\max}(A) \leq \sqrt{N} + C\sqrt{n} + t. \qquad (5.22)$$

Here $C = C_K$, $c = c_K > 0$ depend only on the subgaussian norm $K = \max_i \|A_i\|_{\psi_2}$ of the rows.

This result is a general version of Corollary 5.35 (up to absolute constants); instead of independent Gaussian entries we allow independent sub-gaussian rows. This of course covers all matrices with independent sub-gaussian entries such as Gaussian and Bernoulli. It also applies to some natural matrices whose entries are not independent. One such example is a matrix whose rows are independent spherical random vectors (Example 5.25).

Proof. The proof is a basic version of a *covering argument*, and it has three steps. We need to control $\|Ax\|_2$ for all vectors x on the unit sphere S^{n-1}. To this end, we discretize the sphere using a net $\mathcal{N}$ (the approximation step), establish a tight control of $\|Ax\|_2$ for every fixed vector $x \in \mathcal{N}$ with high probability (the concentration step), and finish off by taking a union bound over all x in the net. The concentration step will be based on the deviation inequality for sub-exponential random variables, Corollary 5.17.

Step 1: Approximation. Recalling Lemma 5.36 for the matrix $B = A/\sqrt{N}$ we see that the conclusion of the theorem is equivalent to

$$\left\|\frac{1}{N}A^*A - I\right\| \leq \max(\delta, \delta^2) =: \varepsilon \quad \text{where} \quad \delta = C\sqrt{\frac{n}{N}} + \frac{t}{\sqrt{N}}. \tag{5.23}$$

Using Lemma 5.4, we can evaluate the operator norm in (5.23) on a $\frac{1}{4}$-net $\mathcal{N}$ of the unit sphere S^{n-1}:

$$\left\|\frac{1}{N}A^*A - I\right\| \leq 2\max_{x \in \mathcal{N}}\left|\left\langle\left(\frac{1}{N}A^*A - I\right)x, x\right\rangle\right| = 2\max_{x \in \mathcal{N}}\left|\frac{1}{N}\|Ax\|_2^2 - 1\right|.$$

So to complete the proof it suffices to show that, with the required probability,

$$\max_{x \in \mathcal{N}}\left|\frac{1}{N}\|Ax\|_2^2 - 1\right| \leq \frac{\varepsilon}{2}.$$

By Lemma 5.2, we can choose the net $\mathcal{N}$ so that it has cardinality $|\mathcal{N}| \leq 9^n$.

Step 2: Concentration. Let us fix any vector $x \in S^{n-1}$. We can express $\|Ax\|_2^2$ as a sum of independent random variables

$$\|Ax\|_2^2 = \sum_{i=1}^{N}\langle A_i, x\rangle^2 =: \sum_{i=1}^{N} Z_i^2 \tag{5.24}$$

where A_i denote the rows of the matrix A. By assumption, $Z_i = \langle A_i, x\rangle$ are independent sub-gaussian random variables with $\mathbb{E}Z_i^2 = 1$ and $\|Z_i\|_{\psi_2} \leq K$. Therefore, by Remark 5.18 and Lemma 5.14, $Z_i^2 - 1$ are independent centered sub-exponential random variables with $\|Z_i^2 - 1\|_{\psi_1} \leq 2\|Z_i^2\|_{\psi_1} \leq 4\|Z_i\|_{\psi_2}^2 \leq 4K^2$.

We can therefore use an exponential deviation inequality, Corollary 5.17, to control the sum (5.24). Since $K \geq \|Z_i\|_{\psi_2} \geq \frac{1}{\sqrt{2}}(\mathbb{E}|Z_i|^2)^{1/2} = \frac{1}{\sqrt{2}}$, this gives

$$\mathbb{P}\left\{\left|\frac{1}{N}\|Ax\|_2^2 - 1\right| \geq \frac{\varepsilon}{2}\right\} = \mathbb{P}\left\{\left|\frac{1}{N}\sum_{i=1}^{N} Z_i^2 - 1\right| \geq \frac{\varepsilon}{2}\right\} \leq 2\exp\left[-\frac{c_1}{K^4}\min(\varepsilon^2, \varepsilon)N\right]$$

$$= 2\exp\left[-\frac{c_1}{K^4}\delta^2 N\right] \leq 2\exp\left[-\frac{c_1}{K^4}(C^2 n + t^2)\right]$$

where the last inequality follows by the definition of δ and using the inequality $(a+b)^2 \geq a^2 + b^2$ for $a, b \geq 0$.

Step 3: Union bound. Taking the union bound over all vectors x in the net $\mathcal{N}$ of cardinality $|\mathcal{N}| \leq 9^n$, we obtain

$$\mathbb{P}\Big\{ \max_{x \in \mathcal{N}} \big| \frac{1}{N} \|Ax\|_2^2 - 1 \big| \geq \frac{\varepsilon}{2} \Big\} \leq 9^n \cdot 2 \exp\Big[-\frac{c_1}{K^4}(C^2 n + t^2) \Big] \leq 2 \exp\Big(-\frac{c_1 t^2}{K^4} \Big)$$

where the second inequality follows for $C = C_K$ sufficiently large, e.g. $C = K^2\sqrt{\ln 9/c_1}$. As we noted in Step 1, this completes the proof of the theorem. $\square$

Remark 5.40 (Non-isotropic distributions).   1. A version of Theorem 5.39 holds for general, non-isotropic sub-gaussian distributions. Assume that A is an $N \times n$ matrix whose rows A_i are independent sub-gaussian random vectors in $\mathbb{R}^n$ with second moment matrix Σ. Then for every $t \geq 0$, the following inequality holds with probability at least $1 - 2\exp(-ct^2)$:

$$\big\| \frac{1}{N} A^* A - \Sigma \big\| \leq \max(\delta, \delta^2) \quad \text{where} \quad \delta = C\sqrt{\frac{n}{N}} + \frac{t}{\sqrt{N}}. \tag{5.25}$$

   Here as before $C = C_K$, $c = c_K > 0$ depend only on the subgaussian norm $K = \max_i \|A_i\|_{\psi_2}$ of the rows. This result is a general version of (5.23). It follows by a straighforward modification of the argument of Theorem 5.39.

2. A more natural, multiplicative form of (5.25) is the following. Assume that $\Sigma^{-1/2} A_i$ are isotropic sub-gaussian random vectors, and let K be the maximum of their sub-gaussian norms. Then for every $t \geq 0$, the following inequality holds with probability at least $1 - 2\exp(-ct^2)$:

$$\big\| \frac{1}{N} A^* A - \Sigma \big\| \leq \max(\delta, \delta^2) \, \|\Sigma\| \quad \text{where} \quad \delta = C\sqrt{\frac{n}{N}} + \frac{t}{\sqrt{N}} \tag{5.26}$$

   Here again $C = C_K$, $c = c_K > 0$. This result follows from Theorem 5.39 applied to the isotropic random vectors $\Sigma^{-1/2} A_i$.

5.4.2   Heavy-tailed rows

The class of sub-gaussian random variables in Theorem 5.39 may sometimes be too restrictive in applications. For example, if the rows of A are independent coordinate or frame random vectors (Examples 5.21 and 5.25), they are poorly sub-gaussian and Theorem 5.39 is too weak. In such cases, one would use the following result instead, which operates in remarkable generality.

Theorem 5.41 (Heavy-tailed rows). *Let A be an $N \times n$ matrix whose rows A_i are independent isotropic random vectors in $\mathbb{R}^n$. Let m be a number such that $\|A_i\|_2 \leq \sqrt{m}$ almost surely for all i. Then for every $t \geq 0$, one has*

$$\sqrt{N} - t\sqrt{m} \leq s_{\min}(A) \leq s_{\max}(A) \leq \sqrt{N} + t\sqrt{m} \tag{5.27}$$

with probability at least $1 - 2n \cdot \exp(-ct^2)$, where $c > 0$ is an absolute constant.

Recall that $(\mathbb{E}\|A_i\|_2^2)^{1/2} = \sqrt{n}$ by Lemma 5.20. This indicates that one would typically use Theorem 5.41 with $m = O(n)$. In this case the result takes the form

$$\sqrt{N} - t\sqrt{n} \leq s_{\min}(A) \leq s_{\max}(A) \leq \sqrt{N} + t\sqrt{n} \tag{5.28}$$

with probability at least $1 - 2n \cdot \exp(-c't^2)$. This is a form of our desired inequality (5.2) for heavy-tailed matrices. We shall discuss this more after the proof.

Proof. We shall use the non-commutative Bernstein's inequality, Theorem 5.29.

Step 1: Reduction to a sum of independent random matrices. We first note that $m \geq n \geq 1$ since by Lemma 5.20 we have $\mathbb{E}\|A_i\|_2^2 = n$. Now we start an argument parallel to Step 1 of Theorem 5.39. Recalling Lemma 5.36 for the matrix $B = A/\sqrt{N}$ we see that the desired inequalities (5.27) are equivalent to

$$\left\|\frac{1}{N}A^*A - I\right\| \leq \max(\delta, \delta^2) =: \varepsilon \quad \text{where} \quad \delta = t\sqrt{\frac{m}{N}}. \tag{5.29}$$

We express this random matrix as a sum of independent random matrices:

$$\frac{1}{N}A^*A - I = \frac{1}{N}\sum_{i=1}^{N} A_i \otimes A_i - I = \sum_{i=1}^{N} X_i, \quad \text{where} \quad X_i := \frac{1}{N}(A_i \otimes A_i - I);$$

note that X_i are independent centered $n \times n$ random matrices.

Step 2: Estimating the mean, range and variance. We are going to apply the non-commutative Bernstein inequality, Theorem 5.29, for the sum $\sum_i X_i$. Since A_i are isotropic random vectors, we have $\mathbb{E}A_i \otimes A_i = I$ which implies that $\mathbb{E}X_i = 0$ as required in the non-commutative Bernstein inequality.

We estimate the range of X_i using that $\|A_i\|_2 \leq \sqrt{m}$ and $m \geq 1$:

$$\|X_i\| \leq \frac{1}{N}(\|A_i \otimes A_i\| + 1) = \frac{1}{N}(\|A_i\|_2^2 + 1) \leq \frac{1}{N}(m + 1) \leq \frac{2m}{N} =: K$$

To estimate the total variance $\|\sum_i \mathbb{E}X_i^2\|$, we first compute

$$X_i^2 = \frac{1}{N^2}\big[(A_i \otimes A_i)^2 - 2(A_i \otimes A_i) + I\big],$$

so using that the isotropy assumption $\mathbb{E}A_i \otimes A_i = I$ we obtain

$$\mathbb{E}X_i^2 = \frac{1}{N^2}\big[\mathbb{E}(A_i \otimes A_i)^2 - I\big]. \tag{5.30}$$

Since $(A_i \otimes A_i)^2 = \|A_i\|_2^2 A_i \otimes A_i$ is a positive semi-definite matrix and $\|A_i\|_2^2 \leq m$ by assumption, we have $\big\|\mathbb{E}(A_i \otimes A_i)^2\big\| \leq m \cdot \|\mathbb{E}A_i \otimes A_i\| = m$. Putting this into (5.30) we obtain

$$\|\mathbb{E}X_i^2\| \leq \frac{1}{N^2}(m + 1) \leq \frac{2m}{N^2}$$

where we again used that $m \geq 1$. This yields[17]

$$\Big\| \sum_{i=1}^{N} \mathbb{E}X_i^2 \Big\| \leq N \cdot \max_i \|\mathbb{E}X_i^2\| = \frac{2m}{N} =: \sigma^2.$$

Step 3: Application of the non-commutative Bernstein's inequality. Applying Theorem 5.29 (see Remark 5.30) and recalling the definitions of ε and δ in (5.29), we we bound the probability in question as

$$\mathbb{P}\Big\{ \Big\| \frac{1}{N} A^*A - I \Big\| \geq \varepsilon \Big\} = \mathbb{P}\Big\{ \Big\| \sum_{i=1}^{N} X_i \Big\| \geq \varepsilon \Big\} \leq 2n \cdot \exp\Big[-c\min\Big(\frac{\varepsilon^2}{\sigma^2}, \frac{\varepsilon}{K} \Big) \Big]$$

$$\leq 2n \cdot \exp\Big[-c\min(\varepsilon^2, \varepsilon) \cdot \frac{N}{2m} \Big] = 2n \cdot \exp\Big(-\frac{c\delta^2 N}{2m} \Big) = 2n \cdot \exp(-ct^2/2).$$

This completes the proof. □

Theorem 5.41 for heavy-tailed rows is different from Theorem 5.39 for sub-gaussian rows in two ways: the boundedness assumption[18] $\|A_i\|_2^2 \leq m$ appears, and the probability bound is weaker. We will now comment on both differences.

Remark 5.42 (Boundedness assumption). Observe that some boundedness assumption on the distribution is needed in Theorem 5.41. Let us see this on the following example. Choose $\delta \geq 0$ arbitrarily small, and consider a random vector $X = \delta^{-1/2}\xi Y$ in $\mathbb{R}^n$ where ξ is a $\{0,1\}$-valued random variable with $\mathbb{E}\xi = \delta$ (a "selector") and Y is an independent isotropic random vector in $\mathbb{R}^n$ with an arbitrary distribution. Then X is also an isotropic random vector. Consider an $N \times n$ random matrix A whose rows A_i are independent copies of X. However, if $\delta \geq 0$ is suitably small then $A = 0$ with high probability, hence no nontrivial lower bound on $s_{\min}(A)$ is possible.

Inequality (5.28) fits our goal (5.2), but not quite. The reason is that the probability bound is only non-trivial if $t \geq C\sqrt{\log n}$. Therefore, in reality Theorem 5.41 asserts that

$$\sqrt{N} - C\sqrt{n \log n} \leq s_{\min}(A) \leq s_{\max}(A) \leq \sqrt{N} + C\sqrt{n \log n} \tag{5.31}$$

with probability, say 0.9. This achieves our goal (5.2) up to a logarithmic factor.

Remark 5.43 (Logarithmic factor). The logarithmic factor can not be removed from (5.31) for some heavy-tailed distributions. Consider for instance the coordinate distribution introduced in Example 5.21. In order that $s_{\min}(A) > 0$ there must be no zero columns in A. Equivalently, each coordinate vector $e_1, \ldots, e_n$ must be picked at least once in N independent trials (each row of A picks an independent coordinate vector). Recalling the classical coupon collector's problem, one must make at least $N \geq Cn \log n$ trials to make this occur with high probability. Thus the logarithm is necessary in the left hand side of (5.31).[19]

---

[17]Here the seemingly crude application of triangle inequality is actually not so loose. If the rows A_i are identically distributed, then so are X_i^2, which makes the triangle inequality above into an equality.

[18]Going a little ahead, we would like to point out that the almost sure boundedness can be relaxed to the bound in expectation $\mathbb{E}\max_i \|A_i\|_2^2 \leq m$, see Theorem 5.45.

[19]This argument moreover shows the optimality of the probability bound in Theorem 5.41. For example, for $t = \sqrt{N}/2\sqrt{n}$ the conclusion (5.28) implies that A is well conditioned (i.e. $\sqrt{N}/2 \leq s_{\min}(A) \leq s_{\max}(A) \leq 2\sqrt{N}$) with probability $1 - n \cdot \exp(-cN/n)$. On the other hand, by the coupon collector's problem we estimate the probability that $s_{\min}(A) > 0$ as $1 - n \cdot (1 - \frac{1}{n})^N \approx 1 - n \cdot \exp(-N/n)$.

A version of Theorem 5.41 holds for general, non-isotropic distributions. It is convenient to state it in terms of the equivalent estimate (5.29):

Theorem 5.44 (Heavy-tailed rows, non-isotropic). *Let A be an $N \times n$ matrix whose rows A_i are independent random vectors in $\mathbb{R}^n$ with the common second moment matrix $\Sigma = \mathbb{E} A_i \otimes A_i$. Let m be a number such that $\|A_i\|_2 \leq \sqrt{m}$ almost surely for all i. Then for every $t \geq 0$, the following inequality holds with probability at least $1 - n \cdot \exp(-ct^2)$:*

$$\big\| \frac{1}{N} A^* A - \Sigma \big\| \leq \max(\|\Sigma\|^{1/2} \delta, \delta^2) \quad where \quad \delta = t \sqrt{\frac{m}{N}}. \tag{5.32}$$

Here $c > 0$ is an absolute constant. In particular, this inequality yields

$$\|A\| \leq \|\Sigma\|^{1/2} \sqrt{N} + t \sqrt{m}. \tag{5.33}$$

Proof. We note that $m \geq \|\Sigma\|$ because $\|\Sigma\| = \|\mathbb{E} A_i \otimes A_i\| \leq \mathbb{E}\|A_i \otimes A_i\| = \mathbb{E}\|A_i\|_2^2 \leq m$. Then (5.32) follows by a straightforward modification of the argument of Theorem 5.41. Furthermore, if (5.32) holds then by triangle inequality

$$\frac{1}{N} \|A\|^2 = \big\| \frac{1}{N} A^* A \big\| \leq \|\Sigma\| + \big\| \frac{1}{N} A^* A - \Sigma \big\|$$
$$\leq \|\Sigma\| + \|\Sigma\|^{1/2} \delta + \delta^2 \leq (\|\Sigma\|^{1/2} + \delta)^2.$$

Taking square roots and multiplying both sides by $\sqrt{N}$, we obtain (5.33). $\qquad\square$

The *almost sure* boundedness requirement in Theorem 5.41 may sometimes be too restrictive in applications, and it can be relaxed to a bound *in expectation*:

Theorem 5.45 (Heavy-tailed rows; expected singular values). *Let A be an $N \times n$ matrix whose rows A_i are independent isotropic random vectors in $\mathbb{R}^n$. Let $m := \mathbb{E} \max_{i \leq N} \|A_i\|_2^2$. Then*

$$\mathbb{E} \max_{j \leq n} |s_j(A) - \sqrt{N}| \leq C \sqrt{m \log \min(N, n)}$$

where C is an absolute constant.

The proof of this result is similar to that of Theorem 5.41, except that this time we will use Rudelson's Corollary 5.28 instead of matrix Bernstein's inequality. To this end, we need a link to symmetric Bernoulli random variables. This is provided by a general *symmetrization argument*:

Lemma 5.46 (Symmetrization). *Let (X_i) be a finite sequence of independent random vectors valued in some Banach space, and (ε_i) be independent symmetric Bernoulli random variables. Then*

$$\mathbb{E} \big\| \sum_i (X_i - \mathbb{E} X_i) \big\| \leq 2\mathbb{E} \big\| \sum_i \varepsilon_i X_i \big\|. \tag{5.34}$$

Proof. We define random variables $\tilde{X}_i = X_i - X_i'$ where (X_i') is an independent copy of the sequence (X_i). Then $\tilde{X}_i$ are independent symmetric random variables, i.e. the sequence $(\tilde{X}_i)$ is distributed identically with $(-\tilde{X}_i)$ and thus also with $(\varepsilon_i \tilde{X}_i)$. Replacing $\mathbb{E}X_i$ by $\mathbb{E}X_i'$ in (5.34) and using Jensen's inequality, symmetry, and triangle inequality, we obtain the required inequality

$$\mathbb{E}\Big\| \sum_i (X_i - \mathbb{E}X_i) \Big\| \le \mathbb{E}\Big\| \sum_i \tilde{X}_i \Big\| = \mathbb{E}\Big\| \sum_i \varepsilon_i \tilde{X}_i \Big\|$$

$$\le \mathbb{E}\Big\| \sum_i \varepsilon_i X_i \Big\| + \mathbb{E}\Big\| \sum_i \varepsilon_i X_i' \Big\| = 2\mathbb{E}\Big\| \sum_i \varepsilon_i X_i \Big\|. \qquad \square$$

We will also need a probabilistic version of Lemma 5.36 on approximate isometries. The proof of that lemma was based on the elementary inequality $|z^2 - 1| \ge \max(|z - 1|, |z - 1|^2)$ for $z \ge 0$. Here is a probabilistic version:

Lemma 5.47. *Let Z be a non-negative random variable. Then $\mathbb{E}|Z^2 - 1| \ge \max(\mathbb{E}|Z - 1|, (\mathbb{E}|Z - 1|)^2)$.*

Proof. Since $|Z - 1| \le |Z^2 - 1|$ pointwise, we have $\mathbb{E}|Z - 1| \le \mathbb{E}|Z^2 - 1|$. Next, since $|Z - 1|^2 \le |Z^2 - 1|$ pointwise, taking square roots and expectations we obtain $\mathbb{E}|Z - 1| \le \mathbb{E}|Z^2 - 1|^{1/2} \le (\mathbb{E}|Z^2 - 1|)^{1/2}$, where the last bound follows by Jensen's inequality. Squaring both sides completes the proof. $\square$

Proof of Theorem 5.45. **Step 1: Application of Rudelson's inequality.** As in the proof of Theorem 5.41, we are going to control

$$E := \mathbb{E}\Big\| \frac{1}{N}A^*A - I \Big\| = \mathbb{E}\Big\| \frac{1}{N}\sum_{i=1}^N A_i \otimes A_i - I \Big\| \le \frac{2}{N}\mathbb{E}\Big\| \sum_{i=1}^N \varepsilon_i A_i \otimes A_i \Big\|$$

where we used Symmetrization Lemma 5.46 with independent symmetric Bernoulli random variables ε_i (which are independent of A as well). The expectation in the right hand side is taken both with respect to the random matrix A and the signs (ε_i). Taking first the expectation with respect to (ε_i) (conditionally on A) and afterwards the expectation with respect to A, we obtain by Rudelson's inequality (Corollary 5.28) that

$$E \le \frac{C\sqrt{l}}{N}\,\mathbb{E}\Big(\max_{i \le N}\|A_i\|_2 \cdot \Big\| \sum_{i=1}^N A_i \otimes A_i \Big\|^{1/2} \Big)$$

where $l = \log\min(N, n)$. We now apply the Cauchy-Schwarz inequality. Since by the triangle inequality $\mathbb{E}\big\| \frac{1}{N}\sum_{i=1}^N A_i \otimes A_i \big\| = \mathbb{E}\big\| \frac{1}{N}A^*A \big\| \le E + 1$, it follows that

$$E \le C\sqrt{\frac{ml}{N}}(E + 1)^{1/2}.$$

This inequality is easy to solve in E. Indeed, considering the cases $E \le 1$ and $E > 1$ separately, we conclude that

$$E = \mathbb{E}\Big\| \frac{1}{N}A^*A - I \Big\| \le \max(\delta, \delta^2) \quad \text{where } \delta := C\sqrt{\frac{2ml}{N}}.$$

Step 2: Diagonalization. Diagonalizing the matrix A^*A one checks that

$$\left\|\frac{1}{N}A^*A - I\right\| = \max_{j \leq n}\left|\frac{s_j(A)^2}{N} - 1\right| = \max\left(\left|\frac{s_{\min}(A)^2}{N} - 1\right|, \left|\frac{s_{\max}(A)^2}{N} - 1\right|\right).$$

It follows that

$$\max\left(\mathbb{E}\left|\frac{s_{\min}(A)^2}{N} - 1\right|, \mathbb{E}\left|\frac{s_{\max}(A)^2}{N} - 1\right|\right) \leq \max(\delta, \delta^2).$$

(we replaced the expectation of maximum by the maximum of expectations). Using Lemma 5.47 separately for the two terms on the left hand side, we obtain

$$\max\left(\mathbb{E}\left|\frac{s_{\min}(A)}{\sqrt{N}} - 1\right|, \mathbb{E}\left|\frac{s_{\max}(A)}{\sqrt{N}} - 1\right|\right) \leq \delta.$$

Therefore

$$\mathbb{E}\max_{j \leq n}\left|\frac{s_j(A)}{\sqrt{N}} - 1\right| = \mathbb{E}\max\left(\left|\frac{s_{\min}(A)}{\sqrt{N}} - 1\right|, \left|\frac{s_{\max}(A)}{\sqrt{N}} - 1\right|\right)$$

$$\leq \mathbb{E}\left(\left|\frac{s_{\min}(A)}{\sqrt{N}} - 1\right| + \left|\frac{s_{\max}(A)}{\sqrt{N}} - 1\right|\right) \leq 2\delta.$$

Multiplying both sides by $\sqrt{N}$ completes the proof. $\qquad\square$

In a way similar to Theorem 5.44 we note that a version of Theorem 5.45 holds for general, non-isotropic distributions.

Theorem 5.48 (Heavy-tailed rows, non-isotropic, expectation). *Let A be an $N \times n$ matrix whose rows A_i are independent random vectors in $\mathbb{R}^n$ with the common second moment matrix $\Sigma = \mathbb{E}A_i \otimes A_i$. Let $m := \mathbb{E}\max_{i \leq N}\|A_i\|_2^2$. Then*

$$\mathbb{E}\left\|\frac{1}{N}A^*A - \Sigma\right\| \leq \max(\|\Sigma\|^{1/2}\delta, \delta^2) \quad \text{where} \quad \delta = C\sqrt{\frac{m\log\min(N, n)}{N}}.$$

Here C is an absolute constant. In particular, this inequality yields

$$\left(\mathbb{E}\|A\|^2\right)^{1/2} \leq \|\Sigma\|^{1/2}\sqrt{N} + C\sqrt{m\log\min(N, n)}.$$

Proof. The first part follows by a simple modification of the proof of Theorem 5.45. The second part follows from the first like in Theorem 5.44. $\qquad\square$

Remark 5.49 (Non-identical second moments). The assumption that the rows A_i have a common second moment matrix Σ is not essential in Theorems 5.44 and 5.48. The reader will be able to formulate more general versions of these results. For example, if A_i have arbitrary second moment matrices $\Sigma_i = \mathbb{E}A_i \otimes A_i$ then the conclusion of Theorem 5.48 holds with $\Sigma = \frac{1}{N}\sum_{i=1}^N \Sigma_i$.

5.4.3   Applications to estimating covariance matrices

One immediate application of our analysis of random matrices is in statistics, for the fundamental problem of *estimating covariance matrices*. Let X be a random vector in $\mathbb{R}^n$; for simplicity we assume that X is centered,[20] $\mathbb{E}X = 0$. Recall that the covariance matrix of X is the $n \times n$ matrix $\Sigma = \mathbb{E}X \otimes X$, see Section 5.2.5.

The simplest way to estimate Σ is to take some N independent samples X_i from the distribution and form the *sample covariance matrix* $\Sigma_N = \frac{1}{N} \sum_{i=1}^N X_i \otimes X_i$. By the law of large numbers, $\Sigma_N \to \Sigma$ almost surely as $N \to \infty$. So, taking sufficiently many samples we are guaranteed to estimate the covariance matrix as well as we want. This, however, does not address the quantitative aspect: what is the minimal *sample size N* that guarantees approximation with a given accuracy?

The relation of this question to random matrix theory becomes clear when we arrange the samples $X_i =: A_i$ as rows of the $N \times n$ random matrix A. Then the sample covariance matrix is expressed as $\Sigma_N = \frac{1}{N} A^* A$. Note that A is a matrix with independent rows but usually not independent entries (unless we sample from a product distribution). We worked out the analysis of such matrices in Section 5.4, separately for sub-gaussian and general distributions. As an immediate consequence of Theorem 5.39, we obtain:

Corollary 5.50 (Covariance estimation for sub-gaussian distributions). *Consider a sub-gaussian distribution in $\mathbb{R}^n$ with covariance matrix Σ, and let $\varepsilon \in (0,1)$, $t \geq 1$. Then with probability at least $1 - 2\exp(-t^2 n)$ one has*

$$\text{If } N \geq C(t/\varepsilon)^2 n \quad \text{then } \|\Sigma_N - \Sigma\| \leq \varepsilon.$$

Here $C = C_K$ depends only on the sub-gaussian norm $K = \|X\|_{\psi_2}$ of a random vector taken from this distribution.

Proof. It follows from (5.25) that for every $s \geq 0$, with probability at least $1 - 2\exp(-cs^2)$ we have $\|\Sigma_N - \Sigma\| \leq \max(\delta, \delta^2)$ where $\delta = C\sqrt{n/N} + s/\sqrt{N}$. The conclusion follows for $s = C't\sqrt{n}$ where $C' = C'_K$ is sufficiently large. □

Summarizing, Corollary 5.50 shows that the sample size

$$N = O(n)$$

suffices to approximate the covariance matrix of a sub-gaussian distribution in $\mathbb{R}^n$ by the sample covariance matrix.

Remark 5.51 (Multiplicative estimates, Gaussian distributions). A weak point of Corollary 5.50 is that the sub-gaussian norm K may in turn depend on $\|\Sigma\|$.

To overcome this drawback, instead of using (5.25) in the proof of this result one can use the multiplicative version (5.26). The reader is encouraged to state a general result that follows from this argument. We just give one special example for arbitrary *centered Gaussian distributions* in $\mathbb{R}^n$. For every $\varepsilon \in (0,1)$, $t \geq 1$, the following holds with probability at least $1 - 2\exp(-t^2 n)$:

$$\text{If } N \geq C(t/\varepsilon)^2 n \quad \text{then } \|\Sigma_N - \Sigma\| \leq \varepsilon\|\Sigma\|.$$

---

[20]More generally, in this section we estimate the *second moment matrix* $\mathbb{E}X \otimes X$ of an arbitrary random vector X (not necessarily centered).

Here C is an absolute constant.

Finally, Theorem 5.44 yields a similar estimation result for arbitrary distributions, possibly heavy-tailed:

Corollary 5.52 (Covariance estimation for arbitrary distributions)**.** *Consider a distribution in $\mathbb{R}^n$ with covariance matrix Σ and supported in some centered Euclidean ball whose radius we denote $\sqrt{m}$. Let $\varepsilon \in (0, 1)$ and $t \geq 1$. Then the following holds with probability at least $1 - n^{-t^2}$:*

$$\text{If } N \geq C(t/\varepsilon)^2 \|\Sigma\|^{-1} m \log n \quad \text{then } \|\Sigma_N - \Sigma\| \leq \varepsilon \|\Sigma\|.$$

Here C is an absolute constant.

Proof. It follows from Theorem 5.44 that for every $s \geq 0$, with probability at least $1 - n \cdot \exp(-cs^2)$ we have $\|\Sigma_N - \Sigma\| \leq \max(\|\Sigma\|^{1/2}\delta, \delta^2)$ where $\delta = s\sqrt{m/N}$. Therefore, if $N \geq (s/\varepsilon)^2 \|\Sigma\|^{-1} m$ then $\|\Sigma_N - \Sigma\| \leq \varepsilon \|\Sigma\|$. The conclusion follows with $s = C't\sqrt{\log n}$ where C' is a sufficiently large absolute constant. $\qquad\square$

Corollary 5.52 is typically used with $m = O(\|\Sigma\| n)$. Indeed, if X is a random vector chosen from the distribution in question, then its expected norm is easy to estimate: $\mathbb{E}\|X\|_2^2 = \text{tr}(\Sigma) \leq n\|\Sigma\|$. So, by Markov's inequality, most of the distribution is supported in a centered ball of radius $\sqrt{m}$ where $m = O(n\|\Sigma\|)$. If all distribution is supported there, i.e. if $\|X\| = O(\sqrt{n\|\Sigma\|})$ almost surely, then the conclusion of Corollary 5.52 holds with sample size $N \geq C(t/\varepsilon)^2 n \log n$.

Remark 5.53 (Low-rank estimation)*.* In certain applications, the distribution in $\mathbb{R}^n$ lies close to a low dimensional subspace. In this case, a smaller sample suffices for covariance estimation. The intrinsic dimension of the distribution can be measured with the *effective rank* of the matrix Σ, defined as

$$r(\Sigma) = \frac{\text{tr}(\Sigma)}{\|\Sigma\|}.$$

One always has $r(\Sigma) \leq \text{rank}(\Sigma) \leq n$, and this bound is sharp. For example, if X is an isotropic random vector in $\mathbb{R}^n$ then $\Sigma = I$ and $r(\Sigma) = n$. A more interesting example is where X takes values in some r-dimensional subspace E, and the restriction of the distribution of X onto E is isotropic. The latter means that $\Sigma = P_E$, where P_E denotes the orthogonal projection in $\mathbb{R}^n$ onto E. Therefore in this case $r(\Sigma) = r$. The effective rank is a stable quantity compared with the usual rank. For distributions that are approximately low-dimenional, the effective rank is still small.

The effective rank $r = r(\Sigma)$ always controls the typical norm of X, as $\mathbb{E}\|X\|_2^2 = \text{tr}(\Sigma) = r\|\Sigma\|$. It follows by Markov's inequality that most of the distribution is supported in a ball of radius $\sqrt{m}$ where $m = O(r\|\Sigma\|)$. Assume that all of the distribution is supported there, i.e. if $\|X\| = O(\sqrt{r\|\Sigma\|})$ almost surely. Then the conclusion of Corollary 5.52 holds with sample size $N \geq C(t/\varepsilon)^2 r \log n$.

We can summarize this discussion in the following way: the sample size

$$N = O(n \log n)$$

suffices to approximate the covariance matrix of a general distribution in $\mathbb{R}^n$ by the sample covariance matrix. Furthermore, for distributions that are approximately low-dimensional, a smaller sample size is sufficient. Namely, if the effective rank of Σ equals r then a sufficient sample size is

$$N = O(r \log n).$$

Remark 5.54 (Boundedness assumption). Without the boundedness assumption on the distribution, Corollary 5.52 may fail. The reasoning is the same as in Remark 5.42: for an isotropic distribution which is highly concentrated at the origin, the sample covariance matrix will likely equal 0.

Still, one can weaken the boundedness assumption using Theorem 5.48 instead of Theorem 5.44 in the proof of Corollary 5.52. The weaker requirement is that $\mathbb{E} \max_{i \leq N} \|X_i\|_2^2 \leq m$ where X_i denote the sample points. In this case, the covariance estimation will be guaranteed in expectation rather than with high probability; we leave the details for the interested reader.

A different way to enforce the boundedness assumption is to reject any sample points X_i that fall outside the centered ball of radius $\sqrt{m}$. This is equivalent to sampling from the conditional distribution inside the ball. The conditional distribution satisfies the boundedness requirement, so the results discussed above provide a good covariance estimation for it. In many cases, this estimate works even for the original distribution – namely, if only a small part of the distribution lies outside the ball of radius $\sqrt{m}$. We leave the details for the interested reader; see e.g. [81].

5.4.4 Applications to random sub-matrices and sub-frames

The absence of any moment hypotheses on the distribution in Section 5.4.2 (except finite variance) makes these results especially relevant for discrete distributions. One such situation arises when one wishes to sample entries or rows from a given matrix B, thereby creating a *random sub-matrix A*. It is a big program to understand what we can learn about B by seeing A, see [34, 25, 66]. In other words, we ask – what properties of B pass onto A? Here we shall only scratch the surface of this problem: we notice that random sub-matrices of certain size preserve the property of being an *approximate isometry*.

Corollary 5.55 (Random sub-matrices). *Consider an $M \times n$ matrix B such that[21] $s_{\min}(B) = s_{\max}(B) = \sqrt{M}$. Let m be such that all rows B_i of B satisfy $\|B_i\|_2 \leq \sqrt{m}$. Let A be an $N \times n$ matrix obtained by sampling N random rows from B uniformly and independently. Then for every $t \geq 0$, with probability at least $1 - 2n \cdot \exp(-ct^2)$ one has*

$$\sqrt{N} - t\sqrt{m} \leq s_{\min}(A) \leq s_{\max}(A) \leq \sqrt{N} + t\sqrt{m}.$$

Here $c > 0$ is an absolute constant.

---

[21]The first hypothesis says $B^*B = MI$. Equivalently, $\bar{B} := \frac{1}{\sqrt{M}}B$ is an isometry, i.e. $\|\bar{B}x\|_2 = \|x\|_2$ for all x. Equivalently, the columns of $\bar{B}$ are orthonormal.

Proof. By assumption, $I = \frac{1}{M}B^*B = \frac{1}{M}\sum_{i=1}^{M} B_i \otimes B_i$. Therefore, the uniform distribution on the set of the rows $\{B_1, \dots, B_M\}$ is an isotropic distribution in $\mathbb{R}^n$. The conclusion then follows from Theorem 5.41. $\square$

Note that the conclusion of Corollary 5.55 does not depend on the dimension M of the ambient matrix B. This happens because this result is a specific version of sampling from a discrete isotropic distribution (uniform on the rows of B), where size M of the support of the distribution is irrelevant.

The hypothesis of Corollary 5.55 implies[22] that $\frac{1}{M}\sum_{i=1}^{M}\|B_i\|_2^2 = n$. Hence by Markov's inequality, most of the rows B_i satisfy $\|B_i\|_2 = O(\sqrt{n})$. This indicates that Corollary 5.55 would be often used with $m = O(n)$. Also, to ensure a positive probability of success, the useful magnitude of t would be $t \sim \sqrt{\log n}$. With this in mind, the extremal singular values of A will be close to each other (and to $\sqrt{N}$) if $N \gg t^2 m \sim n \log n$.

Summarizing, Corollary 5.55 states that a random $O(n \log n) \times n$ sub-matrix of an $M \times n$ isometry is an approximate isometry.[23]

Another application of random matrices with heavy-tailed isotropic rows is for *sampling from frames*. Recall that frames are generalizations of bases without linear independence, see Example 5.21. Consider a tight frame $\{u_i\}_{i=1}^{M}$ in $\mathbb{R}^n$, and for the sake of convenient normalization, assume that it has bounds $A = B = M$. We are interested in whether a small random subset of $\{u_i\}_{i=1}^{M}$ is still a nice frame in $\mathbb{R}^n$. Such question arises naturally because frames are used in signal processing to create *redundant representations* of signals. Indeed, every signal $x \in \mathbb{R}^n$ admits frame expansion $x = \frac{1}{M}\sum_{i=1}^{M}\langle u_i, x\rangle u_i$. Redundancy makes frame representations more robust to errors and losses than basis representations. Indeed, we will show that if one loses all except $N = O(n \log n)$ random coefficients $\langle u_i, x\rangle$ one is still able to reconstruct x from the received coefficients $\langle u_{i_k}, x\rangle$ as $x \approx \frac{1}{N}\sum_{k=1}^{N}\langle u_{i_k}, x\rangle u_{i_k}$. This boils down to showing that a random subset of size $N = O(n \log n)$ of a tight frame in $\mathbb{R}^n$ is an approximate tight frame.

Corollary 5.56 (Random sub-frames, see [80])**.** *Consider a tight frame $\{u_i\}_{i=1}^{M}$ in $\mathbb{R}^n$ with frame bounds $A = B = M$. Let number m be such that all frame elements satisfy $\|u_i\|_2 \le \sqrt{m}$. Let $\{v_i\}_{i=1}^{N}$ be a set of vectors obtained by sampling N random elements from the frame $\{u_i\}_{i=1}^{M}$ uniformly and independently. Let $\varepsilon \in (0,1)$ and $t \ge 1$. Then the following holds with probability at least $1 - 2n^{-t^2}$:*

$$\text{If } N \ge C(t/\varepsilon)^2 m \log n \quad \text{then } \{v_i\}_{i=1}^{N} \text{ is a frame in } \mathbb{R}^n$$

with bounds $A = (1 - \varepsilon)N$, $B = (1 + \varepsilon)N$. Here C is an absolute constant.

In particular, if this event holds, then every $x \in \mathbb{R}^n$ admits an approximate representation using only the sampled frame elements:

$$\Big\| \frac{1}{N}\sum_{i=1}^{N}\langle v_i, x\rangle v_i - x \Big\| \le \varepsilon \|x\|.$$

---

[22]To recall why this is true, take trace of both sides in the identity $I = \frac{1}{M}\sum_{i=1}^{M} B_i \otimes B_i$.

[23]For the purposes of compressed sensing, we shall study the more difficult *uniform* problem for random sub-matrices in Section 5.6. There B itself will be chosen as a column sub-matrix of a given $M \times M$ matrix (such as DFT), and one will need to control all such B simultaneously, see Example 5.73.

Proof. The assumption implies that $I = \frac{1}{M} \sum_{i=1}^{M} u_i \otimes u_i$. Therefore, the uniform distribution on the set $\{u_i\}_{i=1}^{M}$ is an isotropic distribution in $\mathbb{R}^n$. Applying Corollary 5.52 with $\Sigma = I$ and $\Sigma_N = \frac{1}{N} \sum_{i=1}^{N} v_i \otimes v_i$ we conclude that $\|\Sigma_N - I\| \le \varepsilon$ with the required probability. This clearly completes the proof. $\qquad\square$

As before, we note that $\frac{1}{M} \sum_{i=1}^{M} \|u_i\|_2^2 = n$, so Corollary 5.56 would be often used with $m = O(n)$. This shows, liberally speaking, that a random subset of a frame in $\mathbb{R}^n$ of size $N = O(n \log n)$ is again a frame.

Remark 5.57 (Non-uniform sampling). The boundedness assumption $\|u_i\|_2 \le \sqrt{m}$, although needed in Corollary 5.56, can be removed by non-uniform sampling. To this end, one would sample from the set of normalized vectors $\bar{u}_i := \sqrt{n} \frac{u_i}{\|u_i\|_2}$ with probabilities proportional to $\|u_i\|_2^2$. This defines an isotropic distribution in $\mathbb{R}^n$, and clearly $\|\bar{u}_i\|_2 = \sqrt{n}$. Therefore, by Theorem 5.56, a random sample of $N = O(n \log n)$ vectors obtained this way forms an almost tight frame in $\mathbb{R}^n$. This result does not require any bound on $\|u_i\|_2$.

5.5 Random matrices with independent columns

In this section we study the extreme singular values of $N \times n$ random matrices A with independent columns A_j. We are guided by our ideal bounds (5.2) as before. The same phenomenon occurs in the column independent model as in the row independent model – sufficiently tall random matrices A are approximate isometries. As before, being tall will mean $N \gg n$ for sub-gaussian distributions and $N \gg n \log n$ for arbitrary distributions.

The problem is equivalent to studying *Gram matrices* $G = A^*A = (\langle A_j, A_k \rangle)_{j,k=1}^{n}$ of independent isotropic random vectors $A_1, \ldots, A_n$ in $\mathbb{R}^N$. Our results can be interpreted using Lemma 5.36 as showing that the normalized Gram matrix $\frac{1}{N}G$ is an *approximate identity* for N, n as above.

Let us first try to prove this with a heuristic argument. By Lemma 5.20 we know that the diagonal entries of $\frac{1}{N}G$ have mean $\frac{1}{N}\mathbb{E}\|A_j\|_2^2 = 1$ and off-diagonal ones have zero mean and standard deviation $\frac{1}{N}(\mathbb{E}\langle A_j, A_k \rangle^2)^{1/2} = \frac{1}{\sqrt{N}}$. If, hypothetically, the off-diagonal entries were independent, then we could use the results of matrices with independent entries (or even rows) developed in Section 5.4. The off-diagonal part of $\frac{1}{N}G$ would have norm $O(\sqrt{\frac{n}{N}})$ while the diagonal part would approximately equal I. Hence we would have

$$\left\|\frac{1}{N}G - I\right\| = O\left(\sqrt{\frac{n}{N}}\right), \tag{5.35}$$

i.e. $\frac{1}{N}G$ is an approximate identity for $N \gg n$. Equivalently, by Lemma 5.36, (5.35) would yield the ideal bounds (5.2) on the extreme singular values of A.

Unfortunately, the entries of the Gram matrix G are obviously not independent. To overcome this obstacle we shall use the *decoupling* technique of probability theory [22]. We observe that there is still enough independence encoded in G. Consider a principal sub-matrix $(A_S)^*(A_T)$ of $G = A^*A$ with disjoint index sets S and T. If we condition on $(A_k)_{k \in T}$ then this sub-matrix has independent rows. Using an elementary decoupling

technique, we will indeed seek to replace the full Gram matrix G by one such decoupled $S \times T$ matrix with independent rows, and finish off by applying results of Section 5.4.

By transposition one can try to reduce our problem to studying the $n \times N$ matrix A^*. It has independent rows and the same singular values as A, so one can apply results of Section 5.4. The conclusion would be that, with high probability,

$$\sqrt{n} - C\sqrt{N} \le s_{\min}(A) \le s_{\max}(A) \le \sqrt{n} + C\sqrt{N}.$$

Such estimate is only good for *flat* matrices ($N \le n$). For *tall* matrices ($N \ge n$) the lower bound would be trivial because of the (possibly large) constant C. So, from now on we can focus on tall matrices ($N \ge n$) with independent columns.

5.5.1 Sub-gaussian columns

Here we prove a version of Theorem 5.39 for matrices with independent columns.

Theorem 5.58 (Sub-gaussian columns). *Let A be an $N \times n$ matrix ($N \ge n$) whose columns A_i are independent sub-gaussian isotropic random vectors in $\mathbb{R}^N$ with $\|A_j\|_2 = \sqrt{N}$ a. s. Then for every $t \ge 0$, the inequality holds*

$$\sqrt{N} - C\sqrt{n} - t \le s_{\min}(A) \le s_{\max}(A) \le \sqrt{N} + C\sqrt{n} + t \qquad (5.36)$$

with probability at least $1 - 2\exp(-ct^2)$, where $C = C'_K$, $c = c'_K > 0$ depend only on the subgaussian norm $K = \max_j \|A_j\|_{\psi_2}$ of the columns.

The only significant difference between Theorem 5.39 for independent rows and Theorem 5.58 for independent columns is that the latter requires *normalization of columns*, $\|A_j\|_2 = \sqrt{N}$ almost surely. Recall that by isotropy of A_j (see Lemma 5.20) one always has $(\mathbb{E}\|A_j\|_2^2)^{1/2} = \sqrt{N}$, but the normalization is a bit stronger requirement. We will discuss this more after the proof of Theorem 5.58.

Remark 5.59 (Gram matrices are an approximate identity). By Lemma 5.36, the conclusion of Theorem 5.58 is equivalent to

$$\Big\| \frac{1}{N} A^* A - I \Big\| \le C\sqrt{\frac{n}{N}} + \frac{t}{\sqrt{N}}$$

with the same probability $1 - 2\exp(-ct^2)$. This establishes our ideal inequality (5.35). In words, the normalized Gram matrix of n independent sub-gaussian isotropic random vectors in $\mathbb{R}^N$ is an approximate identity whenever $N \gg n$.

The proof of Theorem 5.58 is based on the decoupling technique [22]. What we will need here is an elementary decoupling lemma for double arrays. Its statement involves the notion of a *random subset* of a given finite set. To be specific, we define a random set T of $[n]$ with a given average size $m \in [0, n]$ as follows. Consider independent $\{0, 1\}$ valued random variables $\delta_1, \ldots, \delta_n$ with $\mathbb{E}\delta_i = m/n$; these are sometimes called *independent selectors*. Then we define the random subset $T = \{i \in [n] : \delta_i = 1\}$. Its average size equals $\mathbb{E}|T| = \mathbb{E}\sum_{i=1}^{n} \delta_i = m$.

Lemma 5.60 (Decoupling). *Consider a double array of real numbers $(a_{ij})_{i,j=1}^n$ such that $a_{ii} = 0$ for all i. Then*

$$\sum_{i,j\in[n]} a_{ij} = 4\mathbb{E} \sum_{i\in T,\, j\in T^c} a_{ij}$$

where T is a random subset of $[n]$ with average size $n/2$. In particular,

$$4 \min_{T\subseteq[n]} \sum_{i\in T,\, j\in T^c} a_{ij} \leq \sum_{i,j\in[n]} a_{ij} \leq 4 \max_{T\subseteq[n]} \sum_{i\in T,\, j\in T^c} a_{ij}$$

where the minimum and maximum are over all subsets T of $[n]$.

Proof. Expressing the random subset as $T = \{i \in [n]: \delta_i = 1\}$ where δ_i are independent selectors with $\mathbb{E}\delta_i = 1/2$, we see that

$$\mathbb{E} \sum_{i\in T,\, j\in T^c} a_{ij} = \mathbb{E} \sum_{i,j\in[n]} \delta_i(1-\delta_j)a_{ij} = \frac{1}{4} \sum_{i,j\in[n]} a_{ij},$$

where we used that $\mathbb{E}\delta_i(1-\delta_j) = \frac{1}{4}$ for $i \neq j$ and the assumption $a_{ii} = 0$. This proves the first part of the lemma. The second part follows trivially by estimating expectation by maximum and minimum. $\square$

Proof of Theorem 5.58. **Step 1: Reductions.** Without loss of generality we can assume that the columns A_i have zero mean. Indeed, multiplying each column A_i by ± 1 arbitrarily preserves the extreme singular values of A, the isotropy of A_i and the sub-gaussian norms of A_i. Therefore, by multiplying A_i by independent symmetric Bernoulli random variables we achieve that A_i have zero mean.

For $t = O(\sqrt{N})$ the conclusion of Theorem 5.58 follows from Theorem 5.39 by transposition. Indeed, the $n \times N$ random matrix A^* has independent rows, so for $t \geq 0$ we have

$$s_{\max}(A) = s_{\max}(A^*) \leq \sqrt{n} + C_K\sqrt{N} + t \tag{5.37}$$

with probability at least $1 - 2\exp(-c_K t^2)$. Here $c_K > 0$ and we can obviously assume that $C_K \geq 1$. For $t \geq C_K\sqrt{N}$ it follows that $s_{\max}(A) \leq \sqrt{N} + \sqrt{n} + 2t$, which yields the conclusion of Theorem 5.58 (the left hand side of (5.36) being trivial). So, it suffices to prove the conclusion for $t \leq C_K\sqrt{N}$. Let us fix such t.

It would be useful to have some a priori control of $s_{\max}(A) = \|A\|$. We thus consider the desired event

$$\mathcal{E} := \big\{s_{\max}(A) \leq 3C_K\sqrt{N}\big\}.$$

Since $3C_K\sqrt{N} \geq \sqrt{n} + C_K\sqrt{N} + t$, by (5.37) we see that $\mathcal{E}$ is likely to occur:

$$\mathbb{P}(\mathcal{E}^c) \leq 2\exp(-c_K t^2). \tag{5.38}$$

Step 2: Approximation. This step is parallel to Step 1 in the proof of Theorem 5.39, except now we shall choose $\varepsilon := \delta$. This way we reduce our task to the

following. Let $\mathcal{N}$ be a $\frac{1}{4}$-net of the unit sphere S^{n-1} such that $|\mathcal{N}| \leq 9^n$. It suffices to show that with probability at least $1 - 2\exp(-c_K' t^2)$ one has

$$\max_{x \in \mathcal{N}} \left| \frac{1}{N} \|Ax\|_2^2 - 1 \right| \leq \frac{\delta}{2}, \quad \text{where } \delta = C\sqrt{\frac{n}{N}} + \frac{t}{\sqrt{N}}.$$

By (5.38), it is enough to show that the probability

$$p := \mathbb{P}\left\{ \max_{x \in \mathcal{N}} \left| \frac{1}{N} \|Ax\|_2^2 - 1 \right| > \frac{\delta}{2} \text{ and } \mathcal{E} \right\} \tag{5.39}$$

satisfies $p \leq 2\exp(-c_K'' t^2)$, where $c_K'' > 0$ may depend only on K.

Step 3: Decoupling. As in the proof of Theorem 5.39, we will obtain the required bound for a fixed $x \in \mathcal{N}$ with high probability, and then take a union bound over x. So let us fix any $x = (x_1, \ldots, x_n) \in S^{n-1}$. We expand

$$\|Ax\|_2^2 = \left\| \sum_{j=1}^n x_j A_j \right\|_2^2 = \sum_{j=1}^n x_j^2 \|A_j\|_2^2 + \sum_{j,k \in [n],\, j \neq k} x_j x_k \langle A_j, A_k \rangle. \tag{5.40}$$

Since $\|A_j\|_2^2 = N$ by assumption and $\|x\|_2 = 1$, the first sum equals N. Therefore, subtracting N from both sides and dividing by N, we obtain the bound

$$\left| \frac{1}{N} \|Ax\|_2^2 - 1 \right| \leq \left| \frac{1}{N} \sum_{j,k \in [n],\, j \neq k} x_j x_k \langle A_j, A_k \rangle \right|.$$

The sum in the right hand side is $\langle G_0 x, x \rangle$ where G_0 is the off-diagonal part of the Gram matrix $G = A^* A$. As we indicated in the beginning of Section 5.5, we are going to replace G_0 by its decoupled version whose rows and columns are indexed by disjoint sets. This is achieved by Decoupling Lemma 5.60: we obtain

$$\left| \frac{1}{N} \|Ax\|_2^2 - 1 \right| \leq \frac{4}{N} \max_{T \subseteq [n]} |R_T(x)|, \quad \text{where } R_T(x) = \sum_{j \in T,\, k \in T^c} x_j x_k \langle A_j, A_k \rangle.$$

We substitute this into (5.39) and take union bound over all choices of $x \in \mathcal{N}$ and $T \subseteq [n]$. As we know, $|\mathcal{N}| \leq 9^n$, and there are 2^n subsets T in $[n]$. This gives

$$p \leq \mathbb{P}\left\{ \max_{x \in \mathcal{N},\, T \subseteq [n]} |R_T(x)| > \frac{\delta N}{8} \text{ and } \mathcal{E} \right\}$$

$$\leq 9^n \cdot 2^n \cdot \max_{x \in \mathcal{N},\, T \subseteq [n]} \mathbb{P}\left\{ |R_T(x)| > \frac{\delta N}{8} \text{ and } \mathcal{E} \right\}. \tag{5.41}$$

Step 4: Conditioning and concentration. To estimate the probability in (5.41), we fix a vector $x \in \mathcal{N}$ and a subset $T \subseteq [n]$ and we condition on a realization of random vectors $(A_k)_{k \in T^c}$. We express

$$R_T(x) = \sum_{j \in T} x_j \langle A_j, z \rangle \quad \text{where } z = \sum_{k \in T^c} x_k A_k. \tag{5.42}$$

Under our conditioning z is a fixed vector, so $R_T(x)$ is a sum of independent random variables. Moreover, if event $\mathcal{E}$ holds then z is nicely bounded:

$$\|z\|_2 \le \|A\|\|x\|_2 \le 3C_K\sqrt{N}. \tag{5.43}$$

If in turn (5.43) holds then the terms $\langle A_j, z\rangle$ in (5.42) are independent centered sub-gaussian random variables with $\|\langle A_j, z\rangle\|_{\psi_2} \le 3KC_K\sqrt{N}$. By Lemma 5.9, their linear combination $R_T(x)$ is also a sub-gaussian random variable with

$$\|R_T(x)\|_{\psi_2} \le C_1\Big(\sum_{j\in T} x_j^2\|\langle A_j, z\rangle\|_{\psi_2}^2\Big)^{1/2} \le \widehat{C}_K\sqrt{N} \tag{5.44}$$

where $\widehat{C}_K$ depends only on K.

We can summarize these observations as follows. Denoting the conditional probability by $\mathbb{P}_T = \mathbb{P}\{\,\cdot\,|(A_k)_{k\in T^c}\}$ and the expectation with respect to $(A_k)_{k\in T^c}$ by $\mathbb{E}_{T^c}$, we obtain by (5.43) and (5.44) that

$$\mathbb{P}\Big\{|R_T(x)| > \frac{\delta N}{8} \text{ and } \mathcal{E}\Big\} \le \mathbb{E}_{T^c}\mathbb{P}_T\Big\{|R_T(x)| > \frac{\delta N}{8} \text{ and } \|z\|_2 \le 3C_K\sqrt{N}\Big\}$$

$$\le 2\exp\Big[-c_1\Big(\frac{\delta N/8}{\widehat{C}_K\sqrt{N}}\Big)^2\Big] = 2\exp\Big(-\frac{c_2\delta^2 N}{\widehat{C}_K^2}\Big) \le 2\exp\Big(-\frac{c_2 C^2 n}{\widehat{C}_K^2} - \frac{c_2 t^2}{\widehat{C}_K^2}\Big).$$

The second inequality follows because $R_T(x)$ is a sub-gaussian random variable (5.44) whose tail decay is given by (5.10). Here $c_1, c_2 > 0$ are absolute constants. The last inequality follows from the definition of δ. Substituting this into (5.41) and choosing C sufficiently large (so that $\ln 36 \le c_2 C^2/\widehat{C}_K^2$), we conclude that

$$p \le 2\exp\big(-c_2 t^2/\widehat{C}_K^2\big).$$

This proves an estimate that we desired in Step 2. The proof is complete. $\qquad\square$

Remark 5.61 (Normalization assumption). Some a priori control of the norms of the columns $\|A_j\|_2$ is necessary for estimating the extreme singular values, since

$$s_{\min}(A) \le \min_{i\le n}\|A_j\|_2 \le \max_{i\le n}\|A_j\|_2 \le s_{\max}(A).$$

With this in mind, it is easy to construct an example showing that a normalization assumption $\|A_i\|_2 = \sqrt{N}$ is essential in Theorem 5.58; it can not even be replaced by a boundedness assumption $\|A_i\|_2 = O(\sqrt{N})$.

Indeed, consider a random vector $X = \sqrt{2}\xi Y$ in $\mathbb{R}^N$ where ξ is a $\{0,1\}$-valued random variable with $\mathbb{E}\xi = 1/2$ (a "selector") and X is an independent spherical random vector in $\mathbb{R}^n$ (see Example 5.25). Let A be a random matrix whose columns A_j are independent copies of X. Then A_j are independent centered sub-gaussian isotropic random vectors in $\mathbb{R}^n$ with $\|A_j\|_{\psi_2} = O(1)$ and $\|A_j\|_2 \le \sqrt{2N}$ a.s. So all assumptions of Theorem 5.58 except normalization are satisfied. On the other hand $\mathbb{P}\{X = 0\} = 1/2$, so matrix A has a zero column with overwhelming probability $1 - 2^{-n}$. This implies that $s_{\min}(A) = 0$ with this probability, so the lower estimate in (5.36) is false for all nontrivial N, n, t.

5.5.2 Heavy-tailed columns

Here we prove a version of Theorem 5.45 for independent heavy-tailed columns.

We thus consider $N \times n$ random matrices A with independent columns A_j. In addition to the normalization assumption $\|A_j\|_2 = \sqrt{N}$ already present in Theorem 5.58 for subgaussian columns, our new result must also require an a priori control of the off-diagonal part of the Gram matrix $G = A^*A = (\langle A_j, A_k \rangle)_{j,k=1}^n$.

Theorem 5.62 (Heavy-tailed columns). *Let A be an $N \times n$ matrix $(N \geq n)$ whose columns A_j are independent isotropic random vectors in $\mathbb{R}^N$ with $\|A_j\|_2 = \sqrt{N}$ a. s. Consider the incoherence parameter*

$$m := \frac{1}{N} \mathbb{E} \max_{j \leq n} \sum_{k \in [n],\, k \neq j} \langle A_j, A_k \rangle^2.$$

Then $\mathbb{E}\big\| \frac{1}{N} A^*A - I \big\| \leq C_0 \sqrt{\frac{m \log n}{N}}$. In particular,

$$\mathbb{E} \max_{j \leq n} |s_j(A) - \sqrt{N}| \leq C\sqrt{m \log n}. \tag{5.45}$$

Let us briefly clarify the role of the incoherence parameter m, which controls the lengths of the rows of the off-diagonal part of G. After the proof we will see that a control of m is essential in Theorem 5.41. But for now, let us get a feel of the typical size of m. We have $\mathbb{E}\langle A_j, A_k \rangle^2 = N$ by Lemma 5.20, so for every row j we see that $\frac{1}{N} \sum_{k \in [n],\, k \neq j} \langle A_j, A_k \rangle^2 = n - 1$. This indicates that Theorem 5.62 would be often used with $m = O(n)$.

In this case, Theorem 5.41 establishes our ideal inequality (5.35) up to a logarithmic factor. In words, the normalized Gram matrix of n independent isotropic random vectors in $\mathbb{R}^N$ is an approximate identity whenever $N \gg n \log n$.

Our proof of Theorem 5.62 will be based on decoupling, symmetrization and an application of Theorem 5.48 for a decoupled Gram matrix with independent rows. The decoupling is done similarly to Theorem 5.58. However, this time we will benefit from formalizing the decoupling inequality for Gram matrices:

Lemma 5.63 (Matrix decoupling). *Let B be a $N \times n$ random matrix whose columns B_j satisfy $\|B_j\|_2 = 1$. Then*

$$\mathbb{E}\|B^*B - I\| \leq 4 \max_{T \subseteq [n]} \mathbb{E}\|(B_T)^* B_{T^c}\|.$$

Proof. We first note that $\|B^*B - I\| = \sup_{x \in S^{n-1}} \big| \|Bx\|_2^2 - 1 \big|$. We fix $x = (x_1, \ldots, x_n) \in S^{n-1}$ and, expanding as in (5.40), observe that

$$\|Bx\|_2^2 = \sum_{j=1}^n x_j^2 \|B_j\|_2^2 + \sum_{j,k \in [n],\, j \neq k} x_j x_k \langle B_j, B_k \rangle.$$

The first sum equals 1 since $\|B_j\|_2 = \|x\|_2 = 1$. So by Decoupling Lemma 5.60, a random subset T of $[n]$ with average cardinality $n/2$ satisfies

$$\|Bx\|_2^2 - 1 = 4\mathbb{E}_T \sum_{j \in T, k \in T^c} x_j x_k \langle B_j, B_k \rangle.$$

Let us denote by $\mathbb{E}_T$ and $\mathbb{E}_B$ the expectations with respect to the random set T and the random matrix B respectively. Using Jensen's inequality we obtain

$$\mathbb{E}_B \|B^* B - I\| = \mathbb{E}_B \sup_{x \in S^{n-1}} \left| \|Bx\|_2^2 - 1 \right|$$

$$\leq 4\mathbb{E}_B \mathbb{E}_T \sup_{x \in S^{n-1}} \left| \sum_{j \in T, k \in T^c} x_j x_k \langle B_j, B_k \rangle \right| = 4\mathbb{E}_T \mathbb{E}_B \|(B_T)^* B_{T^c}\|.$$

The conclusion follows by replacing the expectation by the maximum over T. $\qquad\square$

Proof of Theorem 5.62. **Step 1: Reductions and decoupling.** It would be useful to have an a priori bound on $s_{\max}(A) = \|A\|$. We can obtain this by transposing A and applying one of the results of Section 5.4. Indeed, the random $n \times N$ matrix A^* has independent rows A_i^* which by our assumption are normalized as $\|A_i^*\|_2 = \|A_i\|_2 = \sqrt{N}$. Applying Theorem 5.45 with the roles of n and N switched, we obtain by the triangle inequality that

$$\mathbb{E}\|A\| = \mathbb{E}\|A^*\| = \mathbb{E} s_{\max}(A^*) \leq \sqrt{n} + C\sqrt{N \log n} \leq C\sqrt{N \log n}. \qquad (5.46)$$

Observe that $n \leq m$ since by Lemma 5.20 we have $\frac{1}{N}\mathbb{E}\langle A_j, A_k \rangle^2 = 1$ for $j \neq k$. We use Matrix Decoupling Lemma 5.63 for $B = \frac{1}{\sqrt{N}} A$ and obtain

$$E \leq \frac{4}{N} \max_{T \subseteq [n]} \mathbb{E}\|(A_T)^* A_{T^c}\| = \frac{4}{N} \max_{T \subseteq [n]} \mathbb{E}\|\Gamma\| \qquad (5.47)$$

where $\Gamma = \Gamma(T)$ denotes the decoupled Gram matrix

$$\Gamma = (A_T)^* A_{T^c} = \big(\langle A_j, A_k \rangle\big)_{j \in T, k \in T^c}.$$

Let us fix T; our problem then reduces to bounding the expected norm of Γ.

Step 2: The rows of the decoupled Gram matrix. For a subset $S \subseteq [n]$, we denote by $\mathbb{E}_{A_S}$ the conditional expectation given A_{S^c}, i.e. with respect to $A_S = (A_j)_{j \in S}$. Hence $\mathbb{E} = \mathbb{E}_{A_{T^c}} \mathbb{E}_{A_T}$.

Let us condition on A_{T^c}. Treating $(A_k)_{k \in T^c}$ as fixed vectors we see that, conditionally, the random matrix Γ has independent rows

$$\Gamma_j = \big(\langle A_j, A_k \rangle\big)_{k \in T^c}, \quad j \in T.$$

So we are going to use Theorem 5.48 to bound the norm of Γ. To do this we need estimates on (a) the norms and (b) the second moment matrices of the rows Γ_j.

(a) Since for $j \in T$, Γ_j is a random vector valued in $\mathbb{R}^{T^c}$, we estimate its second moment matrix by choosing $x \in \mathbb{R}^{T^c}$ and evaluating the scalar second moment

$$\mathbb{E}_{A_T} \langle \Gamma_j, x \rangle^2 = \mathbb{E}_{A_T} \Big(\sum_{k \in T^c} \langle A_j, A_k \rangle x_k \Big)^2 = \mathbb{E}_{A_T} \Big\langle A_j, \sum_{k \in T^c} x_k A_k \Big\rangle^2$$

$$= \Big\| \sum_{k \in T^c} x_k A_k \Big\|^2 = \|A_{T^c} x\|_2^2 \leq \|A_{T^c}\|_2^2 \|x\|_2^2.$$

In the third equality we used isotropy of A_j. Taking maximum over all $j \in T$ and $x \in \mathbb{R}^{T^c}$, we see that the second moment matrix $\Sigma(\Gamma_j) = \mathbb{E}_{A_T} \Gamma_j \otimes \Gamma_j$ satisfies

$$\max_{j \in T} \|\Sigma(\Gamma_j)\| \leq \|A_{T^c}\|^2. \tag{5.48}$$

(b) To evaluate the norms of Γ_j, $j \in T$, note that $\|\Gamma_j\|_2^2 = \sum_{k \in T^c} \langle A_j, A_k \rangle^2$. This is easy to bound, because the assumption says that the random variable

$$M := \frac{1}{N} \max_{j \in [n]} \sum_{k \in [n], \, k \neq j} \langle A_j, A_k \rangle^2 \quad \text{satisfies } \mathbb{E}M = m.$$

This produces the bound $\mathbb{E} \max_{j \in T} \|\Gamma_j\|_2^2 \leq N \cdot \mathbb{E}M = Nm$. But at this moment we need to work conditionally on A_{T^c}, so for now we will be satisfied with

$$\mathbb{E}_{A_T} \max_{j \in T} \|\Gamma_j\|_2^2 \leq N \cdot \mathbb{E}_{A_T} M. \tag{5.49}$$

Step 3: The norm of the decoupled Gram matrix. We bound the norm of the random $T \times T^c$ Gram matrix Γ with (conditionally) independent rows using Theorem 5.48 and Remark 5.49. Since by (5.48) we have $\big\| \frac{1}{|T|} \sum_{j \in T} \Sigma(\Gamma_j) \big\| \leq \frac{1}{|T|} \sum_{j \in T} \|\Sigma(\Gamma_j)\| \leq \|A_{T^c}\|^2$, we obtain using (5.49) that

$$\mathbb{E}_{A_T} \|\Gamma\| \leq (\mathbb{E}_{A_T} \|\Gamma\|^2)^{1/2} \leq \|A_{T^c}\| \sqrt{|T|} + C\sqrt{N \cdot \mathbb{E}_{A_T}(M) \log |T^c|}$$
$$\leq \|A_{T^c}\| \sqrt{n} + C\sqrt{N \cdot \mathbb{E}_{A_T}(M) \log n}. \tag{5.50}$$

Let us take expectation of both sides with respect to A_{T^c}. The left side becomes the quantity we seek to bound, $\mathbb{E}\|\Gamma\|$. The right side will contain the term which we can estimate by (5.46):

$$\mathbb{E}_{A_{T^c}} \|A_{T^c}\| = \mathbb{E}\|A_{T^c}\| \leq \mathbb{E}\|A\| \leq C\sqrt{N \log n}.$$

The other term that will appear in the expectation of (5.50) is

$$\mathbb{E}_{A_{T^c}} \sqrt{\mathbb{E}_{A_T}(M)} \leq \sqrt{\mathbb{E}_{A_{T^c}} \mathbb{E}_{A_T}(M)} \leq \sqrt{\mathbb{E}M} = \sqrt{m}.$$

So, taking the expectation in (5.50) and using these bounds, we obtain

$$\mathbb{E}\|\Gamma\| = \mathbb{E}_{A_{T^c}} \mathbb{E}_{A_T} \|\Gamma\| \leq C\sqrt{N \log n}\sqrt{n} + C\sqrt{Nm \log n} \leq 2C\sqrt{Nm \log n}$$

where we used that $n \leq m$. Finally, using this estimate in (5.47) we conclude

$$E \leq 8C\sqrt{\frac{m \log n}{N}}.$$

This establishes the first part of Theorem 5.62. The second part follow by the diagonalization argument as in Step 2 of the proof of Theorem 5.45. $\square$

Remark 5.64 (Incoherence). A priori control on the *incoherence* is essential in Theorem 5.62. Consider for instance an $N \times n$ random matrix A whose columns are independent coordinate random vectors in $\mathbb{R}^N$. Clearly $s_{\max}(A) \geq \max_j \|A_i\|_2 = \sqrt{N}$. On the other hand, if the matrix is not too tall, $n \gg \sqrt{N}$, then A has two identical columns with high probability, which yields $s_{\min}(A) = 0$.

5.6 Restricted isometries

In this section we consider an application of the non-asymptotic random matrix theory in compressed sensing. For a thorough introduction to compressed sensing, see the introductory chapter of this book and [28, 20].

In this area, $m \times n$ matrices A are considered as measurement devices, taking as input a signal $x \in \mathbb{R}^n$ and returning its measurement $y = Ax \in \mathbb{R}^m$. One would like to take measurements economically, thus keeping m as small as possible, and still to be able to recover the signal x from its measurement y.

The interesting regime for compressed sensing is where we take very few measurements, $m \ll n$. Such matrices A are not one-to-one, so recovery of x from y is not possible for all signals x. But in practical applications, the amount of "information" contained in the signal is often small. Mathematically this is expressed as *sparsity* of x. In the simplest case, one assumes that x has few non-zero coordinates, say $|\operatorname{supp}(x)| \leq k \ll n$. In this case, using any non-degenerate matrix A one can check that x can be recovered whenever $m > 2k$ using the optimization problem $\min\{|\operatorname{supp}(x)| : Ax = y\}$.

This optimization problem is highly non-convex and generally NP-complete. So instead one considers a convex relaxation of this problem, $\min\{\|x\|_1 : Ax = y\}$. A basic result in compressed sensing, due to Candès and Tao [17, 16], is that for sparse signals $|\operatorname{supp}(x)| \leq k$, the convex problem recovers the signal x from its measurement y exactly, provided that the measurement matrix A is quantitatively non-degenerate. Precisely, the non-degeneracy of A means that it satisfies the following *restricted isometry property* with $\delta_{2k}(A) \leq 0.1$.

Definition (Restricted isometries). *An $m \times n$ matrix A satisfies the* restricted isometry property *of order $k \geq 1$ if there exists $\delta_k \geq 0$ such that the inequality*

$$(1 - \delta_k)\|x\|_2^2 \leq \|Ax\|_2^2 \leq (1 + \delta_k)\|x\|_2^2 \tag{5.51}$$

holds for all $x \in \mathbb{R}^n$ with $|\operatorname{supp}(x)| \leq k$. The smallest number $\delta_k = \delta_k(A)$ is called the restricted isometry constant *of A.*

In words, A has a restricted isometry property if A acts as an approximate isometry on all sparse vectors. Clearly,

$$\delta_k(A) = \max_{|T| \le k} \|A_T^* A_T - I_{\mathbb{R}^T}\| = \max_{|T| = \lfloor k \rfloor} \|A_T^* A_T - I_{\mathbb{R}^T}\| \tag{5.52}$$

where the maximum is over all subsets $T \subseteq [n]$ with $|T| \le k$ or $|T| = \lfloor k \rfloor$.

The concept of restricted isometry can also be expressed via extreme singular values, which brings us to the topic we studied in the previous sections. A is a restricted isometry if and only if all $m \times k$ sub-matrices A_T of A (obtained by selecting arbitrary k columns from A) are approximate isometries. Indeed, for every $\delta \ge 0$, Lemma 5.36 shows that the following two inequalities are equivalent up to an absolute constant:

$$\delta_k(A) \le \max(\delta, \delta^2); \tag{5.53}$$
$$1 - \delta \le s_{\min}(A_T) \le s_{\max}(A_T) \le 1 + \delta \quad \text{for all } |T| \le k. \tag{5.54}$$

More precisely, (5.53) implies (5.54) and (5.54) implies $\delta_k(A) \le 3\max(\delta, \delta^2)$.

Our goal is thus to find matrices that are good restricted isometries. What good means is clear from the goals of compressed sensing described above. First, we need to keep the restricted isometry constant $\delta_k(A)$ below some small absolute constant, say 0.1. Most importantly, we would like the number of measurements m to be small, ideally proportional to the sparsity $k \ll n$.

This is where non-asymptotic random matrix theory enters. We shall indeed show that, with high probability, $m \times n$ random matrices A are good restricted isometries of order k with $m = O^*(k)$. Here the O^* notation hides some logarithmic factors of n. Specifically, in Theorem 5.65 we will show that

$$m = O(k \log(n/k))$$

for sub-gaussian random matrices A (with independent rows or columns). This is due to the strong concentration properties of such matrices. A general observation of this kind is Proposition 5.66. It says that if for a given x, a random matrix A (taken from any distribution) satisfies inequality (5.51) with high probability, then A is a good restricted isometry.

In Theorem 5.71 we will extend these results to random matrices without concentration properties. Using a uniform extension of Rudelson's inequality, Corollary 5.28, we shall show that

$$m = O(k \log^4 n) \tag{5.55}$$

for heavy-tailed random matrices A (with independent rows). This includes the important example of random Fourier matrices.

5.6.1 Sub-gaussian restricted isometries

In this section we show that $m \times n$ sub-gaussian random matrices A are good restricted isometries. We have in mind either of the following two models, which we analyzed in Sections 5.4.1 and 5.5.1 respectively:

Row-independent model: the rows of A are independent sub-gaussian isotropic random vectors in $\mathbb{R}^n$;

Column-independent model: the columns A_i of A are independent sub-gaussian isotropic random vectors in $\mathbb{R}^m$ with $\|A_i\|_2 = \sqrt{m}$ a.s.

Recall that these models cover many natural examples, including Gaussian and Bernoulli matrices (whose entries are independent standard normal or symmetric Bernoulli random variables), general sub-gaussian random matrices (whose entries are independent sub-gaussian random variables with mean zero and unit variance), "column spherical" matrices whose columns are independent vectors uniformly distributed on the centered Euclidean sphere in $\mathbb{R}^m$ with radius $\sqrt{m}$, "row spherical" matrices whose rows are independent vectors uniformly distributed on the centered Euclidean sphere in $\mathbb{R}^d$ with radius $\sqrt{d}$, etc.

Theorem 5.65 (Sub-gaussian restricted isometries)**.** *Let A be an $m \times n$ sub-gaussian random matrix with independent rows or columns, which follows either of the two models above. Then the normalized matrix $\bar{A} = \frac{1}{\sqrt{m}} A$ satisfies the following for every sparsity level $1 \le k \le n$ and every number $\delta \in (0, 1)$:*

$$\text{if } m \ge C\delta^{-2} k \log(en/k) \quad \text{then } \delta_k(\bar{A}) \le \delta$$

with probability at least $1 - 2\exp(-c\delta^2 m)$. Here $C = C_K$, $c = c_K > 0$ depend only on the subgaussian norm $K = \max_i \|A_i\|_{\psi_2}$ of the rows or columns of A.

Proof. Let us check that the conclusion follows from Theorem 5.39 for the row-independent model, and from Theorem 5.58 for the column-independent model. We shall control the restricted isometry constant using its equivalent description (5.52). We can clearly assume that k is a positive integer.

Let us fix a subset $T \subseteq [n]$, $|T| = k$ and consider the $m \times k$ random matrix A_T. If A folows the row-independent model, then the rows of A_T are orthogonal projections of the rows of A onto $\mathbb{R}^T$, so they are still independent sub-gaussian isotropic random vectors in $\mathbb{R}^T$. If alternatively, A follows the column-independent model, then trivially the columns of A_T satisfy the same assumptions as the columns of A. In either case, Theorem 5.39 or Theorem 5.58 applies to A_T. Hence for every $s \ge 0$, with probability at least $1 - 2\exp(-cs^2)$ one has

$$\sqrt{m} - C_0\sqrt{k} - s \le s_{\min}(A_T) \le s_{\max}(A_T) \le \sqrt{m} + C_0\sqrt{k} + s. \qquad (5.56)$$

Using Lemma 5.36 for $\bar{A}_T = \frac{1}{\sqrt{m}} A_T$, we see that (5.56) implies that

$$\|\bar{A}_T^* \bar{A}_T - I_{\mathbb{R}^T}\| \le 3\max(\delta_0, \delta_0^2) \quad \text{where } \delta_0 = C_0\sqrt{\frac{k}{m}} + \frac{s}{\sqrt{m}}.$$

Now we take a union bound over all subsets $T \subset [n]$, $|T| = k$. Since there are $\binom{n}{k} \le (en/k)^k$ ways to choose T, we conclude that

$$\max_{|T|=k} \|\bar{A}_T^* \bar{A}_T - I_{\mathbb{R}^T}\| \le 3\max(\delta_0, \delta_0^2)$$

with probability at least $1 - \binom{n}{k} \cdot 2\exp(-cs^2) \geq 1 - 2\exp\big(k\log(en/k) - cs^2\big)$. Then, once we choose $\varepsilon > 0$ arbitrarily and let $s = C_1\sqrt{k\log(en/k)} + \varepsilon\sqrt{m}$, we conclude with probability at least $1 - 2\exp(-c\varepsilon^2 m)$ that

$$\delta_k(\bar{A}) \leq 3\max(\delta_0, \delta_0^2) \quad \text{where } \delta_0 = C_0\sqrt{\frac{k}{m}} + C_1\sqrt{\frac{k\log(en/k)}{m}} + \varepsilon.$$

Finally, we apply this statement for $\varepsilon := \delta/6$. By choosing constant C in the statement of the theorem sufficiently large, we make m large enough so that $\delta_0 \leq \delta/3$, which yields $3\max(\delta_0, \delta_0^2) \leq \delta$. The proof is complete. $\qquad\square$

The main reason Theorem 5.65 holds is that the random matrix A has a strong concentration property, i.e. that $\|\bar{A}x\|_2 \approx \|x\|_2$ with high probability for every fixed sparse vector x. This concentration property alone implies the restricted isometry property, regardless of the specific random matrix model:

Proposition 5.66 (Concentration implies restricted isometry, see [10]). *Let A be an $m \times n$ random matrix, and let $k \geq 1$, $\delta \geq 0$, $\varepsilon > 0$. Assume that for every fixed $x \in \mathbb{R}^n$, $|\operatorname{supp}(x)| \leq k$, the inequality*

$$(1-\delta)\|x\|_2^2 \leq \|Ax\|_2^2 \leq (1+\delta)\|x\|_2^2$$

holds with probability at least $1 - \exp(-\varepsilon m)$. Then we have the following:

$$\text{if } m \geq C\varepsilon^{-1}k\log(en/k) \quad \text{then } \delta_k(\bar{A}) \leq 2\delta$$

with probability at least $1 - \exp(-\varepsilon m/2)$. Here C is an absolute constant.

In words, the restricted isometry property can be checked on each individual vector x with high probability.

Proof. We shall use the expression (5.52) to estimate the restricted isometry constant. We can clearly assume that k is an integer, and focus on the sets $T \subseteq [n]$, $|T| = k$. By Lemma 5.2, we can find a net $\mathcal{N}_T$ of the unit sphere $S^{n-1} \cap \mathbb{R}^T$ with cardinality $|\mathcal{N}_T| \leq 9^k$. By Lemma 5.4, we estimate the operator norm as

$$\big\|A_T^* A_T - I_{\mathbb{R}^T}\big\| \leq 2\max_{x \in \mathcal{N}_T}\big|\big\langle (A_T^* A_T - I_{\mathbb{R}^T})x, x\big\rangle\big| = 2\max_{x \in \mathcal{N}_T}\big|\|Ax\|_2^2 - 1\big|.$$

Taking maximum over all subsets $T \subseteq [n]$, $|T| = k$, we conclude that

$$\delta_k(A) \leq 2\max_{|T|=k}\max_{x \in \mathcal{N}_T}\big|\|Ax\|_2^2 - 1\big|.$$

On the other hand, by assumption we have for every $x \in \mathcal{N}_T$ that

$$\mathbb{P}\big\{\big|\|Ax\|_2^2 - 1\big| > \delta\big\} \leq \exp(-\varepsilon m).$$

Therefore, taking a union bound over $\binom{n}{k} \leq (en/k)^k$ choices of the set T and over 9^k elements $x \in \mathcal{N}_T$, we obtain that

$$\mathbb{P}\{\delta_k(A) > 2\delta\} \leq \binom{n}{k}9^k\exp(-\varepsilon m) \leq \exp\big(k\ln(en/k) + k\ln 9 - \varepsilon m\big)$$

$$\leq \exp(-\varepsilon m/2)$$

where the last line follows by the assumption on m. The proof is complete. $\qquad\square$

5.6.2 Heavy-tailed restricted isometries

In this section we show that $m \times n$ random matrices A with independent heavy-tailed rows (and uniformly bounded coefficients) are good restricted isometries. This result will be established in Theorem 5.71. As before, we will prove this by controlling the extreme singular values of all $m \times k$ sub-matrices A_T. For each individual subset T, this can be achieved using Theorem 5.41: one has

$$\sqrt{m} - t\sqrt{k} \leq s_{\min}(A_T) \leq s_{\max}(A_T) \leq \sqrt{m} + t\sqrt{k} \tag{5.57}$$

with probability at least $1 - 2k \cdot \exp(-ct^2)$. Although this optimal probability estimate has optimal order, it is too weak to allow for a union bound over all $\binom{n}{k} = (O(1)n/k)^k$ choices of the subset T. Indeed, in order that $1 - \binom{n}{k} 2k \cdot \exp(-ct^2) > 0$ one would need to take $t > \sqrt{k \log(n/k)}$. So in order to achieve a nontrivial lower bound in (5.57), one would be forced to take $m \geq k^2$. This is too many measurements; recall that our hope is $m = O^*(k)$.

This observation suggests that instead of controlling each sub-matrix A_T separately, we should learn how to control all A_T at once. This is indeed possible with the following uniform version of Theorem 5.45:

Theorem 5.67 (Heavy-tailed rows; uniform)**.** *Let $A = (a_{ij})$ be an $N \times d$ matrix ($1 < N \leq d$) whose rows A_i are independent isotropic random vectors in $\mathbb{R}^d$. Let K be a number such that all entries $|a_{ij}| \leq K$ almost surely. Then for every $1 < n \leq d$, we have*

$$\mathbb{E} \max_{|T| \leq n} \max_{j \leq |T|} |s_j(A_T) - \sqrt{N}| \leq Cl\sqrt{n}$$

where $l = \log(n)\sqrt{\log d}\sqrt{\log N}$ and where $C = C_K$ may depend on K only. The maximum is, as usual, over all subsets $T \subseteq [d]$, $|T| \leq n$.

The non-uniform prototype of this result, Theorem 5.45, was based on Rudelson's inequality, Corollary 5.28. In a very similar way, Theorem 5.67 is based on the following uniform version of Rudelon's inequality.

Proposition 5.68 (Uniform Rudelson's inequality [67])**.** *Let $x_1, \ldots, x_N$ be vectors in $\mathbb{R}^d$, $1 < N \leq d$, and let K be a number such that all $\|x_i\|_\infty \leq K$. Let $\varepsilon_1, \ldots, \varepsilon_N$ be independent symmetric Bernoulli random variables. Then for every $1 < n \leq d$ one has*

$$\mathbb{E} \max_{|T| \leq n} \Big\| \sum_{i=1}^{N} \varepsilon_i (x_i)_T \otimes (x_i)_T \Big\| \leq Cl\sqrt{n} \cdot \max_{|T| \leq n} \Big\| \sum_{i=1}^{N} (x_i)_T \otimes (x_i)_T \Big\|^{1/2}$$

where $l = \log(n)\sqrt{\log d}\sqrt{\log N}$ and where $C = C_K$ may depend on K only.

The non-uniform Rudelson's inequality (Corollary 5.28) was a consequence of a non-commutative Khintchine inequality. Unfortunately, there does not seem to exist a way to deduce Proposition 5.68 from any known result. Instead, this proposition is proved using Dudley's integral inequality for Gaussian processes and estimates of covering numbers going back to Carl, see [67]. It is known however that such usage of Dudley's inequality

is not optimal (see e.g. [75]). As a result, the logarithmic factors in Proposition 5.68 are probably not optimal.

In contrast to these difficulties with Rudelson's inequality, proving uniform versions of the other two ingredients of Theorem 5.45 – the deviation Lemma 5.47 and Symmetrization Lemma 5.46 – is straightforward.

Lemma 5.69. *Let $(Z_t)_{t \in \mathcal{T}}$ be a stochastic process[24] such that all $Z_t \geq 0$. Then $\mathbb{E} \sup_{t \in \mathcal{T}} |Z_t^2 - 1| \geq \max(\mathbb{E} \sup_{t \in \mathcal{T}} |Z_t - 1|, (\mathbb{E} \sup_{t \in \mathcal{T}} |Z_t - 1|)^2)$.*

Proof. The argument is entirely parallel to that of Lemma 5.47. □

Lemma 5.70 (Symmetrization for stochastic processes). *Let X_{it}, $1 \leq i \leq N$, $t \in \mathcal{T}$, be random vectors valued in some Banach space B, where $\mathcal{T}$ is a finite index set. Assume that the random vectors $X_i = (X_{ti})_{t \in \mathcal{T}}$ (valued in the product space $B^{\mathcal{T}}$) are independent. Let $\varepsilon_1, \ldots, \varepsilon_N$ be independent symmetric Bernoulli random variables. Then*

$$\mathbb{E} \sup_{t \in \mathcal{T}} \Big\| \sum_{i=1}^N (X_{it} - \mathbb{E} X_{it}) \Big\| \leq 2 \mathbb{E} \sup_{t \in \mathcal{T}} \Big\| \sum_{i=1}^N \varepsilon_i X_{it} \Big\|.$$

Proof. The conclusion follows from Lemma 5.46 applied to random vectors X_i valued in the product Banach space $B^{\mathcal{T}}$ equipped with the norm $\||(Z_t)_{t \in \mathcal{T}}\|| = \sup_{t \in \mathcal{T}} \|Z_t\|$. The reader should also be able to prove the result directly, following the proof of Lemma 5.46. □

Proof of Theorem 5.67. Since the random vectors A_i are isotropic in $\mathbb{R}^d$, for every fixed subset $T \subseteq [d]$ the random vectors $(A_i)_T$ are also isotropic in $\mathbb{R}^T$, so $\mathbb{E}(A_i)_T \otimes (A_i)_T = I_{\mathbb{R}^T}$. As in the proof of Theorem 5.45, we are going to control

$$E := \mathbb{E} \max_{|T| \leq n} \Big\| \frac{1}{N} A_T^* A_T - I_{\mathbb{R}^T} \Big\| = \mathbb{E} \max_{|T| \leq n} \Big\| \frac{1}{N} \sum_{i=1}^N (A_i)_T \otimes (A_i)_T - I_{\mathbb{R}^T} \Big\|$$

$$\leq \frac{2}{N} \mathbb{E} \max_{|T| \leq n} \Big\| \sum_{i=1}^N \varepsilon_i (A_i)_T \otimes (A_i)_T \Big\|$$

where we used Symmetrization Lemma 5.70 with independent symmetric Bernoulli random variables $\varepsilon_1, \ldots, \varepsilon_N$. The expectation in the right hand side is taken both with respect to the random matrix A and the signs (ε_i). First taking the expectation with respect to (ε_i) (conditionally on A) and afterwards the expectation with respect to A, we obtain by Proposition 5.68 that

$$E \leq \frac{C_K l \sqrt{n}}{N} \mathbb{E} \max_{|T| \leq n} \Big\| \sum_{i=1}^N (A_i)_T \otimes (A_i)_T \Big\|^{1/2} = \frac{C_K l \sqrt{n}}{\sqrt{N}} \mathbb{E} \max_{|T| \leq n} \Big\| \frac{1}{N} A_T^* A_T \Big\|^{1/2}$$

---

[24]A stochastic process (Z_t) is simply a collection of random variables on a common probability space indexed by elements t of some abstract set $\mathcal{T}$. In our particular application, $\mathcal{T}$ will consist of all subsets $T \subseteq [d]$, $|T| \leq n$.

By the triangle inequality, $\mathbb{E}\max_{|T|\leq n}\left\|\frac{1}{N}A_T^*A_T\right\| \leq E + 1$. Hence we obtain

$$E \leq C_K l\sqrt{\frac{n}{N}}(E+1)^{1/2}$$

by Hölder's inequality. Solving this inequality in E we conclude that

$$E = \mathbb{E}\max_{|T|\leq n}\left\|\frac{1}{N}A_T^*A_T - I_{\mathbb{R}^T}\right\| \leq \max(\delta, \delta^2) \quad \text{where } \delta = C_K l\sqrt{\frac{2n}{N}}. \tag{5.58}$$

The proof is completed by a diagonalization argument similar to Step 2 in the proof of Theorem 5.45. One uses there a uniform version of deviation inequality given in Lemma 5.69 for stochastic processes indexed by the sets $|T| \leq n$. We leave the details to the reader. $\qquad\square$

Theorem 5.71 (Heavy-tailed restricted isometries). *Let $A = (a_{ij})$ be an $m \times n$ matrix whose rows A_i are independent isotropic random vectors in $\mathbb{R}^n$. Let K be a number such that all entries $|a_{ij}| \leq K$ almost surely. Then the normalized matrix $\bar{A} = \frac{1}{\sqrt{m}}A$ satisfies the following for $m \leq n$, for every sparsity level $1 < k \leq n$ and every number $\delta \in (0,1)$:*

$$\text{if } m \geq C\delta^{-2}k\log n\log^2(k)\log(\delta^{-2}k\log n\log^2 k) \quad \text{then } \mathbb{E}\delta_k(\bar{A}) \leq \delta. \tag{5.59}$$

Here $C = C_K > 0$ may depend only on K.

Proof. The result follows from Theorem 5.67, more precisely from its equivalent statement (5.58). In our notation, it says that

$$\mathbb{E}\delta_k(\bar{A}) \leq \max(\delta, \delta^2) \quad \text{where } \delta = C_K l\sqrt{\frac{k}{m}} = C_K\sqrt{\frac{k\log m}{m}}\log(k)\sqrt{\log n}.$$

The conclusion of the theorem easily follows. $\qquad\square$

In the interesting sparsity range $k \geq \log n$ and $k \geq \delta^{-2}$, the condition in Theorem 5.71 clearly reduces to

$$m \geq C\delta^{-2}k\log(n)\log^3 k.$$

Remark 5.72 (Boundedness requirement). The *boundedness assumption* on the entries of A is essential in Theorem 5.71. Indeed, if the rows of A are independent coordinate vectors in $\mathbb{R}^n$, then A necessarily has a zero column (in fact $n - m$ of them). This clearly contradicts the restricted isometry property.

Example 5.73.   1. **(Random Fourier measurements):** An important example for Theorem 5.41 is where A realizes random Fourier measurements. Consider the $n \times n$ Discrete Fourier Transform (DFT) matrix W with entries

$$W_{\omega,t} = \exp\left(-\frac{2\pi i\omega t}{n}\right), \quad \omega, t \in \{0, \ldots, n-1\}.$$

Consider a random vector X in $\mathbb{C}^n$ which picks a random row of W (with uniform distribution). It follows from Parseval's inequality that X is isotropic.[25] Therefore

---

[25] For convenience we have developed the theory over $\mathbb{R}$, while this example is over $\mathbb{C}$. As we noted earlier, all our definitions and results can be carried over to the complex numbers. So in this example we use the obvious complex versions of the notion of isotropy and of Theorem 5.71.

the $m \times n$ random matrix A whose rows are independent copies of X satisfies the assumptions of Theorem 5.41 with $K = 1$. Algebraically, we can view A as a *random row sub-matrix of the DFT matrix.*

In compressed sensing, such matrix A has a remarkable meaning – it realizes m *random Fourier measurements* of a signal $x \in \mathbb{R}^n$. Indeed, $y = Ax$ is the DFT of x evaluated at m random points; in words, y consists of m random frequencies of x. Recall that in compressed sensing, we would like to guarantee that with high probability every sparse signal $x \in \mathbb{R}^n$ (say, $|\operatorname{supp}(x)| \le k$) can be effectively recovered from its m random frequencies $y = Ax$. Theorem 5.71 together with Candès-Tao's result (recalled in the beginning of Section 5.6) imply that an exact recovery is given by the convex optimization problem $\min\{\|x\|_1 : Ax = y\}$ provided that we observe *slightly more frequencies than the sparsity of a signal*: $m \gtrsim\geq C\delta^{-2}k\log(n)\log^3 k$.

2. **(Random sub-matrices of orthogonal matrices):** In a similar way, Theorem 5.71 applies to a random row sub-matrix A of an *arbitrary bounded orthogonal matrix W.* Precisely, A may consist of m randomly chosen rows, uniformly and without replacement,[26] from an arbitrary $n \times n$ matrix $W = (w_{ij})$ such that $W^*W = nI$ and with uniformly bounded coefficients, $\max_{ij} |w_{ij}| = O(1)$. The examples of such W include the class of *Hadamard matrices* – orthogonal matrices in which all entries equal ± 1.

5.7   Notes

For Section 5.1   We work with two kinds of moment assumptions for random matrices: sub-gaussian and heavy-tailed. These are the two extremes. By the central limit theorem, the sub-gaussian tail decay is the strongest condition one can demand from an isotropic distribution. In contrast, our heavy-tailed model is completely general – no moment assumptions (except the variance) are required. It would be interesting to analyze random matrices with independent rows or columns in the intermediate regime, *between sub-gaussian and heavy-tailed* moment assumptions. We hope that for distributions with an appropriate finite moment (say, $(2+\varepsilon)$th or 4th), the results should be the same as for sub-gaussian distributions, i.e. no $\log n$ factors should occur. In particular, tall random matrices ($N \gg n$) should still be approximate isometries. This indeed holds for sub-exponential distributions [2]; see [82] for an attempt to go down to finite moment assumptions.

For Section 5.2   The material presented here is well known. The volume argument presented in Lemma 5.2 is quite flexible. It easily generalizes to covering numbers of more general metric spaces, including convex bodies in Banach spaces. See [60, Lemma 4.16] and other parts of [60] for various methods to control covering numbers.

---

[26]Since in the interesting regime very few rows are selected, $m \ll n$, sampling with or without replacement are formally equivalent. For example, see [67] which deals with the model of sampling without replacement.

For Section 5.2.3 The concept of sub-gaussian random variables is due to Kahane [39]. His definition was based on the moment generating function (Property 4 in Lemma 5.5), which automatically required sub-gaussian random variables to be centered. We found it more convenient to use the equivalent Property 3 instead. The characterization of sub-gaussian random variables in terms of tail decay and moment growth in Lemma 5.5 also goes back to [39].

The rotation invariance of sub-gaussian random variables (Lemma 5.9) is an old observation [15]. Its consequence, Proposition 5.10, is a general form of *Hoeffding's inequality*, which is usually stated for bounded random variables. For more on large deviation inequalities, see also notes for Section 5.2.4.

Khintchine inequality is usually stated for the particular case of symmetric Bernoulli random variables. It can be extended for $0 < p < 2$ using a simple extrapolation argument based on Hölder's inequality, see [45, Lemma 4.1].

For Section 5.2.4 Sub-gaussian and sub-exponential random variables can be studied together in a general framework. For a given exponent $0 < \alpha < \infty$, one defines general ψ_α random variables, those with moment growth $(\mathbb{E}|X|^p)^{1/p} = O(p^{1/\alpha})$. Sub-gaussian random variables correspond to $\alpha = 2$ and sub-exponentials to $\alpha = 1$. The reader is encouraged to extend the results of Sections 5.2.3 and 5.2.4 to this general class.

Proposition 5.16 is a form of *Bernstein's inequality*, which is usually stated for bounded random variables in the literature. These forms of Hoeffding's and Bernstein's inequalities (Propositions 5.10 and 5.16) are partial cases of a large deviation inequality for general ψ_α norms, which can be found in [72, Corollary 2.10] with a similar proof. For a thorough introduction to large deviation inequalities for sums of independent random variables (and more), see the books [59, 45, 24] and the tutorial [11].

For Section 5.2.5 Sub-gaussian distributions in $\mathbb{R}^n$ are well studied in geometric functional analysis; see [53] for a link with compressed sensing. General ψ_α distributions in $\mathbb{R}^n$ are discussed e.g. in [32].

Isotropic distributions on convex bodies, and more generally isotropic log-concave distributions, are central to asymptotic convex geometry (see [31, 57]) and computational geometry [78]. A completely different way in which isotropic distributions appear in convex geometry is from *John's decompositions* for contact points of convex bodies, see [9, 63, 79]. Such distributions are finitely supported and therefore are usually heavy-tailed.

For an introduction to the concept of *frames* (Example 5.21), see [41, 19].

For Section 5.2.6 The non-commutative Khintchine inequality, Theorem 5.26, was first proved by Lust-Piquard [48] with an unspecified constant B_p in place of $C\sqrt{p}$. The optimal value of B_p was computed by Buchholz [13, 14]; see [62, Section 6.5] for an thorough introduction to Buchholz's argument. For the complementary range $1 \le p \le 2$, a corresponding version of non-commutative Khintchine inequality was obtained by Lust-Piquard and Pisier [47]. By a duality argument implicitly contained in [47] and independently observed by Marius Junge, this latter inequality also implies the optimal order $B_p = O(\sqrt{p})$, see [65] and [61, Section 9.8].

Rudelson's Corollary 5.28 was initially proved using a majorizing measure technique; our proof follows Pisier's argument from [65] based on the non-commutative Khintchine inequality.

For Section 5.3 The "Bai-Yin law" (Theorem 5.31) was established for $s_{\max}(A)$ by Geman [30] and Yin, Bai and Krishnaiah [84]. The part for $s_{\min}(A)$ is due to Silverstein [70] for Gaussian random matrices. Bai and Yin [8] gave a unified treatment of both extreme singular values for general distributions. The fourth moment assumption in Bai-Yin's law is known to be necessary [7].

Theorem 5.32 and its argument is due to Gordon [35, 36, 37]. Our exposition of this result and of Corollary 5.35 follows [21].

Proposition 5.34 is just a tip of an iceberg called *concentration of measure phenomenon*. We do not discuss it here because there are many excellent sources, some of which were mentioned in Section 5.1. Instead we give just one example related to Corollary 5.35. For a general random matrix A with independent centered entries bounded by 1, one can use Talagrand's concentration inequality for convex Lipschitz functions on the cube [73, 74]. Since $s_{\max}(A) = \|A\|$ is a convex function of A, Talagrand's concentration inequality implies $\mathbb{P}\big\{|s_{\max}(A) - \text{Median}(s_{\max}(A))| \geq t\big\} \leq 2e^{-ct^2}$. Although the precise value of the median may be unknown, integration of this inequality shows that $|\mathbb{E}s_{\max}(A) - \text{Median}(s_{\max}(A))| \leq C$.

For the recent developments related to the *hard edge* problem for almost square and square matrices (including Theorem 5.38) see the survey [69].

For Section 5.4 Theorem 5.39 on random matrices with sub-gaussian rows, as well as its proof by a covering argument, is a folklore in geometric functional analysis. The use of covering arguments in a similar context goes back to Milman's proof of Dvoretzky's theorem [55]; see e.g. [9] and [60, Chapter 4] for an introduction. In the more narrow context of extreme singular values of random matrices, this type of argument appears recently e.g. in [2].

The breakthrough work on heavy-tailed isotropic distributions is due to Rudelson [65]. He used Corollary 5.28 in the way we described in the proof of Theorem 5.45 to show that $\frac{1}{N}A^*A$ is an approximate isometry. Probably Theorem 5.41 can also be deduced by a modification of this argument; however it is simpler to use the non-commutative Bernstein's inequality.

The symmetrization technique is well known. For a slightly more general two-sided inequality than Lemma 5.46, see [45, Lemma 6.3].

The problem of estimating covariance matrices described in Section 5.4.3 is a basic problem in statistics, see e.g. [38]. However, most work in the statistical literature is focused on the normal distribution or general product distributions (up to linear transformations), which corresponds to studying random matrices with independent entries. For non-product distributions, an interesting example is for uniform distributions on convex sets [40]. As we mentioned in Example 5.25, such distributions are sub-exponential but not necessarily sub-gaussian, so Corollary 5.50 does not apply. Still, the sample size $N = O(n)$ suffices to estimate the covariance matrix in this case [2]. It is conjectured that the same should hold for general distributions with finite (e. g. 4th) moment

assumption [82].

Corollary 5.55 on random sub-matrices is a variant of the Rudelson's result from [64]. The study of random sub-matrices was continued in [66]. Random sub-frames were studied in [80] where a variant of Corollary 5.56 was proved.

For Section 5.5 Theorem 5.58 for sub-gaussian columns seems to be new. However, historically the efforts of geometric functional analysts were immediately focused on the more difficult case of sub-exponential tail decay (given by uniform distributions on convex bodies). An indication to prove results like Theorem 5.58 by decoupling and covering is present in [12] and is followed in [32, 2].

The normalization condition $\|A_j\|_2 = \sqrt{N}$ in Theorem 5.58 can not be dropped but can be relaxed. Namely, consider the random variable $\delta := \max_{i \leq n} \left| \frac{\|A_j\|_2^2}{N} - 1 \right|$. Then the conclusion of Theorem 5.58 holds with (5.36) replaced by

$$(1 - \delta)\sqrt{N} - C\sqrt{n} - t \leq s_{\min}(A) \leq s_{\max}(A) \leq (1 + \delta)\sqrt{N} + C\sqrt{n} + t.$$

Theorem 5.62 for heavy-tailed columns also seems to be new. The incoherence parameter m is meant to prevent collisions of the columns of A in a quantitative way. It is not clear whether the *logarithmic factor* is needed in the conclusion of Theorem 5.62, or whether the incoherence parameter alone takes care of the logarithmic factors whenever they appear. The same question can be raised for all other results for heavy-tailed matrices in Section 5.4.2 and their applications – can we replace the logarithmic factors by more sensitive quantities (e.g. the logarithm of the incoherence parameter)?

For Section 5.6 For a mathematical introduction to compressed sensing, see the introductory chapter of this book and [28, 20].

A version of Theorem 5.65 was proved in [54] for the row-independent model; an extension from sub-gaussian to sub-exponential distributions is given in [3]. A general framework of stochastic processes with sub-exponential tails is discussed in [52]. For the column-independent model, Theorem 5.65 seems to be new.

Proposition 5.66 that formalizes a simple approach to restricted isometry property based on concentration is taken from [10]. Like Theorem 5.65, it can also be used to show that Gaussian and Bernoulli random matrices are restricted isometries. Indeed, it is not difficult to check that these matrices satisfy a concentration inequality as required in Proposition 5.66 [1].

Section 5.6.2 on heavy-tailed restricted isometries is an exposition of the results from [67]. Using concentration of measure techniques, one can prove a version of Theorem 5.71 with high probability $1 - n^{-c \log^3 k}$ rather than in expectation [62]. Earlier, Candes and Tao [18] proved a similar result for random Fourier matrices, although with a slightly higher exponent in the logarithm for the number of measurements in (5.55), $m = O(k \log^6 n)$. The survey [62] offers a thorough exposition of the material presented in Section 5.6.2 and more.

Bibliography

[1] Achlioptas, D. (2003). Database-friendly random projections: Johnson-Lindenstrauss with binary coins, in: Special issue on PODS 2001 (Santa Barbara, CA). *J. Comput. System Sci.*, **66**, 671–687.

[2] Adamczak, R., Litvak, A., Pajor, A., Tomczak-Jaegermann, N. (2010). Quantitative estimates of the convergence of the empirical covariance matrix in log-concave ensembles, *J. Amer. Math. Soc.*, **23**, 535–561.

[3] Adamczak, R., Litvak, A., Pajor, A., Tomczak-Jaegermann, N. (2010). Restricted isometry property of matrices with independent columns and neighborly polytopes by random sampling, *Constructive Approximation*, to appear.

[4] Ahlswede, R. and Winter, A. (2002). Strong converse for identification via quantum channels, *IEEE Trans. Inform. Theory*, **48**, 569–579.

[5] Anderson, G., Guionnet, A. and Zeitouni, O. (2009). *An Introduction to Random Matrices*. Cambridge: Cambridge University Press.

[6] Bai, Z. and Silverstein, J. (2010). *Spectral analysis of large dimensional random matrices*. Second edition. New York: Springer.

[7] Bai, Z., Silverstein, J. and Yin, Y. (1988). A note on the largest eigenvalue of a large-dimensional sample covariance matrix, *J. Multivariate Anal.*, **26**, 166–168.

[8] Bai, Z. and Yin, Y. (1993). Limit of the smallest eigenvalue of a large-dimensional sample covariance matrix, *Annals of Probability*, **21**, 1275–1294.

[9] Ball, K. (1997). An elementary introduction to modern convex geometry, in *Flavors of geometry*, pp. 1–58. Math. Sci. Res. Inst. Publ., 31, Cambridge: Cambridge University Press.

[10] Baraniuk, R., Davenport, M., DeVore, R. and Wakin, M. (2008). A simple proof of the restricted isometry property for random matrices, *Constructive Approximation*, **28**, 253–263.

[11] Boucheron, S. Bousquet, O. and Lugosi, G. (2004). Concentration inequalities, in *Advanced Lectures in Machine Learning*, edited by Bousquet, O., Luxburg, U. and Rätsch, G. Springer, pp. 208–240.

[12] Bourgain, J. (1999). Random points in isotropic convex sets, in: *Convex geometric analysis (Berkeley, CA, 1996)*, pp. 53–58. Math. Sci. Res. Inst. Publ., 34. Cambridge: Cambridge University Press.

[13] Buchholz, A. (2001). Operator Khintchine inequality in non-commutative probability, *Math. Ann.*, **319**, 1–16.

[14] Buchholz, A. (2005). Optimal constants in Khintchine type inequalities for fermions, Rademachers and q-Gaussian operators, *Bull. Pol. Acad. Sci. Math.*, **53**, 315–321.

[15] Buldygin, V. V. and Kozachenko, Ju. V. (1980). Sub-Gaussian random variables, *Ukrainian Mathematical Journal*, **32**, 483–489.

[16] Candès, E. The restricted isometry property and its implications for compressed sensing, *Compte Rendus de l'Academie des Sciences, Paris, Serie I*, **346**, 589–592.

[17] Candès, E. and Tao, T. (2005). Decoding by linear programming, *IEEE Trans. Inform. Theory*, **51**, 4203–4215.

[18] Candès, E. and Tao, T. (2006). Near-optimal signal recovery from random projections: universal encoding strategies? *IEEE Trans. Inform. Theory*, **52**, 5406–5425.

[19] Christensen, O. (2008). *Frames and bases. An introductory course.* Applied and Numerical Harmonic Analysis. Boston, MA: Birkhäuser Boston, Inc.

[20] Compressive Sensing Resources, `http://dsp.rice.edu/cs`

[21] Davidson, K. R. and Szarek, S. J. (2001). Local operator theory, random matrices and Banach spaces, in *Handbook of the geometry of Banach spaces*, Vol. I, pp. 317–366. Amsterdam: North-Holland.

[22] de la Peña, V. and Giné, E. (1999). *Decoupling. From dependence to independence. Randomly stopped processes. U-statistics and processes. Martingales and beyond.* New York: Springer-Verlag.

[23] Deift, P. and Gioev, D. (2009). *Random matrix theory: invariant ensembles and universality.* Courant Lecture Notes in Mathematics, 18. Courant Institute of Mathematical Sciences, New York; Providence, RI: American Mathematical Society.

[24] Dembo, A. and Zeitouni, O. (1993). *Large deviations techniques and applications.* Boston, MA: Jones and Bartlett Publishers.

[25] Drineas, P., Kannan, R. and Mahoney, M. (2006). Fast Monte Carlo algorithms for matrices. I, II III, *SIAM J. Comput.*, **36** (2006), 132–206.

[26] Durrett, R. (2005). *Probability: theory and examples.* Belmont: Duxbury Press.

[27] Feldheim, O. and Sodin, S. (2008). A universality result for the smal lest eigenvalues of certain sample covariance matrices, *Geometric and Functional Analysis*, to appear.

[28] Fornasier, M. and Rauhut, H. (2010). Compressive Sensing, in *Handbook of Mathematical Methods in Imaging*, edited by Scherzer, O. Springer, to appear.

[29] Füredi, Z.; Komlós, J. (1981). The eigenvalues of random symmetric matrices, *Combinatorica*, **1**, 233–241.

[30] Geman, S. (1980). A limit theorem for the norm of random matrices, *Annals of Probability*, **8**, 252–261.

[31] Giannopoulos, A. (2003). *Notes on isotropic convex bodies*, Warsaw.

[32] Giannopoulos, A. and Milman, V. (2000). Concentration property on probability spaces, *Advances in Mathematics*, **156**, 77–106.

[33] Giannopoulos, A. and Milman, V. (2001). Euclidean structure in finite dimensional normed spaces, in *Handbook of the geometry of Banach spaces*, Vol. I, pp. 707–779. Amsterdam: North-Holland.

[34] Golub, G., Mahoney, M., Drineas, P. and Lim, L.-H. (2006). Bridging the gap between numerical linear algebra, theoretical computer science, and data applications, *SIAM News*, **9**, Number 8.

[35] Gordon, Y. (1984). On Dvoretzky's theorem and extensions of Slepian's lemma, in *Israel seminar on geometrical aspects of functional analysis (1983/84), II.* Tel Aviv: Tel Aviv University.

[36] Gordon, Y. (1985). Some inequalities for Gaussian processes and applications, *Israel Journal of Mathematics*, **50**, 265–289.

[37] Gordon, Y. (1992). Majorization of Gaussian processes and geometric applications, *Probab. Theory Related Fields*, **91**, 251–267.

[38] Johnstone, I. (2001). On the distribution of the largest eigenvalue in principal components analysis, *Ann. Statist.*, **29**, 295–327.

[39] Kahane, J.-P. (1960). Propriétés locales des fonctions à séries de Fourier aléatoires, *Studia Mathematica*, **19**, 1–25.

[40] Kannan, R., Lovász, L. and Simonovits, M. (1995). Isoperimetric problems for convex bodies and a localization lemma, *Discrete Comput. Geom.*, **13**, 541–559.

[41] Kovačević, J. and Chebira, A. (2008). *An Introduction to Frames.* Foundations and Trends in Signal Processing. Now Publishers.

[42] Latala, R. (2005). Some estimates of norms of random matrices, *Proc. Amer. Math. Soc.*, **133**, 1273-1282.

[43] Ledoux, M. (2005). *The concentration of measure phenomenon.* Mathematical Surveys and Monographs, 89. Providence: American Mathematical Society.

[44] Ledoux, M. (2007). Deviation inequalities on largest eigenvalues, in *Geometric aspects of functional analysis*, pp. 167–219. Lecture Notes in Math., 1910. Berlin: Springer.

[45] Ledoux, M. and Talagrand, M. (1991). *Probability in Banach spaces.* Berlin: Springer-Verlag.

[46] Litvak, A., Pajor, A., Rudelson, M. and Tomczak-Jaegermann, N. (2005). Smallest singular value of random matrices and geometry of random polytopes, *Adv. Math.*, **195**, 491–523.

[47] Lust-Piquard, F. and Pisier, G. (1991). Noncommutative Khintchine and Paley inequalities, *Ark. Mat.*, **29**, 241–260.

[48] Lust-Piquard, F. (1986). Inégalités de Khintchine dans $C_p(1 < p < \infty)$, *C. R. Acad. Sci. Paris Sér. I Math.*, **303**, 289–292.

[49] Matoušek, J. (2002). *Lectures on discrete geometry.* Graduate Texts in Mathematics, 212. New York: Springer-Verlag.

[50] Meckes, M. (2004). Concentration of norms and eigenvalues of random matrices. *J. Funct. Anal.*, **211**, 508–524.

[51] Mehta, M. L. (2004). *Random matrices.* Pure and Applied Mathematics (Amsterdam), 142. Amsterdam: Elsevier/Academic Press.

[52] Mendelson, S. (2008). On weakly bounded empirical processes, *Math. Ann.*, **340**, 293–314.

[53] Mendelson, S., Pajor, A. and Tomczak-Jaegermann, N. (2007). Reconstruction and subgaussian operators in asymptotic geometric analysis, *Geom. Funct. Anal.*, **17**, 1248–1282.

[54] Mendelson, S., Pajor, A. and Tomczak-Jaegermann, N. (2008). Uniform uncertainty principle for Bernoulli and subgaussian ensembles, *Constr. Approx.*, **28** (2008), 277–289.

[55] Milman, V. D. (1974). A new proof of A. Dvoretzky's theorem on cross-sections of convex bodies. *Funkcional. Anal. i Prilozhen.*, **5**, 28–37.

[56] Milman, V. and Schechtman, G. (1986). *Asymptotic theory of finite-dimensional normed spaces. With an appendix by M. Gromov.* Lecture Notes in Mathematics, 1200. Berlin: Springer-Verlag.

[57] Paouris, G. (2006). Concentration of mass on convex bodies, *Geom. Funct. Anal.*, **16**, 1021–1049.

[58] Péché, S. and Soshnikov, A. (2008). On the lower bound of the spectral norm of symmetric random matrices with independent entries, *Electron. Commun. Probab.*, **13**, 280–290.

[59] Petrov, V. V. (1975). *Sums of independent random variables.* New York-Heidelberg: Springer-Verlag.

[60] Pisier, G. (1989). *The volume of convex bodies and Banach space geometry.* Cambridge Tracts in Mathematics, 94. Cambridge: Cambridge University Press.

[61] Pisier, G. (2003). *Introduction to operator space theory.* London Mathematical Society Lecture Note Series, 294. Cambridge: Cambridge University Press.

[62] Rauhut, H. (2010). Compressive sensing and structured random matrices, in *Theoretical Foundations and Numerical Methods for Sparse Recovery*, edited by Fornasier, M. Radon Series Comp. Appl. Math., Volume 9, pp. 1–92. deGruyter.

[63] Rudelson, M. (1997). Contact points of convex bodies, *Israel Journal of Mathematics*, **101**, 93–124.

[64] Rudelson, M. (1999). Almost orthogonal submatrices of an orthogonal matrix, *Israel J. of Math.*, **111**, 143–155.

[65] Rudelson, M. (1999). Random vectors in the isotropic position, *Journal of Functional Analysis*, **164**, 60–72.

[66] Rudelson, M. and Vershynin, R. (2007). Sampling from large matrices: an approach through geometric functional analysis, *J. ACM*, **54**, Art. 21, 19 pp.

[67] Rudelson, M. and Vershynin, R. (2008). On sparse reconstruction from Fourier and Gaussian measurements, *Comm. Pure Appl. Math.*, **61**, 1025–1045.

[68] Rudelson, M. and Vershynin, R. (2009). Smallest singular value of a random rectangular matrix, *Comm. Pure Appl. Math.*, **62**, 1707–1739.

[69] Rudelson, M. and Vershynin, R. (2010). Non-asymptotic theory of random matrices: extreme singular values, *Proceedings of the International Congress of Mathematicians*, Hyderabad, India, to appear.

[70] Silverstein, J. (1985). The smallest eigenvalue of a large-dimensional Wishart matrix, *Annals of Probability*, **13**, 1364–1368.

[71] Soshnikov, A. (2002). A note on universality of the distribution of the largest eigenvalues in certain sample covariance matrices, *J. Statist. Phys.*, **108**, 1033–1056.

[72] Talagrand, M. (1994). The supremum of some canonical processes, *American Journal of Mathematics*, **116**, 283–325.

[73] Talagrand, M. (1995). Concentration of measure and isoperimetric inequalities in product spaces, *Inst. Hautes Études Sci. Publ. Math.*, **81**, 73–205.

[74] Talagrand, M. (1996). A new look at independence, *Annals of Probability*, **24**, 1–34.

[75] Talagrand, M. (2005). *The generic chaining. Upper and lower bounds of stochastic processes.* Springer Monographs in Mathematics. Berlin: Springer-Verlag.

[76] Tao, T. and Vu, V. (2009). From the Littlewood-Offord problem to the circular law: universality of the spectral distribution of random matrices, *Bull. Amer. Math. Soc. (N.S.)*, **46**, 377–396.

[77] Tropp, J. (2010). User-friendly tail bounds for sums of random matrices, submitted.

[78] Vempala, S. (2005). Geometric random walks: a survey, in *Combinatorial and computational geometry*, pp. 577–616. Math. Sci. Res. Inst. Publ., 52. Cambridge: Cambridge University Press.

[79] Vershynin, R. (2001). John's decompositions: selecting a large part, *Israel Journal of Mathematics*, **122**, 253–277.

[80] Vershynin, R. (2005). Frame expansions with erasures: an approach through the non-commutative operator theory, *Appl. Comput. Harmon. Anal.*, **18**, 167–176.

[81] Vershynin, R. (2010). Approximating the moments of marginals of high-dimensional distributions, *Annals of Probability*, to appear.

[82] Vershynin, R. (2010). How close is the sample covariance matrix to the actual covariance matrix?, *Journal of Theoretical Probability*, to appear.

[83] Vu, V. (2007). Spectral norm of random matrices, *Combinatorica*, **27**, 721–736.

[84] Yin, Y. Q., Bai, Z. D. and Krishnaiah, P. R. (1998). On the limit of the largest eigenvalue of the large-dimensional sample covariance matrix, *Probab. Theory Related Fields*, **78**, 509–521.

Index