



Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Covariance Estimation for High-Dimensional Distributions

Master Thesis

Angelo Gnazzo

July 25, 2023

Advisors: Dr. March Boedihardjo, Prof. Dr. Afonso Bandeira

Department of Mathematics, ETH Zürich

Acknowledgements

I would like to thank my supervisors Dr. March Tian Boedihardjo and Prof. Dr. Afonso Bandeira from ETH Zurich for helping me throughout this work. They offered numerous great insights, explanations, and support.

I would like to thank various friends for their ideas and stimulating discussions.

I also would like to thank everyone that has been part of these years at ETH Zurich: professors, colleagues, and employees.

Abstract

This thesis focuses on the estimation of covariance matrices for a variety of high-dimensional distributions, a problem pivotal in several areas such as multivariate statistics, machine learning, or Markowitz's modern portfolio theory. The dissertation aims to study, through a random matrix theory approach, the theoretical guarantees for the convergence of the sample covariance matrix to the true covariance matrix.

The results presented in this thesis provide valuable insights into the non-asymptotic theory of random matrices and several key ideas and concepts from high-dimensional probability.

Contents

Contents	iii
1 Introduction	1
1.1 Asymptotic and Non-asymptotic regimes in Random Matrix Theory	1
1.2 Estimation of covariance matrices of high-dimensional distributions	2
1.3 Scope and content of this dissertation	3
1.4 Basic Notation	3
2 Setting and Overview of the Methods	7
2.1 Setting	7
2.2 The moment method	9
2.3 The Laplace transform method	10
2.4 The random process method	12
2.5 Bounds for Singular Values via Net Arguments	12
2.6 Sub-gaussian, sub-exponential, and Orlicz Norms of random vectors	17
3 Case Study: Gaussian Random Vectors	27
3.1 Preliminaries	27
3.2 Main Result	35
3.3 Proof of Main Result	36
4 Case Study: Isotropic Subgaussian Random Vectors	47
4.1 Preliminaries	47
4.2 Main Result	49
4.3 The main technical step	49
4.4 Proof of Main Result	51
5 Case Study: Heavy-tailed Random Vectors	53

5.1	Matrix Bernstein Inequality	53
5.2	Main Result	55
5.3	Proof of Matrix Bernstein Inequality	58
5.4	Appendix of this Chapter	64
6	Case Study: Isotropic Random Vectors sampled from a convex body	71
6.1	Preliminaries	71
6.2	Main Result	74
6.3	Proof of Main Result	75
6.4	Proof of the other results	77
6.4.1	Proof of Lemma 6.11	77
6.4.2	Proof of Lemma 6.13	82
6.5	Appendix of this Chapter	88
7	Case Study: Isotropic Log-concave Random Vectors	99
7.1	Preliminaries	99
7.2	Main Result	100
7.3	The main technical step	101
7.4	Proof of the Main Result	113
7.5	Appendix of this Chapter	118
8	Case Study: Random Vectors with Sub-exponential marginals	121
8.1	Main Result	121
8.2	Proof of the Main Result	122
9	Case Study: Random Vectors with bounded L_p marginal moments	129
9.1	Main Result	129
9.2	Proof of the Main Result	130
10	Conclusion	143
A	Appendix	145
A.1	Basic Inequalities	145
A.2	Basic Probability Theory	145
	Bibliography	147

Introduction

1.1 Asymptotic and Non-asymptotic regimes in Random Matrix Theory

Random matrix theory studies properties of $n \times d$ matrices A chosen from some distribution on the set of all matrices. These random objects were first introduced by Wishart in 1928 in the field of statistics [51]. Their popularity increased following the groundbreaking work of physicist E.P. Wigner, who innovatively applied them to represent atomic levels in nuclear energy [50]. Indeed, the asymptotic regime, where the dimensions $n, d \rightarrow \infty$, was well suited for statistical physics, e.g. when random matrices served as finite-dimensional models of infinite-dimensional operators. Since then, the asymptotic theory of random matrices has established connections with numerous diverse mathematical disciplines. These include but are not limited to, analytic number theory [40], combinatorics [8], operator theory and orthogonal polynomials [21], and mathematical physics. The books [42], [4], [22], [6] offer an introduction to the classical problems of random matrix theory and its fascinating connections.

The asymptotic regime has been very well studied because, as dimensions n and d grow to infinity, it is observed 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 are Wigner's semicircle law for the eigenvalues of symmetric Gaussian matrices, the Marchenko-Pastur law, Bai-Yin's law, and many more. Here is one of those examples.

Theorem 1.1 (*Bai-Yin's law [7]*). *Let A be an $n \times d$ random matrix with entries given by independent copies of a distribution with zero mean, unit variance, and finite fourth moment. Let $s_1(A) \geq s_d(A)$ be the largest and smallest singular values of A . As $\frac{d}{n} \rightarrow c \in [0, 1]$,*

$$\frac{s_1(A)}{\sqrt{n}} \rightarrow 1 + \sqrt{c}, \quad \frac{s_d(A)}{\sqrt{n}} \rightarrow 1 - \sqrt{c} \quad a. s.$$

This theorem does not provide information about matrices with given fixed dimensions $d, n < \infty$. However, in some other areas including statistics, convex geometry, and compressed sensing, the limiting regime displayed above may not be very useful.

In contrast to this, a non-asymptotic answer to the same question of Theorem 1.1 is the following.

Theorem 1.2 (Gaussian deviation [20]). *Let A be an $n \times d$ random matrix with independent entries satisfying the Gaussian $\mathcal{N}(0, 1)$ distribution. Then, for every $t > 0$, with probability at least $1 - 2e^{-t^2}$,*

$$\sqrt{n} - C\sqrt{d} - t \leq s_d(A) \leq s_1(A) \leq \sqrt{n} + C\sqrt{d} + t,$$

where C is an absolute constant.

This answer is less precise (because of an absolute constant C) but more quantitative because for fixed dimensions it gives an exponential probability of success. For introductory texts in the so-called high dimensional probability, we refer the reader to [49], [46], [44], [10] and [15].

1.2 Estimation of covariance matrices of high-dimensional distributions

For a high-dimensional distribution μ , its covariance structure is captured by its covariance matrix Σ . Let X be a random vector in $\mathbb{R}^d$ distributed according to μ . We may assume that X is centered (by estimating and subtracting $\mathbb{E}X$). The covariance matrix of X is defined as:

$$\Sigma := \mathbb{E}XX^T.$$

$\Sigma = \Sigma(X)$ is a symmetric, positive semi-definite $d \times d$ matrix. It is a multivariate version of the variance $\text{Var}(X)$.

A fundamental problem in multivariate statistics is to obtain an accurate estimate of the covariance matrix of a multivariate distribution given independent samples from the distribution. This challenge plays an important role in statistical analysis in many fields, including portfolio allocation and risk management in finance, graphical modeling, and clustering for gene discovery in bioinformatics. Moreover, covariance estimation serves as a building block for many important statistical learning methods. It is extremely important whenever we need to understand the spread of the data and its

marginals, for example, when we perform regression analysis [26] or principal component analysis [27]. Recent developments in the field showed that, under certain assumptions on the moments of 1-dimensional projections of X , and also some rather strong structural assumptions on Σ , it is possible to obtain a satisfactory estimator of Σ even when the size of the sample n is much smaller than the dimension d . In this regime, covariance estimation is generally impossible (for dimension reasons), but usually, in practice, one knows an a priori some structure of the covariance matrix Σ . For example, Σ or Σ^{-1} is often known to be sparse, having few non-zero entries, such as in graphical models. There is a vast line of research dealing with these questions ([11], [18]), which, however, do not have a direct connection with the results we will present.

Indeed, in this dissertation, to proceed with the estimation of the covariance matrix, we take a sample of n independent points $X_1, \dots, X_n$ from the distribution and form the sample covariance matrix:

$$\Sigma_n := \frac{1}{n} \sum_{i=1}^n X_i X_i^T$$

Σ_n is a random matrix, which hopefully approximates Σ well. Then, the Covariance Estimation Problem can be formulated as follows:

Question 1.3 *Determine the minimal sample size $n = n(d)$ that guarantees with high probability (say, 0.99) that the sample covariance matrix Σ_n estimates the actual covariance matrix Σ with fixed accuracy (say, $\varepsilon = 0.01$) in the operator norm:*

$$\|\Sigma_n - \Sigma\| \leq \varepsilon \|\Sigma\|$$

1.3 Scope and content of this dissertation

We consider a general framework for the Covariance Estimation Problem. In Chapter 2 we will present an overview of the main methods from Random Matrix Theory used to approach the problem, from a probabilistic point of view. In the following chapters, we will apply such methods to obtain guarantees on the convergence of the sample covariance matrix estimator Σ_n for n independent points sampled from several distributions, ranging from Gaussian to heavy-tailed.

1.4 Basic Notation

For a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and a random variable $X : \Omega \rightarrow \mathbb{R}$, we denote the expectation or mean by $\mathbb{E}X$ and the variance by $\text{Var}(X)$. In this

dissertation, we will assume all sets to be measurable.

Let $L_p = L_p(\Omega, \mathcal{A}, \mathbb{P})$ be the standard L_p spaces, i.e. the vector space consisting of all random variables X on Ω with finite L_p -norm, and $\|X\|_p = (\mathbb{E}[|X|^p])^{1/p}$ the associate norm or quasi-norm (depending on the value of $p > 0$).

A random variable X is (essentially) bounded if $\|X\|_\infty < \infty$.

The moment generating function (MGF) $M_X(\lambda) = \mathbb{E}e^{\lambda X}$ is the (real) Laplace transform of X with respect to the measure $\mathbb{P}$. The cumulant generating function (CGF) $K_X(\lambda)$ is given by: $K_X(\lambda) = \log \mathbb{E}e^{\lambda X}$.

We work in $\mathbb{R}^d$, which is equipped with a Euclidean structure $\langle \cdot, \cdot \rangle$. We denote by $\|\cdot\|_2$ the corresponding Euclidean norm. The volume (d -dimensional Lebesgue measure) is denoted by $|\cdot|$.

Random vectors are denoted by $X = (x_1, \dots, x_d) \in \mathbb{R}^d$. They have mean $\mu = \mathbb{E}X = (\mathbb{E}x_1, \dots, \mathbb{E}x_d)$ and covariance matrix $\Sigma = \mathbb{E}[(X - \mathbb{E}X)(X - \mathbb{E}X)^T]$.

The d -dimensional hyperball of radius R is defined by:

$$B_2^d(R) = \{x \in \mathbb{R}^d : x_1^2 + \dots + x_d^2 \leq R^2\}$$

The d -dimensional hypersphere (or d -sphere) of radius R is given by:

$$S^{d-1}(R) = \{x \in \mathbb{R}^d : x_1^2 + \dots + x_d^2 = R^2\}$$

If there is no conflict with other notation, we may write B_2^d for $B_2^d(1)$ and S^{d-1} for $S^{d-1}(1)$.

The canonical basis of $\mathbb{R}^d$ will be denoted by $(e_1, \dots, e_d)$.

Consider a matrix $A \in \mathbb{R}^{d_1 \times d_2}$. We will denote by $s_1(A) \geq s_{d_2}(A)$ the largest and smallest singular values of A . The values $s_1(A)$ and $s_{d_2}(A)$ are respectively the smallest number M and the largest number m such that:

$$m\|x\|_2 \leq \|Ax\|_2 \leq M\|x\|_2 \text{ for all } x \in \mathbb{R}^{d_2}$$

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

$$\|A\| = \sup_{x \in \mathbb{R}^{d_2} \setminus \{0\}} \frac{\|Ax\|_2}{\|x\|_2} = \sup_{x \in S^{d_2-1}} \|Ax\|_2 \quad (1.1)$$

Then, the link between the extreme singular values and the spectral norm is the following:

$$s_1(A) = \|A\|$$

Moreover, observe that it is often convenient to study A through the $d_2 \times d_2$ symmetric positive-semidefinite matrix $A^T A$. By studying this matrix we can obtain the eigenvalues of $|A| := \sqrt{A^T A}$, which will be non-negative real numbers and, when arranged in non-decreasing order, constitute the singular values of A .

The trace of a square matrix A is defined as:

$$\text{Tr}(A) = \sum_i A_{ii}$$

If A is a random symmetric matrix, the definitions of (MGF) and (CGF) translate to the matrix moment generating function M_A and the matrix cumulant generating function K_A .

Chapter 2

Setting and Overview of the Methods

As we already discussed and we will see in practice and more detail in the next chapters, the goal of our study is to obtain probabilistic inequalities, i.e. upper bounds and lower bounds, on spectral properties of the matrices in question.

These probabilistic inequalities, clearly, should be able to capture faithfully the underlying structure.

This, in general, can be very well achieved by studying the *spectral norms* of random matrices.

Before we can obtain any meaningful bounds, we must first introduce some basic approaches for bounding the spectral norms of random matrices. In this section, we will collect the main methods that will be used in this work.

2.1 Setting

In this dissertation, we will try to solve the Covariance Estimation Problem, presented with Question 1.3. In what follows, we will provide a detailed presentation of the key objects and assumptions that will serve as the foundation for our work throughout this manuscript.

Definition 2.1 *Let X be a random vector in $\mathbb{R}^d$ distributed according to a d -dimensional probability distribution μ . We may assume that X is centered (by estimating and subtracting $\mathbb{E}X$). The covariance matrix of X is defined as:*

$$\Sigma := \mathbb{E}XX^T$$

$\Sigma = \Sigma(X)$ is a symmetric, positive semi-definite $d \times d$ matrix.

To estimate the true covariance matrix, we will make use of the sample covariance matrix estimator.

2. SETTING AND OVERVIEW OF THE METHODS

Definition 2.2 Let $X_1, \dots, X_n$ be i.i.d. copies of X , a random vector in $\mathbb{R}^d$ with covariance matrix Σ . The sample covariance matrix estimator is defined as:

$$\Sigma_n := \frac{1}{n} \sum_{i=1}^n X_i X_i^T$$

Observe that Σ_n is a random matrix since it is a function of the random vectors $X_1, \dots, X_n$.

Moreover, Σ_n is an unbiased estimator of Σ and, by the Law of Large Numbers: $\Sigma_n \rightarrow \Sigma$ as $n \rightarrow \infty$. So, by taking n large enough, we are guaranteed to estimate the covariance matrix as we want. However, we are interested in a quantitative aspect, as discussed in Chapter 1 when we introduced non-asymptotic random matrix theory and we posed Question 1.3. We ask ourselves: what is the minimal sample size n that guarantees the approximation with a given accuracy?

Since Σ_n is a random matrix, we will use methods from the subject of random matrix theory to study the rate of convergence of Σ_n to Σ . Now, let $\mathbf{X}$ be a random $d \times n$ matrix whose columns are the X_i 's, i.e. $X_{ik} = (X_k)_i$, for $i \in 1, \dots, d$, $k \in 1, \dots, n$.

Then, $\Sigma_n = \frac{1}{n} \sum_{k=1}^n X_k X_k^T = \frac{\mathbf{X} \mathbf{X}^T}{n}$. This observation makes the relation of Question 1.3 to random matrix theory clear. To answer Question 1.3, we need to investigate if, under appropriate moment assumptions on the distribution (of the columns of $\mathbf{X}$), there are results similar to Bai-Yin, Theorem 1.1.

Note that, since we want to achieve a relative error ε in the spectral norm as $\|\Sigma_n - \Sigma\| \leq \varepsilon \|\Sigma\|$, we can assume, to simplify the study, that $\Sigma = I$. This means that the random vector X is isotropic.

Definition 2.3 A random vector $X \in \mathbb{R}^d$ is called isotropic if:

$$\Sigma = \mathbb{E} X X^T = I,$$

where I is the identity matrix. Equivalently, X is isotropic if:

$$\mathbb{E} \langle X, x \rangle^2 = \|x\|_2^2 \quad \forall x \in \mathbb{R}^d$$

The notion of isotropy generalizes the assumption of unit variance in higher dimensions.

Also, observe that every full-dimensional random vector X can be made into an isotropic one by the linear transformation: $\Sigma^{-1/2} X$. This allows us to mostly focus on studying isotropic random vectors in the following of the dissertation.

Lemma 2.4 Let X, Y be independent isotropic random vectors in $\mathbb{R}^d$. Then $\mathbb{E} \|X\|_2^2 = d$ and $\mathbb{E} \langle X, Y \rangle^2 = d$.

Proof (of Lemma 2.4) The first part follows from $\mathbb{E}\|X\|_2^2 = \mathbb{E} \operatorname{Tr}(XX^T) = \operatorname{Tr}(\mathbb{E}XX^T) = \operatorname{Tr}(I) = d$. The second part follows by conditioning on Y , using isotropy of X and using the first part for Y : this way we obtain $\mathbb{E}\langle X, Y \rangle^2 = \mathbb{E}\|Y\|_2^2 = d$. $\square$

Remark 2.5 *In most cases, observe that we will assume in the rest of this dissertation that the distribution is supported in a centered ball of radius $\mathcal{O}(\sqrt{d})$. Most of the (isotropic) distribution always lies in that ball, as $\mathbb{E}\|X\|_2^2 = d$.*

2.2 The moment method

Let X be a $n \times n$ symmetric random matrix. If we want to bound the spectral norm $\|X\|$, we immediately realize one obstacle is the fact that the map $X \mapsto \|X\|$ is highly nonlinear.

Therefore, it is not obvious how to efficiently relate the distribution of $\|X\|$ to the distribution of the entries X_{ij} . One of the most effective approaches to simplifying this relationship is obtained by applying the following elementary observation.

Lemma 2.6 *Let X be a $n \times n$ symmetric matrix. Then:*

$$\|X\| \asymp \operatorname{Tr}[X^{2p}]^{1/2p} \quad \text{for } p \asymp \log n$$

This observation is very powerful because, unlike $\|X\|$, which is a very complicated function of the entries of X , the quantity $\operatorname{Tr}[X^{2p}]$ is a polynomial in the matrix entries. Therefore, if we turn our attention to estimating the trace, the problem becomes a combinatorial one, much easier to solve. Indeed, the quantity $\mathbb{E}[\operatorname{Tr}[X^{2p}]]$, the $2p$ -th moment of the matrix X , can be evaluated explicitly and studied in details.

So, given this observation, what is the route to study properties of the spectral norm of matrices?

Lemma 2.6 implies that

$$\mathbb{E}[\|X\|^{2p}]^{1/2p} \asymp \mathbb{E}[\operatorname{Tr}[X^{2p}]]^{1/2p} \quad \text{for } p \asymp \log n,$$

This gives a direct route to controlling the spectral norm of a random matrix. This route, essentially, consists of bounding from above the expected norm of a self-adjoint matrix A by $[\mathbb{E}(\operatorname{Tr} A^k)]^{1/k}$, for an even integer k . Various forms of this idea are referred to as the moment method.

Lemma 2.6 can be proved starting from the following question:

Given $x \in \mathbb{R}^n$, we have $\|x\|_p \rightarrow \|x\|_\infty$ as $p \rightarrow \infty$.

Therefore, $\|x\|_p \approx \|x\|_\infty$ for p large.

But, what is the right order to be talking of “ p large enough”?

Lemma 2.7 *If $p \asymp \log n$, then $\|x\|_p \asymp \|x\|_\infty \forall x \in \mathbb{R}^n$.*

Proof (of Lemma 2.7) Observe that:

$$\max_{1 \leq i \leq n} |x_i|^p \leq \sum_{i=1}^n |x_i|^p \leq n \cdot \max_{1 \leq i \leq n} |x_i|^p$$

Therefore, $\|x\|_\infty \leq \|x\|_p \leq n^{1/p} \cdot \|x\|_\infty$. If we impose $n^{1/p} \asymp 1$, we discover that we can achieve this when $\log p \asymp n$. $\square$

Then, we can prove Lemma 2.6 by applying Lemma 2.7.

Proof (of Lemma 2.6) Let $\lambda = (\lambda_1, \dots, \lambda_n)$ be the eigenvalues of X . Then, $\|X\| = \|\lambda\|_\infty$ and $\text{Tr}[X^{2p}]^{1/2p} = \|\lambda\|_{2p} = \sum_{1 \leq i \leq n} |\lambda_i|^{2p}$. The proof follows by applying Lemma 2.7 to λ , the spectrum of the matrix X . $\square$

We will make, now, some additional observations regarding the moment method.

Remark 2.8 *Observe that this method is inherently dimension-dependent (as one must choose $p \sim \log n$), so that, if one is searching for dimension-free bounds, it is generally not well suited.*

2.3 The Laplace transform method

The transform method is a useful technique in developing concentration inequalities for sums of independent random variables in the scalar setting. It is often referred to as the "Bernstein trick" or "Chernoff bounding" due to its simplicity and effectiveness.

In the matrix setting, a good extension of this technique exists, allowing us to establish concentration inequalities for sums of independent random matrices. Similar to the scalar case, the matrix Laplace transform method is straightforward to apply and highly valuable. In this section, we will explain how it works. First, we obtain tail bounds for a random matrix in terms of its (MGF). We can control the tail probabilities of the extreme eigenvalues of a random matrix by producing a bound for the trace of the matrix (MGF). This will constitute the main observation for the Laplace transform method.

Proposition 2.9 *Let X be a random symmetric matrix. For all $t \in \mathbb{R}$:*

$$\mathbb{P}(s_1(X) \geq t) \leq \inf_{\theta > 0} e^{-\theta t} \mathbb{E} \text{Tr}(e^{\theta X})$$

and

$$\mathbb{P}(s_d(X) \leq -t) \leq \inf_{\theta < 0} e^{-\theta t} \mathbb{E} \text{Tr}(e^{\theta X})$$

Proof (of Proposition 2.9) First, we proceed with the bound regarding the maximum eigenvalue $s_1(X)$. Fix $\theta > 0$, and observe that:

$$\begin{aligned}
 \mathbb{P}(s_1(X) \geq t) &= \mathbb{P}\left(e^{\theta s_1(X)} \geq e^{\theta t}\right) \\
 &\leq \frac{\mathbb{E} e^{\theta s_1(X)}}{e^{\theta t}} && \text{by Markov's inequality} \\
 &= \frac{\mathbb{E} e^{s_1(\theta X)}}{e^{\theta t}} \\
 &\leq \frac{\mathbb{E} \operatorname{Tr}(e^{\theta X})}{e^{\theta t}} && \text{since } e^{s_1(\theta X)} = s_1(e^{\theta X}) \leq \operatorname{Tr} e^{\theta X}
 \end{aligned}$$

Now, we can take the infimum over all positive θ , to achieve the tightest possible bound, obtaining the desired result. To prove the result concerning the minimum eigenvalue, we use a similar approach. Fix $\theta < 0$:

$$\begin{aligned}
 \mathbb{P}(s_d(X) \leq t) &= \mathbb{P}\left(e^{\theta s_d(X)} \geq e^{\theta t}\right) \\
 &\leq \frac{\mathbb{E} e^{\theta s_d(X)}}{e^{\theta t}} && \text{by Markov's inequality} \\
 &= \frac{\mathbb{E} e^{s_1(\theta X)}}{e^{\theta t}} \leq \frac{\mathbb{E} \operatorname{Tr}(e^{\theta X})}{e^{\theta t}} && \begin{array}{l} \text{since } s_d(-A) = s_1(A) \\ \text{and } e^{s_1(\theta X)} \leq \operatorname{Tr} e^{\theta X} \end{array}
 \end{aligned}$$

To conclude, we minimize over $\theta < 0$. □

We can adapt the proof of Proposition 2.9 to obtain bounds for the expectation of the maximum eigenvalue of a random symmetric matrix.

Proposition 2.10 *Let X be a random symmetric matrix. Then:*

$$\mathbb{E} s_1(X) \leq \inf_{\theta > 0} \frac{1}{\theta} \log \mathbb{E} \operatorname{Tr} e^{\theta X}$$

and

$$\mathbb{E} s_d(X) \geq \sup_{\theta < 0} \frac{1}{\theta} \log \mathbb{E} \operatorname{Tr} e^{\theta X}$$

Proof (of Proposition 2.10) Fix $\theta > 0$:

$$\begin{aligned}
 \mathbb{E} s_1(X) &= \frac{1}{\theta} \mathbb{E} \log e^{s_1(\theta X)} \\
 &\leq \frac{1}{\theta} \log \mathbb{E} e^{s_1(\theta X)} && \text{by Jensen's inequality} \\
 &\leq \frac{1}{\theta} \log \mathbb{E} \operatorname{Tr} e^{\theta X} && \text{since } e^{s_1(\theta X)} = s_1(e^{\theta X}) \leq \operatorname{Tr} e^{\theta X}
 \end{aligned}$$

Then, we take the infimum over $\theta > 0$. The expectation bound for the minimum eigenvalues is obtained in an analogous way. □

2.4 The random process method

The so-called random process method relies on the following elementary statement.

Lemma 2.11 *Let X be a $n \times n$ symmetric matrix. Then:*

$$\|X\| = \sup_{v \in B_2^n} |\langle v, Xv \rangle|,$$

where B_2^n denotes the Euclidean unit ball in $\mathbb{R}^n$.

When X is a symmetric random matrix, we can view $v \mapsto \langle v, Xv \rangle$ as a random process, indexed by the Euclidean unit ball. Thus controlling the expected spectral norm of X is none other than a special instance of the general probabilistic problem of controlling the expected supremum of a random process.

The simplest approach to bounding the suprema of random processes is the ε -net method.

ε -net method

The idea behind this approach is to approximate the supremum over the unit ball B by the maximum over a finite discretization B_ε of the unit ball, which reduces the problem to computing the maximum of a finite number of random variables. Indeed, we know that when one wants to estimate, for instance, $\mathbb{P}\left(\sup_{v \in B_2^n} f(v) \geq t\right)$, for some function $f(v)$, which in this case it would be $f(v) = |\langle v, Xv \rangle|$, it so evaluate the union of such event, $\forall v$. When other approaches don't apply, one strategy for estimating such a union is to use countable subadditivity and upper bound it by a sum of the individual probabilities. However, for that, you need the union to be countable, which is why it is necessary to introduce the finite discretization B_ε .

As direct control of the random processes that arise from structured random matrices is largely intractable, a different approach is often needed. To this end, the key idea is to use comparison theorems to bound the expected supremum of one random process by that of another random process. The basic idea is to design a suitable comparison process that dominates the random process $\|X\| = \sup_{v \in B_2^n} |\langle v, Xv \rangle|$ but that is easier to control. Classical comparison theorems for Gaussian processes, for instance, are due to Slepian and Fernique and we will see them in more detail in the next chapters.

2.5 Bounds for Singular Values via Net Arguments

In this section, we will elaborate further on the concept of an ε -net and explore how it can be utilized to establish bounds on singular values. The

content discussed in this section primarily relies on the material presented in [49].

Definition 2.12 Let (T, d) be a metric space. Consider a compact set $K \subset T$ and fix $\varepsilon > 0$. A subset $\mathcal{N}_\varepsilon \subset K$ is an ε -net of K if:

$$\forall k \in K, \exists k_0 \in \mathcal{N}_\varepsilon : d(k, k_0) \leq \varepsilon$$

Definition 2.13 The covering number $\mathcal{N}(K, d, \varepsilon)$ is the smallest possible cardinality of an ε -net $\mathcal{N}_\varepsilon$ of K .

Alternatively, $\mathcal{N}(K, d, \varepsilon)$ can be defined as the minimum number of closed balls, each with a radius of ε and centered in K , needed to completely envelop K . From now on, we will omit the metric d when we use the standard Euclidean distance.

Definition 2.14 A subset $\mathcal{P}$ of a metric space (T, d) is ε -separated if, for $x, y \in \mathcal{P}$, $d(x, y) \geq \varepsilon$ whenever $x \neq y$.

The packing number $\mathcal{P}(K, d, \varepsilon)$ is the largest cardinality of a ε -separated set of a given subset $K \subset T$.

Observe that covering numbers and packing numbers are finite. As noted in subsection (2.4), where we discussed the reasons for introducing a finite discretization of a set K , it is crucial to be able to bound the size of this finite discretization.

Lemma 2.15 The unit Euclidean sphere S^{d-1} equipped with the Euclidean metric satisfies for every $\varepsilon > 0$ that

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

Proof (of Lemma 2.15) Let us fix $\varepsilon > 0$ and choose $\mathcal{N}_\varepsilon$ to be a maximal ε -separated subset of S^{d-1} .

Maximal ε -separated net, means that $\mathcal{N}_\varepsilon$ is such that $d(x, y) \geq \varepsilon$ for all $x, y \in \mathcal{N}_\varepsilon$, $x \neq y$, and no subset of S^{d-1} containing $\mathcal{N}_\varepsilon$ satisfies this property. The maximality property implies that $\mathcal{N}_\varepsilon$ is an ε -net of S^{d-1} . Indeed, if this was not the case, there would exist $x \in S^{d-1}$ 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, by the triangle inequality, the separation property implies that the balls of radii $\frac{\varepsilon}{2}$ centered at the points in $\mathcal{N}_\varepsilon$ are disjoint. On the other hand, all such balls lie in $(1 + \frac{\varepsilon}{2})B_2^d$. Then, comparing the volume gives:

$$\text{vol}\left(\frac{\varepsilon}{2}B_2^d\right) \cdot |\mathcal{N}_\varepsilon| \leq \text{vol}\left(\left(1 + \frac{\varepsilon}{2}\right)B_2^d\right)$$

2. SETTING AND OVERVIEW OF THE METHODS

Since $\text{vol}(rB_2^d) = r^d \text{vol}(B_2^d)$ for all $r \geq 0$, we conclude that

$$|\mathcal{N}_\varepsilon| \leq \frac{(1+\frac{\varepsilon}{2})^d}{(\frac{\varepsilon}{2})^d} = \left(1 + \frac{2}{\varepsilon}\right)^d \text{ as wanted.} \quad \square$$

Now, we will see how we can compute the spectral norm using an ε -net. Indeed, recalling the definition of the spectral norm (1.1), to evaluate it one needs to take the supremum over the whole sphere $\mathbb{S}^{d-1}$. However, one can essentially replace the sphere by its ε -net. We will consider the case of symmetric matrices as in this dissertation we will always deal with such a type of matrices. Indeed, we can always adapt to this case using Hermitian dilation.

Definition 2.16 *Given a general matrix $A \in \mathbb{R}^{d_1 \times d_2}$, the Hermitian dilation is the map from a general matrix to a symmetric matrix defined by*

$$\tilde{A} = \begin{bmatrix} 0 & A \\ A^T & 0 \end{bmatrix}$$

Observe that the Hermitian dilation is a real-linear map.

Furthermore, the dilation retains important spectral information. Indeed, note that the square of the dilation satisfies:

$$(\tilde{A})^2 = \begin{bmatrix} AA^T & 0 \\ 0 & A^T A \end{bmatrix}$$

Therefore, the squared eigenvalues of $\tilde{A}$ coincide with the squared singular values of A , along with an appropriate number of zeros. As a consequence, $\|\tilde{A}\| = \|A\|$.

Lemma 2.17 *Let A be a symmetric $d \times d$ matrix, and let $\mathcal{N}_\varepsilon$ be an ε -net of $\mathbb{S}^{d-1}$ for some $\varepsilon \in [0, 1)$. Then:*

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

Proof (of Lemma 2.17) Take $x \in \mathbb{S}^{d-1}$ 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:

$$\begin{aligned} |\langle Ax, x \rangle - \langle Ay, y \rangle| &= |\langle Ax, x - y \rangle + \langle A(x - y), y \rangle| \\ &\leq \|A\| \|x\|_2 \|x - y\|_2 + \|A\| \|x - y\|_2 \|y\|_2 \leq 2\varepsilon \|A\| \end{aligned}$$

Therefore: $|\langle Ay, y \rangle| \geq |\langle Ax, x \rangle| - 2\varepsilon \|A\| = (1 - 2\varepsilon) \|A\|$.

Finally, we take the maximum over all $y \in \mathcal{N}_\varepsilon$ in this inequality, thus completing the proof. $\square$

We will see, now, an example of how to use the net argument to find a bound on the singular values of a matrix with independent $\mathcal{N}(0, 1)$ entries.

Example 2.18 (Bound on the spectral norm of a matrix with i.i.d. gaussian entries)
 Let G be a random $n \times m$ matrix with entries being independent $\mathcal{N}(0, 1)$ random variables. It holds that:

$$\mathbb{E}\|G\| \leq C(\sqrt{n} + \sqrt{m})$$

Indeed, observe $\|G\| = \sup_{v \in \mathbb{S}^{m-1}} \|Gv\|_2 = \sup_{\substack{u \in \mathbb{S}^{n-1} \\ v \in \mathbb{S}^{m-1}}} u^T Gv$. Computing the spectral norm involves the maximization of a certain quantity over the sphere $\mathbb{S}^{m-1}$. Now, as we discussed above, we want to perform the maximization over a finite subset $\mathcal{N} \subset \mathbb{S}^{m-1}$.

Therefore, let $\mathcal{N}$ be a ε -net of $\mathbb{S}^{m-1}$. Using Lemma 2.17:

$$\left\| \frac{1}{n} G^T G - I \right\| \leq \frac{1}{1-2\varepsilon} \sup_{u \in \mathcal{N}} \left| u^T \frac{1}{n} (G^T G - I) u \right| = \frac{1}{1-2\varepsilon} \sup_{u \in \mathcal{N}} \left| \frac{1}{n} \|Gu\|_2^2 - 1 \right|$$

Recall that G_{ij} are i.i.d $\mathcal{N}(0, 1)$ for all $i \in \{1, \dots, n\}$, $j \in \{1, \dots, m\}$ and that $\|u\|_2 = 1 \forall u \in \mathcal{N}$.

Thus, $\frac{1}{n} \|Gu\|_2^2 - 1 = \frac{1}{n} \sum (g_i^2 - 1)$, where g_i 's are i.i.d. $\mathcal{N}(0, 1)$. Moreover, observe that $\sum_{i=1}^n g_i^2$ follows a χ^2 distribution with n degrees of freedom. Therefore, we will need a concentration inequality for χ^2 random variables.

Proposition 2.19 Let $g_1, \dots, g_n$ be i.i.d $\mathcal{N}(0, 1)$. For all $t \in [0, 1]$:

$$\mathbb{P} \left(\left| \frac{1}{n} \sum_i (g_i^2 - 1) \right| \geq t \right) \leq 2 \exp \left(-n \frac{t^2}{8} \right)$$

Proof (of Proposition 2.19) Note that:

$$\begin{aligned} \mathbb{P} \left(\left| \frac{1}{n} \sum_{i=1}^n (g_i^2 - 1) \right| \geq t \right) &\leq \mathbb{P} \left(\frac{1}{n} \sum_{i=1}^n (g_i^2 - 1) \geq t \right) \\ &\quad + \mathbb{P} \left(\frac{1}{n} \sum_{i=1}^n (g_i^2 - 1) \leq -t \right) \end{aligned}$$

Thus, it suffices to bound the two terms on the right-hand side of the inequality above separately. To bound the first, observe that by Markov's inequality, for any $\lambda \geq 0$, we have:

$$\begin{aligned} \mathbb{P} \left(\frac{1}{n} \sum_{i=1}^n (g_i^2 - 1) \geq t \right) &= \mathbb{P} \left[\exp \left(\lambda \sum_{i=1}^n (g_i^2 - 1) \right) \geq \exp(\lambda n t) \right] \\ &\leq \exp(-\lambda n t) \cdot \mathbb{E} \left[\exp \left(\lambda \sum_{i=1}^n (g_i^2 - 1) \right) \right] \end{aligned}$$

2. SETTING AND OVERVIEW OF THE METHODS

Since $g_1, \dots, g_n$ are *i.i.d.*, we have:

$$\begin{aligned} \mathbb{E} \left[\exp \left(\lambda \sum_{i=1}^n (g_i^2 - 1) \right) \right] &= \prod_{i=1}^n \mathbb{E} [\exp (\lambda (g_i^2 - 1))] \\ &= (\mathbb{E} [\exp (\lambda (g_1^2 - 1))])^n \end{aligned}$$

Using the density function of a standard Gaussian random variable, for any $\lambda < \frac{1}{2}$, we have:

$$\begin{aligned} \mathbb{E} [\exp (\lambda (g_1^2 - 1))] &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp (\lambda (x^2 - 1)) \cdot \exp \left(-\frac{x^2}{2} \right) dx \\ &= \frac{\exp(-\lambda)}{\sqrt{1-2\lambda}} \cdot \frac{\sqrt{1-2\lambda}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp \left(-\frac{x^2}{2} \sqrt{1-2\lambda} \right) dx \end{aligned}$$

Observe that $x \mapsto \frac{1}{\sqrt{2\pi}} \sqrt{1-2\lambda} \exp \left(-\frac{x^2}{2} \sqrt{1-2\lambda} \right)$ is the density function of a centered Gaussian random variable with variance $\frac{1}{1-2\lambda}$, for any $\lambda < \frac{1}{2}$. Using the inequality $-x - x^2 \leq \ln(1-x)$, which is valid for any $x \in [0, \frac{1}{2}]$, we obtain, for any $\lambda \in [0, \frac{1}{4}]$:

$$\mathbb{E} [\exp (\lambda (g_1^2 - 1))] = \frac{\exp(-\lambda)}{\sqrt{1-2\lambda}} \leq \exp (2\lambda^2)$$

Therefore, we get, for any $\lambda \in [0, \frac{1}{4}]$:

$$\mathbb{P} \left(\frac{1}{n} \sum_{i=1}^n (g_i^2 - 1) \geq t \right) \leq \exp (n (2\lambda^2 - \lambda t))$$

By setting $\lambda = \frac{t}{4}$, which is an admissible choice since $\lambda \in [0, \frac{1}{4}]$ is satisfied whenever $t \in [0, 1]$, we obtain:

$$\mathbb{P} \left(\frac{1}{n} \sum_{i=1}^n (g_i^2 - 1) \geq t \right) \leq \exp \left(-n \frac{t^2}{8} \right)$$

To bound the second term, we proceed in a similar manner. Specifically, by Markov's inequality, for any $\lambda \geq 0$, we have:

$$\mathbb{P} \left(\frac{1}{n} \sum_{i=1}^n (g_i^2 - 1) \leq -t \right) \leq \exp(-\lambda dt) \cdot (\mathbb{E} [\exp (\lambda (1 - g_1^2))])^n$$

From the inequality $(1 + 2x)^{-1/2} \leq \exp (2x^2 - x)$, which holds for all $x \geq 0$, a simple calculation yields:

$$\mathbb{E} [\exp (\lambda (1 - g_1^2))] = \frac{\exp(\lambda)}{\sqrt{1+2\lambda}} \leq \exp (2\lambda^2)$$

Using the same argument as before, we obtain, for any $t \in [0, 1]$:

$$\mathbb{P} \left(\frac{1}{n} \sum_{i=1}^n (g_i^2 - 1) \leq -t \right) \leq \exp \left(-m \frac{t^2}{8} \right)$$

This completes the proof. $\square$

Now, going back to the original problem, finding upper and lower bounds on the singular values of a Gaussian random matrix, we proceed in the following way. Let $\mathcal{N}$ be a $\frac{1}{4}$ -net of $\mathbb{S}^{n-1}$. By the decomposition in χ^2 random variables, and using Proposition 2.19 and Lemma 2.15, we have:

$$\begin{aligned} \mathbb{P} \left(\left\| \frac{1}{n} G^T G - I \right\| \geq 2t \right) &\leq \mathbb{P} \left(\sup_{u \in \mathcal{N}} \left| \frac{1}{n} \|Gu\|_2^2 - 1 \right| \geq t \right) \\ &\leq \sum_{u \in \mathcal{N}} \mathbb{P} \left(\left| \frac{1}{n} \|Gu\|_2^2 - 1 \right| \geq t \right) \\ &\leq 2 \cdot 9^n \cdot \exp \left(-m \frac{t^2}{8} \right) \end{aligned}$$

By setting $t^2 = \frac{8 \ln 9}{n} (m + \eta^2)$ for some $\eta \geq 0$ and noting that $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for any $a, b \geq 0$, we get:

$$\mathbb{P} \left(\left\| \frac{1}{n} G^T G - I \right\| \geq 2 \cdot \sqrt{\frac{8 \ln 9}{n}} \cdot (\sqrt{m} + \eta) \right) \leq 2 \exp(-\ln 9 \cdot \eta^2)$$

It follows that when $\eta > 0$ is sufficiently large, we have

$$\sqrt{n} - c(\sqrt{m} + \eta) \leq s_1(G) \leq \sqrt{n} + c(\sqrt{m} + \eta)$$

with high probability, where $c = 2\sqrt{8 \ln 9}$.

2.6 Sub-gaussian, sub-exponential, and Orlicz Norms of random vectors

We search for bounds that inform how a random variable X concentrates near the mean, in other words, an expression of the form:

$$\mathbb{P} (|X - \mu| > t) \leq f(t),$$

where μ is the mean of the random variable X and $f(t)$ is an error function that converges fast, possibly exponentially, to zero.

The Gaussian distribution exhibits a notably rapid decay.

Proposition 2.20 (*Tails for the Standard Gaussian Distribution*). *Let g be a standard Gaussian variable. Then, for every $t > 0$:*

$$\mathbb{P}(|g| \geq t) \leq \min \left\{ 1, \sqrt{\frac{2}{\pi}} \frac{1}{t} \right\} e^{-t^2/2}$$

Proof (of Proposition 2.20) By symmetry, we have:

$$\mathbb{P}(|g| \geq t) = \frac{2}{\sqrt{2\pi}} \int_t^\infty e^{-x^2/2} dx$$

A change of variables yields:

$$\int_t^\infty e^{-x^2/2} dx = \int_0^\infty e^{-(x+t)^2/2} dx = e^{-t^2/2} \int_0^\infty e^{-xt} e^{-x^2/2} dx$$

From the estimates $e^{-xt} \leq 1$ and $e^{-x^2/2} \leq 1$ we obtain either:

$$\int_t^\infty e^{-x^2/2} dx \leq e^{-t^2/2} \int_0^\infty e^{-x^2/2} dx = \sqrt{\frac{\pi}{2}} e^{-t^2/2}$$

or

$$\int_t^\infty e^{-x^2/2} dx \leq e^{-t^2/2} \int_0^\infty e^{-xt} dx = \frac{1}{t} e^{-t^2/2}. \quad \square$$

Therefore, the standard Gaussian distribution concentrates heavily around the mean and an analogous result holds for Gaussian random variables. Motivated by the strong concentration of Gaussian variables, we focus our study on a specific category of random variables that display fast decay. It would be beneficial to broaden these findings to encompass a more diverse range of distributions. At a minimum, we expect that the normal distribution should be included in this class.

The sub-gaussian distribution

The following proposition states that some properties of such variables: the proofs follow [49].

Proposition 2.21 *Let X be a random variable. Then the following properties are equivalent. Each holds for a constant $K_i > 0$.*

(i) *The tails of X satisfy:*

$$\forall t \geq 0, \quad \mathbb{P}(|X| \geq t) \leq 2e^{-t^2/K_1}$$

(ii) *The moments of X satisfy:*

$$\forall p \geq 1, \quad \|X\|_p \leq K_2 \sqrt{p}$$

(iii) *The (MGF) of X^2 satisfies:*

2.6. Sub-gaussian, sub-exponential, and Orlicz Norms of random vectors

$$\forall \lambda \in \mathbb{R}, |\lambda| \leq K_3^{-1}, \quad \mathbb{E}e^{\lambda^2 X^2} \leq e^{K_3^2 \lambda^2}$$

(iv) The (MGF) of X^2 is bounded at some point:

$$\mathbb{E}e^{X^2/K_4^2} \leq 2$$

(v) If X has zero mean, the following is also equivalent to all above:

$$\forall \lambda \in \mathbb{R}, \quad \mathbb{E}e^{\lambda X} \leq e^{K_5^2 \lambda^2}$$

Proof (of Proposition 2.21)

- (i) $\Rightarrow$ (ii) Suppose property (i) holds. By homogeneity, we may assume $K_1 = 1$. Then:

$$\begin{aligned} \mathbb{E}|X|^p &= \int_0^\infty \mathbb{P}(|X|^p \geq u) du && \text{by integral layer cake identity} \\ &= \int_0^\infty \mathbb{P}(|X| \geq t) p t^{p-1} dt && \text{from the change of variables } u = t^p \\ &\leq \int_0^\infty 2e^{-t^2} p t^{p-1} dt && \text{from Property (i)} \\ &= p\Gamma(p/2) && \text{by setting } t^2 = s \text{ in the expression for the Gamma function} \\ &\leq p(p/2)^{p/2} && \text{as } \Gamma(x) \leq x^x \end{aligned}$$

Now take the p -th root to get (ii) with $K_2 \leq 2$.

- (ii) $\Rightarrow$ (iii) Assume again by homogeneity that $K_2 = 1$. From the Taylor series expansion of the exponential:

$$\mathbb{E}e^{\lambda^2 X^2} = 1 + \sum_{p=1}^\infty \frac{\lambda^{2p} \mathbb{E}[X^{2p}]}{p!},$$

from property (ii), $\mathbb{E}[X^{2p}] \leq (2p)^p$ and by the expansion of e^p , $p! \geq (p/e)^p$. Now, substitute these bounds into the Taylor expansion:

$$\mathbb{E}e^{\lambda^2 X^2} \leq 1 + \sum_{p=1}^\infty \frac{(2\lambda^2 p)^p}{(p/e)^p} = \sum_{p=0}^\infty (2e\lambda^2)^p = \frac{1}{1-2e\lambda^2}, \quad \text{if } |\lambda| < \frac{1}{\sqrt{2e}}$$

In the last step, we used the convergence of the geometric series. For $x \in [0, \frac{1}{2}]$, the inequality $\frac{1}{1-x} \leq e^{2x}$ is valid and then:

$$\mathbb{E}e^{\lambda^2 X^2} \leq e^{4e\lambda^2}, \quad \text{if } |\lambda| \leq \frac{1}{\sqrt{2e}}$$

We thus obtain (iii) for $K_3 = \sqrt{2e}$.

- (iii) $\Rightarrow$ (iv) This implication is straightforward, just take λ sufficient small such that $e^{K_3^2 \lambda^2} \leq 2$.
- (iv) $\Rightarrow$ (i) By homogeneity assume $K_4 = 1$. Then:

$$\begin{aligned}
 \mathbb{P}(|X| \geq t) &= \mathbb{P}\left(e^{X^2} \geq et^2\right) \\
 &\leq e^{-t^2} \mathbb{E}\left[e^{X^2}\right] && \text{by Markov's inequality} \\
 &\leq 2e^{-t^2} && \text{by property (iv)}
 \end{aligned}$$

We obtain (i) with $K_1 = 1$.

• (iii) $\Rightarrow$ (v)

Assume (iii) holds with $K_3 = 1$. Since $e^x \leq x + e^{x^2}$ for all real x :

$$\begin{aligned}
 \mathbb{E}e^{\lambda X} &\leq \mathbb{E}\left[\lambda X + e^{\lambda^2 X^2}\right] \\
 &= \mathbb{E}e^{\lambda^2 X^2} && \text{since we assume } \mathbb{E}X = 0 \\
 &\leq e^{\lambda^2}, && \text{if } |\lambda| \leq 1
 \end{aligned}$$

Therefore, for $|\lambda| \leq 1$, we get the desired bound. Now suppose that $|\lambda| > 1$. Use $2\lambda x \leq \lambda^2 + x^2$ for all λ and x to obtain:

$$\mathbb{E}e^{\lambda X} \leq e^{\lambda^2/2} \mathbb{E}e^{X^2/2} \leq e^{\lambda^2/2} e^{1/2} \leq e^{\lambda^2}$$

The second inequality follows from property (iii) and the last one from the fact that λ has absolute value greater than one. Thus (iii) $\Rightarrow$ (v) with $K_5 = 1$.

• (v) $\Rightarrow$ (i) For this, we use a smart trick, the so-called Laplace transform method (Section (2.3)). Let $\lambda > 0$ be a parameter to be chosen later. Write:

$$\begin{aligned}
 \mathbb{P}(X \geq t) &= \mathbb{P}\left(e^{\lambda X} \geq e^{\lambda t}\right) \\
 &\leq e^{-\lambda t} \mathbb{E}e^{\lambda X} && \text{by Markov's inequality} \\
 &\leq e^{-\lambda t} e^{\lambda^2} && \text{by Property (v)} \\
 &= e^{-\lambda t + \lambda^2} \\
 &= e^{-t^2/4} && \text{minimizing for } \lambda : \lambda = \frac{t}{2}
 \end{aligned}$$

Apply the same method to $-X$ to obtain $\mathbb{P}(X \leq -t) \leq e^{-t^2/4}$. Using the union bound we finally obtain (i) with $K_1 = 2$. $\square$

Definition 2.22 (*Sub-gaussian random variables, subgaussian norm*) Any random variable that satisfies properties (i)-(iv) is called sub-gaussian random variable. The sub-gaussian norm $\|X\|_{\psi_2}$ is:

$$\|X\|_{\psi_2} = \inf \left\{ t > 0 : \mathbb{E}e^{X^2/t^2} \leq 2 \right\}$$

From property (iv), X is sub-gaussian if and only if $\|X\|_{\psi_2}$ is finite.

Example 2.23 Some classical examples of sub-gaussian random variables are:

1. (Gaussians) If $X \sim N(0, \sigma^2)$, Proposition 2.20 and homogeneity implies that $\|X\|_{\psi_2} \leq C\sigma$, for an absolute constant C . By the triangular inequality, $X \sim N(\mu, \sigma^2)$, $\|X\|_{\psi_2} \leq C\sigma + |\mu| \frac{1}{\sqrt{\ln 2}}$
2. (Rademacher) Let X be a random variable with Rademacher distribution. Since $|X| = 1$, it follows that X is a sub-gaussian random variable with $\|X\|_{\psi_2} = \frac{1}{\sqrt{\ln 2}}$
3. (Bounded variables) Any bounded random variable X is subgaussian with $\|X\|_{\psi_2} \leq C\|X\|_{\infty}$, for $C = \frac{1}{\sqrt{\ln 2}}$.

Remark 2.24 The definition of a subgaussian random variable X does not require X to be centered. In any case, one can always center X using the simple fact that if X is subgaussian, then so is $X - \mathbb{E}X$. Moreover, $\|X - \mathbb{E}X\|_{\psi_2} \leq 2\|X\|_{\psi_2}$. This follows by the triangle inequality $\|X - \mathbb{E}X\|_{\psi_2} \leq \|X\|_{\psi_2} + \|\mathbb{E}X\|_{\psi_2}$, with $\|\mathbb{E}X\|_{\psi_2} = \|\mathbb{E}X\| \leq \mathbb{E}|X| \leq \|X\|_{\psi_2}$.

Moreover, the sum of sub-gaussian random variables is also sub-gaussian. This property is also called "rotation invariance". This is a remarkable property that, from the Gaussian distribution, passes onto sub-gaussian random variables.

Proposition 2.25 (Sum of sub-gaussian random variables) Let $X_1, \dots, X_n$ be independent, mean zero, sub-gaussian random variables. Then:

$$\|\sum_{i=1}^n X_i\|_{\psi_2}^2 \leq C \sum_{i=1}^n \|X_i\|_{\psi_2}^2$$

Proof (of Proposition 2.25) We consider again the (MGF). For $\forall \lambda > 0$:

$$\begin{aligned} \mathbb{E}e^{\lambda \sum_{i=1}^n X_i} &= \prod_{i=1}^n \mathbb{E}e^{\lambda X_i} && \text{by independence} \\ &\leq \prod_{i=1}^n \mathbb{E}e^{C\lambda^2 \|X_i\|_{\psi_2}^2} && \text{from Properties (ii)-(v)} \\ &= e^{K\lambda^2}, \end{aligned}$$

for $K = C \sum_{i=1}^n \|X_i\|_{\psi_2}^2$.

The sum $\sum_{i=1}^n X_i$ has zero mean by the linearity of the expected value. By Properties (ii)-(v) and the equivalences in Proposition 2.21,

$$\|\sum_{i=1}^n X_i\|_{\psi_2}^2 \leq K. \quad \square$$

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

2. SETTING AND OVERVIEW OF THE METHODS

Proposition 2.26 (Hoeffding-type inequality) *Let $X_1, \dots, X_n$ be independent centered sub-gaussian random variables, and let $K = \max_i \|X_i\|_{\psi_2}$. Then for every $a = (a_1, \dots, a_n) \in \mathbb{R}^n$ and every $t \geq 0$, we have:*

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

where $c > 0$ is an absolute constant.

Proof (of Proposition 2.26) Lemma 2.25 implies the bound $\|\sum_i a_i X_i\|_{\psi_2}^2 \leq C \sum_i a_i^2 \|X_i\|_{\psi_2}^2 \leq CK^2 \|a\|_2^2$. The first claim of Proposition 2.21 yields the required tail decay. $\square$

Remark 2.27 *Observe that the number 2 in the Definition 2.22 can be replaced by another constant. It arises naturally in many scenarios.*

The sub-exponential distribution

When the tails are slightly heavier than the Gaussian tails, we can still mime the characterizations of sub-gaussian distributions as in Proposition 2.21 to obtain concentration bounds in the so-called sub-exponential case.

Proposition 2.28 *Let X be a random variable. Then the following properties are equivalent (each holds for some $K_i > 0$).*

(i) *The tails of X satisfy:*

$$\forall t > 0, \quad \mathbb{P}(|X| \geq t) \leq 2e^{-t/K_1}$$

(ii) *The moments of X satisfy:*

$$\forall p \geq 1, \quad \|X\|_p \leq K_2 p$$

(iii) *The (MGF) of $|X|$ satisfies:*

$$\forall \lambda \in \mathbb{R}, \quad 0 \leq \lambda \leq K_3^{-1}, \quad \mathbb{E}e^{\lambda|X|} \leq e^{K_3 \lambda}$$

(iv) *The (MGF) of $|X|$ is bounded at some point:*

$$\mathbb{E}e^{|X|/K_4} \leq 2$$

(v) *If X is mean zero, the following is also equivalent to all the above:*

$$\forall \lambda \in \mathbb{R}, \quad |\lambda| \leq K_5^{-1}, \quad \mathbb{E}e^{\lambda X} \leq e^{K_5^2 \lambda^2}$$

Proof (of Proposition 2.28) We only prove that (ii) is equivalent to (v), the other properties can be proved using the same arguments in the proof of Proposition 2.21. Suppose that (ii) holds with $K_2 = 1$ and λ is positive (if it is zero the inequality is trivial). By the Taylor expansion:

$$\mathbb{E}e^{\lambda X} = 1 + \lambda \mathbb{E}X + \sum_{p=2}^{\infty} \frac{\lambda^p \mathbb{E}[X^p]}{p!} = 1 + \sum_{p=2}^{\infty} \frac{\lambda^p \mathbb{E}[X^p]}{p!},$$

since X has zero mean. Use the estimates $\mathbb{E}[X^p] \leq (p)^p$ and $p! \geq (p/e)^p$ and:

$$\begin{aligned}
 \mathbb{E}e^{\lambda X} &\leq 1 + \sum_{p=2}^{\infty} \frac{(\lambda p)^p}{(p/e)^p} \\
 &= 1 + \frac{(e\lambda)^2}{1 - e\lambda} && \text{provided that } |e\lambda| < 1 \\
 &\leq 1 + 2e^2\lambda^2 && \text{if } |e\lambda| \leq \frac{1}{2} \\
 &\leq e^{2e^2\lambda^2} && \text{from the elementary inequality } 1 + x \leq e^x
 \end{aligned}$$

Now apply the bound above for $-X$ and conclude that it holds for all λ with absolute value less than $\frac{1}{2e}$ (the right-hand side does not change). We now assume that (v) holds with $K_5 = 1$. Consider the numeric inequality that is valid for every x and $p > 1$:

$$|x|^p \leq p^p (e^x + e^{-x})$$

To see that the inequality holds, divide both sides by p^p and compare the Taylor series. Applying this inequality:

$$\mathbb{E}|X|^p \leq p^p (\mathbb{E}e^X + \mathbb{E}e^{-X}) \leq 2p^p$$

The last inequality follows from Property (v) applied twice with $\lambda = 1$ and $\lambda = -1$. Taking the p -root, concludes the proof. $\square$

Definition 2.29 (*Sub-exponential random variable, sub-exponential norm*). A random variable X is sub-exponential if it obeys (i)-(iv). By Proposition 2.28, X is sub-exponential if and only if the following norm is finite:

$$\|X\|_{\psi_1} = \inf \left\{ t > 0 : \mathbb{E}e^{|X|/t} \leq 2 \right\}$$

There is the following relation between sub-exponential and sub-gaussian random variables.

Proposition 2.30 (*Sub-exponentials are squares of sub-gaussians*). A random variable X is sub-gaussian if and only if X^2 is sub-exponential. Moreover:

$$\|X\|_{\psi_2}^2 = \|X^2\|_{\psi_1} \leq 2\|X\|_{\psi_2}^2$$

Proof (of Proposition 2.30) Indeed, $\|X^2\|_{\psi_1}$ is the infimum of $t_1 > 0$ such that $\mathbb{E}e^{X^2/t_1} \leq 2$ and $\|X\|_{\psi_2}$ is the infimum of $t_2 > 0$ such that $\mathbb{E}e^{X^2/t_2^2} \leq 2$. We get that $t_1 = t_2^2$. The inequality follows from the definition. $\square$

2. SETTING AND OVERVIEW OF THE METHODS

Sub-exponential random variables satisfy a large deviation inequality similar to the one for sub-gaussians stated in Proposition 2.26.

Proposition 2.31 (*Bernstein-type inequality*). *Let $X_1, \dots, X_n$ be independent centered sub-exponential random variables, and $K = \max_i \|X_i\|_{\psi_1}$. Then for every $a = (a_1, \dots, a_n) \in \mathbb{R}^n$ and every $t \geq 0$, we have:*

$$\mathbb{P} \left\{ \left| \sum_{i=1}^n a_i X_i \right| \geq t \right\} \leq 2 \exp \left[-c \min \left(\frac{t^2}{K^2 \|a\|_2^2}, \frac{t}{K \|a\|_\infty} \right) \right],$$

where $c > 0$ is an absolute constant.

Proof (of Proposition 2.31) Without loss of generality, we can assume that $K = 1$ by replacing X_i with $\frac{X_i}{K}$ and t with $\frac{t}{K}$. We use Markov's inequality for the sum $S = \sum_i a_i X_i$ and for $\lambda > 0$:

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

Note that, if $|\lambda| \leq c/\|a\|_\infty$ then $|\lambda a_i| \leq c$ for all i , so by (v) of Proposition 2.28, we get:

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

Therefore, by choosing $\lambda = \min\left(\frac{t}{2C\|a\|_2^2}, \frac{c}{\|a\|_\infty}\right)$, we obtain that:

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

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. $\square$

For convenience, we restate the Bernstein bound in a different form.

Theorem 2.32 (*Bernstein inequality - Usual form*) *Let $X_1, \dots, X_n$ be independent centered subexponential random variables with $\|X_i\|_{\psi_1} \leq K$. Then, for every $\varepsilon \geq 0$:*

$$\mathbb{P}(|\sum_{i=1}^n X_i| \geq \varepsilon n) \leq 2 \exp \left(-c \min \left(\frac{\varepsilon^2}{K^2}, \frac{\varepsilon}{K} \right) n \right),$$

where $c > 0$ is an absolute constant.

Proof (of Theorem 2.32) We use Proposition 2.31 with $a_i = 1$ and $t = \varepsilon n$ and the result follows. $\square$

The Orlicz norms

We have defined sub-gaussian and sub-exponential variables in terms of bounds on the (MGF). Following Remark 2.27, sub-gaussian and sub-exponential distributions can be introduced within a more general framework of Orlicz spaces.

Definition 2.33 A function $\psi : [0, \infty) \mapsto [0, \infty)$ is called an Orlicz function if ψ is convex, increasing and it satisfies:

$$\psi(0) = 0, \quad \psi(x) \rightarrow \infty \text{ as } x \rightarrow \infty$$

Definition 2.34 For a given Orlicz function ψ , the Orlicz norm of a random variable X is defined as:

$$\|X\|_\psi = \inf \left\{ t > 0 : \mathbb{E} \psi \left(\frac{|X|}{t} \right) \leq 1 \right\}$$

The Orlicz space $L_\psi = L_\psi(\Omega, \mathcal{A}, \mathbb{P})$ consists of all random variables X on the probability space $(\Omega, \mathcal{A}, \mathbb{P})$ with finite Orlicz norm, i.e.:

$$L_\psi = \{X : \|X\|_\psi < \infty\}$$

To see how this definition includes the sub-gaussian and sub-exponential random variables, consider the function:

$$\psi_\alpha = e^{x^\alpha} - 1$$

This is an Orlicz function, from Definition 2.33 and observe that for $\alpha = 2$ and $\alpha = 1$ we recover, respectively, Definition 2.22 and Definition 2.29. Equivalently, the Orlicz norm for the function ψ_α can be defined as:

$$\|X\|_{\psi_\alpha} = \sup_{p \geq 1} \frac{\|X\|_p}{p^{1/\alpha}} \quad (2.1)$$

Generalization to Random Vectors

It is possible to generalize the above definitions for random vectors $X \in \mathbb{R}^d$. It is well-known that the multivariate normal distribution can be characterized through its one-dimensional marginals: a random vector X has a normal distribution in $\mathbb{R}^d$ if and only if the one-dimensional marginals $\langle X, x \rangle$ are normal for all $x \in \mathbb{R}^d$.

Guided by this characterization, it becomes intuitive to establish the definition of multivariate sub-gaussian, sub-exponential distributions as follows.

Definition 2.35 (Sub-gaussian random vectors). A random vector X in $\mathbb{R}^d$ is called sub-gaussian if the one-dimensional marginals $\langle X, x \rangle$ are sub-gaussian random variables for all $x \in \mathbb{R}^d$. The sub-gaussian norm of X is defined as:

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

A good example of a sub-gaussian random vector is a random vector with independent, sub-gaussian coordinates.

Lemma 2.36 (Sub-gaussian distributions with independent coordinates). Let $X = (X_1, \dots, X_n) \in \mathbb{R}^n$ be a random vector with independent, mean zero, subgaussian coordinates X_i . Then X is a sub-gaussian random vector, and:

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

Proof (of Lemma 2.36) This is a consequence of Proposition 2.30. Indeed, for a fixed unit vector $x = (x_1, \dots, x_d) \in \mathbb{S}^{d-1}$ we have:

$$\begin{aligned} \|\langle X, x \rangle\|_{\psi_2}^2 &= \left\| \sum_{i=1}^d x_i X_i \right\|_{\psi_2}^2 \leq C \sum_{i=1}^d x_i^2 \|X_i\|_{\psi_2}^2 && \text{by Proposition 2.30} \\ &\leq C \max_{i \leq n} \|X_i\|_{\psi_2}^2 && \text{using that } \sum_{i=1}^d x_i^2 = 1 \end{aligned}$$

This completes the proof. □

Numerous random vectors belong to this class.

Example 2.37 Some examples of sub-gaussian random vectors are:

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 2.36
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 2.36 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 $\{x \in \mathbb{S}^{d-1} : x_1 > \varepsilon\}$ makes up at most $\exp\left(-\frac{\varepsilon^2 d}{2}\right)$ proportion of the total area on the sphere.

Equivalently, we can define sub-exponential random vectors.

Definition 2.38 (Sub-exponential random vectors). A random vector X in $\mathbb{R}^d$ is called sub-exponential if the one-dimensional marginals $\langle X, x \rangle$ are sub-exponential random variables for all $x \in \mathbb{R}^d$. The sub-gaussian norm of X is defined as:

2. SETTING AND OVERVIEW OF THE METHODS

$$\|X\|_{\psi_1} = \sup_{x \in S^{d-1}} \|\langle X, x \rangle\|_{\psi_1}$$

Chapter 3

Case Study: Gaussian Random Vectors

In this Chapter, to be able to answer Question 1.3, we will mainly use tools related to the "random process method" (2.4), which we introduced in the previous chapter, and to the theory of Gaussian processes.

3.1 Preliminaries

Random processes are collections of random variables $(X_t)_{t \in T}$ that are not necessarily independent. In many classical examples of probability theory such as Brownian motion, t stands for time, meaning $T \subset \mathbb{R}$. But in high-dimensional probability, it is important to go beyond this case and allow T to be a general abstract set: in our cases, it will be mainly an arbitrary subset of $\mathbb{R}^d$.

In this section, we will prove remarkably sharp comparison inequalities for Gaussian processes, such as Sudakov-Fernique's, and will introduce useful techniques, such as the one of Gaussian interpolation.

The content discussed in this section primarily relies on the material presented in [46], [49].

We start by formally introducing Gaussian processes.

Definition 3.1 *A random process $(X_t)_{t \in T}$ is called a Gaussian process if, for any finite subset $T_0 \subset T$, the random vector $(X_t)_{t \in T_0}$ has a normal distribution.*

Next, we introduce some arguments that will be fundamental to develop comparison inequalities and that are of general interest for the Gaussian case.

3. CASE STUDY: GAUSSIAN RANDOM VECTORS

Lemma 3.2 (*Gaussian integration by parts*) Let $X \sim \mathcal{N}(0, 1)$. Then, for any differentiable function $f : \mathbb{R} \mapsto \mathbb{R}$ we have:

$$\mathbb{E} f'(X) = \mathbb{E} [Xf(X)]$$

Proof (of Lemma 3.2) Assume that f has bounded support. Then, we can extend the result to general integrable functions using an approximation argument. We denote by $p(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$ the gaussian density. Then, it follows, by expressing the expectation as an integral and using the classical integration by parts:

$$\begin{aligned} \mathbb{E} f'(X) &= \int_{\mathbb{R}} f'(x) p(x) dx = - \int_{\mathbb{R}} f(x) p'(x) dx \\ &= \int_{\mathbb{R}} f(x) \cdot x \cdot p(x) dx && \text{using } p'(x) = -xp(x) \\ &= \mathbb{E} [Xf(X)], \end{aligned}$$

proving the lemma. □

We will, now, extend this result to general Gaussian variables and, later, to high dimensions.

First extension Let $X \sim \mathcal{N}(0, \sigma^2)$. Then, for any differentiable function $f : \mathbb{R} \mapsto \mathbb{R}$ we have:

$$\mathbb{E} [Xf(X)] = \sigma^2 \mathbb{E} f'(X)$$

Proof (of First extension) We can write X as $X = \sigma \cdot Z$, where $Z \sim \mathcal{N}(0, 1)$. The density function of X is: $p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2}{2\sigma^2}\right)$. Therefore:
 $p'(x) = -\frac{xp(x)}{\sigma^2}$. It follows:

$$\begin{aligned} \mathbb{E} [Xf(X)] &= \int_{\mathbb{R}} xf(x)p(x)dx = - \int_{\mathbb{R}} \sigma^2 f(x)p'(x)dx \\ &= +\sigma^2 \int_{\mathbb{R}} f'(x)p(x)dx = \sigma^2 \mathbb{E} [f'(X)], \end{aligned}$$

thus proving the result. □

We generalize, now, Gaussian integration by parts to high dimensions:

Second extension (*Multivariate Gaussian integration by parts*) Let $X \sim \mathcal{N}(0, \Sigma)$. Then, for any differentiable function $f : \mathbb{R}^d \mapsto \mathbb{R}$ we have:

$$\mathbb{E} [Xf(X)] = \Sigma \cdot \mathbb{E} \nabla f(X)$$

Proof (of Second extension) Observe that, by matrix-vector multiplication, the conclusion of the lemma is equivalent to:

$$\mathbb{E} [X_i f(X)] = \sum_{k=1}^d \Sigma_{ik}^{1/2} \mathbb{E} [Z_k f(\Sigma^{1/2} Z)] \quad \text{for } i = 1, \dots, d$$

Now, let $Z \sim \mathcal{N}(0, I)$. Then, clearly X has the same distribution as $\Sigma^{1/2} Z$. Thus:

$$\mathbb{E} [X_i f(X)] = \sum_{k=1}^d \Sigma_{ik}^{1/2} \mathbb{E} [Z_k f(\Sigma^{1/2} Z)]$$

Since $Z \sim \mathcal{N}(0, I)$, the $\{Z_k\}$ are independent and $Z_k \sim \mathcal{N}(0, 1) \forall k = 1, \dots, d$. We can apply the univariate Gaussian integration by parts (Lemma 3.2) for $\mathbb{E} [Z_k f(\Sigma^{1/2} Z)]$, conditionally on all variables except $Z_k \sim \mathcal{N}(0, 1)$. Then:

$$\begin{aligned} \mathbb{E} [X_i f(X)] &= \sum_{k=1}^d \Sigma_{ik}^{1/2} \sum_{j=1}^d \Sigma_{jk}^{1/2} \mathbb{E} \left[\frac{\partial f}{\partial x_j} (\Sigma^{1/2} Z) \right] \\ &= \sum_{j=1}^d \sum_{k=1}^d \Sigma_{ik}^{1/2} \Sigma_{jk}^{1/2} \mathbb{E} \left[\frac{\partial f}{\partial x_j} (\Sigma^{1/2} Z) \right] \\ &= \sum_{j=1}^d \Sigma_{ij} \mathbb{E} \left[\frac{\partial f}{\partial x_j} (X) \right], \end{aligned}$$

thus proving the result. $\square$

Having developed this argument, we are now ready to develop the technique of Gaussian interpolation. First, we describe briefly what it consists of.

Assume that T is finite. Then $X = (X_t)_{t \in T}$ and $Y = (Y_t)_{t \in T}$ are Gaussian random vectors in $\mathbb{R}^d$. Moreover, we also assume that X and Y are independent. We, then, define the Gaussian random vector $Z(u) \in \mathbb{R}^d$ that continuously interpolates between $Z(0) = Y$ and $Z(1) = X$:

$$Z(u) := \sqrt{u}X + \sqrt{1-u}Y, \quad u \in [0, 1]$$

Lemma 3.3 (Gaussian interpolation) *Consider two independent Gaussian vectors $X \sim \mathcal{N}(0, \Sigma^X)$ and $Y \sim \mathcal{N}(0, \Sigma^Y)$. The interpolation vector is defined as: $Z(u) := \sqrt{u}X + \sqrt{1-u}Y$, $u \in [0, 1]$. Then, for any twice-differentiable function $f : \mathbb{R}^d \mapsto \mathbb{R}$, we have:*

$$\frac{d}{du} \mathbb{E} f(Z(u)) = \frac{1}{2} \sum_{i,j} (\Sigma_{ij}^X - \Sigma_{ij}^Y) \mathbb{E} \left[\frac{\partial^2 f}{\partial x_i \partial x_j} (Z(u)) \right]$$

Proof (of Lemma 3.3) We will use the chain rule:

$$\frac{d}{du} \mathbb{E} f(Z(u)) = \sum_{i=1}^d \mathbb{E} \frac{\partial f}{\partial x_i} (Z(u)) \frac{dZ_i}{du} = \frac{1}{2} \sum_{i=1}^d \mathbb{E} \frac{\partial f}{\partial x_i} (Z(u)) \left(\frac{X_i}{\sqrt{u}} - \frac{Y_i}{\sqrt{1-u}} \right)$$

3. CASE STUDY: GAUSSIAN RANDOM VECTORS

We break this sum into two, and first compute the contribution of the terms containing X_i . To do so, we condition on Y and:

$$\begin{aligned} \sum_{i=1}^d \frac{1}{\sqrt{u}} \mathbb{E} \left[X_i \frac{\partial f}{\partial x_i} (Z(u)) \right] &= \text{define } g_i(x) = \frac{\partial f}{\partial x_i} (Z(u)) \\ &= \sum_{i=1}^d \frac{1}{\sqrt{u}} \left(\sum_{j=1}^d \Sigma_{ij}^X \mathbb{E} \left[\frac{\partial g_i}{\partial x_j} (x) \right] \right) = \text{by multivariate Gaussian integration by parts} \\ &= \sum_{i=1}^d \frac{1}{\sqrt{u}} \left\{ \sum_{j=1}^d \Sigma_{ij}^X \sqrt{u} \mathbb{E} \left[\frac{\partial^2 f}{\partial x_i \partial x_j} (\sqrt{u}X + \sqrt{1-u}Y) \right] \right\} \end{aligned}$$

Therefore: $\sum_{i=1}^d \frac{1}{\sqrt{u}} \mathbb{E} \left[X_i \frac{\partial f}{\partial x_i} (Z(u)) \right] = \sum_{i,j} \Sigma_{ij}^X \mathbb{E} \frac{\partial^2 f}{\partial x_i \partial x_j} (Z(u))$. Finally, we take the expectation of both sides with respect to Y , therefore lifting the conditioning on Y and we repeat the same steps for the terms containing Y_i , thus obtaining the desired result. $\square$

Finally, we can now state a classical comparison theorem for Gaussian processes due to Slepian and Fernique.

Lemma 3.4 (*Slepian-Fernique inequality*) Let $Y \sim \mathcal{N}(0, \Sigma^Y)$, $Z \sim \mathcal{N}(0, \Sigma^Z)$ be centered Gaussian vectors in $\mathbb{R}^d$. Suppose that $\mathbb{E} (Y_i - Y_j)^2 \leq \mathbb{E} (Z_i - Z_j)^2 \forall 1 \leq i, j \leq d$. Then:

$$\mathbb{E} \sup_{1 \leq i \leq d} Y_i \leq \mathbb{E} \sup_{1 \leq i \leq d} Z_i$$

Proof (of Lemma 3.4) Let g, g' be independent standard Gaussian vectors. We can assume that $Y = (\Sigma^Y)^{1/2} g$ and $Z = (\Sigma^Z)^{1/2} g'$. Let $W(t) = \sqrt{t}Z + \sqrt{1-t}Y$ for $t \in [0, 1]$. Then:

$$\begin{aligned} \frac{d}{dt} \mathbb{E} [f(W(t))] &= \frac{1}{2} \mathbb{E} \left[\left\langle \nabla f(W(t)), \frac{Z}{\sqrt{t}} - \frac{Y}{\sqrt{1-t}} \right\rangle \right] && \text{by multivariate chain rule} \\ &= \frac{1}{2} \mathbb{E} \left[\frac{1}{\sqrt{t}} \left\langle (\Sigma^Z)^{1/2} \nabla f(W(t)), g' \right\rangle \right. \\ &\quad \left. - \frac{1}{\sqrt{1-t}} \left\langle (\Sigma^Y)^{1/2} \nabla f(W(t)), g \right\rangle \right] && \text{by Gaussian interpolation} \\ &= \frac{1}{2} \sum_{i=1}^d (\Sigma_{ij}^Z - \Sigma_{ij}^Y) \mathbb{E} \left[\frac{\partial^2 f}{\partial x_i \partial x_j} (W(t)) \right] && \text{by Gaussian integration by parts} \end{aligned}$$

Now that we have this identity, we would like to apply it to $f(x) = \sup_i x_i$. If we can show that $\frac{d}{dt} \mathbb{E} [\sup_i W_i(t)] \leq 0$, that would imply our goal:

$$\mathbb{E} [\sup_i Z_i] = \mathbb{E} [\sup_i W_i(1)] \geq \mathbb{E} [\sup_i W_i(0)] = \mathbb{E} [\sup_i Y_i]$$

One problem is that the function $x \mapsto \sup_i x_i$ is not sufficiently smooth, as it doesn't possess second derivatives. Therefore, we are required to work with a smooth approximation. As a smooth approximation to the maximum function, we will use the LogSumExp.

Definition 3.5 *The LogSumExp is defined as the logarithm of the sum of the exponentials of the arguments:*

$$f_\beta(x) = \frac{1}{\beta} \log \left(\sum_{i=1}^d \exp(\beta x_i) \right)$$

Let's briefly discuss why the LogSumExp is a good sup approximation.

Claim 3.6 *We have:*

$$\sup_i x_i \leq f_\beta(x) \leq \sup_i x_i + \frac{1}{\beta} \log d$$

Therefore: $f_\beta(x) \mapsto \sup_i x_i$ as $\beta \rightarrow \infty$.

Proof (of Claim 3.6) $\sup_i x_i = \frac{1}{\beta} \log \sup e^{\beta x_i} \leq f_\beta(x) \leq \frac{1}{\beta} \log (d \cdot \sup_i e^{\beta x_i}) = \sup_i x_i + \frac{\log d}{\beta}$, thus easily proving the result. $\square$

Observe that LogSumExp is convex, and strictly increasing everywhere in its domain. Writing $x = (x_1, \dots, x_d)$, the partial derivatives of the function are: $\frac{\partial}{\partial x_i} f_\beta(x) = \frac{\exp(\beta x_i)}{\sum_j \exp(\beta x_j)}$. Therefore, the gradient of LogSumExp is the softmax of the function.

In our case we define: $\frac{\partial}{\partial x_i} f_\beta(x) =: p_i(x)$, where the p_i are such that: $p_i(x) \geq 0$ and $\sum_i p_i(x) = 1$.

Then, we can evaluate:

$$\begin{aligned}
 \frac{d}{dt} \mathbb{E} [f_\beta(W(t))] &= \frac{1}{2} \sum_{i,j} \left(\Sigma_{ij}^Z - \Sigma_{ij}^Y \right) \mathbb{E} [\beta (p_i(W(t)) \delta_{ij} - p_i(W(t)) p_j(W(t)))] \\
 &= \frac{1}{2} \beta \left[\sum_{i,j} (\mathbb{E} Z_i Z_j - \mathbb{E} Y_i Y_j) \mathbb{E} [p_i(W(t)) \delta_{ij}] \quad \begin{array}{l} \text{since} \\ \mathbb{E} Z_i Z_j = \Sigma_{ij}^Z, \\ \mathbb{E} Y_i Y_j = \Sigma_{ij}^Y \end{array} \right. \\
 &\quad \left. - \sum_{i,j} (\mathbb{E} Z_i Z_j - \mathbb{E} Y_i Y_j) \mathbb{E} [p_i(W(t)) p_j(W(t))] \right] \\
 &= \frac{1}{2} \beta \left[\sum_i (\mathbb{E} Z_i^2 - \mathbb{E} Y_i^2) p_i(W(t)) \quad \text{by independence} \right. \\
 &\quad \left. - \sum_{i,j} (\mathbb{E} Z_i Z_j - \mathbb{E} Y_i Y_j) \mathbb{E} [p_i(W(t)) p_j(W(t))] \right] \\
 &= \frac{1}{2} \beta \left[\sum_{i,j} (\mathbb{E} Z_i^2 - \mathbb{E} Y_i^2 \right. \quad \begin{array}{l} \text{since } \sum p_i = 1 \\ \implies p_i = \sum_j p_i p_j \end{array} \\
 &\quad \left. - \mathbb{E} Z_i Z_j + \mathbb{E} Y_i Y_j) \mathbb{E} [p_i(W(t)) p_j(W(t))] \right] \\
 &= \frac{1}{4} \beta \left[\sum_{i < j} \mathbb{E} [p_i(W(t)) p_j(W(t))] \cdot \right. \\
 &\quad \left. (\mathbb{E} (Z_i - Z_j)^2 - \mathbb{E} (Y_i - Y_j)^2) \right]
 \end{aligned}$$

Now, observe that:

- $\mathbb{E} p_i(W(t)) p_j(W(t)) \geq 0$ by the definition of $p_i(x)$
- $(\mathbb{E} (Z_i - Z_j)^2 - \mathbb{E} (Y_i - Y_j)^2) \geq 0$ by the hypothesis of the lemma

Therefore $\frac{d}{dt} \mathbb{E} [f_\beta(W(t))] \geq 0$ and we get: $\mathbb{E} [f_\beta(Y)] \leq \mathbb{E} [f_\beta(Z)]$.

By letting $\beta \rightarrow \infty$ and using dominated convergence, we get: $\mathbb{E} \sup_i Y_i \leq \mathbb{E} \sup_i Z_i$, concluding the proof. $\square$

We will, now, state and prove a very powerful inequality, useful to bound the fluctuations of the spectral norm and that we will use in the proof of the main result of this chapter.

Theorem 3.7 (*Kahane-Khintchine inequality* [24]) *Consider a family $(X_i)_{i=1}^s \in \mathbb{R}^d$ of fixed vectors and a family $(g_i)_{i=1}^s$ of i.i.d. standard Gaussian random variables. Then:*

$$\text{Var} (\| \sum_{i=1}^s g_i X_i \|_2) \leq (\mathbb{E} \| \sum_{i=1}^s g_i X_i \|_2)^2$$

Remark 3.8 *We make the following remark. If W is a centered random vector in $\mathbb{R}^d$, i.e. $\mathbb{E} W = 0$, then W has joint Gaussian entries $\iff$ it can be written as*

$W = \sum_{i=1}^s g_i w_i$, where s is some positive integer, g_i are standard Gaussian random variables, and w_i are fixed vectors.

Therefore, this Remark clarifies that the Kahane-Khintchine inequality holds for every Gaussian vector, in full generality.

Proof of Kahane-Khintchine inequality

To prove this inequality, we will prove a theorem valid for centered Gaussian vectors and from which the Kahane-Khintchine inequality will follow.

Theorem 3.9 *Let W be a centered Gaussian vector in $\mathbb{R}^d$ with covariance matrix Σ . Let*

$$J_d = \frac{1}{2\sqrt{\pi}} \int_0^\infty \frac{1}{v^{3/2}} \left(1 - \left(1 + \frac{2v}{d}\right)^{-d/2}\right) dv$$

Then:

$$\frac{\mathbb{E}(\|W\|_2)}{\sqrt{\mathbb{E}(\|W\|_2^2)}} \geq J_d,$$

where the equality holds if and only if W is isotropic, i.e. Σ is proportional to the identity matrix $I_{d \times d}$.

Moreover, J_d is an increasing function with d and it increases to 1.

Thanks to this theorem, we can prove the Kahane-Khintchine inequality. Indeed, in particular, the Theorem affirms that:

$$\frac{\mathbb{E}(\|W\|_2)}{\sqrt{\mathbb{E}(\|W\|_2^2)}} \geq J_d \geq J_1 = \frac{\sqrt{2}}{\sqrt{\pi}}$$

This implies the Kahane-Khintchine inequality:

$$\text{Var}(\|W\|_2) = \mathbb{E}(\|W\|_2^2) - \mathbb{E}(\|W\|_2)^2 \leq \left(\frac{\pi}{2} - 1\right) \mathbb{E}(\|W\|_2)^2 \leq \mathbb{E}(\|W\|_2)^2$$

Therefore, we will now turn our attention to the proof of Theorem 3.9.

Proof (of Theorem 3.9) First, write $\Sigma = U^T D U$, where U is orthogonal and $D = \text{diag}(D_{11}, \dots, D_{dd})$ is diagonal. Therefore if $Z \sim \mathcal{N}(0, I_{d \times d})$, we have $W \sim \Sigma^{1/2} Z \sim D^{1/2} Z$, since $Z \sim UZ$ due to rotational invariance of the Gaussian distribution.

Therefore:

$$\|W\|_2^2 \sim D_{11}z_1^2 + \dots + D_{dd}z_d^2, \quad (3.1)$$

where $z_1, \dots, z_d$ are i.i.d $\mathcal{N}(0, 1)$. Thus, this implies that:

3. CASE STUDY: GAUSSIAN RANDOM VECTORS

$$\mathbb{E}(\|W\|_2^2) = \sum_{i=1}^d D_{ii} = \text{Tr}(\Sigma)$$

Now, we will try to write in an alternative form the quantity $\mathbb{E}(\|W\|)$. We already know from (3.1) that we can write:

$$\mathbb{E}(\|W\|_2) = \mathbb{E}[(D_{11}z_1^2 + \dots + D_{dd}z_d^2)^{1/2}]$$

Therefore, we will use the following integral representation of the square root of the positive number x .

Claim 3.10 *Given a positive number x . Then:*

$$\sqrt{x} = \frac{1}{2\sqrt{\pi}} \int_0^\infty \frac{1 - e^{-sx}}{s^{3/2}} ds \quad (3.2)$$

Observe that the equation is correct for $x = 0$ and that the derivatives of both sides coincide, but we will not prove it here. As a consequence of the claim, we have:

$$\begin{aligned} \mathbb{E}(\|W\|_2) &= \frac{1}{2\sqrt{\pi}} \int_0^\infty (1 - \mathbb{E}(e^{-s(D_{11}z_1^2 + \dots + D_{dd}z_d^2)})) \frac{ds}{s^{3/2}} \\ &= \frac{1}{2\sqrt{\pi}} \int_0^\infty \left(1 - \prod_{i=1}^d \mathbb{E}(e^{-sD_{ii}z_i^2})\right) \frac{ds}{s^{3/2}} \\ &= \frac{1}{2\sqrt{\pi}} \int_0^\infty \left(1 - \prod_{i=1}^d (1 + 2sD_{ii})^{-1/2}\right) \frac{ds}{s^{3/2}} \end{aligned}$$

By equation (3.1) we can assume, without loss of generality that $\sum_{i=1}^d D_{ii} = 1$. We will make use, also of the elementary inequality:

- for $a_i > 0$, we have: $a_1 \cdot \dots \cdot a_d \leq \left(\frac{1}{d}(a_1 + \dots + a_d)\right)^d$

Then, if we apply this inequality to $a_i = 1 + 2sD_{ii}$, together with the assumption $\sum_{i=1}^d D_{ii} = 1$ we get:

$$\prod_{i=1}^d (1 + 2sD_{ii}) \leq \left(1 + \frac{2s}{d}\right)^d$$

Recall that we wanted to prove that:

$$J_d \leq \frac{\mathbb{E}(\|W\|_2)}{\sqrt{\mathbb{E}(\|W\|_2^2)}} = \mathbb{E}\left(\frac{\|W\|_2}{(D_{11} + \dots + D_{dd})^{1/2}}\right) = \mathbb{E}(\|W\|_2),$$

therefore we have:

$$\begin{aligned}
 \mathbb{E}(\|W\|_2) &= \frac{1}{2\sqrt{\pi}} \int_0^\infty \left(1 - \prod_{i=1}^d (1 + 2sD_{ii})^{-1/2}\right) \frac{ds}{s^{3/2}} \\
 &\geq \frac{1}{2\sqrt{\pi}} \int_0^\infty \frac{1}{v^{3/2}} \left(1 - \left(1 + \frac{2v}{d}\right)^{-d/2}\right) dv \\
 &= J_d
 \end{aligned}$$

Notice that the equality holds, when, instead of an inequality, we use the equality: $a_1 \cdot \dots \cdot a_d \leq (\frac{1}{d}(a_1 + \dots + a_d))^d$. Thus, we have equality in the result of Theorem 3.9 if and only if $D_{11} = \dots = D_{dd} = \frac{1}{d}$.

Next, we prove that J_d is an increasing function.

Consider the function:

$$f(x) = \frac{1}{2\sqrt{\pi}} \int_0^\infty \left(1 - \left(1 + \frac{s}{x}\right)^{-x}\right) \frac{ds}{s^{3/2}},$$

which satisfies $f(\frac{d}{2}) = J_d$. Observe that, from (Claim 3.10) one has:

$$\lim_{x \rightarrow \infty} f(x) = \frac{1}{2\sqrt{\pi}} \int_0^\infty (1 - e^{-s}) \frac{ds}{s^{3/2}} = 1$$

Therefore, to show that f is increasing on $(0, \infty)$ we need to show that, for a fixed s the function $x \mapsto \left(1 + \frac{s}{x}\right)^{-x}$ is decreasing. This is equivalent to showing that the function $x \mapsto g(x) = x \log \left(1 + \frac{s}{x}\right)$ is increasing.

This can be verified by computing the first derivative:

$$g'(x) = \log \left(1 + \frac{s}{x}\right) - \frac{s}{s+x}$$

The sign of $g'(x)$ is the same as the sign of $h(\frac{s}{x}) = \left(1 + \frac{s}{x}\right) \log \left(1 + \frac{s}{x}\right) - \frac{s}{x}$. Observe that $h(t) = (1+t) \log(1+t) - t$ is positive for $t > 0$, given that $h(0) = 0$ and $h'(t) = \log(1+t) > 0$. This concludes the proof since we can now say that $J_1 \leq J_d \quad \forall d$. We just note that $J_1 = \mathbb{E}(|Z|) = \sqrt{\frac{2}{\pi}}$, when $Z \sim \mathcal{N}(0, 1)$. $\square$

3.2 Main Result

We are now ready to develop the main result that will help us obtain a complete understanding of the structural parameters that control the expected deviation $\mathbb{E}\|\Sigma_n - \Sigma\|$ in the setting of this chapter.

Theorem 3.11 (Koltchinskii, Lounici [29], [47]) *Let $X_1, \dots, X_n$ be i.i.d. centered Gaussian random vectors in $\mathbb{R}^d$ with covariance matrix Σ . Let Σ_n be the sample covariance matrix. Then:*

$$\mathbb{E} \|\Sigma_n - \Sigma\| \asymp \|\Sigma\| \left(\sqrt{\frac{r(\Sigma)}{n}} + \frac{r(\Sigma)}{n} \right),$$

where $r(\Sigma) := \frac{\text{Tr}(\Sigma)}{\|\Sigma\|}$ is the effective rank of Σ .

Remark 3.12 The effective rank of the matrix Σ measures the intrinsic dimension of the distribution. One has always $r(\Sigma) \leq \text{rank}(\Sigma) \leq d$ and this bound is sharp. Indeed, if X is an isotropic random vector in $\mathbb{R}^d$, then the covariance matrix is $\Sigma = I$, where I is the identity matrix and $r(\Sigma) = \text{rank}(\Sigma) = d$.

3.3 Proof of Main Result

We will first prove the upper bound of Theorem 3.11.

Upper bound

To proceed with the proof of the upper bound for Theorem 3.11, we will use the random process method (introduced in section 2.4). It would be clear how to proceed if we had to bound $\|\Sigma_n\|$, as in our setting $\|\Sigma_n\| = \frac{\|\mathbf{X}\|^2}{n}$. In this case, indeed, it would suffice to bound $\|\mathbf{X}\|$, which is the supremum of a Gaussian process, using in a straightforward way Slepian's inequality or another Gaussian comparison method. Instead, the quantity we need to bound is $\|\Sigma_n - \Sigma\|$ and the previous observation does not apply as this latter quantity is not the supremum of a centered Gaussian process, but rather of a squared Gaussian process:

$$\|\Sigma_n - \Sigma\| = \sup_{v \in B_2^d} \left| \frac{1}{n} \sum_{k=1}^n (\langle v, X_k \rangle^2 - \mathbb{E} \langle v, X_k \rangle^2) \right|$$

To study squared Gaussian processes, we will use the following technique called decoupling:

Lemma 3.13 (Decoupling, [47]) Let $\tilde{\mathbf{X}}$ be an independent copy of $\mathbf{X}$. Then:

$$\mathbb{E} \|\Sigma_n - \Sigma\| \leq \frac{2}{n} \mathbb{E} \|\mathbf{X} \tilde{\mathbf{X}}\|.$$

Proof (of Lemma 3.13) Recall that we defined $\Sigma_n = \frac{\mathbf{X} \mathbf{X}^T}{n}$. Observe that:

$$(\Sigma_n - \Sigma) = \mathbb{E} \left[(\mathbf{X} + \tilde{\mathbf{X}}) (\mathbf{X} - \tilde{\mathbf{X}})^T \mid \mathbf{X} \right]$$

Using this observation:

$$\begin{aligned}
 \mathbb{E} \|\Sigma_n - \Sigma\| &= \frac{1}{n} \mathbb{E} \left[\left\| \mathbb{E} \left[(\mathbf{X} + \tilde{\mathbf{X}}) (\mathbf{X} - \tilde{\mathbf{X}})^T \mid \mathbf{X} \right] \right\| \right] && \text{by the previous observation} \\
 &\leq \frac{1}{n} \mathbb{E} \left[\mathbb{E} \left(\left\| (\mathbf{X} + \tilde{\mathbf{X}}) (\mathbf{X} - \tilde{\mathbf{X}})^T \right\| \mid \mathbf{X} \right) \right] && \text{Jensen's inequality conditionally} \\
 &\leq \frac{1}{n} \mathbb{E} \left\| (\mathbf{X} + \tilde{\mathbf{X}}) (\mathbf{X} - \tilde{\mathbf{X}})^T \right\| && \text{by Law of total expectation}
 \end{aligned}$$

To obtain the desired bound, it remains to note that $(\mathbf{X} + \tilde{\mathbf{X}}, \mathbf{X} - \tilde{\mathbf{X}}) \sim \sqrt{2} (\mathbf{X}, \tilde{\mathbf{X}})$, as $\mathbf{X} + \tilde{\mathbf{X}}$ and $\mathbf{X} - \tilde{\mathbf{X}}$ are both normally distributed centered random vectors, with covariance matrix $2 \cdot \Sigma$, and they are independent. $\square$

Decoupling allows us to bound $\|\Sigma_n - \Sigma\|$ by the supremum of a Gaussian process. Indeed, by the Lemma, we can replace the square $\mathbf{X}\mathbf{X}^T$ of a Gaussian matrix by the product of the two independent copies $\mathbf{X}\tilde{\mathbf{X}}^T$. This product is not Gaussian but we can treat it as such if we condition on one of the copies, e.g. $\tilde{\mathbf{X}}$.

In conclusion, we can affirm that $\|\mathbf{X}\tilde{\mathbf{X}}\|$ is the supremum of a Gaussian process conditionally on $\tilde{\mathbf{X}}$. This will be the key observation leading to our strategy: use Slepian-Fernique inequality conditionally on $\tilde{\mathbf{X}}$ to obtain interesting bounds. Indeed, one example of such interesting bounds is the following.

Lemma 3.14 ([47]) *Let $X_1, \dots, X_n$ be i.i.d. centered Gaussian random vectors in $\mathbb{R}^d$ with covariance matrix Σ . Let Σ_n be the sample covariance matrix. Then:*

$$\mathbb{E} \|\Sigma_n - \Sigma\| \lesssim \mathbb{E} \|\mathbf{X}\| \cdot \frac{\sqrt{\text{Tr}(\Sigma)}}{n} + \|\Sigma\| \cdot \sqrt{\frac{\text{r}(\Sigma)}{n}}$$

Observe that Lemma 3.14 is very powerful because it will reduce the problem of bounding $\mathbb{E} \|\Sigma_n - \Sigma\|$ to the much more straightforward problem of bounding $\mathbb{E} \|\mathbf{X}\|$. As we saw at the beginning of the section, $\|\mathbf{X}\|$ is the supremum of a Gaussian process, so it's possible to obtain a bound of this quantity from a direct application of the Slepian-Fernique inequality.

Now, we will turn our attention to how to prove Lemma 3.14.

Proof (of Lemma 3.14) We know already, by Lemma 3.13, that:

$$\mathbb{E} \|\Sigma_n - \Sigma\| \leq \frac{2}{n} \mathbb{E} \|\mathbf{X}\tilde{\mathbf{X}}^T\| = \frac{2}{n} \mathbb{E} \left[\sup_{v, w \in B} Z_{v, w} \right],$$

where $Z_{v, w} := \sum_{k=1}^n \langle v, X_k \rangle \langle w, \tilde{X}_k \rangle$.

Now, we want to apply Slepian-Fernique inequality, Lemma 3.4, on $Z_{v, w}$, conditionally on $\tilde{\mathbf{X}}$, so we want to verify that the conditions stated in the theorem hold. First, we verify that for some other Gaussian process $Y_{v, w}$:

3. CASE STUDY: GAUSSIAN RANDOM VECTORS

$$\mathbb{E} \left[(Z_{v,w} - Z_{v',w'})^2 \mid \tilde{\mathbf{X}} \right] \leq \mathbb{E} \left[(Y_{v,w} - Y_{v',w'})^2 \mid \tilde{\mathbf{X}} \right]$$

The steps to do so will be, roughly:

1. add and subtract some "mixed" quantity with v' and w , then group the terms to have the square of two quantities inside of the expectation
2. use the elementary inequality: $(a + b)^2 \leq 2a^2 + 2b^2$ and then develop the square of the sum using the assumption $\mathbb{E}X_k = 0$
3. use of the trace trick, which consists of seeing the scalar product as the trace of a matricial product

$$\begin{aligned} \mathbb{E} \left[(Z_{v,w} - Z_{v',w'})^2 \mid \tilde{\mathbf{X}} \right] &= \mathbb{E} \left[\left(\sum_{k=1}^n \langle v, X_k \rangle \langle w, \tilde{X}_k \rangle - \sum_{k=1}^n \langle v', X_k \rangle \langle w', \tilde{X}_k \rangle \right)^2 \mid \tilde{\mathbf{X}} \right] \\ &= \mathbb{E} \left[\left(\sum_{k=1}^n \langle v, X_k \rangle \langle w, \tilde{X}_k \rangle - \sum_{k=1}^n \langle v', X_k \rangle \langle w, \tilde{X}_k \rangle \right. \right. \\ &\quad \left. \left. + \sum_{k=1}^n \langle v', X_k \rangle \langle w, \tilde{X}_k \rangle - \sum_{k=1}^n \langle v, X_k \rangle \langle w, \tilde{X}_k \rangle \right)^2 \mid \tilde{\mathbf{X}} \right] \\ &= \mathbb{E} \left[\left(\sum_{k=1}^n \langle v - v', X_k \rangle \langle w, \tilde{X}_k \rangle + \sum_{k=1}^n \langle v', X_k \rangle \langle w - w', \tilde{X}_k \rangle \right)^2 \mid \tilde{\mathbf{X}} \right] \\ &\leq 2\mathbb{E} \left[\left(\sum_{k=1}^n \langle v - v', X_k \rangle \langle w, \tilde{X}_k \rangle \right)^2 \mid \tilde{\mathbf{X}} \right] \quad \text{by the elementary inequality:} \\ &\quad + 2\mathbb{E} \left[\left(\sum_{k=1}^n \langle v', X_k \rangle \langle w - w', \tilde{X}_k \rangle \right)^2 \mid \tilde{\mathbf{X}} \right] \end{aligned}$$

Now, recalling the independence assumption of the X_k 's and that $\mathbb{E}X_k = 0$ and by the elementary equality:

$$(\sum_{k=1}^n a_i)^2 = \sum_{k=1}^n a_i^2 + 2 \sum_{i < j} a_i \cdot a_j$$

We can continue the verification of the hypothesis to apply Slepian-Fernique inequality:

$$\begin{aligned}
 & 2\mathbb{E} \left[\left(\sum_{k=1}^n \langle v - v', X_k \rangle \langle w, \tilde{X}_k \rangle \right)^2 \mid \tilde{\mathbf{X}} \right] + 2\mathbb{E} \left[\left(\sum_{k=1}^n \langle v', X_k \rangle \langle w - w', \tilde{X}_k \rangle \right)^2 \mid \tilde{\mathbf{X}} \right] \\
 &= 2\mathbb{E} \left[\left(\sum_{k=1}^n \langle v - v', X_k \rangle^2 \langle w, \tilde{X}_k \rangle^2 \right) \mid \tilde{\mathbf{X}} \right] + 2\mathbb{E} \left[\left(\sum_{k=1}^n \langle v', X_k \rangle^2 \langle w - w', \tilde{X}_k \rangle^2 \right) \mid \tilde{\mathbf{X}} \right] \\
 &= 2 \sum_{k=1}^n \mathbb{E} \left[(\langle v - v', X_k \rangle^2 \langle w, \tilde{X}_k \rangle^2) \mid \tilde{\mathbf{X}} \right] + 2 \sum_{k=1}^n \mathbb{E} \left[(\langle v', X_k \rangle^2 \langle w - w', \tilde{X}_k \rangle^2) \mid \tilde{\mathbf{X}} \right] \\
 &= 2 \sum_{k=1}^n \langle w, \tilde{X}_k \rangle^2 \mathbb{E} \left[\text{Tr} \left((v - v') (v - v')^T X_k X_k^T \right) \mid \tilde{\mathbf{X}} \right] \quad \text{by the trace trick} \\
 &\quad + 2 \sum_{k=1}^n \langle w - w', \tilde{X}_k \rangle^2 \mathbb{E} \left[\text{Tr} \left(v' v'^T X_k X_k^T \right) \mid \tilde{\mathbf{X}} \right] \\
 &= 2 \sum_{k=1}^n \langle w, \tilde{X}_k \rangle^2 \text{Tr} \left((v - v') (v - v')^T \Sigma \right) \quad \text{recall that: } \mathbb{E}[X_k X_k^T] = \Sigma \\
 &\quad + 2 \sum_{k=1}^n \langle w - w', \tilde{X}_k \rangle^2 \text{Tr} \left(v v^T \Sigma \right) \\
 &= 2 \langle v - v', \Sigma (v - v') \rangle \sum_{k=1}^n \langle w, \tilde{X}_k \rangle^2 \quad \text{by the trace trick} \\
 &\quad + 2 \langle v', \Sigma v' \rangle \sum_{k=1}^n \langle w - w', \tilde{X}_k \rangle^2
 \end{aligned}$$

Now, we can take the supremum over $v, v', w, w' \in B_2^d$, obtaining:

$$\begin{aligned}
 \mathbb{E} \left[(Z_{v,w} - Z_{v',w'})^2 \mid \tilde{\mathbf{X}} \right] &\leq 2 \|\tilde{\mathbf{X}}\|^2 \left\| \Sigma^{1/2} (v - v') \right\|^2 + 2 \|\Sigma\| \left\| \tilde{\mathbf{X}}^T (w - w') \right\|^2 \\
 &\leq \mathbb{E} \left[(Y_{v,w} - Y_{v',w'})^2 \mid \tilde{\mathbf{X}} \right]
 \end{aligned}$$

where we defined $Y_{v,w} := \sqrt{2} \|\tilde{\mathbf{X}}\| \langle v, \Sigma^{1/2} g \rangle + (2\|\Sigma\|)^{1/2} \langle w, \tilde{\mathbf{X}} g' \rangle$, with g, g' two independent standard Gaussian vectors in $\mathbb{R}^d$ and $\mathbb{R}^n$, respectively.

Now, we can finally use Slepian-Fernique inequality (conditionally), obtaining:

$$\begin{aligned}
 \mathbb{E} \left[\sup_{v,w \in B} Z_{v,w} \mid \tilde{\mathbf{X}} \right] &\leq \mathbb{E} \left[\sup_{v,w \in B} Y_{v,w} \mid \tilde{\mathbf{X}} \right] \lesssim \\
 &\|\tilde{\mathbf{X}}\| \mathbb{E} \left\| \Sigma^{1/2} g \right\|_2 + \|\Sigma\|^{1/2} \mathbb{E} \left[\|\tilde{\mathbf{X}} g'\|_2 \mid \tilde{\mathbf{X}} \right]
 \end{aligned}$$

3. CASE STUDY: GAUSSIAN RANDOM VECTORS

Observe that:

$$\begin{aligned}
 1. \mathbb{E} [\|\tilde{\mathbf{X}}g'\|_2 \mid \tilde{\mathbf{X}}] &= \mathbb{E} [\|\tilde{\mathbf{X}}g'\|_2^2 \mid \tilde{\mathbf{X}}]^{1/2} \leq (\mathbb{E} [\langle \tilde{\mathbf{X}}g', \tilde{\mathbf{X}}g' \rangle \mid \tilde{\mathbf{X}}])^{1/2} \\
 &= \left(\text{Tr} (\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T \mathbb{E} g' g'^T) \right)^{1/2} = \sqrt{\text{Tr} (\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T)} \\
 2. \mathbb{E} \|\Sigma^{1/2}g\|_2 &= \mathbb{E} \left[\left(\|\Sigma^{1/2}g\|_2^2 \right)^{1/2} \right] \leq \left(\mathbb{E} [\langle \Sigma^{1/2}g, \Sigma^{1/2}g \rangle] \right)^{1/2} \\
 &= \left(\text{Tr} [\Sigma \mathbb{E} g g^T] \right)^{1/2} = \sqrt{\text{Tr} (\Sigma)}
 \end{aligned}$$

Therefore, we have:

$$\mathbb{E} \left[\sup_{v,w \in B} Z_{v,w} \mid \tilde{\mathbf{X}} \right] \leq \|\tilde{\mathbf{X}}\| \sqrt{\text{Tr} (\Sigma)} + \|\Sigma\|^{1/2} \sqrt{\text{Tr} (\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T)}$$

We conclude the proof by taking the expectation with respect to $\tilde{\mathbf{X}}$, and noting that:

- $\mathbb{E} \|\tilde{\mathbf{X}}\| = \mathbb{E} \|\mathbf{X}\|$
- $\mathbb{E} \left[\sqrt{\text{Tr} (\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T)} \right] \leq \sqrt{\mathbb{E} [\text{Tr} (\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T)]} = \sqrt{\text{Tr} (\mathbb{E} [\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T])} = \sqrt{\text{Tr} (n \cdot \Sigma)} = \sqrt{n \cdot \text{Tr} (\Sigma)}$

Finally, we have:

$$\begin{aligned}
 \mathbb{E} \|\Sigma_n - \Sigma\| &\leq \frac{2}{n} \mathbb{E} \left[\sup_{v,w \in B} Z_{v,w} \right] \lesssim \frac{1}{n} \mathbb{E} \|\mathbf{X}\| \sqrt{\text{Tr} (\Sigma)} + \frac{1}{n} \|\Sigma\|^{1/2} \sqrt{n \cdot \text{Tr} (\Sigma)} \\
 &= \mathbb{E} \|\mathbf{X}\| \frac{\sqrt{\text{Tr} (\Sigma)}}{n} + \|\Sigma\| \sqrt{\frac{\text{r} (\Sigma)}{n}}
 \end{aligned}$$

This concludes the proof. □

Now that we saw how to reduce the problem of bounding $\mathbb{E} \|\Sigma_n - \Sigma\|$ to the problem of bounding $\mathbb{E} \|\mathbf{X}\|$, let's see how we can actually bound from above the latter quantity.

Lemma 3.15 ([47]) *Let $X_1, \dots, X_n$ be i.i.d. centered Gaussian random vectors in $\mathbb{R}^d$ with covariance matrix Σ . Let Σ_n be the sample covariance matrix. Then:*

$$\mathbb{E} \|\mathbf{X}\| \lesssim \sqrt{\text{Tr} (\Sigma)} + \sqrt{n \cdot \|\Sigma\|}$$

Proof (of Lemma 3.15) As noted above, $\|\mathbf{X}\|$ is the supremum of a Gaussian process and, to bound the supremum of Gaussian processes the most used technique is to apply the Slepian-Fernique inequality.

Note that:

$$\begin{aligned}
 \mathbb{E} (\langle v, \mathbf{X}w \rangle - \langle v', \mathbf{X}w' \rangle)^2 &= \mathbb{E} (\langle v, \mathbf{X}w \rangle - \langle v', \mathbf{X}w \rangle + \langle v', \mathbf{X}w \rangle - \langle v', \mathbf{X}w' \rangle)^2 \\
 &= \mathbb{E} (\langle (v - v'), \mathbf{X}w \rangle + \langle v', \mathbf{X}(w - w') \rangle)^2 \\
 &\leq 2\mathbb{E} \langle (v - v'), \mathbf{X}w \rangle^2 + 2\mathbb{E} \langle v', \mathbf{X}(w - w') \rangle^2 \\
 &= 2\mathbb{E} \text{Tr} \left((v - v') (v - v')^T (\mathbf{X}w) (\mathbf{X}w)^T \right) \\
 &\quad + 2\mathbb{E} \text{Tr} \left(v' v'^T (\mathbf{X}(w - w')) (\mathbf{X}(w - w'))^T \right) \\
 &= 2\mathbb{E} \text{Tr} \left((v - v') (v - v')^T \mathbf{X} \mathbf{X}^T \|w\|^2 \right) \\
 &\quad + 2\mathbb{E} \text{Tr} \left(v' v'^T \mathbf{X} \mathbf{X}^T \|w - w'\|^2 \right) \\
 &= 2\|\Sigma^{1/2} (v - v')\|_2^2 \|w\|_2^2 + 2\|\Sigma^{1/2} v'\|_2^2 \|w - w'\|_2^2 \\
 &= \mathbb{E} (X'_{v,w} - X'_{v',w'})^2
 \end{aligned}$$

when $\|v\|_2, \|w\|_2 \leq 1$, where we defined $X'_{v,w} = \sqrt{2}\langle v, \Sigma^{1/2}g \rangle + \sqrt{2}\|\Sigma\|^{1/2}\langle w, g' \rangle$, with g, g' independent standard gaussian vectors in $\mathbb{R}^d$ and $\mathbb{R}^n$, respectively.

Then, it follows, by the Slepian-Fernique inequality:

$$\mathbb{E}\|\mathbf{X}\| = \mathbb{E} \left[\sup_{v,w \in B} \langle v, \mathbf{X}w \rangle \right] \leq \mathbb{E} \left[\sup_{v,w \in B} X'_{v,w} \right] \lesssim \mathbb{E}\|\Sigma^{1/2}g\| + \|\Sigma\|^{1/2}\mathbb{E}\|g'\|$$

To conclude the proof of the lemma, recall:

- $\mathbb{E}\|\Sigma^{1/2}g\|_2 \leq \sqrt{\text{Tr}(\Sigma)}$
- $\mathbb{E}\|g\|_2 \leq \sqrt{n}$, since g' is a standard gaussian vector in $\mathbb{R}^n$.

Therefore, we achieve:

$$\mathbb{E}\|\mathbf{X}\| \lesssim \sqrt{\text{Tr}(\Sigma)} + \sqrt{n \cdot \|\Sigma\|},$$

completing the proof. $\square$

We are now ready to complete the proof of the upper bound for Theorem 3.11.

Indeed, combining the results of Lemma 3.14 and Lemma 3.15, we get:

$$\begin{aligned}
 \mathbb{E}\|\Sigma_n - \Sigma\| &\lesssim \mathbb{E}\|\mathbf{X}\| \frac{\sqrt{\text{Tr}(\Sigma)}}{n} + \|\Sigma\| \sqrt{\frac{\mathbf{r}(\Sigma)}{n}} \\
 &\lesssim \left(\sqrt{\text{Tr}(\Sigma)} + \sqrt{n\|\Sigma\|} \right) \frac{\sqrt{\text{Tr}(\Sigma)}}{n} + \|\Sigma\| \sqrt{\frac{\mathbf{r}(\Sigma)}{n}}
 \end{aligned}$$

Equivalently:

$$\mathbb{E}\|\Sigma_n - \Sigma\| \lesssim \|\Sigma\| \frac{\mathbf{r}(\Sigma)}{n} + \|\Sigma\| \sqrt{\frac{\mathbf{r}(\Sigma)}{n}}$$

Lower bound

Now, we will continue the proof of Theorem 3.11 by turning to the lower bound.

The main idea behind the proof for the lower bound is that the decoupling inequality of Lemma 3.13 can be partially reversed.

Lemma 3.16 ([47]) *Let $\tilde{\mathbf{X}}$ be an independent copy of $\mathbf{X}$. Then, for every $v \in \mathbb{R}^d$:*

$$\mathbb{E} \|(\Sigma_n - \Sigma) v\| \geq \frac{1}{n} \mathbb{E} \|\mathbf{X} \tilde{\mathbf{X}}^T v\| - \frac{\|\Sigma v\|}{\sqrt{n}}$$

Proof (of Lemma 3.16) In order to prove the lemma, we first verify that the random matrix defined as $X' := \left(I_{d \times d} - \frac{\Sigma v v^T}{\langle v, \Sigma v \rangle} \right) \mathbf{X}$ is independent of the random vector $\mathbf{X}^T v$.

Claim 3.17 X' is independent of $\mathbf{X}^T v$.

Proof (of Claim 3.17) The random matrix X' is a matrix in $\mathbb{R}^{d \times n}$, and the random vector $\mathbf{X}^T v$ is a vector in $\mathbb{R}^n$. Then, the matrix X' is independent of $\mathbf{X}^T v$ if and only if the vector $\mathbf{X}^T v$ is independent of the columns of X' . Our goal is to prove exactly this. We can write the matrix $\mathbf{X}$ in terms of its columns as:

$$\mathbf{X} = [X_1, \dots, X_n] = \Sigma^{1/2} [g_1, \dots, g_n] = \Sigma^{1/2} \cdot G$$

where g_i are independent vectors in $\mathbb{R}^d$ of *i.i.d.* standard Gaussians. Then, we can write the vector $\mathbf{X}^T v$ as:

$$\mathbf{X}^T v = G^T \Sigma^{1/2} v = G^T w$$

defining $w := \Sigma^{1/2} v$. Then, the matrix X' can be written as:

$$X' = \Sigma^{1/2} \left(I_{d \times d} - \frac{w w^T}{\|w\|_2^2} \right) G$$

Note that since what we care about is independence, we can "forget" about the factor $\Sigma^{1/2}$. Due to the property that Gaussians continue to be Gaussians after a change of basis, without loss of generality, we can assume the vector w to be a vector from the canonical basis $\{e_1, \dots, e_d\} \in \mathbb{R}^d$, e.g. $w = e_1$. Then, the vector $\mathbf{X}^T v = G^T w = G^T e_1 = [g_1^{(1)}, g_2^{(1)}, \dots, g_n^{(1)}]^T$, i.e. the first column of G^T . Let's consider an arbitrary (una a caso) columns i of the matrix X' , then this is equal to: $X'_i = [0, g_i^{(2)}, g_i^{(3)}, \dots, g_i^{(d)}]^T$. Since the columns of G are independent and $\sim \mathcal{N}(0, I_{d \times d})$, it's clear that the vector X'_i is independent of the vector $\mathbf{X}^T v$. $\square$

Now that we established the independence between the vector $\mathbf{X}^T v$ and the columns of the random matrix X' , we can continue with the proof of the lemma. Observe that:

$$(\Sigma_n - \Sigma) v = \frac{XX^T v}{n} - \Sigma v = \frac{X'X^T v}{n} + \left(\frac{\langle v, Zv \rangle}{\langle v, \Sigma v \rangle} - 1 \right) \Sigma v$$

Notice that, since the columns of X' are *i.i.d.* and independent of the vector $\mathbf{X}^T v$, the pair $(X'X^T v, \mathbf{X}^T v)$ has the same distribution as the pair $(X'_1 \|\mathbf{X}^T v\|_2, \mathbf{X}^T v)$, where X'_1 denotes the first column of X' . It follows:

$$\begin{aligned} \mathbb{E} \|(\Sigma_n - \Sigma) v\|_2 &= \mathbb{E} \left\| \frac{X'_1 \|\mathbf{X}^T v\|_2}{n} + \left(\frac{\langle v, Zv \rangle}{\langle v, \Sigma v \rangle} - 1 \right) \Sigma v \right\|_2 \\ &= \mathbb{E}_{X'} \left[\mathbb{E} \left\| \frac{X'_1 \|\mathbf{X}^T v\|_2}{n} + \left(\frac{\langle v, Zv \rangle}{\langle v, \Sigma v \rangle} - 1 \right) \Sigma v \right\|_2 \mid X' \right] \\ &\geq \mathbb{E}_{X'} \left[\left\| \mathbb{E} \left(\frac{X'_1 \|\mathbf{X}^T v\|_2}{n} + \left(\frac{\langle v, Zv \rangle}{\langle v, \Sigma v \rangle} - 1 \right) \Sigma v \right) \mid X' \right\|_2 \right] \\ &= \mathbb{E}_{X'} \left[\left\| \frac{1}{n} X'_1 \mathbb{E} (\|\mathbf{X}^T v\|_2) + 0 \right\|_2 \right] = \frac{1}{n} \mathbb{E} \|\mathbf{X}^T v\|_2 \cdot \mathbb{E} \|X'_1\|_2 \end{aligned}$$

where we used Jensen's inequality conditionally on X' .

Now, we can achieve the following lower bound:

$$\mathbb{E} \|X'_1\|_2 = \mathbb{E} \left\| X_1 - \frac{\Sigma v v^T X_1}{\langle v, \Sigma v \rangle} \right\|_2 \geq \mathbb{E} \|X_1\|_2 - \|\Sigma v\|_2 \frac{\mathbb{E} |\langle v, X_1 \rangle|}{\langle v, \Sigma v \rangle} \geq \mathbb{E} \|X_1\|_2 - \frac{\|\Sigma v\|_2}{\langle v, \Sigma v \rangle^{1/2}},$$

where we used:

$$\begin{aligned} \bullet \mathbb{E} |\langle v, X_1 \rangle| &= \mathbb{E} \left[(|\langle v, X_1 \rangle|^2)^{1/2} \right] = \mathbb{E} \left[(\langle v, X_1 \rangle \langle v, X_1 \rangle)^{1/2} \right] \\ &= \mathbb{E} \left[(v^T X_1 X_1^T v)^{1/2} \right] \leq (\mathbb{E} [v^T X_1 X_1^T v])^{1/2} = (v^T \Sigma v)^{1/2} = \langle v, \Sigma v \rangle^{1/2}, \end{aligned}$$

which follows by Jensen's inequality. Therefore, we have:

$$\begin{aligned} \mathbb{E} \|(\Sigma_n - \Sigma) v\|_2 &\geq \frac{1}{n} \mathbb{E} \|\mathbf{X}^T v\|_2 \cdot \mathbb{E} \|X'_1\|_2 = \frac{1}{n} \mathbb{E} \|\tilde{\mathbf{X}}^T v\|_2 \cdot \mathbb{E} \|X'_1\|_2 \\ &\geq \frac{1}{n} \mathbb{E} \|X_1\|_1 \cdot \mathbb{E} \|\tilde{\mathbf{X}}^T v\|_2 - \frac{1}{n} \mathbb{E} \|\mathbf{X}^T v\|_2 \frac{\|\Sigma v\|_2}{\langle v, \Sigma v \rangle^{1/2}} \\ &\geq \frac{1}{n} \mathbb{E} \|\tilde{\mathbf{X}}^T v\|_2 - \frac{\|\Sigma v\|_2}{\sqrt{n}}, \end{aligned}$$

where we used that $X_1 \|\tilde{\mathbf{X}} v\|_2$ has the same distribution as $\mathbf{X} \|\tilde{\mathbf{X}} v\|_2$ and that (usual trick):

$$\begin{aligned} \bullet \mathbb{E} \|\mathbf{X}^T v\|_2 &= \mathbb{E} (\|\mathbf{X}^T v\|_2^2)^{1/2} \leq (\mathbb{E} \langle \mathbf{X}^T v, \mathbf{X}^T v \rangle)^{1/2} = (\mathbb{E} v^T \mathbf{X} \mathbf{X}^T v)^{1/2} = \\ &= (v^T \Sigma \cdot n v)^{1/2} = \sqrt{n} \langle v, \Sigma v \rangle^{1/2} \end{aligned}$$

Therefore, we proved a "reverse version" of the decoupling Lemma 3.13. $\square$

As a Corollary to Lemma 3.16, we obtain the first term in the lower bound of Theorem 3.11.

Corollary 3.18 ([47]) *Let $X_1, \dots, X_n$ be i.i.d. centered Gaussian random vectors in $\mathbb{R}^d$ with covariance matrix Σ . Let Σ_n be the sample covariance matrix. Then:*

$$\mathbb{E} \|\Sigma_n - \Sigma\| \gtrsim \|\Sigma\| \sqrt{\frac{\mathbf{r}(\Sigma)}{n}}$$

Proof (of Corollary 3.18) From Lemma 3.16, we have, $\forall v \in \mathbb{R}^d$:

$$\mathbb{E} \|(\Sigma_n - \Sigma)v\| \geq \frac{1}{n} \mathbb{E} \|\mathbf{X} \tilde{\mathbf{X}}^T v\| - \frac{\|\Sigma v\|}{n}$$

By taking the supremum over $v \in B$:

$$\begin{aligned} \mathbb{E} \|\Sigma_n - \Sigma\| + \frac{\|\Sigma\|}{n} &\geq \frac{1}{n} \sup_{v \in B} \mathbb{E} \|\mathbf{X} \tilde{\mathbf{X}}^T v\| = \sup_{v \in B_2^d} \mathbb{E} \|X_1\| \|\tilde{\mathbf{X}}^T v\|_2 \\ &= \frac{1}{n} \mathbb{E} \|X_1\|_2 \cdot \sup_{v \in B_2^d} \mathbb{E} \|\tilde{\mathbf{X}}^T v\|_2 \end{aligned}$$

Now, we reduced the problem to find an appropriate lower bound for the quantities $\mathbb{E} \|X_1\|_2$ and $\sup_{v \in B_2^d} \mathbb{E} \|\tilde{\mathbf{X}}^T v\|_2$. We start with the first one.

Claim 3.19 *We have that: $\mathbb{E} \|X_1\|_2 \gtrsim (\mathbb{E} \|X_1\|_2^2)^{1/2} = \sqrt{\text{Tr}(\Sigma)}$*

Proof (of Claim 3.19) First, note that proving $\mathbb{E} \|X_1\|_2 \gtrsim (\mathbb{E} \|X_1\|_2^2)^{1/2}$ is equivalent to proving that $\text{Var}(\|X_1\|_2) \lesssim (\mathbb{E} \|X_1\|_2^2)^2$. To show this equivalent statement, we simply apply the Kahane-Khintchine inequality (Theorem 3.7), thus achieving:

$$\mathbb{E} \|X_1\|_2 \gtrsim (\mathbb{E} \|X_1\|_2^2)^{1/2}$$

Next, we need to prove that $(\mathbb{E} \|X_1\|_2^2)^{1/2} = \sqrt{\text{Tr}(\Sigma)}$, which is simply a consequence of the "trace trick":

$$(\mathbb{E} \|X_1\|_2^2)^{1/2} = (\mathbb{E} \langle X_1, X_1 \rangle)^{1/2} = [\text{Tr}(\mathbb{E} X_1^T X_1)]^{1/2} = \sqrt{\text{Tr}(\Sigma)} \quad \square$$

Afterwards, we compute an appropriate lower bound for $\sup_{v \in B} \mathbb{E} \|\tilde{\mathbf{X}}^T v\|_2$.

Claim 3.20 *We have that: $\mathbb{E} \|\tilde{\mathbf{X}}^T v\|_2 \gtrsim \mathbb{E} [\|\tilde{\mathbf{X}}^T v\|_2^2]^{1/2} = \sqrt{n \langle v, \Sigma v \rangle}$*

Proof (of Claim 3.20) The proof follows the exact steps we used to prove the previous claim. $\square$

Therefore, using these two lower bounds, we get:

$$\mathbb{E} \|\Sigma_n - \Sigma\| + \frac{\|\Sigma\|}{\sqrt{n}} \gtrsim \|\Sigma\| \sqrt{\frac{r(\Sigma)}{n}}$$

On the other hand, we can estimate by the central limit theorem:

$$\frac{\|\Sigma\|}{\sqrt{n}} \lesssim \sup_{v \in B} \mathbb{E} |\langle v, (\Sigma_n - \Sigma)v \rangle| \leq \mathbb{E} \|\Sigma_n - \Sigma\|,$$

as $\langle v, (\Sigma_n - \Sigma)v \rangle = \langle v, \Sigma v \rangle \frac{1}{n} \sum_{k=1}^n (Y_k^2 - 1)$ with $Y_k = \frac{\langle v, X_k \rangle}{\langle v, \Sigma v \rangle^{1/2}} \sim \mathcal{N}(0, 1)$.

Therefore, we obtain:

$$\mathbb{E} \|\Sigma_n - \Sigma\| \gtrsim \mathbb{E} \|\Sigma_n - \Sigma\| + \frac{\|\Sigma\|}{\sqrt{n}} \gtrsim \|\Sigma\| \sqrt{\frac{r(\Sigma)}{n}},$$

thus proving the Corollary. $\square$

Proof of the main result

We can now, finally complete the proof of Theorem 3.11, having developed all the necessary ingredients to achieve lower and upper bound.

Indeed, the upper bound follows immediately from Lemmas 3.14 and 3.15.

For the lower bound, instead:

- first, suppose that $r(\Sigma) \leq 2n$. Then $\sqrt{\frac{r(\Sigma)}{n}} \gtrsim \frac{r(\Sigma)}{n}$, and the result follows from Corollary 5.7.
- if $r(\Sigma) > 2n$,

$$\mathbb{E} \|\Sigma_n - \Sigma\| \geq \mathbb{E} \|\Sigma_n\| - \|\Sigma\| \geq \frac{\mathbb{E} \|X_1\|_2^2}{n} - \|\Sigma\| \frac{r(\Sigma)}{2n} = \|\Sigma\| \frac{r(\Sigma)}{2n},$$

where we used that $\Sigma_n = \frac{1}{n} \sum_{k=1}^n X_k X_k^T \geq \frac{1}{n} X_1 X_1^T$.

Case Study: Isotropic Subgaussian Random Vectors

4.1 Preliminaries

Note that, from what we defined in Section 2.1, $\mathbf{X}$ is a matrix with independent columns, but usually not independent entries (unless we sample from a product distribution).

For simplicity, in order to develop the results of this Chapter, we'll consider a model with independent rows, i.e. we'll consider $\mathbf{X}^T$.

Therefore, we will now consider the following general model of random matrices: we assume the independence of the rows. We already established that such matrices are naturally generated by high dimensional distributions. In fact, we recall that we can think about it as, given an arbitrary probability distribution in $\mathbb{R}^d$, the process of taking n independent samples from it and arranging them as the rows of the matrix $\mathbf{X}$.

In this Chapter, we aim to obtain results of the form of Bai-Yin's law, Theorem 1.1:

$$\sqrt{n} - C \cdot \sqrt{d} \leq s_{\min}(\mathbf{X}) \leq s_{\max}(\mathbf{X}) \leq \sqrt{n} + C \cdot \sqrt{d} \quad (4.1)$$

with high probability. For tall matrices where $n \gg d$, both sides of the inequality would be close to each other, which would guarantee that $\mathbf{X}$ acts as an approximate isometry.

Definition 4.1 *A matrix $\mathbf{X}$ of dimension $n \times d$ is said to be an approximate isometry if acts almost like an isometric embedding of ℓ_2^d into ℓ_2^n :*

$$(1 - \delta) K \|x\|_2 \leq \|\mathbf{X}x\|_2 \leq (1 + \delta) K \|x\|_2 \quad \forall x \in \mathbb{R}^d,$$

where K is an appropriate normalization factor and $\delta \ll 1$. This is equivalent to saying that all the singular values of $\mathbf{X}$ are close to each other:

4. CASE STUDY: ISOTROPIC SUBGAUSSIAN RANDOM VECTORS

$$(1 - \delta) K \leq s_d(X) \leq s_1(X) \leq (1 + \delta) K,$$

where $s_d(X)$ and $s_1(X)$ are the smallest and largest singular values of X .

Now, we observe that the normalized matrix $\frac{1}{\sqrt{n}}X$ is an approximate isometry if and only if $\frac{1}{n}X^T X$ is an approximate identity.

Lemma 4.2 (Approximate isometries, [48]) Consider a matrix $A \in \mathbb{R}^{n \times d}$ that satisfies $\|A^T A - I\| \leq \max(\delta, \delta^2)$, for some $\delta > 0$. Then:

$$1 - \delta \leq s_d(A) \leq s_1(A) \leq 1 + \delta$$

Conversely, if A satisfies $s_d(A) \geq 1 - \delta$ and $s_1(A) \leq 1 + \delta$, then $\|A^T A - I\| \leq 3 \max(\delta, \delta^2)$.

Proof (of Lemma 4.2) First, we prove the first part of the lemma. Observe that $\|A^T A - I\| \leq \max(\delta, \delta^2)$ holds if and only if $|\|Ax\|_2^2 - 1| \leq \max(\delta, \delta^2) \forall x \in \mathbb{S}^{d-1}$.

Indeed, without loss of generality, we can assume $\|x\|_2 = 1$. Then, from the hypothesis: $\max(\delta, \delta^2) \geq |\langle (A^T A - I)x, x \rangle| = |\|Ax\|_2^2 - 1|$. We can apply, then, the elementary inequality: $\max(|z - 1|, |z - 1|^2) \leq |z^2 - 1| \forall z \geq 0$. Take $z = \|Ax\|_2$ and we get the desired inequality. Indeed:

$$\max(\delta, \delta^2) \geq |\|Ax\|_2^2 - 1| \geq \max(|\|Ax\|_2 - 1|, |\|Ax\|_2 - 1|^2)$$

Then, if the $\max(\delta, \delta^2)$ is attained for δ , we get:

$$\delta = \max(\delta, \delta^2) \geq \max(|\|Ax\|_2 - 1|, |\|Ax\|_2 - 1|^2) \geq |\|Ax\|_2 - 1|$$

Similarly, if the $\max(\delta, \delta^2)$ is attained for δ^2 , we get:

$$\delta^2 = \max(\delta, \delta^2) \geq \max(|\|Ax\|_2 - 1|, |\|Ax\|_2 - 1|^2) \geq |\|Ax\|_2 - 1|^2$$

Hence, we get the $|\|Ax\|_2 - 1| \leq \delta$. This condition implies, from the definition of singular values, the desired inequality:

$$1 - \delta \leq s_d(A) \leq s_1(A) \leq 1 + \delta$$

The converse part of the lemma is proved in a similar way by using the elementary inequality: $|z^2 - 1| \leq 3 \max(|z - 1|, |z - 1|^2) \forall z \geq 0$. Indeed, by hypothesis we have: $\|Ax\|_2^2 \in (1 - \delta, 1 + \delta)$. Then, since $\|A^T A - I\| = \max(|\|Ax\|_2^2 - 1|)$, we get, by choosing $z = \|Ax\|_2$ and using the elementary inequality and the bounds on $\|Ax\|_2$:

$$|\|Ax\|_2^2 - 1| \leq 3 \max(|\|Ax\|_2 - 1|, |\|Ax\|_2 - 1|^2) \leq 3 \max(\delta, \delta^2) \quad \square$$

4.2 Main Result

The main result of this Chapter is the following.

Theorem 4.3 ([48]) *Consider an isotropic subgaussian distribution in $\mathbb{R}^d$. Let $X_1, \dots, X_n$ be independent copies of such distribution, and let Σ_n be the sample covariance matrix. Let $\varepsilon \in (0, 1)$, $t \geq 1$. Then, with probability at least $1 - 2 \exp(-t^2 d)$ one has:*

$$\text{if } n \geq C \left(\frac{t}{\varepsilon}\right)^2 d \quad \text{then} \quad \|\Sigma_n - I\| \leq \varepsilon$$

4.3 The main technical step

The following result goes in the direction of finding a result of the type (4.1), for random matrices with independent subgaussian rows.

Theorem 4.4 ([48]) *Let $\mathbf{X}$ be a $n \times d$ matrix whose rows X_i are independent subgaussian isotropic random vectors in $\mathbb{R}^d$. Then, for every $t \geq 0$, with probability at least $1 - 2 \exp(-c \cdot t^2)$, one has:*

$$\sqrt{n} - C\sqrt{d} - t \leq s_d(\mathbf{X}) \leq s_1(\mathbf{X}) \leq \sqrt{n} + C\sqrt{d} + t,$$

where $C = C_K$ and $c = c_K > 0$ are two constants depending only on the subgaussian constant $K = \max_i \|X_i\|_{\psi_2}$ of the rows.

Proof (of Theorem 4.4) The proof will show a classical covering argument and will be articulated in three steps. Since we want control on the smallest and largest singular values, we need to control $\|\mathbf{X}x\|_2$ for all vectors x on the unit sphere $\mathbb{S}^{d-1}$. To do so, we discretize the sphere using a net $\mathcal{N}$ (approximation step); then we establish a tight control of $\|\mathbf{X}x\|_2$ for every fixed vector $x \in \mathcal{N}$ (concentration step); finally, we finish off by taking a union bound over all the vectors $x \in \mathcal{N}$ (union bound step).

1. **Approximation step:** First, observe that Lemma 4.2 allows us to reduce the goal of proving results of the type (4.1) to showing an equivalent bound:

$$\left\| \frac{1}{n} \mathbf{X}^T \mathbf{X} - I \right\| \leq \max(\delta, \delta^2),$$

$$\text{where } \delta = \mathcal{O} \left(\sqrt{\frac{d}{n}} \right).$$

Now that we established what is the item we want to prove, since we want to compute the approximation of the spectral norm on a ε -net of $\mathbb{S}^{d-1}$ for some $\varepsilon \in (0, 1)$, we will use Lemma 2.17.

Now, we will apply Lemma 4.2 for the matrix $\frac{\mathbf{X}}{\sqrt{n}}$. Then, we can state the conclusion of Theorem 4.4 in the equivalent form:

$$\left\| \frac{1}{n} \mathbf{X}^T \mathbf{X} - I \right\| \leq \max(\delta, \delta^2) := \varepsilon, \quad (4.2)$$

where $\delta = C \sqrt{\frac{d}{n}} + \frac{t}{\sqrt{n}}$.

Using Lemma 2.17, we can evaluate the operator norm on a $\frac{1}{4}$ -net $\mathcal{N}$ of the unit sphere $\mathbb{S}^{d-1}$:

$$\begin{aligned} \left\| \frac{1}{n} \mathbf{X}^T \mathbf{X} - I \right\| &\leq 2 \max_{x \in \mathcal{N}} \left| \left\langle \left(\frac{1}{n} \mathbf{X}^T \mathbf{X} - I \right) x, x \right\rangle \right| \\ &= 2 \max_{x \in \mathcal{N}} \left| \frac{1}{n} \|\mathbf{X}x\|_2^2 - 1 \right| \end{aligned}$$

To obtain the desired inequality and complete the proof of the theorem, it suffices to show that, with the required probability:

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

We will be able to achieve so in the following steps.

2. **Concentration step:** Fix a vector $x \in \mathbb{S}^{d-1}$. We can express $\|\mathbf{X}x\|_2^2$ as a sum of independent random variables.

$$\|\mathbf{X}x\|_2^2 = \sum_{i=1}^n \langle X_i, x \rangle^2 =: \sum_{i=1}^n Z_i^2,$$

where we recall that X_i denote the rows of the matrix $\mathbf{X}$. By assumption, $Z_i := \langle X_i, x \rangle$ are independent subgaussian random variables with $\mathbb{E}Z_i^2 = 1$ and $\|Z_i\|_{\psi_2} \leq K$. Then, using Remark 2.24 and Lemma 2.30, we can affirm that $Z_i^2 - 1$ are independent centered subexponential random variables with $\|Z_i^2 - 1\|_{\psi_1} \leq \|Z_i^2\|_{\psi_1} \leq 4\|Z_i\|_{\psi_2}^2 \leq 4K^2$.

Then, to obtain concentration, we will use the exponential deviation inequality we obtained in Section 2.6, Theorem 2.32, to control the sum $\|\mathbf{X}x\|_2^2 = \sum_{i=1}^n Z_i^2$. Then:

$$\begin{aligned} \mathbb{P} \left(\left| \frac{1}{n} \|\mathbf{X}x\|_2^2 - 1 \right| \geq \frac{\varepsilon}{n} \right) &= \mathbb{P} \left(\left| \frac{1}{n} \sum_{i=1}^n Z_i^2 - 1 \right| \geq \frac{\varepsilon}{n} \right) \\ &\leq 2 \exp \left(-c \min \left(\frac{\varepsilon^2}{(4K^2)^2}, \frac{\varepsilon}{4K^2} \right) n \right) \\ &\leq 2 \exp \left(-c \min \left(\frac{\varepsilon^2}{16K^4}, \frac{\varepsilon}{16K^4} \right) n \right) \quad \begin{array}{l} \text{since } \min \left(\frac{x}{K^2}, \frac{y}{K} \right) \geq \min \left(\frac{x}{K^2}, \frac{y}{K^2} \right) \\ \text{if } K > 1 \text{ and } \exp(-x) \text{ is decreasing} \end{array} \\ &\leq 2 \exp \left(-\frac{c_1}{K^4} \min(\varepsilon^2, \varepsilon) n \right) \\ &= 2 \exp \left(-\frac{c_1}{K^4} \delta^2 n \right) \\ &\leq 2 \exp \left(-\frac{c_1}{K^4} (C^2 \cdot d + t^2) \right) \quad \begin{array}{l} \text{since } (a+b)^2 \geq a^2 + b^2 \\ \text{if } a, b > 0 \end{array} \end{aligned}$$

where we used that $K \geq \|Z_i\|_{\psi_2} \geq \frac{1}{\sqrt{2}} (\mathbb{E}|Z_i|^2)^{1/2} = \frac{1}{\sqrt{2}}$, implying that $4K^2 > 1$ (the subexponential norm of the centered subexponential variables $Z_i^2 - 1$).

3. **Union bound step:** Before taking the union bound, we need to bound the size of the net $\mathcal{N}$. We can use Lemma 2.15.

By this lemma, since we chose in Step 1 a $\frac{1}{4}$ -net $\mathcal{N}$, the latter has cardinality $|\mathcal{N}| \leq 9^d$. Then, taking the union bound over all the vectors x in the net $\mathcal{N}$, we get:

$$\begin{aligned} \mathbb{P} \left(\max_{x \in \mathcal{N}} \left| \frac{1}{n} \|\mathbf{X}x\|_2^2 - 1 \right| \geq \frac{\varepsilon}{n} \right) &\leq 9^d \cdot 2 \exp \left(-\frac{c_1}{K^4} (C^2 \cdot d + t^2) \right) \\ &\leq 2 \exp \left(-\frac{c_1 t^2}{K^4} \right) \end{aligned}$$

where the last inequality follows for $C = C_K$ sufficiently large, e.g. $C = K^2 \sqrt{\frac{\ln 9}{c_1}}$. This completes the proof of the theorem, as noted in Step 1, thanks to Lemma 4.2. $\square$

Remark 4.5 *A version of this theorem holds for general, non-isotropic subgaussian distributions. Indeed, assume that $\mathbf{X}$ is a $n \times d$ matrix whose rows X_i are independent subgaussian random vectors in $\mathbb{R}^d$ with covariance matrix Σ . Then, for every $t \geq 0$, the following inequality holds with probability at least $1 - 2 \exp(-ct^2)$:*

$$\left\| \frac{1}{n} \mathbf{X}^T \mathbf{X} - \Sigma \right\| \leq \max(\delta, \delta^2),$$

where $\delta = C \sqrt{\frac{d}{n}} + \frac{t}{\sqrt{n}}$. Here, again, $C = C_K$, $c = c_K$

4.4 Proof of Main Result

Now that we proved Theorem 4.4, we can obtain as a consequence, the main result of this Chapter, Theorem 4.3, which gives a bound for the covariance estimation of subgaussian distributions.

Proof (of Theorem 4.3) It follows from Theorem 4.4 that, for every $t \geq 0$, with probability at least $1 - 2 \exp(-ct^2)$ we have $\|\Sigma_n - I\| \leq \max(\delta, \delta^2)$, where $\delta = C \sqrt{\frac{d}{n}} + \frac{t}{\sqrt{n}}$. The conclusion follows for $t = C' t \sqrt{d}$ where $C' = C'_K$ is sufficiently large. $\square$

Chapter 5

Case Study: Heavy-tailed Random Vectors

In this chapter, we describe matrix concentration inequalities that generalize the classical Bernstein bound. Matrix Bernstein inequalities concern random matrices formed as a sum of independent random matrices, assumed to be bounded in the spectral norm. Inequalities of this type are very useful to study how much a certain random matrix deviates from its mean value in the spectral norm.

Matrix Bernstein inequality is a powerful tool with a huge number of applications. In particular, it can be applied to study problems that use random sampling to approximate a specific matrix. This range of problems applies to the study of the sample covariance matrix.

In the following, we will study in detail Matrix Bernstein inequality and see how it can be applied to study the problem of estimation of covariance matrices that belongs to heavy-tailed distributions.

5.1 Matrix Bernstein Inequality

In the scalar case, Bernstein inequality, in the literature, refers to several concentration results. We aim to present the best-known of these results, related to the deviation of the sum of centered, independent random variables from their mean, provided that these variables are subject to a uniform bound.

Theorem 5.1 (Bernstein's Inequality [45]) *Let $X_1, \dots, X_n$ be independent centered bounded random variables satisfying $|X_i| \leq a$ and $\mathbb{E}X_i^2 = \sigma^2$. Then:*

$$\mathbb{P} \{ |\sum_{i=1}^n X_i| > t \} \leq 1 \exp \left(-\frac{t^2}{2n\sigma^2 + \frac{2}{3}at} \right)$$

5. CASE STUDY: HEAVY-TAILED RANDOM VECTORS

This inequality shows that $\sum_{i=1}^n X_i$ concentrates around zero (, i.e. its mean). Moreover, the tails of $\sum_{i=1}^n X_i$ make a transition from subgaussian decay at moderate deviations to subexponential decay at large deviations. Similarly, a very useful generalization of scalar Bernstein's inequality exists for bounding the largest eigenvalue of the sum of independent random matrices.

Theorem 5.2 ([45]) *Let $\{X_i\}_{i=1}^n$ be a sequence of independent random matrices with common dimension $d_1 \times d_2$. Assume that each X_i satisfies:*

$$\mathbb{E}X_i = 0 \quad \text{and} \quad \|X_i\|_2 \leq L \quad \forall i$$

Let $S_n = \sum_{i=1}^n X_i$ and let $v(S_n)$ be the matrix variance statistic of the sum:

$$v(S_n) = \max \left\{ \|\mathbb{E}(S_n S_n^T)\|, \|\mathbb{E}(S_n^T S_n)\| \right\}$$

Then:

$$\mathbb{E}\|S_n\| \leq \sqrt{2v(S_n) \log(d_1 + d_2)} + \frac{1}{3}L \log(d_1 + d_2)$$

Furthermore, for all $t \geq 0$:

$$\mathbb{P}(\|S_n\| \geq t) \leq (d_1 + d_2) \exp \left(-\frac{t^2/2}{v(S_n) + L \frac{t}{3}} \right)$$

Observe that, comparing this result with the scalar one of Theorem 5.1, the reader will notice an extra-dimensional factor due to the ambient dimension $d_1 + d_2$. The expectation bound shows that $\mathbb{E}\|S_n\|$ is on the same scale as $\sqrt{v(S_n)}$, the root of the matrix variance statistic and the upper bound L for the summands. Also, notice that, when the summands X_i are symmetric, the two terms in the maximum of the definition of $v(S_n)$ coincide.

Let's interpret the probability bound. Indeed, similarly to the scalar case, it is worth studying the deviation of the tails of S_n . Indeed, the concentration inequality can also be rewritten as:

$$\mathbb{P}(\|S_n\| \geq t) \leq \begin{cases} (d_1 + d_2) \cdot e^{-\frac{3t^2}{8v(S_n)}} & \text{if } t \leq \frac{v(S_n)}{L}, \\ (d_1 + d_2) \cdot e^{-\frac{3t}{8L}} & \text{if } t \geq \frac{v(S_n)}{L} \end{cases}$$

This means that:

- for moderate values of t , the tail probability decays as fast as the tail of a Gaussian random variable whose variance is comparable with $v(S_n)$
- for larger values of t , the tail probability decays at least as fast as that of an exponential random variable whose mean is comparable with L

Of course, matrix Bernstein's inequality can be extended to the case of the sum of uncentered random matrices. This formulation describes the spectral norm deviation of the sum from its mean value. The following corollary represents this situation and it is an immediate consequence of Theorem 5.2

Corollary 5.3 ([45]) *Let $\{X_i\}_{i=1}^n$ be a sequence of independent random matrices with common dimension $d_1 \times d_2$. Assume that each X_i has uniformly bounded deviation from its mean:*

$$\|X_i - \mathbb{E}X_i\|_2 \leq L \quad \forall i$$

Let $S_n = \sum_{i=1}^n X_i$ and let $v(S_n)$ be the matrix variance statistic of the sum:

$$v(S_n) = \max \left\{ \|\mathbb{E}(S_n - \mathbb{E}S_n)(S_n - \mathbb{E}S_n)^T\|, \|\mathbb{E}(S_n - \mathbb{E}S_n)^T(S_n - \mathbb{E}S_n)\| \right\}$$

Then:

$$\mathbb{E}\|S_n - \mathbb{E}S_n\| \leq \sqrt{2v(S_n) \log(d_1 + d_2)} + \frac{1}{3}L \log(d_1 + d_2)$$

Furthermore, for all $t \geq 0$:

$$\mathbb{P}(\|S_n - \mathbb{E}S_n\| \geq t) \leq (d_1 + d_2) \exp \left(-\frac{t^2/2}{v(S_n) + L \frac{t}{3}} \right)$$

5.2 Main Result

We want to apply Theorem 5.2 to measure how well the sample covariance estimator approximates the true covariance matrix. Indeed, suppose we have access to n independent samples $X_1, \dots, X_n$, each of them distributed as the random vector $X \in \mathbb{R}^d$. Recall that the sample covariance estimator Σ_n is defined as the random matrix:

$$\Sigma_n = \frac{1}{n} \sum_i X_i X_i^T$$

The random matrix Σ_n is an unbiased estimator for the true covariance matrix: $\mathbb{E}\Sigma_n = \Sigma$.

Moreover, we just saw that the sample covariance matrix estimator can be expressed as a sum of independent random matrices $X_i X_i^T$. This is precisely the type of decomposition that allows us to use tools like matrix Bernstein's inequality. With this concentration inequality, we want to study how the spectral norm distance $\|\Sigma_n - \Sigma\|$ between the sample covariance and the true covariance depends on the number n of samples.

For simplicity, assume that the ℓ_2 norm of the random vectors X_i is bounded: $\|X_i\|_2^2 \leq B \quad \forall i$. We define with Y the random deviation of the estimator Σ_n from the true covariance matrix Σ :

5. CASE STUDY: HEAVY-TAILED RANDOM VECTORS

$$Y = \Sigma_n - \Sigma = \sum_i S_i \quad \text{where} \quad S_i = \frac{1}{n} (X_i X_i^T - \Sigma) \quad \forall i$$

The random matrices S_i are independent identically distributed and centered. To apply Bernstein's inequality, we need to find a uniform bound L for the summands, and we need to control the matrix variance statistic $v(Y)$.

First, then, we develop the uniform bound on the spectral norm of each summand:

$$\begin{aligned} \|S_i\| &= \frac{1}{n} \left\| (X_i X_i^T - \Sigma) \right\| \\ &\leq \frac{1}{n} \|X_i X_i^T\| - \|\Sigma\| && \text{by the triangle inequality} \\ &\leq \frac{2B}{n} && \begin{array}{l} \text{since } \|X_i\|_2^2 \leq B \forall i: \|X_i X_i^T\| = \|X_i\|_2^2 \leq B \\ \text{and } \|\Sigma\| = \|\mathbb{E}(X_i X_i^T)\| \leq \mathbb{E}\|X_i X_i^T\| \leq B \end{array} \end{aligned}$$

Now, we need to assure that the matrix variance statistic $v(Y)$ is properly defined:

$$v(Y) = \|\mathbb{E}Y^2\| = \|\mathbb{E}(\Sigma_n - \Sigma)^T (\Sigma_n - \Sigma)\| = \|\sum_{i=1}^n \mathbb{E} S_i^2\|$$

By direct calculation, we determine the variance of each of the summands above:

$$\begin{aligned} \mathbb{E} S_i^2 &= \frac{1}{n^2} \mathbb{E} (X_i X_i^T - \Sigma)^2 \\ &= \frac{1}{n^2} \mathbb{E} (\|X_i\|_2^2 X_i X_i^T - (X_i X_i^T) \Sigma - \Sigma (X_i X_i^T) + \Sigma^2) \\ &\preceq \frac{1}{n^2} (\mathbb{E}(X_i X_i^T) \cdot B - \mathbb{E}(X_i X_i^T) \cdot \Sigma - \Sigma \cdot \mathbb{E}(X_i X_i^T) + \Sigma^2) && \text{since } \|X_i\|_2^2 \leq B \\ &= \frac{1}{n^2} (B \cdot \Sigma - \Sigma^2) \preceq \frac{B}{n^2} \Sigma && \text{since } \mathbb{E}(X_i X_i^T) = \Sigma \end{aligned}$$

Then, we can sum the relation over i , obtaining:

$$0 \preceq \sum_{i=1}^n \mathbb{E} S_i^2 \preceq \frac{B}{n} \cdot \Sigma$$

Finally, we can apply the spectral norm to the last expression, evaluating the matrix variance statistic $v(Y)$:

$$v(Y) = \|\sum_{i=1}^n \mathbb{E} S_i^2\| \leq \frac{B}{n} \cdot \|\Sigma\|$$

Now, we have now verified all the hypotheses necessary to analyze the sample covariance estimator with matrix Bernstein inequality.

Theorem 5.4 ([45]) *Let $\{X_i\}_{i=1}^n$ be a sequence of independent copies of the centered random vector $X \in \mathbb{R}^d$, with covariance $\mathbb{E}X_iX_i^T = \Sigma$. Assume that each X_i satisfies:*

$$\|X_i\|_2^2 \leq B \quad \forall i$$

Let $\Sigma_n = \frac{1}{n} \sum_i X_iX_i^T$ be the sample covariance estimator. Then:

$$\mathbb{E}\|\Sigma_n - \Sigma\| \leq \sqrt{\frac{2B\|\Sigma\|\log(2d)}{n}} + \frac{2B\log(2d)}{3n}$$

Proof (of Theorem 5.4) Using the analysis we did previously, we can apply Bernstein's inequality, Theorem 5.2, in particular, the expectation bound, with the uniform bound $L = \frac{2B}{n}$ and the variance bound $v(Y) \leq \frac{B\|\Sigma\|}{n}$, obtaining the claimed result. $\square$

Observe that this means that the error in approximating the sample covariance matrix is not too large when we have a sufficient number of samples. If we wish to obtain a relative error of the order of a fixed $\varepsilon > 0$, we may take: $n \geq \frac{2B\log(2d)}{\varepsilon^2\|\Sigma\|}$, which yields the result:

$$\mathbb{E}\|\Sigma_n - \Sigma\| \leq (\varepsilon + \varepsilon^2)\|\Sigma\|$$

It is often the case that $B = C \cdot d$, where C is an absolute constant, so we discover that $n = C \cdot \varepsilon^{-2}d \log d$ samples are sufficient for the sample covariance estimator to provide a ε -accurate estimate of the true covariance matrix Σ . This bound is qualitatively sharp for worst-case distributions.

Example 5.5 (Good example) *We saw in the previous chapters (and we will see also in the following ones) that there are several examples where the logarithmic oversampling is not needed.*

For instance, consider a standard Gaussian vector $X \sim \mathcal{N}(0, I)$. For any fixed direction $u \in \mathbb{S}^{d-1}$, we have $\langle u, X \rangle \sim \mathcal{N}(0, 1)$.

Therefore, for independent copies $X_1, \dots, X_n$:

$$u^T \Sigma_n u = \frac{1}{n} \sum_{i=1}^n \langle u, X_i \rangle^2 \sim \chi^2(n)$$

Using Proposition 2.19, we get:

$$\mathbb{P}\left(\left|\frac{1}{n} \sum_{i=1}^n \langle u, X_i \rangle^2 - 1\right| > \varepsilon\right) \leq \exp(-n\varepsilon^2)$$

5. CASE STUDY: HEAVY-TAILED RANDOM VECTORS

Taking $n \geq \frac{d}{\varepsilon^2}$, we get a ε -approximation of the covariance matrix with high probability.

Example 5.6 (Worst-case example) Consider a coordinate random vector X , which is uniformly distributed in the set $\{\sqrt{d}e_i\}_{i=1}^d$, where $\{e_i\}_{i=1}^d$ is the canonical basis of $\mathbb{R}^d$. Equivalently, $X = \sqrt{d}e_i$ with probability $\frac{1}{d}$. It follows that X is an isotropic random vector in $\mathbb{R}^d$:

$$\mathbb{E}XX^T = \frac{1}{d} \sum_{i=1}^d de_ie_i^T = I$$

Now, consider n copies of the random vector X and arrange them as the rows of a $n \times d$ matrix $\mathbf{X}$. Observe that, in order that $s_d(\mathbf{X}) > 0$ there must be no zero columns in $\mathbf{X}$. Equivalently, this means that each coordinate vector $\{e_i\}_{i=1}^d$ must be picked at least once in n independent trials.

Therefore, we look at the following scenario. We want to collect d different coordinates, and in each sample, we pick one uniformly at random. Let N be the random variable that counts the number of samples we need. We first compute $\mathbb{E}N$. Let N_i be the number of coordinates we sample while we already picked exactly $i - 1$ different coordinates. Then: $N = \sum_{i=1}^d N_i$. Observe that the probability of seeing a new coordinate is exactly:

$$p_i = 1 - \frac{i-1}{d} = \frac{d-i+1}{d}$$

We will keep sampling until we have indeed picked one of the remaining $d - i + 1$ coordinates that we have not seen yet. This is a geometric distribution with success probability p_i . Therefore: $\mathbb{E}N_i = \frac{1}{p_i} = \frac{d}{d-i+1}$. Thus, by the linearity of expectation:

$$\mathbb{E}N = \sum_{i=1}^d \mathbb{E}N_i = \sum_{i=1}^d \frac{d}{d-i+1} = d \cdot \sum_{i=1}^d \frac{1}{i}$$

It is well known that the harmonic number $\sum_{i=1}^d \frac{1}{i}$ satisfies:

$$\sum_{i=1}^d \frac{1}{i} = \log d + \mathcal{O}(1).$$

Therefore, we expect to have picked all d different coordinates after about $d \log d$ steps. By Markov's inequality, we get this result also in high probability.

Therefore, one must take at least $n \geq Cd \log d$ samples to obtain a nontrivial covariance approximation for the random vector X , proving that the logarithmic oversampling $\log d$ is needed for some worst-case distributions.

5.3 Proof of Matrix Bernstein Inequality

In this section, we will prove matrix Bernstein inequality, Theorem 5.2. The statement will follow from a more general matrix concentration inequality for a sum of symmetric, bounded random matrices.

Theorem 5.7 (*Matrix Bernstein: Symmetric Case, [45]*) Let $\{X_i\}_{i=1}^n$ be a sequence of independent random symmetric $d \times d$ matrices. Assume that each X_i satisfies:

$$\mathbb{E}X_i = 0 \quad \text{and} \quad \|X_i\| = s_1(X_i) \leq L \quad \forall i$$

Let $S_n = \sum_{i=1}^n X_i$ and let $v(S_n)$ be the matrix variance statistic of the sum:

$$v(S_n) = \|\mathbb{E}S_n^2\| = \left\| \sum_{i=1}^n \mathbb{E}X_i^2 \right\|$$

Then:

$$\mathbb{E}s_1(S_n) \leq \sqrt{2v(S_n) \log d} + \frac{1}{3}L \log d$$

Furthermore, for all $t \geq 0$:

$$\mathbb{P}(s_1(S_n) \geq t) \leq d \exp\left(-\frac{t^2/2}{v(S_n) + L \frac{t}{3}}\right)$$

Observe that Theorem 5.7 also yields information about the minimum eigenvalue of the sum of independent random symmetric $d \times d$ matrices. Indeed, suppose that each of the matrices that compose the sum satisfies:

$$\mathbb{E}X_i = 0 \quad \text{and} \quad s_d(X_i) \geq -L' \quad \forall i$$

Consequently, we can apply Theorem 5.7 to $(-S_n)$, obtaining:

$$\mathbb{E}s_d(S_n) \geq -\sqrt{2v(S_n) \log d} - \frac{1}{3}L' \log d$$

and, for all $t \geq 0$:

$$\mathbb{P}(s_d(S_n) \leq -t) \leq d \exp\left(-\frac{t^2/2}{v(S_n) + L' \frac{t}{3}}\right)$$

One of the biggest challenges in establishing Matrix Bernstein Inequality is to find appropriate bounds for the matrix (MGF) and (CGF) of a matrix that satisfies the hypothesis of Theorem 5.7.

Bounds for matrix (MGF) and (CGF)

We will start, then, with a lemma that serves the purpose of finding an estimate for matrix (MGF) and (CGF).

Lemma 5.8 ([45]) *Suppose X is a random symmetric matrix that satisfies:*

$$\mathbb{E}X = 0 \quad \text{and} \quad \|X\| = s_1(X) \leq L$$

Then, for $0 < \lambda < \frac{3}{L}$:

$$\mathbb{E}e^{\lambda X} \preceq \exp\left(\frac{\lambda^2/2}{1-\frac{\lambda L}{3}} \cdot \mathbb{E}X^2\right) \quad \text{and} \quad \log \mathbb{E}e^{\lambda X} \preceq \frac{\lambda^2/2}{1-\frac{\lambda L}{3}} \cdot \mathbb{E}X^2$$

Proof (of Lemma 5.8) Let's fix $\lambda > 0$. Then:

$$e^{\lambda X} = I + \lambda X + \left(e^{\lambda X} - \lambda X - I\right) = I + \lambda X + X \cdot f(X) \cdot X \quad (5.1)$$

, where $f(x) = \frac{e^{\lambda x} - \lambda x - 1}{x^2}$ for $x \neq 0$ and $f(0) = \frac{\lambda^2}{2}$.

Observe that the function f is increasing, as it can be verified by computing its derivative and discovering that it's positive. Therefore: $f(x) \leq f(L)$ when $x \leq L$. By assumption, the eigenvalues of the random matrix X don't exceed L , so, in the interval $[0, L]$ by the Transfer Rule (Proposition 5.10):

$$f(X) \preceq f(L). \quad (5.2)$$

Now, the Conjugation rule (Proposition 5.9), allows us to introduce the relation 5.2 into the expansion 5.1 we found above:

$$e^{\lambda X} \preceq I + \lambda X + X \cdot (f(L) \cdot I) \cdot X = I + \lambda X + f(L) \cdot X^2$$

To obtain the desired result, though, we need to develop a further estimate for $f(L)$ and we do so using an application of the Taylor series:

$$\begin{aligned} f(L) &= \frac{e^{\lambda L} - \lambda L - 1}{L^2} = \frac{1}{L^2} \sum_{k=2}^{\infty} \frac{(\lambda L)^k}{k!} = \frac{1}{L^2} \sum_{k=2}^{\infty} \lambda^2 L^2 \frac{(\lambda L)^{k-2}}{k!} \\ &\preceq \frac{\lambda^2}{2} \sum_{k=2}^{\infty} \frac{(\lambda L)^{k-2}}{3^{k-2}} \quad \text{since } k! \geq 2 \cdot 3^{k-2} \quad \forall k = 2, 3, 4, \dots \\ &= \frac{\lambda^2/2}{1 - \frac{\lambda L}{3}} \end{aligned}$$

To complete the proof of the (MGF) bound, by combining the last expressions:

$$e^{\lambda X} \preceq I + \lambda X + \frac{\lambda^2/2}{1-\frac{\lambda L}{3}} \cdot X^2$$

Therefore, since expectation preserves the semidefinite order:

$$\begin{aligned} \mathbb{E}e^{\lambda X} &\preceq I + \frac{\lambda^2/2}{1-\frac{\lambda L}{3}} \cdot \mathbb{E}X^2 && \text{since } \mathbb{E}X = 0 \\ &\preceq \exp\left(\frac{\lambda^2/2}{1-\frac{\lambda L}{3}} \cdot \mathbb{E}X^2\right) && \begin{array}{l} \text{by applying Proposition 5.10} \\ \text{to the equality: } 1 + \theta \leq e^\theta \forall \theta \in \mathbb{R} \end{array} \end{aligned}$$

In this way, we proved the matrix (MGF) bound.

To conclude also with the semidefinite bound the matrix (CGF), we extract the logarithm of the (MGF), recalling the fact that the matrix logarithm preserves the semidefinite order, i.e. :

for positive definite matrices A and B with the same dimensions:
if $A \preceq B \implies \log A \preceq \log B$

This concludes the proof of Lemma 5.8. □

Proof of Matrix Bernstein in the Symmetric case

With Lemma 5.8, we are now ready to establish the Matrix Bernstein inequality for random symmetric matrices.

Proof (of Theorem 5.7) Consider a sequence $\{X_i\}_{i=1}^n$ of independent random symmetric $d \times d$ matrices. Each of the X_i satisfies:

$$\mathbb{E}X_i = 0 \quad \text{and} \quad \|X_i\| = s_1(X_i) \leq L \quad \forall i$$

By Lemma 5.8 we get:

$$\log \mathbb{E}e^{\lambda X_i} \leq \frac{\lambda^2/2}{1-\frac{\lambda L}{3}} \cdot \mathbb{E}X_i^2 \quad \forall 0 < \lambda \leq \frac{3}{L}$$

Let $S_n = \sum_i X_i$. First, we prove the bound in expectation from Theorem 5.7. We use the bound for sums of independent random matrices we stated in Theorem 5.20, obtaining:

$$\begin{aligned}
 \mathbb{E}s_1(S_n) &\leq \inf_{\lambda>0} \frac{1}{\lambda} \log \text{Tr} \left(\exp \left(\sum_i \log \mathbb{E} e^{\lambda X_i} \right) \right) \\
 &\leq \inf_{0<\lambda<\frac{3}{L}} \frac{1}{\lambda} \log \text{Tr} \left(\exp \left(\frac{\lambda^2/2}{1-\frac{\lambda L}{3}} \sum_i \mathbb{E} X_i^2 \right) \right) && \text{from the above bound} \\
 &= \inf_{0<\lambda<\frac{3}{L}} \frac{1}{\lambda} \log \text{Tr} \left(\exp \left(\frac{\lambda^2/2}{1-\frac{\lambda L}{3}} \mathbb{E} S_n^2 \right) \right) && \text{obtained from Lemma 5.8} \\
 &\leq \inf_{0<\lambda<\frac{3}{L}} \frac{1}{\lambda} \log \left[d \cdot s_1 \left(\exp \left(\frac{\lambda^2/2}{1-\frac{\lambda L}{3}} \mathbb{E} S_n^2 \right) \right) \right] && \text{by the inequality:} \\
 &= \inf_{0<\lambda<\frac{3}{L}} \frac{1}{\lambda} \log \left[d \cdot \exp \left(\frac{\lambda^2/2}{1-\frac{\lambda L}{3}} \cdot s_1(\mathbb{E} S_n^2) \right) \right] && \sum_{i=1}^d a_i \leq d \cdot \max_i a_i \text{ for } a_i > 0 \\
 &= \inf_{0<\lambda<\frac{3}{L}} \frac{1}{\lambda} \log \left[d \cdot \exp \left(\frac{\lambda^2/2}{1-\frac{\lambda L}{3}} \cdot v(S_n) \right) \right] && \text{from the definition} \\
 &= \inf_{0<\lambda<\frac{3}{L}} \left[\frac{\log d}{\lambda} + \frac{\lambda^2/2}{1-\frac{\lambda L}{3}} \cdot v(S_n) \right] && \text{of } v(S_n)
 \end{aligned}$$

Finally, we find the infimum over $0 < \lambda < \frac{3}{L}$ to complete. The optimal argument is:

$$\lambda = \frac{6L \log d + 9\sqrt{2v(S_n) \log d}}{2L^2 t + 9(S_n) + 6L\sqrt{2(S_n) \log d}}$$

Next, we develop the tail bound for $s_1(S_n)$. We use again the bound for sums of independent random matrices we stated in Theorem 5.20, this time the one in probability, obtaining:

$$\begin{aligned}
 \mathbb{P}(s_1(S_n) \geq t) &\leq \inf_{\lambda>0} e^{-\lambda t} \text{Tr} \exp \left(\sum_i \log \mathbb{E} e^{\lambda X_i} \right) \\
 &\leq \inf_{0<\lambda<\frac{3}{L}} e^{-\lambda t} \text{Tr} \exp \left(\frac{\lambda^2/2}{1-\frac{\lambda L}{3}} \sum_i \mathbb{E} X_i^2 \right) \\
 &\leq \inf_{0<\lambda<\frac{3}{L}} d \cdot e^{-\lambda t} \exp \left(\frac{\lambda^2/2}{1-\frac{\lambda L}{3}} \cdot v(S_n) \right)
 \end{aligned}$$

Then we compute the infimum over $0 < \lambda < \frac{3}{L}$ and with the choice $\lambda = \frac{t}{v(Y) + Lt/3}$, we get the desired bound. $\square$

Now that we have proved Matrix Bernstein for the symmetric case, we are able to prove the general case of Theorem 5.2. We derive Theorem 5.2 from Theorem 5.7 by applying the latter to the Hermitian dilation (Definition 2.16) of a sum of general rectangular matrices.

Proof (of Theorem 5.2) Consider the sequence $\{X_i\}_{i=1}^n$ of independent random matrices with common dimension $d_1 \times d_2$. Assume that each X_i satisfies:

$$\mathbb{E}X_i = 0 \quad \text{and} \quad \|X_i\| \leq L \quad \forall i$$

We define two random matrices:

- $S_n = \sum_i X_i$
- $\tilde{S}_n = \sum_i \tilde{X}_i$, where $\tilde{X}_i$ is the Hermitian dilation of X_i

Now, recall from Definition 2.16 that:

$$\|S_n\| = s_1(\tilde{S}_n)$$

Also, the variance statistic of a dilation coincides with the one of the original random matrix. Indeed:

$$\begin{aligned} \text{Var}(\tilde{S}_n) &= \mathbb{E} \begin{bmatrix} 0 & (S_n - \mathbb{E}S_n) \\ (S_n - \mathbb{E}S_n)^T & 0 \end{bmatrix}^2 \\ &= \mathbb{E} \begin{bmatrix} (S_n - \mathbb{E}S_n)(S_n - \mathbb{E}S_n)^T & 0 \\ 0 & (S_n - \mathbb{E}S_n)^T(S_n - \mathbb{E}S_n) \end{bmatrix} \\ &= \begin{bmatrix} \mathbb{E}(S_n - \mathbb{E}S_n)(S_n - \mathbb{E}S_n)^T & 0 \\ 0 & \mathbb{E}(S_n - \mathbb{E}S_n)^T(S_n - \mathbb{E}S_n) \end{bmatrix} \end{aligned}$$

Therefore, using the definition of the matrix variance statistics:

$$\begin{aligned} v(\tilde{S}_n) &= \|\text{Var}(\tilde{S}_n)\| \\ &= \max \left\{ \left\| \mathbb{E}(S_n - \mathbb{E}S_n)(S_n - \mathbb{E}S_n)^T \right\|, \left\| \mathbb{E}(S_n - \mathbb{E}S_n)^T(S_n - \mathbb{E}S_n) \right\| \right\} \\ &= v(S_n) \end{aligned}$$

Therefore, we just apply Theorem 5.7 to obtain Theorem 5.2. $\square$

5.4 Appendix of this Chapter

We will state the following simple fact: the semidefinite order is preserved by conjugation.

Proposition 5.9 (*Conjugation Rule, [45]*) *Let A and B be symmetric matrices of the same dimensions. Let C be a general matrix with compatible dimensions. Then:*

$$\text{if } A \preceq B \implies CAC^T \preceq CBC^T$$

Proposition 5.10 (*Transfer Rule, [45]*) *Let f, g be real-valued functions defined on an interval I of the real line. Let A be a symmetric matrix whose eigenvalues are contained in the interval I . Then:*

$$\text{if } f(a) \leq g(a) \quad \forall a \in I \implies f(A) \preceq g(A)$$

Proof (of Proposition 5.10) Since A , by assumption, is a symmetric matrix we can use the spectral decomposition $A = Q\Lambda Q^T$, where Q is an orthogonal matrix and Λ is a matrix with non-negative entries on its diagonal and otherwise zero entries. Then, since by assumption $f(a) \leq g(a) \quad \forall a \in I$ and all the eigenvalues of A are contained in the interval I , it follows:

$$f(\Lambda) \preceq g(\Lambda)$$

Finally, we can use the Conjugation rule from Proposition 5.9 to conjugate this relation also by Q . $\square$

Subadditivity of matrix (MGF)

First, we want to develop a bound for the (MGF) and (CGF) of a sum of independent random matrices using information about the summands. For this, we need to state a result of Elliott Lieb on the convexity properties of the trace exponential function. The proof of Lieb's Theorem depends on the properties of a bivariate function called the matrix relative entropy, which finds a lot of use in subjects such as quantum statistical mechanics and quantum information theory.

Definition 5.11 (*Matrix Relative Entropy*) *Let A and B be positive definite matrices of the same dimensions. The entropy of A relative to B is given by:*

$$H(A, B) = \text{Tr} [A(\log A - \log B) - (A - B)]$$

One can think of the relative entropy as a measure of the difference between the matrix A and the matrix B . We will need two facts about Matrix Relative Entropy. First, we will prove that Matrix Relative Entropy is Nonnegative.

Proposition 5.12 ([45]) *For positive definite matrices A and B of the same dimensions, the matrix relative entropy $H(A, B) \geq 0$.*

Proof (of Proposition 5.12) To prove that Matrix Relative Entropy is Nonnegative, we first need a Trace Inequality, called Generalized Klein Inequality.

Lemma 5.13 ([45]) *(Generalized Klein Inequality) Let $f_i : I \mapsto \mathbb{R}$ and $g_i : I \mapsto \mathbb{R}$ be functions defined on an interval I of the real line, and suppose:*

$$\sum_i f_i(x)g_i(y) \geq 0 \text{ for all } x, y \in I$$

If A and B are symmetric matrices whose eigenvalues are contained in I , then:

$$\sum_i \text{Tr} [f_i(A)g_i(B)] \geq 0$$

Proof (of Lemma 5.13) Consider the singular value (eigenvalue in this case) decompositions of the two symmetric matrices A and B :

$$A = U_A \Sigma_A U_A^T = \sum_j \lambda_j u_j u_j^T \text{ and } B = V_B \Sigma_B V_B^T = \sum_k \mu_k v_k v_k^T,$$

where $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d$ and $\mu_1 \geq \mu_2 \geq \dots \geq \mu_d$ the eigenvalues of the matrices A and B , and u_k, v_k are the corresponding eigenvectors. Then:

$$\begin{aligned} \sum_i \text{Tr} [f_i(A)g_i(B)] &= \sum_i \text{Tr} \left[\left(\sum_k f_i(\lambda_k) u_k u_k^T \right) \left(\sum_j g_i(\mu_j) v_j v_j^T \right) \right] \\ &= \sum_{k,j} \left[\sum_i f_i(\lambda_k) g_i(\mu_j) \right] \cdot |\langle u_k, v_j \rangle|^2 \geq 0 \quad \text{by linearity of the trace} \end{aligned}$$

Then, the inequality follows from the assumption of non-negativity on the scalar functions. $\square$

Finally, using the generalized Klein inequality, we can conclude the proof of Proposition 5.12.

Indeed, suppose that $f : \mathbb{R}_{++} \mapsto \mathbb{R}$ is a differentiable, convex function on the positive real line. By the definition of convexity:

$$f(x) \geq f(b) + f'(b)(x - b) \text{ for } x \text{ and } b > 0$$

Using the generalized Klein inequality, Lemma 5.13, we can extend this relationship to matrices:

$$\text{Tr} f(A) \geq \text{Tr} [f(B) + f'(B)(A - B)]$$

for all positive definite A and B . We apply this last inequality to the function $f(x) = x \log x - x$, and rearrange the terms to obtain:

$$H(A, B) = \text{Tr}[A(\log A - \log B) - (A - B)] \geq 0$$

for all positive definite A and B . Therefore, we proved that the matrix relative entropy is nonnegative. $\square$

The next fact we need about Matrix Relative Entropy is that Matrix Relative Entropy is Convex.

Theorem 5.14 ([45]) *Let A and B be positive definite matrices of the same dimensions. The map $(A, B) \mapsto H(A, B)$ is convex. This means that, for positive definite matrices A_1, A_2 and B_1, B_2 of the same dimensions, we have, for $\lambda \in [0, 1]$:*

$$H(\lambda A_1 + (1 - \lambda)A_2, \lambda B_1 + (1 - \lambda)B_2) \leq \lambda H(A_1, B_1) + (1 - \lambda)H(A_2, B_2)$$

The proof of this theorem is intricate and requires substantial space to fully explain. To maintain the flow of our exposition, we refer interested readers to the source [45] for a detailed proof.

We also require a basic fact from convex analysis which states that partial maximization of a concave function produces a concave function. We include a simple proof of the fact.

Claim 5.15 *Let f be a concave function of two variables. Then the function $y \mapsto \sup_x f(x, y)$ obtained by partial maximization is concave.*

Proof (of Claim 5.15) Fix $\varepsilon > 0$. For each pair of points y_1 and y_2 , there are points x_1 and x_2 that satisfy:

$$f(x_1; y_1) \geq \sup_x f(x; y_1) - \varepsilon \text{ and } f(x_2; y_2) \geq \sup_x f(x; y_2) - \varepsilon.$$

For each $\lambda \in [0, 1]$, the concavity of f implies that:

$$\begin{aligned} \sup_x f(x, \lambda y_1 + (1 - \lambda)y_2) &\geq f(\lambda x_1 + (1 - \lambda)x_2, \lambda y_1 + (1 - \lambda)y_2) \\ &\geq \lambda \cdot f(x_1; y_1) + (1 - \lambda) \cdot f(x_2; y_2) \\ &\geq \lambda \cdot \sup_x f(x; y_1) + (1 - \lambda) \cdot \sup_x f(x; y_2) - \varepsilon. \end{aligned}$$

Then, take the limit as $\varepsilon \rightarrow 0$ to see that the partial supremum is a concave function of y . $\square$

We also state a result regarding the variational representation of the trace, which remarks the fact that matrix relative entropy is nonnegative.

Lemma 5.16 *Let A be a positive definite matrix. Then:*

$$\text{Tr}(N) = \sup_{M \succ 0} \text{Tr}(M \log N - M \log M + M)$$

Proof (of Lemma 5.16) By Proposition 5.12, Matrix Relative Entropy is non-negative, therefore $H(N, M) \geq 0$. We rearrange all the terms to obtain:

$$\text{Tr}(N) \geq \text{Tr}(M \log N - M \log M + M)$$

Finally, observe that when $M = N$ both sides are equal, which yields the equality we wanted to prove. $\square$

With the recently established and demonstrated results regarding Matrix Relative Entropy, we can now delve into Lieb's theorem. This theorem describes a new type of phenomenon that emerges in the matrix realm and it's analogous to what happens in the scalar case, where the function $a \mapsto \exp(b + \log a) \quad \forall a, b > 0$ is linear.

Theorem 5.17 (Lieb [32]) *Let A be a symmetric $d \times d$ matrix. Then, the function:*

$$B \mapsto \text{Tr} \exp(A + \log B)$$

is a concave map on the convex cone of $d \times d$ positive definite matrices.

Proof (of Theorem 5.17) Use Lemma 5.16 with $N = \exp(B + \log A)$, obtaining:

$$\text{Tr}(B + \log A) = \sup_{M \succ 0} \text{Tr}(M(B + \log A) - M \log M + M)$$

We can rewrite the last expression using the Matrix Relative Entropy:

$$\text{Tr}(B + \log A) = \sup_{M \succ 0} \text{Tr}(MB) + \text{Tr}(A) - H(T, A)$$

Thanks to Theorem 5.14, for each symmetric matrix B the function of which we are taking the supremum on the right-hand side is a concave function of (T, A) . Therefore, observe that the right-hand side is just the partial maximum of a concave function. Claim 5.15, then, implies that the right-hand side defines a concave function of A . This observation establishes the theorem. $\square$

We can also formulate Lieb's Theorem, Theorem 5.14, in probabilistic terms.

Corollary 5.18 *Let A be a fixed symmetric matrix, and let X be a random symmetric matrix of the same dimension. Then:*

$$\mathbb{E} \operatorname{Tr} \exp(A + X) \leq \operatorname{Tr} (\exp(A + \log \mathbb{E} e^X))$$

Proof (of Corollary 5.18) Let Y be the random matrix defined by: $Y = e^X$. Then:

$$\begin{aligned} \mathbb{E} \operatorname{Tr}(\exp(A + X)) &= \mathbb{E} \operatorname{Tr}(\exp(A + \log Y)) \\ &\leq \operatorname{Tr}(\exp(A + \log \mathbb{E} Y)) && \begin{array}{l} \text{by applying Jensen's inequality,} \\ \text{since by Theorem 5.17} \\ \text{the trace function is concave in } Y \end{array} \\ &= \operatorname{Tr}(\exp(A + \log \mathbb{E} e^X)) \end{aligned}$$

□

We are now prepared to state the rule addition for matrix (CGF).

Lemma 5.19 Consider a sequence $\{X_i\}_{i=1}^n$ of independent symmetric $d \times d$ matrices. Then:

$$\mathbb{E} \operatorname{Tr}(\exp(\sum_i \theta X_i)) \leq \operatorname{Tr}(\exp(\sum_i \log \mathbb{E} e^{\theta X_i})) \quad \text{for } \theta \in \mathbb{R}$$

This is equivalent to state that matrix (CGF) satisfies subadditivity:

$$\operatorname{Tr}(\exp(\log \mathbb{E} e^{\sum_i \theta X_i})) \leq \operatorname{Tr}(\exp(\sum_i \log \mathbb{E} e^{\theta X_i})) \quad \text{for } \theta \in \mathbb{R}$$

Proof (of Lemma 5.19) Without loss of generality, we may assume that $\theta = 1$. We compute:

$$\begin{aligned} \mathbb{E} \operatorname{Tr}(\exp(\sum_{i=1}^n X_i)) &= \mathbb{E} \mathbb{E}_{X_n} \operatorname{Tr}(\exp(\sum_{i=1}^{n-1} X_i + X_n)) \\ &\leq \mathbb{E} \operatorname{Tr}(\exp(\sum_{i=1}^{n-1} X_i + \log \mathbb{E}_{X_n} e^{X_n})) && \text{by Corollary 5.18} \\ &= \mathbb{E} \mathbb{E}_{X_{n-1}} \operatorname{Tr}(\exp(\sum_{i=1}^{n-2} X_i + X_{n-1} + \log \mathbb{E}_{X_n} e^{X_n})) \\ &\leq \dots \leq \operatorname{Tr}(\exp(\sum_{i=1}^n \log \mathbb{E}_{X_i} e^{X_i})) \end{aligned}$$

Therefore, we proved the first of the two results. Also, the second formulation, showing the subadditivity of matrix (CGF), follows from the first one, after some algebraic simplifications. □

Bounds for Sums of Independent Random Matrices

Now, we present some general results on the behavior of a sum of independent random matrices.

Theorem 5.20 ([45]) *Consider a sequence $\{X_i\}_{i=1}^n$ of independent random symmetric $d \times d$ matrices. Then:*

$$\mathbb{E} s_1 (\sum_i X_i) \leq \inf_{\lambda > 0} \frac{1}{\lambda} \log [\text{Tr exp} (\sum_i \log \mathbb{E} e^{\lambda X_i})]$$

and

$$\mathbb{E} s_d (\sum_i X_i) \geq \sup_{\lambda < 0} \frac{1}{\lambda} \log [\text{Tr exp} (\sum_i \log \mathbb{E} e^{\lambda X_i})]$$

Furthermore, for all $t \in \mathbb{R}$:

$$\mathbb{P} \{s_1 (\sum_i X_i) \geq t\} \leq \inf_{\lambda > 0} e^{-\lambda t} \text{Tr exp} (\sum_i \log \mathbb{E} e^{\lambda X_i})$$

and

$$\mathbb{P} \{s_d (\sum_i X_i) \leq t\} \leq \inf_{\lambda < 0} e^{-\lambda t} \text{Tr exp} (\sum_i \log \mathbb{E} e^{\lambda X_i})$$

Proof (of Theorem 5.20) To prove this result, it is sufficient to substitute the subadditivity rule we stated in Lemma 5.19, into the bounds for the extreme eigenvalues of a random matrix, obtained using the Laplace transform method in Proposition 2.9 and Proposition 2.10. $\square$

Case Study: Isotropic Random Vectors sampled from a convex body

In this chapter, we will show that $\forall \varepsilon > 0$, only $C(\varepsilon)d = \mathcal{O}(d)$ sample points are enough to approximate the true covariance matrix of a distribution sampled from an unconditional convex body in $\mathbb{R}^d$ with ε precision. To prove the result, we will introduce and use some results from convex bodies, and we will use the moment method introduced in Section 2.2.

6.1 Preliminaries

First, we introduce some notation to define convex bodies and their properties.

Isotropic convex bodies are fascinating mathematical objects that arise in the field of convex geometry. They possess remarkable properties and exhibit a symmetrical behavior that makes them particularly useful in various areas of mathematics, including optimization, and probability theory.

In recent years, isotropic convex bodies have gained significant attention in the field of random matrix theory. The study of isotropic convex bodies in this context provides insights into the behavior of random matrices. It has led to remarkable results in the estimation of covariance matrices.

We can introduce convex bodies now.

Definition 6.1 *We say that a subset A of $\mathbb{R}^d$ is convex if $(1 - \lambda)x + \lambda y \in A$ for any $x, y \in A$ and any $\lambda \in [0, 1]$*

Definition 6.2 *A convex body is a convex compact subset of $\mathbb{R}^d$ with non-empty interior.*

Let X be a $\mathbb{R}^d$ -valued random vector. In particular, we will consider the case when X is uniformly distributed on a convex body K . Then, the properties of

the random vector X will be reflected on the convex body K . By saying things like " K has the property z " we will mean that a random vector sampled from the uniform distribution on K has the property z .

Definition 6.3 A convex body $K \subset \mathbb{R}^d$ is centered if it satisfies $\mathbb{E}X = 0$

Definition 6.4 K is said to be isotropic if it is centered and: $\mathbb{E}XX^T = I_{d \times d}$. Equivalently, $\forall y \in \mathbb{R}^d: \mathbb{E}\langle X, y \rangle^2 = \|y\|^2$.

If K is in isotropic position, then for all $\theta \in \mathbb{S}^{d-1}$, we have:

$$\mathbb{E}\langle X, \theta \rangle^2 = L_K^2$$

The constant L_K is called the isotropic constant of K .

Definition 6.5 A convex body K is called symmetric if $x \in K \implies -x \in K$

Definition 6.6 A convex body K is said to be unconditional if it is invariant under sign flips of the coordinates:

$$\begin{aligned} \forall \eta = (\eta_1, \dots, \eta_d) \in \{-1, 1\}^d : \\ X = (x_1, \dots, x_d) \in K \iff (\eta_1 x_1, \dots, \eta_d x_d) \in K \end{aligned}$$

Observe that an unconditional convex body can be, equivalently, identified by the fact that:

If $x = (x_1, \dots, x_d) \in K$, then the entire rectangle $\prod_{i=1}^d [-|x_i|, |x_i|]$ is contained in K .

One can prove that for an unconditional isotropic body K and $X = (x_1, \dots, x_d)$ sampled from it, the following property hold:

$$\bullet \mathbb{E}[x_j^2] = L_K^2 \quad \forall j = 1, \dots, d$$

Definition 6.7 We define the normalized part of a convex body $K^+ := 2K \cap \mathbb{R}_+^d$. In other words:

If $x = (x_1, \dots, x_d)$ is uniformly distributed in K , then $(2|x_1|, \dots, 2|x_d|)$ is uniformly distributed in K^+ .

One can check that K^+ has the following properties:

1. if $x = (x_1, \dots, x_d) \in K^+$ and $0 \leq y_j \leq x_j \quad \forall 1 \leq j \leq d$ then $y = (y_1, \dots, y_d) \in K^+$
2. $\mathbb{E}[x_j^2] = 4L_K^2 \quad \forall j = 1, \dots, d$ if $X = (x_1, \dots, x_d)$ is a random vector sampled from K^+

These definitions will become useful in the remaining part of the chapter.

We will state, now, some background material from classical convexity.

We define two important operations:

- The Minkowski sum of two sets $A, B \subset \mathbb{R}^d$ is defined by $A + B := \{a + b : a \in A, b \in B\}$
- for every $\lambda \in \mathbb{R}$ we set $\lambda A := \{\lambda a : a \in A\}$

Note that both operations preserve convexity. The Brunn-Minkowski inequality describes the effect of Minkowski's addition to volume.

Theorem 6.8 (Brunn-Minkowski [41]) *If A and B are two non-empty compact subsets of $\mathbb{R}^d$, then:*

$$|A + B|^{1/d} \geq |A|^{1/d} + |B|^{1/d}$$

This theorem expresses the fact that volume is a concave function with respect to Minkowski's addition. For this reason, we also write it in the following form:

If A and B are non-empty compact subsets of $\mathbb{R}^d$ then for every $\lambda \in (0, 1)$ we have:

$$|\lambda A + (1 - \lambda)B|^{1/d} \leq \lambda |A|^{1/d} + (1 - \lambda) |B|^{1/d}$$

Applying the arithmetic-geometric means inequality to this last form of the Brunn-Minkowski we get:

$$|\lambda A + (1 - \lambda)B| \geq |A|^\lambda \cdot |B|^{1-\lambda}$$

This form of the Brunn-Minkowski inequality has the advantage of being dimension-free.

We will leave the proof of the Brunn-Minkowski inequality in the appendix of this Chapter, Section 6.5.

The Brunn-Minkowski inequality is a fundamental tool for several important geometric inequalities that we will be using often in the context of estimating the covariance matrix of distribution sampled from distributions defined on convex bodies. One of these is the so-called Borell's Lemma. Borell's lemma states that if the intersection of $A \cap K$ of a convex body K with a symmetric convex set A captures more than half of the volume of K , then the proportion of K which stays outside $t \cdot A$, $t > 1$ decreases exponentially with respect to t

as $t \rightarrow \infty$, at a rate which is independent of the body K and the dimension d . This result can be interpreted in terms of probability since we will consider K to be a convex body of volume 1. Then, the volume of the intersection $A \cap K$ is equivalent to $\mathbb{P}(X \in A)$, where X is the usual random vector, sampled uniformly from K .

Theorem 6.9 (Borell's Lemma [14]) *Let K be a convex body of volume 1 in $\mathbb{R}^d$, and let A be a closed, convex and symmetric set such that $|K \cap A| = \theta > \frac{1}{2}$. Then, for every $t > 1$ we have*

$$|K \cap (tA)^c| \leq \theta \left(\frac{1 - \theta}{\theta} \right)^{\frac{t+1}{2}}$$

We will leave also this proof in the appendix of this chapter.

6.2 Main Result

We state, now the main theorem.

Theorem 6.10 ([5]) *There exist absolute constants C, c such that the following holds. Let $\varepsilon \in (0, 1)$. Let X be uniformly distributed on an unconditional isotropic convex body in $\mathbb{R}^d$. For $n \geq \frac{Cd}{\varepsilon^2}$, with probability larger than $1 - C \exp(-cd^{1/5})$:*

$$\|\Sigma_n - I\| \leq \varepsilon$$

Sketch of the proof

To prove the theorem, we will make use of the following simple observation:

$$\|\Sigma_n - I\| \leq \|\Sigma_n - \text{diag}(\Sigma_n)\| + \|\text{diag}(\Sigma_n) - I\|$$

Then, we will bound the two terms that compose the right-hand side in two different ways.

Let's see first, how to bound the first term: $\|\Sigma_n - \text{diag}(\Sigma_n)\|$. We will make use of the following lemma, which will be proved by using a variation of the moment method.

Lemma 6.11 ([5]) *Let X be a random vector uniformly distributed on an unconditional convex body in $\mathbb{R}^d$. Let $n \geq d$ be the number of sample points. For $k = d^{1/5}$, we have:*

$$\left(\mathbb{E} \|\Sigma_n - \text{diag}(\Sigma_n)\|^k \right)^{1/k} \leq C \sqrt{\frac{d}{n}},$$

where C is a universal constant.

Using this Lemma, together with Markov's inequality we will be able to bound the first term.

To bound the second term, we will use, instead, Borell's Lemma (Theorem 6.9), by writing:

$$\|\text{diag}(\Sigma_n) - I\| = \max_{1 \leq j \leq d} \left| \frac{1}{n} \sum_{i=1}^n x_{ij}^2 - 1 \right|$$

where x_{ij} are the components of the i -th copy of the random vector $X \in \mathbb{R}^d$: $X_i = (x_{i1}, \dots, x_{id})$.

With this structure in mind, we proceed with the main proof and the proof of all the complementary results needed.

6.3 Proof of Main Result

As discussed above, we reduce the problem of bounding from above $\|\Sigma_n - I\|$ to the problem of bounding:

- $\|\Sigma_n - \text{diag}(\Sigma_n)\|$
- $\|\text{diag}(\Sigma_n) - I\|$

To bound the first term, we use Lemma 6.11.
We obtain:

$$(\mathbb{E}\|\Sigma_n - \text{diag}(\Sigma_n)\|^k)^{1/k} \leq C\sqrt{\frac{d}{n}}, \quad \text{for } k = d^{1/5}$$

Then, by applying Markov's inequality we get:

$$\begin{aligned} \mathbb{P}\left(\|\Sigma_n - \text{diag}(\Sigma_n)\| \geq \frac{\varepsilon}{2}\right) &= \mathbb{P}\left(\|\Sigma_n - \text{diag}(\Sigma_n)\|^k \geq \left(\frac{\varepsilon}{2}\right)^k\right) \\ &\leq \frac{\mathbb{E}\|\Sigma_n - \text{diag}(\Sigma_n)\|^k}{\left(\frac{\varepsilon}{2}\right)^k} \leq C^k \left(\sqrt{\frac{d}{n}} \cdot \frac{2}{\varepsilon}\right)^k \\ &\leq 2^{-d^{1/5}} \end{aligned}$$

for $n \geq 16C^2 \frac{d}{\varepsilon^2}$.

Now, we turn to the problem of bounding the second term. For $1 \leq i \leq n$, we write $X_i = (x_{i1}, \dots, x_{id})$. Then, the term $\|\text{diag}(\Sigma_n) - I\|$ can be estimated as follows:

$$\|\text{diag}(\Sigma_n) - I\| = \max_{1 \leq j \leq d} \left| \frac{1}{n} \sum_{i=1}^n (x_{ij}^2 - 1) \right|$$

As mentioned above, we will now use Borell's Lemma to prove the following claim:

Claim 6.12 *For some constant c , we have: $\psi_{\frac{1}{2}}(x_{ij}^2 - 1) \leq c$*

Proof (of Claim 6.12) First, we normalize the convex body on which we defined the uniform distribution from which we sample the random vector X . Notice that we can do so since K is unconditional by hypothesis. Specifically, by choosing $a = \frac{1}{|K|^{1/d}}$ and setting $H := aK$ we get an unconditional convex body with unitary volume. Moreover, we choose $A = \{x \text{ s.t. } : x_j = \langle x, e_j \rangle^2 - 1 \leq s\}$. It follows from the unconditionality assumption that: $|H \cap A| \geq \theta > \frac{1}{2}$. Recall that, by the isotropic assumption we have: $\mathbb{E}\langle X, e_j \rangle^2 = 1$. Then, using Markov's inequality and choosing an appropriate value for s :

$$\mathbb{P}(A^C) = \mathbb{P}(\{x \text{ s.t. } \langle x, e_j \rangle^2 > 1 + s\}) \leq \frac{1}{100}$$

It follows: $\mathbb{P}(A) \geq \frac{99}{100}$. Then, applying Borell's lemma, Theorem 6.9:

$$\mathbb{P}(x_{ij}^2 - 1 \geq t) \leq \theta \left(\frac{1-\theta}{\theta} \right)^{\frac{1+t}{2}} \leq 2 \exp\left(-\frac{\sqrt{t}}{K}\right)$$

for appropriate values of $\theta > \frac{1}{2}$. From the definition of Orlicz norm (Definition 2.34), follows the claim. $\square$

Knowing that the $\frac{1}{2}$ -Orlicz norm is bounded from above by a constant, we can use the following Lemma, which will be proved in Section 6.4.2. This lemma is an analogue of Bernstein inequalities for ψ_α random variables with $0 < \alpha \leq 1$.

Lemma 6.13 ([5]) *For any $0 < \alpha \leq 1$, $\exists C(\alpha), c(\alpha)$ such that the following holds: Let $(Y_j)_{1 \leq j \leq m}$ be i.i.d random variables such that: $\mathbb{E}Y_1 = 0$ and $\psi_\alpha(Y_1) \leq K$. Then, $\forall t \geq \frac{C(\alpha)K}{\sqrt{m}}$, we have:*

$$\mathbb{P}\left(\left|\frac{1}{m} \sum_{j=1}^m Y_j\right| > t\right) \leq \exp\left(-c(\alpha) \min\left(m \frac{t^2}{K^2}, m^\alpha \frac{t^\alpha}{K^\alpha}\right)\right)$$

With this Lemma, we can finally finish the proof of Theorem 6.10. Indeed, by using Lemma 6.13, for $0 < \varepsilon \leq 1$:

$$\mathbb{P}(\|\text{diag}(\Sigma_n) - I\| \geq \frac{\varepsilon}{2}) \leq d \exp\left(-c\sqrt{d}\right)$$

Putting together the two bounds in high probability obtained concludes the proof of Theorem 6.10.

6.4 Proof of the other results

In this section, we will now prove the results that were used in Section 6.3 for the proof of Theorem 6.10 and that were only stated above.

6.4.1 Proof of Lemma 6.11

We will start with the proof of Lemma 6.11, which has been used to prove the first term in which the quantity $\|\Sigma_n - I\|$ was decomposed in Theorem 6.10.

To prove this result we will use the moment method, with k as an even integer. For $1 \leq i \leq n$, we write the coordinate of X_i as $(x_{ij})_{1 \leq j \leq d}$

Now:

$$\begin{aligned} \mathbb{E} \|\Sigma_n - \text{diag}(\Sigma_n)\|^k &\leq \mathbb{E} \text{Tr}(\Sigma_n - \text{diag}(\Sigma_n))^k = \text{Tr} \mathbb{E}(\Sigma_n - \text{diag}(\Sigma_n))^k \\ &= \sum_{s_1 \neq s_2 \neq \dots \neq s_k \neq s_1} \mathbb{E} Z_{n s_1 s_2} \dots Z_{n s_{k-1} s_k} Z_{n s_k s_1} \\ &= \frac{1}{n^k} \sum_{\substack{s_1 \neq s_2 \neq \dots \neq s_k \neq s_1 \\ r_1, r_2, \dots, r_k}} \mathbb{E} x_{r_1 s_1} x_{r_1 s_2} x_{r_2 s_2} \dots x_{r_k s_k} x_{r_k s_1} \end{aligned}$$

where the sum is taken over all indices $r_1, r_2, \dots, r_k$ in $\{1, \dots, n\}$ and $s_1, s_1, \dots, s_k$ in $\{1, \dots, d\}$, consecutively distinct. Moreover, since the random vectors $X_1, \dots, X_n$ are independent, each expectation in the sum can be factorized as the product of n factors of the form:

$$\mathbb{E} x_1^{\alpha_1} x_2^{\alpha_2} \dots x_d^{\alpha_d} \quad \text{for some integers } \alpha_1, \dots, \alpha_d \quad (6.1)$$

where $X = (x_1, \dots, x_d)$. Observe that, the fact that the vector X is unconditional also implies that it's invariant under flips of coordinates. This shows that the expectation is zero if one of the α_i is odd. This leads us to the following inequality:

$$\mathbb{E} \|\Sigma_n - \text{diag}(\Sigma_n)\|^k \leq \frac{1}{n^k} \sum_{[(r_1, s_1), \dots, (r_{2k}, s_{2k})] \text{ V-graph}} \mathbb{E} x_{r_1 s_1} \dots x_{r_{2k} s_{2k}},$$

where:

Definition 6.14 A *V-graph* is a $2k$ -tuple of pairs $(r_i, s_i) \in \{1, \dots, n\} \times \{1, \dots, d\}$ such that:

1. $s_{2i+1} = s_{2i}$ (and $s_1 = s_{2k}$)
2. $r_{2i} = r_{2i-1}$
3. each couple (r_i, s_i) appears an even number of times

4. $s_{2i} \neq s_{2i-1}$.

We associate several parameters to a V-graph $G = [(r_1, s_1), \dots, (r_{2k}, s_{2k})]$:

- $r(G)$ = number of distinct row indices from Σ_n that appear in G
- $c(G)$ = number of distinct column indices from Σ_n that appear in G
- $\ell(G) = r(G) + c(G)$
- $d(G)$ = number of distinct couples of indices (r_i, s_i) that appear in G
- $n_2(G)$ = number of indices i such that the couple (r_i, s_i) appears exactly 2 times in G
- $n_+(G)$ = number of indices i such that the couple (r_i, s_i) appears 4 times or more in G .

From this, we clearly have $n_2(G) + n_+(G) = 2k$.

Lemma 6.15 ([5]) *If $G = [(r_1, s_1), \dots, (r_{2k}, s_{2k})]$ is a V-graph, then*

$$\mathbb{E} x_{r_1 s_1} \dots x_{r_{2k} s_{2k}} \leq C^k k^{n_+(G)}.$$

Proof (of Lemma 6.15) Recall that the rows of Σ_n are independent, so this allows us to write the whole expectation as a product of n factors of the form (6.1). Note that the sum of all exponents α_i appearing in the factors of (6.1) is exactly $2k$, while the sum restricted to the exponents satisfying $\alpha_i \geq 4$ is exactly $n_+(G)$. We use the following comparison theorem.

Lemma 6.16 (Bobkov and Nazarov [12])

Let $B_1^n = \{x \in \mathbb{R}^d \text{ s.t. } \sum |x_j| \leq 1\}$ denote the unit ball of ℓ_1^d ; the convex body $\alpha_d B_1^d$ is isotropic for $\alpha_d = \sqrt{\frac{(d+1)(d+2)}{2}}$. If $X = (x_1, \dots, x_d)$ is uniformly distributed on an isotropic unconditional convex body, and if $Y = (y_1, \dots, y_d)$ is uniformly distributed on $Cd B_1^d$, then for any increasing functions $f_i : \mathbb{R}^+ \mapsto \mathbb{R}$ and such that $f_i(0) = 0$:

$$\mathbb{E} \prod_{i=1}^d f_i(|x_i|) \leq \mathbb{E} \prod_{i=1}^d f_i(|y_i|).$$

This Lemma will be proved in the appendix of this Chapter, Section 6.5.

We use Lemma 6.15 with $f_i(x) = x^{\alpha_i}$ on each factor of (6.1). Then, we can estimate the resulting expectation for Y using subindependence.

Lemma 6.17 (Ball and Perissinaki [9]) $\forall$ increasing functions $f_i : \mathbb{R}^+ \mapsto \mathbb{R}$, on ℓ_p^n -balls, we have:

$$\mathbb{E} \prod_{i=1}^d f_i(|y_i|) \leq \prod_{i=1}^d \mathbb{E} f_i(|y_i|).$$

This Lemma will be proved in the appendix of this Chapter, Section 6.5.

Now, by Borell's lemma, Theorem 6.9, there is an absolute constant C such that $\psi_1(y_i) \leq C$, which means that $\mathbb{E}|y_i|^p \leq (Cp)^p$. If $\alpha_i \geq 4$, we use the following bound, which follows from Jensen:

$$\mathbb{E}y_i^{\alpha_i} = \mathbb{E}y_i^{\alpha_i \cdot \frac{n_+(G)}{n_+(G)}} \leq \left(\mathbb{E}y_i^{n_+(G)} \right)^{\frac{\alpha_i}{n_+(G)}} \leq (Cn_+(G))^{\alpha_i} \leq (2Ck)^{\alpha_i}.$$

Using this method to bound separately all the factors of the form (6.1), we are led to

$$\mathbb{E}x_{r_1s_1} \dots x_{r_{2k}s_{2k}} \leq (\mathbb{E}y_1^2)^{n_2(G)/2} (2Ck)^{n_+(G)} \leq C^k K^{n_+(G)}$$

and the lemma is proved. $\square$

Now that we proved Lemma 6.15, we can use it to obtain the inequality:

$$\begin{aligned} \mathbb{E}\|\Sigma_n - \text{diag}(\Sigma_n)\|^k &\leq \frac{1}{n^k} \sum_{[(r_1, s_1), \dots, (r_{2k}, s_{2k})] \text{ V-graph}} \mathbb{E}x_{r_1s_1} \dots x_{r_{2k}s_{2k}} \\ &\leq \frac{C^k}{n^k} \sum_{\text{V-graph}} K^{n_+(G)} \end{aligned} \quad (6.2)$$

The problem has been, now, reduced to the problem of bounding some quantity/parameters associated with the V-graph.

Lemma 6.18 ([5]) *Let $G = ((r_1, s_1), \dots, (r_{2k}, s_{2k}))$ be a V-graph. Then:*

1. $d(G) \leq k$
2. $\ell(G) \leq d(G) + 1 \leq k + 1$
3. $r(G) \leq \frac{k}{2}$
4. $n_+(G) \leq 4(k - \ell(G) + 1)$

Proof (of Lemma 6.18) All these bounds follow from the definition of the properties associated with the definition of a V-graph, Definition 6.14. Indeed:

1. From the third property of Definition 6.14, i.e. the fact that each couple (r_i, s_i) appears an even number of times, then every couple appears at least two times $\implies d(G) \leq k$.
2. Observe that if one reads the graph from (r_1, s_1) to (r_{2k}, s_{2k}) , each of the first $d(G)$ occurrences (r_i, s_i) may bring either a new row index or a new column index. To bring both is not possible due to the first two properties of Definition 6.14. $\implies \ell \leq d(G) + 1$. By the previous point, it's trivial that: $\ell \leq k + 1$.

3. From Definition 6.14 it follows that whenever a row index r appears in G , we can find i and j such that: $r_i = r_j = r$ and $s_i \neq s_j \implies$ by the third property, every row index must appear in G at least four times: $r(G) \leq \frac{2k}{4} = \frac{k}{2}$.
4. Observe that it's true that: $d(G) \leq \frac{1}{2}n_2(G) + \frac{1}{4}n_+(G)$. By what we proved in point 2. we have:

$$\ell - 1 \leq d(G) \leq \frac{1}{2}n_2(G) + \frac{1}{2}n_+(G) - \frac{1}{4}n_+(G)$$

$$\text{Since, } n_2(G) + n_+(G) = 2k \implies n_+(G) \leq 4(k + 1 - \ell). \quad \square$$

Having these bounds on the most important quantities associated with the V-graph, we can proceed in the following way:

- we bound the number of V-graphs with given r and c
- recalling the definition of $\ell(G) = r(G) + c(G)$, we sum over all possible ℓ for a V-graph. Indeed, we fix every time a certain r and c and we use the bound we find with the previous point.

Lemma 6.19 ([5]) *Let $I \subset \{1, \dots, n\}$ and $J \subset \{1, \dots, d\}$, such that $r + c \leq k + 1$, with $r = |I|$ and $c = |J|$. Then the number of V-graphs $G = [(r_1, s_1), \dots, (r_{2k}, s_{2k})]$ such that $\{r_i\} \subset I$ and $\{s_i\} \subset J$ is bounded from above by $(Ck)^k$, for some absolute constant C .*

Proof (of Lemma 6.19) Let $G_{r,c} = (g_{ij})$ be a random $r \times c$ matrix with entries being independent $\mathcal{N}(0, 1)$ random variables. From the example 2.18, for some absolute constant C , we know that: $\mathbb{E} \|G_{r,c}\| \leq C(\sqrt{r} + \sqrt{c}) \leq C'\sqrt{k}$. Moreover, the operator norm is a 1-Lipschitz function with respect to the entries of the matrix, and applying Gaussian concentration implies that for any $t \geq 0$,

$$\mathbb{P}(\|G_{r,c}\| \geq \mathbb{E} \|G_{r,c}\| + t) \leq \exp(-t^2/2).$$

Then, this, in turn, implies that

$$\mathbb{E} \|G_{r,c}\|^{2k} = \int_0^\infty 2kt^{2k-1} \mathbb{P}(\|G_{r,c}\| \geq t) dt \leq (C'\sqrt{k})^{2k}.$$

Now, observe that:

- $\text{Tr}(G_{r,c}^T G_{r,c})^k$ is the sum of the $2k$ -th powers of singular values of $G_{r,c}$:

$$\mathbb{E} \text{Tr}(G_{r,c}^\ell G_{r,c})^k \leq \min(r, c) \mathbb{E} \|G_{r,c}\|^{2k} \leq (C'\sqrt{k})^{2k} \leq (C''k)^k$$

- we can expand $\mathbb{E} \text{Tr}(G_{r,c}^\ell G_{r,c})^k$ as:

$$\mathbb{E} \text{Tr}(G_{r,c}^\ell G_{r,c})^k = \sum_{\substack{2k \text{ tuples } (r_i, s_i) \subset I \times J \\ \text{s.t. Def 1/2/3 are satisfied}}} \mathbb{E} g_{r_1 s_1} \cdots g_{r_{2k} s_{2k}}$$

Notice that all the terms in the sum are ≥ 0 and, actually, ≥ 1 since $\mathbb{E}g_{ij}^p \geq 1$ for p being an even integer. Then, it's clear that, since this setting includes all the V-graphs considered in the statement of the lemma:

$$\begin{aligned}
 \text{number of V-graphs} &= \sum_{\substack{2k \text{ tuples } (r_i, s_i) \subset I \times J \\ \text{s.t. Def 1/2/3/4 are satisfied}}} 1 \leq \sum_{\substack{2k \text{ tuples } (r_i, s_i) \subset I \times J \\ \text{s.t. Def 1/2/3 are satisfied}}} 1 \\
 &\leq \sum_{\substack{2k \text{ tuples } (r_i, s_i) \subset I \times J \\ \text{s.t. Def 1/2/3 are satisfied}}} \mathbb{E}g_{r_1, s_1} \cdots g_{r_{2k}, s_{2k}} \\
 &= \mathbb{E} \text{Tr} \left(G_{r, c}^\ell G_{r, c} \right)^k \leq (C''k)^k
 \end{aligned}$$

Thus, the lemma is proved. $\square$

Now, with the result of Lemma 6.19, the number of V-graph G such that $r(G) = r$, $c(G) = c$ can be bounded by:

$$\binom{n}{r} \binom{d}{c} (Ck)^k \leq \frac{(ne)^r (de)^c}{r^c c^c} (Ck)^k \leq \frac{(C_1)^k n^r d^c}{\ell^\ell} k^k \leq (C_2)^k n^r d^c k^{k-\ell}$$

where we used the inequalities:

- $\binom{B}{b} \leq \left(\frac{Be}{b}\right)^b$, in the first inequality
- $r^r c^c \geq (\ell/2)^\ell$, in the second step of the inequality
- $\frac{k^k}{\ell^\ell} \leq (ke)^{k-\ell}$, in the third step of the inequality

Moreover, we can use Lemma 6.18, to claim that:

$$\begin{aligned}
 n^{r(G)} d^{c(G)} &\leq \begin{cases} n^{\ell(G)} & \text{if } \ell(G) \leq k/2, \\ n^{k/2} n^{\frac{\ell(G)}{k/2}} & \text{if } \ell(G) \geq k/2 \end{cases} \\
 &\leq n^k \left(\frac{d}{n}\right)^{k/2} \frac{1}{d^{k-\ell(G)}}.
 \end{aligned}$$

Thus, fixing an ℓ , we can plug this bound into the inequality for the "number of V-graphs" above, and the number of V-graphs with the same ℓ can be upper-bounded by:

$$(C_2)^k k^{k-\ell(G)} n^k \left(\frac{d}{n}\right)^{k/2} \frac{1}{d^{k-\ell(G)}}$$

We can apply this result to inequality (6.2), using that $n_+(G) \leq 4(k - \ell + 1)$ and setting $m = \ell - 1$:

$$\begin{aligned}
 \mathbb{E} \|\Sigma_n - \text{diag}(\Sigma_n)\|^k &\leq \frac{C^k}{n^k} \sum_{\text{v-graph}} K^{n_+(G)} \\
 &\leq \left(\frac{C}{n}\right)^k \sum_{\ell=2}^{k+1} k^{4(k-\ell+1)} \cdot k \cdot C_2^k \cdot k^{k-\ell} \cdot n^k \left(\frac{d}{n}\right)^{k/2} \frac{1}{d^{k-\ell}} \\
 &\leq C^k \cdot \left(\frac{d}{n}\right)^{k/2} d \sum_{m=1}^k \left(\frac{k^5}{d}\right)^{k-m}
 \end{aligned}$$

We now choose k to be the smallest even integer such that $k \geq d^{1/5}$. We then use the inequality $d \sum (k^5/d)^{k-m} \leq C^k$ to finish the proof, obtaining:

$$\left(\mathbb{E} \|\Sigma_n - \text{diag}(\Sigma_n)\|^k\right)^{1/k} \leq C \sqrt{\frac{d}{n}}$$

Thus, we have completed the proof of Lemma 6.11.

We come back, now, to the proof of the main theorem. Indeed, to bound the second term in which the quantity $\|\Sigma_n - I\|$ was decomposed, at the end of the proof we used a Bernstein type inequality for ψ_α functions, Lemma 6.13. We will turn, now, to the proof of this Lemma.

6.4.2 Proof of Lemma 6.13

We need an interesting result by Latała for this.

Corollary 6.20 ([30]) *If $p \geq 2$ and $Y, Y_1, \dots, Y_m$ are i.i.d symmetric random variables then we have*

$$\|Y_1 + \dots + Y_m\|_p \leq \sup \left\{ \frac{p}{s} \left(\frac{m}{p}\right)^{1/s} \|Y\|_s : \max\left(2, \frac{p}{m}\right) \leq s \leq p \right\}.$$

Proof of Corollary 6.20

We will, now, prove the result stated above. To do so, we will need to introduce a couple of additional definitions and lemmas.

Definition 6.21 *Let us define the following function on $\mathbb{R}$ for $p > 0$:*

1. $\varphi_p(x) = |1 + x|^p$
2. $\tilde{\varphi}_p(x) = \frac{\varphi_p(x) + \varphi_p(-x)}{2}$

Definition 6.22 *For a random variable X we define:*

$$\Phi_p(x) = \mathbb{E} \varphi_p(x)$$

From this, we can define, for a sequence $(X_i)_{i \in I}$ of independent symmetric random variables, indexed by a set I , we define the following Orlicz norm:

$$\| (X_i)_{i \in I} \|_{Orlicz, p} = \inf \left\{ t > 0 : \sum \ln \left(\Phi_p \left(\frac{X_i}{t} \right) \right) \leq p \right\}$$

Lemma 6.23 ([30]) For any $p \geq 2$ and real numbers $a < b < c < d$, satisfying the condition $a + d = b + c = 2$, the function

$$f(t) = |a + t|^p + |b - t|^p + |c - t|^p + |d + t|^p$$

is non-decreasing for $t \geq 0$.

Proof (of Lemma 6.23) Notice that f is convex, therefore it is enough to check that $f'(0) \geq 0$. By doing the derivative, one can see that $\frac{1}{p}f'(0) = |a|^{p-2}a - |b|^{p-2}b - |c|^{p-2}c + |d|^{p-2}d = g(d-1) - g(c-1)$, where

$$g(s) = |1 + s|^{p-2}(1 + s) + |1 - s|^{p-2}(1 - s).$$

So it is enough to show that the function g is nondecreasing on $[0, \infty)$. This is true since

$$g'(s) = (p-1) \left((1+s)^{p-2} - (1-s)^{p-2} \right) \geq 0 \quad \text{for } s \in (0, 1)$$

and

$$g'(s) = (p-1) \left((1+s)^{p-2} - (s-1)^{p-2} \right) \geq 0 \quad \text{for } s \in (1, \infty).$$

Thus, the lemma is proved. $\square$

Lemma 6.24 ([30]) For $Y_1, \dots, Y_m$ independent symmetric random variables and $p \geq 2$ we have

$$\Phi_p(Y_1 + \dots + Y_m) \leq \Phi_p(Y_1) \cdot \dots \cdot \Phi_p(Y_m).$$

Proof (of Lemma 6.24) The proof can be reduced to the case of $m = 2$ and to the case $cY_1 = x\varepsilon_1, Y_2 = y\varepsilon_2$, with $0 \leq y \leq x$. With this new formulation, the problem is reduced to proving the inequality:

$$\tilde{\Phi}_p(x+y) + \tilde{\Phi}_p(x-y) \leq 2\tilde{\Phi}_p(x)\tilde{\Phi}_p(y).$$

In this case, we can apply Lemma 6.23, by setting $a = 1 - x - y, b = 1 - x + y, c = 1 + x - y$ and $d = 1 + x + y$ and the result is proved. $\square$

Lemma 6.25 ([30]) If $t \geq 1, |x| \geq 1$ and $p \geq 1$, then

$$\tilde{\Phi}_p(tx) \geq t^{p/2}\tilde{\Phi}_p(x)$$

Proof (of Lemma 6.25) Let us fix $x \geq 1$ and define for $t \geq 1$, and define:

$$f(t) = \ln \tilde{\varphi}_p(tx) - \frac{p}{2} \ln t.$$

Then the claim we want to prove is equivalent to showing that $f(t) \geq f(1)$. This is true, since f is non-decreasing on $[1, \infty)$. Indeed:

$$f'(t) = \frac{p}{2t} \frac{(tx-1)(tx+1)^{p-1} + (tx+1)(tx-1)^{p-1}}{(tx-1)^p + (tx+1)^p} \geq 0,$$

which proves the result. $\square$

Lemma 6.26 ([30]) *If X, Y are independent symmetric random variables, then:*

$$\Phi_p \left(2X + \Phi_p(X)^{2/p} Y \right) \geq \Phi_p(X) \Phi_p(Y)$$

Proof (of Lemma 6.26) By the convexity of φ_p , we obtain $E\varphi_p(a+b\varepsilon) \leq E\varphi_p(a+c\varepsilon)$ for real numbers a, b, c , such that $|b| \leq |c|$. Therefore, since $\Phi_p(X) \geq 1$, we have for any real numbers x, y with $|y| \leq 1$,

$$\begin{aligned} \mathbb{E} \tilde{\varphi}_p(\varepsilon_1 x) \tilde{\varphi}_p(\varepsilon_2 y) &= \tilde{\varphi}_p(x) \tilde{\varphi}_p(y) = \mathbb{E} \varphi_p(\varepsilon_2 y + \varepsilon_1(x + \varepsilon_2 x y)) \\ &\leq \mathbb{E} \varphi_p(\varepsilon_2 y + \varepsilon_1 2x) \leq \mathbb{E} \varphi_p(2\varepsilon_1 x + \Phi_p(X)^{2/p} \varepsilon_2 y) \\ &= \mathbb{E} \tilde{\varphi}_p(2\varepsilon_1 x + \Phi_p(X)^{2/p} \varepsilon_2 y). \end{aligned}$$

For $|y| \geq 1$, instead, we have, by Lemma 6.25:

$$\tilde{\varphi}_p(tx) \geq t^{p/2} \tilde{\varphi}_p(x)$$

Therefore, it follows:

$$\mathbb{E} \tilde{\varphi}_p(2\varepsilon_1 x + \Phi_p(X)^{2/p} \varepsilon_2 y) \geq \mathbb{E} \tilde{\varphi}_p(\Phi_p(X)^{2/p} \varepsilon_2 y) \geq \Phi_p(X) \mathbb{E} \tilde{\varphi}_p(\varepsilon_2 y)$$

Putting together the cases $|y| \geq 1$ and $|y| \leq 1$ gives the proof of the lemma. $\square$

Lemma 6.27 ([30]) *If $Y_1, Y_2, \dots, Y_m$ are independent symmetric random variables such that $\Phi_p(Y_1) \cdots \Phi_p(Y_m) \leq e^p$, then for $p \geq 1$,*

$$\Phi_p(2e^2(Y_1 + \dots + Y_m)) \geq \Phi_p(Y_1) \cdots \Phi_p(Y_m)$$

Proof (of Lemma 6.27) Let $V_k = 2(\Phi_p(Y_1) \cdots \Phi_p(Y_k))^{2/p}(Y_1 + \dots + Y_k)$. We prove by induction that

$$\Phi_p(V_k) \geq \Phi_p(Y_1) \cdots \Phi_p(Y_k).$$

- **Base case:** for $k = 1$ it is obvious
- **Induction step:** assume that the thesis holds for some k . We want to prove that this implies that the thesis holds also for $k + 1$. Observe that by monotonicity of φ_p Lemma 6.26 and the induction hypothesis we have:

$$\begin{aligned}\Phi_p(V_{k+1}) &\geq \Phi_p\left(2Y_{k+1} + \Phi_p^{2/p}(Y_{k+1})V_k\right) \geq \Phi_p(Y_{k+1})\Phi_p(V_k) \\ &\geq \Phi_p(Y_1) \cdots \Phi_p(Y_k) \cdot \Phi_p(Y_{k+1}).\end{aligned}$$

Thus, we proved by induction our thesis. $\square$

We can, finally, state the theorem from which Corollary 6.20 will follow.

Theorem 6.28 ([30]) *Let $Y_1, Y_2, \dots, Y_m$ be a sequence of independent symmetric random variables, and $p \geq 2$. Then the following inequalities hold:*

$$\frac{e-1}{2e^2} \|(Y_i)_{i \in I}\|_{\text{Orlicz}, p} \leq \|Y_1 + \dots + Y_m\|_p \leq e \|(Y_i)_{i \in I}\|_{\text{Orlicz}, p}$$

Proof (of Theorem 6.28) We first prove the right part of the inequality:

- $\|Y_1 + \dots + Y_m\|_p \leq e \|(Y_i)_{i \in I}\|_{\text{Orlicz}, p}$

To begin, take $t > 0$ and assume that:

$$\sum \ln\left(\Phi_p\left(\frac{Y_i}{t}\right)\right) = p$$

By this assumption, it follows that $\Phi_p\left(\frac{Y_1}{t}\right) \cdots \Phi_p\left(\frac{Y_m}{t}\right) = e^p$.

By Lemma 6.24:

$$\Phi_p\left(\frac{Y_1 + \dots + Y_m}{t}\right) \leq e^p$$

Notice that, by definition, $\varphi_p(x) \geq |x|^p \ \forall x$, so for any symmetric random variable W , $\Phi_p(W) \geq \|W\|_p^p$ and therefore:

$$\|Y_1 + \dots + Y_m\|_p \leq et$$

We finish by taking the infimum over t , thus achieving the Orlicz norm of Definition 6.22.

Now we will prove the other side of the inequality:

- $\frac{e-1}{2e^2} \|(Y_i)_{i \in I}\|_{\text{Orlicz}, p} \leq \|Y_1 + \dots + Y_m\|_p$

To show this inequality, let us observe that by Lemma 6.27, since at the beginning of the proof we assumed $\Phi_p\left(\frac{Y_1}{t}\right) \dots \Phi_p\left(\frac{Y_m}{t}\right) = e^p$, we get:

$$\Phi_p\left(2e^2 \frac{Y_1 + \dots + Y_m}{t}\right) \geq e^p$$

By the triangle inequality, for any random variable W we have the following inequality:

$$\Phi_p(W) \leq (1 + \|W\|_p)^p \quad \text{for } p \geq 2$$

Therefore, for $p \geq 2$, we obtain:

$$e^p \leq \Phi_p\left(2e^2 \frac{Y_1 + \dots + Y_m}{t}\right) \leq \left(1 + \left\|2e^2 \frac{Y_1 + \dots + Y_m}{t}\right\|_p\right)^p$$

By a simple calculation and taking again the infimum over t , we obtain the desired lower estimates, thus completing the proof. $\square$

Finally, we have all the ingredients to prove the result we need, Corollary 6.20.

Proof (of Corollary 6.20) By Theorem 6.28, which we just proved, we have:

$$\|Y_1 + \dots + Y_m\|_p \sim \|Y_i\|_{\text{Orlicz},p} = \inf \left\{ t > 0 : \Phi_p\left(\frac{Y}{t}\right) \leq e^{p/m} \right\}.$$

Therefore, we will relate $\|Y_i\|_{\text{Orlicz},p}$ with the upper bound we want to achieve.

To prove the upper bound, we may assume that:

$$\sup \left\{ \frac{p}{s} \left(\frac{m}{p}\right)^{1/s} \|Y\|_s : \max\left(1, \frac{p}{m}\right) \leq s \leq p \right\} = t.$$

First, we apply triangle inequality, to the function $\varphi_p(x)$:

$$\varphi_p(w) = |1 + w|^p \leq (1 + |w|)^p$$

By Newton binomial (appendix), we can write:

$$\varphi_p(w) \leq \sum_{k \leq p} \binom{p}{k} |w|^k + |w|^p$$

and $\binom{p}{k} \leq (ep/k)^k$, if $n \geq p$ we have that:

$$\Phi_p\left(\frac{Y}{2et}\right) \leq \sum_{k \leq p} \frac{p^k}{(2tk)^k} \|X\|_k^k + \frac{\|Y\|_p^p}{(2et)^p} \leq 1 + \frac{p}{n} \leq e^{p/n}.$$

If $p \geq n$, we have $(p/n)^{1/k} \leq k^{1/k} < e$ for $k \geq p/n$. Also $\|Y\|_k \leq \|Y\|_{p/n}$ for $k \leq p/n$. Therefore, again by Newton binomial, we obtain:

$$\begin{aligned} \Phi_p \left(\frac{Y}{2et} \right) &\leq \exp(p\|Y/2et\|_{p/n}) + \sum_{p/n < k < p} \frac{p^k}{(2tk)^k} \|Y\|_k^k + \frac{\|Y\|_p^p}{(2et)^p} \\ &\leq e^{p/2n} + \frac{p}{n} \leq e^{p/n} \end{aligned}$$

Thus, the result follows. $\square$

Having this result in mind, we can continue the proof of Lemma 6.13. First, we notice the following:

- If Y_i are independent, mean zero random variables, and (ε_i) is a Bernoulli sequence independent of $(Y_i)_{i \in I}$ then

$$\frac{1}{2} \left\| \sum Y_i \right\|_p \leq \left\| \sum \varepsilon_i Y_i \right\|_p \leq 2 \left\| \sum Y_i \right\|_p$$

Therefore, we can apply the above theorems and, specifically, Corollary 6.20 for the sequence of symmetric random variables $(\varepsilon_i Y_i)$, thus translating the same result to the mean zero random variables (Y_i) , with slightly worse constants

To prove Lemma 6.13, we notice that, by Definition 2.34 and the equivalent formulation (2.1) of the Orlicz norm, we have that:

$$\frac{\|Y_1\|_s}{s^{1/\alpha}} \leq \psi_\alpha(Y_1)$$

With the hypotheses of the lemma, $\psi_\alpha(Y_1) \leq K$, therefore: $\|Y_1\|_s \leq Ks^{1/\alpha}$. By using Corollary 6.20:

$$\begin{aligned} \|Y_1 + \dots + Y_m\|_p &\leq \sup_{s \in [\max(2, \frac{p}{m}), p]} \frac{p}{s} \left(\frac{m}{p} \right)^{1/s} \cdot K \cdot s^{1/\alpha} \\ &\leq \begin{cases} C(\alpha) K \sqrt{mp} & \text{if } 2 \leq p \leq m^{\frac{\alpha}{2-\alpha}} \\ C(\alpha) K p^{1/\alpha} & \text{if } p \geq m^{\frac{\alpha}{2-\alpha}} \end{cases} \end{aligned}$$

Now, let $X = \frac{1}{m} (Y_1 + \dots + Y_m)$. Using Markov's inequality we get:

$$\begin{aligned} \mathbb{P}(|X| \geq t) &= \mathbb{P}(|Y_1 + \dots + Y_m|^p \geq (t \cdot m)^p) \leq \frac{\mathbb{E} |Y_1 + \dots + Y_m|^p}{(t \cdot m)^p} \\ &\leq \begin{cases} \left(\frac{C(\alpha) K \sqrt{p}}{t \sqrt{m}} \right)^p & \text{if } 2 \leq p \leq m^{\frac{\alpha}{2-\alpha}} \\ \left(\frac{C(\alpha) K p^{1/\alpha}}{tm} \right)^p & \text{if } p \geq m^{\frac{\alpha}{2-\alpha}} \end{cases} \end{aligned}$$

If $C'K/\sqrt{m} \leq t \leq 2CKm^{-\frac{1-\alpha}{2-\alpha}}$ we use the first bound with $p = \frac{mt^2}{4C^2K^2}$; if $t \geq 2CKm^{-\frac{1-\alpha}{2-\alpha}}$ we use the second bound with $p = \frac{mt}{(2CK)^\alpha}$. In both cases, we obtain $\mathbb{P}(|X| \geq t) \leq 1/2^p$, and the lemma is proved.

6.5 Appendix of this Chapter

Proof of Brunn-Minkowski inequality

There are many proofs of the Brunn-Minkowski inequality. We will present one valid exclusively for the class of convex bodies and based on Brunn's concavity principle, which we introduce now.

Theorem 6.29 (*Brunn's concavity principle*) *Let K be a convex body in $\mathbb{R}^d$ and let F be a k -dimensional subspace. Then, the function $f : F^\perp \mapsto \mathbb{R}$ defined by $f(x) = |K \cap (F + x)|^{1/k}$ is concave on its support.*

In order to be able to prove Brunn's principle, we introduce Steiner's symmetrization:

- for any convex body K in $\mathbb{R}^n$ and any $\theta \in S^{n-1}$ we consider the set $S_\theta(K)$ consisting of all points of the form $x + \lambda\theta$, where x is in the projection $P_{\theta^\perp}(K)$ of K onto $\theta^\perp$ and $|\lambda| \leq \frac{1}{2} \cdot \text{length}[(x + \mathbb{R}\theta) \cap K]$. In other words, we obtain $S_\theta(K)$ by transforming the chords (or line segments) of K . We transform them by, for each chord in K , sliding or moving it along the line perpendicular to θ . Then, the new position of each chord is determined by ensuring that the midpoint of the chord lies on the space perpendicular to θ . Finally, we take the union of all resulting chords.

The set $S_\theta(K)$ is the Steiner symmetrization of K in the direction of θ .

From the definition, some basic properties of Steiner symmetrization follow:

1. Steiner symmetrization preserves convexity: if K is a convex body then $S_\theta(K)$ is also a convex body.
2. Steiner symmetrization preserves volume: $|S_\theta(K)| = |K|$.
3. If K_1 and K_2 are two convex bodies then, for every $\lambda \in (0, 1)$,

$$S_\theta(\lambda K_1 + (1 - \lambda)K_2) \supseteq \lambda S_\theta(K_1) + (1 - \lambda)S_\theta(K_2).$$

A very useful well-known fact is that, if F is a k -dimensional subspace of $\mathbb{R}^d$, $1 \leq k \leq n$, for every convex body K one can find a sequence of successive Steiner's symmetrizations in directions $\theta \in F$ so that the limiting convex body $\tilde{K}$ has the following property:

for every $x \in F^\perp$, $\tilde{K} \cap (F + x)$ is a ball with center at x and radius $r(x)$ such that $|\tilde{K} \cap (F + x)| = |K \cap (F + x)|$.

Then, we can effectively prove Brunn's concavity principle.

Proof (of Theorem 6.29) Using the above observation, the proof is easily completed: using the convexity of $\tilde{K}$ we see that the function of the radius r is concave on its support, and hence f is also concave. $\square$

Having this result, we can, now, prove Brunn-Minkowski inequality for convex bodies.

This result follows from Brunn's concavity principle as follows: if K and T are convex bodies in $\mathbb{R}^d$, we define

$$K_1 = K \times \{0\} \quad \text{and} \quad T_1 = T \times \{1\}$$

in $\mathbb{R}^{d+1}$ and consider their convex hull which we will call H . If

$$H(t) = \left\{ x \in \mathbb{R}^d : (x, t) \in L \right\} \quad \text{for } t \in [0, 1]$$

From this definition it follows that $H(0) = K$, $H(1) = T$ and

$$H\left(\frac{1}{2}\right) = \frac{K + T}{2}.$$

Then, Brunn's concavity principle for $F = \mathbb{R}^d$ shows that

$$\left| \frac{K + T}{2} \right|^{1/d} \geq \frac{1}{2} |K|^{1/d} + \frac{1}{2} |T|^{1/d},$$

thus proving the desired result.

Proof of Borell's Lemma

We will start with a simple claim.

Claim 6.30 *Let A be a closed, convex and symmetric set:*

$$A^c \supseteq \frac{2}{t+1} (tA)^c + \frac{t-1}{t+1} A$$

.

Proof (of Claim 6.30) This claim will be proved by contradiction. Indeed, suppose that the statement of the claim does not hold. Then, there exists $a \in A$ which can be written in the form $a = \frac{2}{t+1}y + \frac{t-1}{t+1}a_1$ for some $a_1 \in A$ and $y \notin tA$. We can use, then, the convexity and symmetry of A , and we can write

$$\frac{1}{t}y = \frac{t+1}{2t}a + \frac{t-1}{2t}(-a_1) \in A,$$

which implies that $y \in tA$. This shows a contradiction. $\square$

Continuing the proof of Borell's lemma, since K is convex, we then have:

$$A^c \cap K \supseteq \frac{2}{t+1} ((tA)^c \cap K) + \frac{t-1}{t+1} (A \cap K).$$

We can now apply the Brunn-Minkowski inequality and we get:

$$1 - \theta = |A^c \cap K| \geq |(tA)^c \cap K|^{\frac{2}{t+1}} \cdot |A \cap K|^{\frac{t-1}{t+1}} = |(tA)^c \cap K|^{\frac{2}{t+1}} \cdot \theta^{\frac{t-1}{t+1}},$$

and the result follows.

Additional useful results for isotropic unconditional convex bodies

We first state a couple of results regarding log-concave functions that will be used in this section. For the sake of clarity, we start with a couple of definitions.

Definition 6.31 Let $f : \mathbb{R}^d \mapsto [0, \infty)$ be an integrable function with finite, positive integral. The barycenter of f is defined as:

$$\text{bar}(f) = \frac{\int_{\mathbb{R}^d} x f(x) dx}{\int_{\mathbb{R}^d} f(x) dx}$$

We say that f is centered if it has barycenter at the origin, i.e. :

$$\int_{\mathbb{R}^d} \langle x, \theta \rangle f(x) dx = 0 \quad \forall \theta \in \mathbb{S}^{d-1}$$

Definition 6.32 A function $f : \mathbb{R}^d \mapsto [0, \infty)$ is called log-concave if:

$$f((1-\lambda)x + \lambda y) \geq f(x)^{1-\lambda} f(y)^\lambda$$

The first result, due to Fradelizi, shows that the value of a centered log-concave function at the origin is comparable to its maximum (up to a constant depending on the dimension).

Theorem 6.33 ([25]) Let $f : \mathbb{R}^d \mapsto [0, \infty)$ be a centered log-concave function. Then:

$$f(0) \leq \|f\|_\infty \leq e^d f(0).$$

Proof (of Theorem 6.33) Without loss of generality, assume that $f > 0$, continuously differentiable and that $\int_{\mathbb{R}^d} f(x) dx = 1$. Then:

$$\begin{aligned} \log f(0) &= \log f \left(\int_{\mathbb{R}^d} y f(y) dy \right) && \text{since } f \text{ is centered} \\ &\geq \int_{\mathbb{R}^d} f(y) \log f(y) dy && \text{by Jensen's inequality} \end{aligned}$$

Let $x \in \mathbb{R}^d$. Since f is log-concave:

$$-\log f(x) \geq -\log f(y) + \langle x - y, \nabla (-\log f)(y) \rangle$$

Then, multiply by $f(y)$ and integrate over y :

$$\begin{aligned} -\log f(x) &= -\log f(x) \int_{\mathbb{R}^d} f(y) dy \\ &\geq -\int_{\mathbb{R}^d} f(y) \log f(y) + \int_{\mathbb{R}^d} \langle x - y, -\nabla f(y) \rangle dy \quad \text{since } f(y) \nabla (-\log f) = -\nabla f \\ &\geq -\int_{\mathbb{R}^d} f(y) \log f(y) - d \quad \text{by integration by parts} \end{aligned}$$

Then, we get, from the first relation for $f(0)$ and the above inequality:

$$\log f(0) \geq \int_{\mathbb{R}^d} f(y) \log f(y) \geq \log f(x) - d \quad \forall x$$

By taking the supremum over x , we get the result. $\square$

Lemma 6.34 ([16]) Let $f : [0, \infty) \mapsto [0, \infty)$ be a log-concave function. Then,

$$F(p) := \left(\frac{p}{\|f\|_\infty} \int_0^\infty x^{p-1} f(x) dx \right)^{1/p}$$

is an increasing function of p on $(0, \infty)$.

Proof (of Lemma 6.34) Without loss of generality, we can assume that $\|f\|_\infty = 1$. Then, for any $0 < p < q$ and $\alpha > 0$:

$$\begin{aligned} \frac{F(q)^q}{q} &= \int_0^\infty x^{q-1} f(x) dx = \int_0^\alpha x^{q-1} f(x) dx + \int_\alpha^\infty x^{q-1} f(x) dx \\ &\geq \int_0^\alpha x^{q-1} f(x) dx + \alpha^{q-p} \int_\alpha^\infty x^{p-1} f(x) dx \\ &= \alpha^{q-p} \int_\alpha^\infty x^{p-1} f(x) dx + \alpha^{q-p} \int_0^\alpha x^{p-1} f(x) dx \\ &\quad - \alpha^{q-p} \int_0^\alpha x^{p-1} f(x) dx + \int_0^\alpha x^{q-1} f(x) dx \\ &= \alpha^{q-p} \frac{F(p)^p}{p} - \alpha^q \int_0^1 (x^{p-1} - x^{q-1}) f(\alpha x) dx \\ &\geq \alpha^{q-p} \frac{F(p)^p}{p} - \alpha^q \left(\frac{1}{p} - \frac{1}{q} \right) \end{aligned}$$

Then, take the minimum over α and see that the choice $\alpha = F(p)$ minimizes the right-hand side, thus giving the desired result. $\square$

Next, we want to study the relation of the moments of the distribution of a random vector sampled from a centered convex body K with the volume of the hyperplane sections of K passing through the origin.

We first observe that the hyperplane sections through the barycenter are, up to an absolute constant, maximal.

Proposition 6.35 ([16]) *Let K be a centered convex body of volume 1 in $\mathbb{R}^d$. Let $\theta \in \mathbb{S}^{d-1}$ and consider the function*

$$f(t) = f_{K,\theta}(t) = \text{vol}_{d-1}(K \cap \{\langle x, \theta \rangle = t\}), \quad t \in \mathbb{R}.$$

Then,

$$\|f\|_\infty \leq ef(0) = e \text{vol}_{d-1}(K \cap \theta^\perp).$$

Proof (of Proposition 6.35) Brunn-Minkowski inequality implies that f is a log-concave density on $\mathbb{R}$ and the assumption that K is centered shows that $\text{bar}(f) = 0$. Therefore, we can apply Theorem 6.33 with dimension $d = 1$, obtaining the desired result. $\square$

Now, we will make the observation that for every $\theta \in \mathbb{S}^{d-1}$ the volume of the $(d-1)$ -dimensional section $\text{vol}_{d-1}(K \cap \theta^\perp)$ of K is closely related to the L_q norms of the linear functional $x \mapsto \langle x, \theta \rangle$. Since this result holds for all centered convex bodies, we will express the L_q norms as an integral over the convex body, without needing to use the probabilistic notation with the expectation, which would imply that we have to deal with a random variable sampled from such convex body. Therefore we will use as norm the following:

$$\|\langle \cdot, \theta \rangle\|_q = \left(\int_K |\langle x, \theta \rangle|^q dx \right)^{1/q}$$

Proposition 6.36 ([35]) *Let K be a centered convex body of volume 1 in $\mathbb{R}^d$. Let $f(t) = f_{K,\theta}(t) = \text{vol}_{d-1}(K \cap \{\langle x, \theta \rangle = t\})$, $t \in \mathbb{R}$ as defined in Proposition 6.35. For every $\theta \in \mathbb{S}^{d-1}$ and every $q > 0$:*

$$\left(\int_K |\langle x, \theta \rangle|^q dx \right)^{1/q} \geq \frac{1}{2e(q+1)^{1/q}} \frac{1}{\text{vol}_{d-1}(K \cap \theta^\perp)}.$$

Proof (of Proposition 6.36) We first apply Lemma 6.34 to the functions $f \cdot \mathbb{1}_{\{t \geq 0\}}$ and $f \cdot \mathbb{1}_{\{t \leq 0\}}$. We have, for $f \cdot \mathbb{1}_{\{t \geq 0\}}$:

$$F(q+1) = \left(\frac{q+1}{\|f\|_\infty} \int_0^\infty t^q f(t) dt \right)^{1/(q+1)} \geq F(1) = \frac{1}{\|f\|_\infty} \int_0^\infty f(t) dt,$$

and similarly, $f \cdot \mathbb{1}_{\{t \leq 0\}}$:

$$\left(\frac{q+1}{\|f\|_\infty} \int_{-\infty}^0 |t|^q f(t) dt \right)^{1/(q+1)} \geq \frac{1}{\|f\|_\infty} \int_{-\infty}^0 f(t) dt$$

Now, we will add the two inequalities, obtaining:

$$\begin{aligned} \frac{1}{\|f\|_\infty} \int_0^\infty f(t) dt + \frac{1}{\|f\|_\infty} \int_{-\infty}^0 f(t) dt &= \frac{1}{\|f\|_\infty} && \text{since } \int_{\mathbb{R}} f(t) dt = \text{vol}_{d-1}(K) = 1 \\ &\leq \left(\frac{q+1}{\|f\|_\infty} \int_{-\infty}^0 |t|^q f(t) dt \right)^{1/(q+1)} + \left(\frac{q+1}{\|f\|_\infty} \int_0^\infty t^q f(t) dt \right)^{1/(q+1)} \\ &\leq 2^{\frac{q}{q+1}} \left(\frac{q+1}{\|f\|_\infty} \int_{-\infty}^\infty |t|^q f(t) dt \right) && \begin{array}{l} \text{from the inequality: } a^s + b^s \leq 2^{1-s}(a+b)^s \\ \forall a, b > 0 \text{ and } 0 < s < 1 \end{array} \end{aligned}$$

This shows that:

$$\int_K |\langle x, \theta \rangle|^q dx = \int_{-\infty}^\infty |t|^q f(t) dt \geq \frac{1}{\|f\|_\infty^q 2^q (1+q)}$$

Finally, we can apply Proposition 6.35 to obtain the result. $\square$

Next, we prove a result regarding the isotropic constant of an isotropic unconditional convex body. This result is due to Bobkov and Nazarov.

Theorem 6.37 ([13]) *Let K be an isotropic unconditional convex body in $\mathbb{R}^d$. Then $L_K^2 \leq \frac{1}{2}$*

Proof (of Theorem 6.37) Recall the definition of the normalized part K^+ of the body K . K^+ has the property, already stated after Definition 6.7 that:

- if $x \in K^+$ and $0 \leq y_j \leq x_j$, then $y \in K^+$

This implies that, if $x = (x_1, \dots, x_d) \in K^+$, then $\prod_{j=1}^d [0, x_j] \subseteq K^+$. It follows that:

$$\prod_{j=1}^d x_j \leq |K^+| = 1 \quad \forall x \in K^+$$

We define the following set:

$$V = \left\{ x \in \mathbb{R}_+^d : \prod_{j=1}^d x_j \geq 1 \right\}$$

V is a convex set and notice that K^+ and V have disjoint interiors. Therefore, there exists a hyperplane H of the form

$$\lambda_1 x_1 + \dots + \lambda_d x_d = \alpha$$

with $\lambda_j > 0$, which touches V at a point $y = (y_1, \dots, y_d)$ and separates it from K^+ . Since H touches V we can choose the λ_j 's so that $\alpha = d$ and $\prod_{j=1}^d \lambda_j = 1$ (we have $\prod_{j=1}^n y_j = 1$ and we can choose $\lambda_j = \prod_{i \neq j} y_i$). Since H separates V from K^+ , we have

$$K^+ \subseteq \left\{ x \in \mathbb{R}_+^d : \lambda_1 x_1 + \dots + \lambda_d x_d \leq d \right\}$$

Next, we can use the arithmetic-geometric means inequality (Theorem A.2), obtaining:

$$\prod_{j=1}^d \left(\int_{K^+} x_j dx \right)^{\lambda_j/d} \leq \int_{K^+} \frac{\lambda_1 x_1 + \dots + \lambda_d x_d}{d} dx \leq 1.$$

Now, applying the standard reverse Hölder inequality for $f(x) = x_j$, we get:

$$4L_K^2 = \int_{K^+} x_j^2 dx \leq 2 \left(\int_{K^+} x_j dx \right)^2 \quad \forall j = 1, \dots, d$$

If we set $w = \frac{\lambda_1 + \dots + \lambda_n}{n}$, then

$$\begin{aligned} (2L_K^2)^w &= \prod_{j=1}^d (2L_K^2)^{\lambda_j/d} \leq \prod_{j=1}^d \left(\int_{K^+} x_j dx \right)^{2\lambda_j/d} \\ &= \left[\prod_{j=1}^d \left(\int_{K^+} x_j dx \right)^{\lambda_j/d} \right]^2 \leq 1, \end{aligned}$$

which shows that $L_K^2 \leq 1/2$. $\square$

Finally, we can combine the results obtained above to get the following result.

Corollary 6.38 *Let K be an isotropic unconditional convex body in $\mathbb{R}^d$. Then:*

$$\text{vol}_{d-1}(K \cap \theta^\perp) \geq \frac{1}{\sqrt{6}} \quad \forall \theta \in \mathbb{S}^{d-1}$$

Proof (of Corollary 6.38) We get this result by combining Theorem 6.37 with Proposition 6.36. $\square$

Comparison theorem and Subindependence

To prove the lemmas 6.16, 6.17, which we used in the proof of Lemma 6.11, we first introduce some results on the geometry of isotropic unconditional convex bodies and linear functionals on unconditional convex bodies. Let K be an isotropic unconditional convex body and let $T = CdB_1^2$. Observe that $K \subseteq T$, for an appropriate constant C . For instance, $C = \sqrt{6}$ works.

We denote by μ_K and μ_T , uniform distributions on K , T respectively. We start with the following distributional inequality about μ_K .

Claim 6.39 ([13]) *If K is isotropic unconditional, then $\forall \alpha_1, \dots, \alpha_d \geq 0$:*

$$\mu_K(|x_1| \geq \alpha_1, \dots, |x_d| \geq \alpha_d) \leq \left(1 - \frac{\alpha_1 + \dots + \alpha_d}{\sqrt{6d}}\right)^d$$

Proof (of Claim 6.39) We define a function $u : K^+ \mapsto [0, +\infty)$ by:

$$u(\alpha_1, \dots, \alpha_d) = \mu_K(|x_1| \geq \alpha_1, \dots, |x_d| \geq \alpha_d)$$

By the Brunn-Minkowski inequality, we can affirm that the function $h = u^{1/d}$ is concave on K^+ . Also, observe that, since the function u defines a probability: $u(0) = 1$. Also:

$$\frac{\partial u(\alpha)}{\partial \alpha_j} \Big|_{\alpha=0} = -\text{vol}_{d-1}(K_j),$$

where K_j are the sections of K defined by: $\{x \in K : x_j = 0\}$. Recall that we know, from Corollary 4.1.4 that: $\text{vol}_{d-1}(K_j) \geq \frac{1}{\sqrt{6}}$. Then, let $\alpha \in K^+$ and consider the function $h_\alpha : [0, 1] \mapsto \mathbb{R}$, defined by $h_\alpha(t) = h(\alpha t)$.

Then:

$$h'_\alpha(0) = \sum_{j=1}^d \alpha_j \frac{\partial h(\alpha)}{\partial \alpha_j}(0) = \sum_{j=1}^d \alpha_j \frac{1}{d} \frac{\partial u(\alpha)}{\partial \alpha_j}(0) \leq -\frac{\alpha_1 + \dots + \alpha_d}{\sqrt{6d}}$$

Moreover, since h is concave, h_α is concave on $[0, 1]$. Then, it follows that h'_α is decreasing on $[0, 1]$. Consequently:

$$h(\alpha) - 1 = h_\alpha(1) - h_\alpha(0) \leq -\frac{\alpha_1 + \dots + \alpha_d}{\sqrt{6d}}$$

From the definition of h follows the claim. $\square$

Now, we will state and prove another claim, based on the previous one, that will be fundamental in the proof of Lemma 6.16 by Bobkov and Nazarov.

Claim 6.40 ([12]) *Let $f : \mathbb{R}^d \mapsto \mathbb{R}^+$ be a function which satisfies:*

- *f is symmetric with respect to each coordinate i: $f(x_1, \dots, x_d) = f(\varepsilon_1 x_1, \dots, \varepsilon_d x_d)$ $\forall x \in \mathbb{R}^d$ and for all the choices of signs*
- *$\exists$ a function $q : \mathbb{R}_+^d \mapsto \mathbb{R}$ such that: $f(x) = \int_0^{x_1} \dots \int_0^{x_d} q(z) dz$ $\forall x = (x_1, \dots, x_d) \in \mathbb{R}_+^d$*

Then, $\forall$ isotropic 1-unconditional convex body K in $\mathbb{R}^d$, $\forall t \geq 0$:

$$\mu_K(\{x \in \mathbb{R}^d : f(x) \geq t\}) \leq \mu_T(\{x \in \mathbb{R}^d : f(x) \geq t\})$$

And, in particular, it follows:

$$\mathbb{E}_{\mu_K} f(x) \leq \mathbb{E}_{\mu_T} f(x)$$

Proof (of Claim 6.40) The proof will be done in two steps. The first step involves the proof of the following inequality:

$$\forall \alpha_1, \dots, \alpha_d \geq 0 :$$

$$\mu_K(|x_1| \geq \alpha_1, \dots, |x_d| \geq \alpha_d) \leq \mu_T(|x_1| \geq \alpha_1, \dots, |x_d| \geq \alpha_d)$$

Notice that, without loss of generality we can assume that $\alpha = (\alpha_1, \dots, \alpha_d) \in \text{interior}(K)$ so that the left-hand side is > 0 . Then, we also have $\alpha \in T$. Notice that:

$$\begin{aligned} \mu_T(|x_1| \geq \alpha_1, \dots, |x_d| \geq \alpha_d) &= \frac{\text{vol}_{d-1}(\{x \in T : |x_1| \geq \alpha_1, \dots, |x_d| \geq \alpha_d\})}{\text{vol}_{d-1}(T)} \\ &= \frac{\text{vol}_{d-1}((\sqrt{6}d - (\alpha_1 + \dots + \alpha_d)B_1^d))}{\text{vol}_{d-1}(\sqrt{6}dB_1^d)} \\ &= \left(1 - \frac{(\alpha_1 + \dots + \alpha_d)}{\sqrt{6}d}\right)^d \end{aligned}$$

Recalling the previous claim, it follows that:

$$\mu_K(|x_1| \geq \alpha_1, \dots, |x_d| \geq \alpha_d) \leq \mu_T(|x_1| \geq \alpha_1, \dots, |x_d| \geq \alpha_d)$$

Now, we are able to turn to the second part of the proof. First, observe that what we just proved can be written as:

$$\mu_K(F(|x|) \geq t) \leq \mu_T(F(|x|) \geq t)$$

for the function $F_\alpha = \mathbb{1}_{\{|x_1| \geq \alpha_1, \dots, |x_d| \geq \alpha_d\}}$. The function F_α corresponds to the Dirac measure which gives unit mass to the point α . The class $\{F_\alpha : \alpha \in \mathbb{R}^d_+\}$ coincides with the extreme points of cone defined by the functions that satisfy the hypothesis of the claim. Then, the result follows. $\square$

Now, we have all the ingredients to prove the comparison lemma 6.16.

Proof (of Lemma 6.16) We can finally prove the comparison lemma by Bobkov and Nazarov. Indeed, to apply Claim 6.40 we need to be able to express f as:

$$f(x) = \int_0^{x_1} \dots \int_0^{x_d} q(z) dz \quad \forall x = (x_1, \dots, x_d) \in \mathbb{R}_+^d$$

By taking $q(x_1, \dots, x_d) = f'_1(x_1) \cdot \dots \cdot f'_d(x_d)$, we can recover the result:

$$\mathbb{E} \prod_{i=1}^d f_i(|x_i|) \leq \mathbb{E} \prod_{i=1}^d f_i(|y_i|). \quad \square$$

We also prove Lemma 6.17 by Ball and Perissinaki which states the subindependence (of expectation).

Proof (of Lemma 6.17) This result is due to Ball and Perissinaki, which stated that the coordinate slabs $\{|x_i| \leq t\}$ of the ℓ_p^d in $\mathbb{R}^d$ are subindependent, if $X = (x_1, \dots, x_d)$ is sampled from a normalized Lebesgue measure, defined on the ℓ_p^d ball. In particular:

Claim 6.41 ([9]) *For any sequence $t_1, \dots, t_d$ of positive numbers,*

$$\mathbb{P} \left(\bigcap_{i=1}^d \{|x_i| \geq t_i\} \right) \leq \prod_{i=1}^d \mathbb{P}(|x_i| \geq t_i)$$

With this claim, then the result in expected value follows by the Layer cake representation, $\forall$ increasing functions $f_i : \mathbb{R}^+ \mapsto \mathbb{R}$, on ℓ_p^d -balls. By defining a function $F : \mathbb{R}^d \mapsto \mathbb{R}$ as: $F(y) = \prod_{i=1}^d f_i(|y_i|)$, then:

$$\mathbb{E} F(Y) = \int_0^\infty \mathbb{P}(\{y \in \mathbb{R}^d \text{ s.t. } : F(Y) \geq t\}) dt$$

Applying Claim 6.41, Lemma 6.17 follows, obtaining:

$$\mathbb{E} \prod_{i=1}^d f_i(|y_i|) \leq \prod_{i=1}^d \mathbb{E} f_i(|y_i|).$$

Therefore, now we turn to the proof of Claim 6.41.

Proof (of Claim 6.41) Observe that, to prove the claim regarding subindependence of coordinate slabs, it is enough to prove that:

$$\mathbb{P} \left(\bigcap_{i=1}^d \{|x_i| \geq t_i\} \right) \leq \mathbb{P}(|x_1| \geq t_1) \mathbb{P} \left(\bigcap_{i=2}^d \{|x_i| \geq t_i\} \right) \quad \square$$

First, we define the event S as: $S := \bigcap_{i=2}^d \{|x_i| \geq t_i\}$. Moreover, for simplicity, we will call A the ℓ_p^d ball in $\mathbb{R}^d$ which we are considering. Therefore, the probability can be written in terms of volume and our goal can be written as:

$$\frac{\text{vol}_{d-1}(A \cap S \cap \{|x_1| \geq t_1\})}{\text{vol}_{d-1}(A)} \leq \frac{\text{vol}_{d-1}(A \cap \{|x_1| \geq t_1\})}{\text{vol}_{d-1}(A)} \cdot \frac{\text{vol}_{d-1}(A \cap S)}{\text{vol}_{d-1}(A)}$$

Consider the following fact:

- if $h : [0, 1] \mapsto \mathbb{R}$ is increasing and satisfies: $\int_0^1 h d\mu = \alpha \mu[0, 1]$, for some positive measure μ and constant $\alpha > 0$, then:

$$\int_0^s h d\mu \leq \alpha \mu[0, s] \quad \text{for all } s \in [0, 1]$$

Thus, we take

$$h(u) = \frac{\text{vol}_{d-1}(A \cap S \cap \{|x_1| = 1 - u\})}{\text{vol}_{d-1}(A \cap \{|x_1| = 1 - u\})}$$

and let

$$g(u) = \frac{\text{vol}_{d-1}(A \cap \{|x_1| = 1 - u\})}{\text{vol}_{d-1}(A)}$$

be the density of our measure μ . Then:

$$\begin{aligned} \int_0^1 h(u) d\mu &= \int_0^1 \frac{\text{vol}_{d-1}(A \cap S \cap \{|x_1| = 1 - u\})}{\text{vol}_{d-1}(A \cap \{|x_1| = 1 - u\})} \frac{\text{vol}_{d-1}(A \cap \{|x_1| = 1 - u\})}{\text{vol}_{d-1}(A)} du \\ &= \frac{\text{vol}_{d-1}(A \cap S)}{\text{vol}_{d-1}(A)} \end{aligned}$$

and

$$\int_0^{1-t} f(u) d\mu = \frac{|A \cap S \cap \{|x_1| \geq t\}|}{|A|}$$

Then, applying the fact above by choosing $\alpha = \frac{|A \cap S|}{|A|}$ and $s = 1 - t_1$ completes the proof. $\square$

Case Study: Isotropic Log-concave Random Vectors

7.1 Preliminaries

The study of spectral properties of random matrices with independent rows or columns that have log-concave distribution is a line of research where much progress has been achieved. The underlying reason is that, because of the high-dimensional nature and of the log-concavity assumption, such matrices behave in, more or less, the same way as matrices with independent entries.

We will prove that, for the class of log-concave measures, a sample size linear in the dimension suffices to obtain an approximation of the true covariance matrix with ε -accuracy. In Chapter 6, we proved that, when a body K is invariant under every reflection with respect to coordinate subspaces, for any $\varepsilon \in (0, 1)$ there exist $C(\varepsilon) > 0$ such that we can closely approximate the covariance matrix whenever $n \geq C(\varepsilon)d$ and with high probability. Finally, the present chapter shows that the same is true for the class of log-concave distributions and, as a consequence, for an arbitrary body K (in the isotropic position).

Indeed, the uniform distribution over a convex set is log-concave.

We already introduced the notion of log-concave function in Chapter 6 (Definition 6.32).

In the following, we recall this definition and we introduce the notion of log-concave measure.

Definition 7.1 *A function $f : \mathbb{R}^d \mapsto [0, \infty)$ is called log-concave if, for any $\lambda \in [0, 1]$:*

$$f((1 - \lambda)x + \lambda y) \geq f(x)^{1-\lambda} f(y)^\lambda$$

Definition 7.2 *A random vector X is sampled from a log-concave measure μ if for any measurable subsets A, B of $\mathbb{R}^d$ and any $\lambda \in [0, 1]$:*

$$\mathbb{P}(X \in \lambda A + (1 - \lambda)B) \geq [\mathbb{P}(X \in A)]^\lambda \cdot [\mathbb{P}(X \in B)]^{1-\lambda}$$

whenever the set

$$\lambda A + (1 - \lambda)B = \{\lambda x_1 + (1 - \lambda)x_2 : x_1 \in A, x_2 \in B\}$$

is measurable.

The Brunn-Minkowski inequality, Theorem 6.8, provides examples of log-concave measures.

A version of Borell's lemma, Theorem 6.9, holds true in the more general context of log-concave probability measures.

Lemma 7.3 *Let X be a random vector sampled from a log-concave measure on $\mathbb{R}^d$. Then, for any symmetric convex set A in $\mathbb{R}^d$ with $\mathbb{P}(X \in A) = \alpha \in (0, 1)$ and any $t > 1$ we have:*

$$1 - \mathbb{P}(X \in tA) \leq \alpha \left(\frac{1-\alpha}{\alpha} \right)^{\frac{t+1}{2}}$$

7.2 Main Result

The question that we discuss in this chapter was posed by Kannan, Lovász, and Simonovits [28] and has its origin in the problem of finding a fast algorithm for the computation of the volume of a given convex body. In a more general probability setting, it can be stated as a covariance estimation problem.

Consider an isotropic log-concave random vector $X \in \mathbb{R}$ and a sequence $\{X_i\}_{i=1}^n$ of independent copies of X . We want to study the rate of convergence of the sample covariance matrix to the identity matrix Σ_n , the true covariance matrix.

Question 7.4 *Given $\varepsilon \in (0, 1)$, what is the minimal value of n for which we can guarantee, with high probability, that:*

$$\|\Sigma_n - I\| \leq \varepsilon ?$$

This question was optimally and comprehensively answered by Adamczak, Litvak, Pajor, and Tomczak-Jaegermann.

Theorem 7.5 ([2]) Let $X_1, \dots, X_n$ be i.i.d. isotropic log-concave random vectors in $\mathbb{R}^d$. $\forall \varepsilon \in (0, 1)$ and $t \geq 1$, $\exists C(\varepsilon, t) > 0$ such that if $n \geq C(\varepsilon, t) \cdot d$ then, with probability greater than $1 - e^{-ct\sqrt{d}}$ we have:

$$\|\Sigma_n - I\| \leq \varepsilon,$$

where $c > 0$ is an absolute constant.

7.3 The main technical step

We will present, now, the main technical step needed to prove Theorem 7.5. First, we introduce the definition of norm of a restriction of the matrix $\mathbf{X} \in \mathbb{R}^{d \times n}$, constructed by setting the vectors $X_1, \dots, X_n$ as columns.

Definition 7.6 Let $X_1, \dots, X_n$ be i.i.d random vectors in $\mathbb{R}^d$. Consider a $d \times n$ random matrix $\mathbf{X}$ whose columns are X_i 's. We denote the operator norm of $\mathbf{X}$ by $\|\mathbf{X}\|$. Then, $\forall 1 \leq m \leq n$, we define:

$$\mathbf{X}_m = \sup_{\substack{z \in \mathbb{S}^{n-1} \text{ s.t.} \\ |\text{supp}(z)| \leq m}} \|\mathbf{X}z\|_2$$

Observe that $\mathbf{X}_m$ increases with m and that for $m = n$ we recover the standard definition of the norm $\|\mathbf{X}\|$.

The main technical step necessary for the proof of Theorem 7.5 is an upper bound for $\mathbf{X}_m$.

Theorem 7.7 ([2]) Let n, d be positive integers with $d \leq n \leq e^{\sqrt{d}}$. Consider n i.i.d. isotropic log-concave random vectors $X_1, \dots, X_n$ in $\mathbb{R}^d$. Then, $\forall t \geq 1$:

$$\mathbb{P} \left(\exists m \leq n : \mathbf{X}_m \geq Ct \left(\sqrt{d} + \sqrt{m} \log \frac{2n}{m} \right) \right) \leq e^{-ct\sqrt{d}},$$

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

From the observation above, then this theorem implies an upper bound for the operator norm $\|\mathbf{X}\|$.

Corollary 7.8 ([2]) Let n, d be positive integers with $d \leq n \leq e^{\sqrt{d}}$. Consider n i.i.d. isotropic log-concave random vectors $X_1, \dots, X_n$ in $\mathbb{R}^d$. Let $\mathbf{X}$ be the $d \times n$ matrix whose columns are the X_i 's. Then, $\forall t \geq 1$, with probability greater than $1 - \exp(-ct\sqrt{d})$ we have:

$$\|\mathbf{X}\| \leq Ct(\sqrt{d} + \sqrt{n}),$$

7. CASE STUDY: ISOTROPIC LOG-CONCAVE RANDOM VECTORS

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

We also need some auxiliary results that we enclose in the following claim.

Claim 7.9 Let n, d be positive integers with $d \leq n \leq e^{\sqrt{d}}$. Consider n i.i.d. isotropic log-concave random vectors $X_1, \dots, X_n$ in $\mathbb{R}^d$. Then:

1. $\exists$ an absolute constant $\delta > 0$ such that: $\|\langle X_i, t \rangle\|_{\psi_1} \leq \delta, \forall t \in \mathbb{S}^{d-1}, \forall 1 \leq i \leq n$. This means that if a measure is log-concave then its linear functionals exhibit a sub-exponential decay
2. For every $n \leq e^{\sqrt{d}}$ and for every $t \geq 1$, we have, with probability greater than $1 - e^{-t\sqrt{d}}$:

$$\max_{1 \leq i \leq n} \|X_i\|_2 \leq Ct\sqrt{d}$$

3. For every $x_1, \dots, x_n \in \mathbb{R}^d, \exists E \subset \{1, \dots, n\}$ such that:

$$\sum_{i \neq j} \langle x_i, x_j \rangle \leq 4 \sum_{i \in E} \sum_{j \in E^c} \langle x_i, x_j \rangle$$

4. Given $E \subset \{1, \dots, n\}$ with $|E| = m$ and $\varepsilon, \alpha \in (0, 1)$, we can choose an ε -net $\mathcal{N}$ of $B_2^n \cap \alpha B_\infty^n \cap \mathbb{R}^E$, with respect to the Euclidean metric, that has cardinality:

$$|\mathcal{N}| \leq \left(\frac{3}{\varepsilon}\right)^m$$

Proof (of Claim 7.9) We prove the first of the four claims.

1. This result follows from Borell's Lemma, Theorem 7.3. Indeed, we can affirm that:

$$\begin{aligned} \mathbb{P}(\langle X_i, t \rangle > \delta) &\leq \frac{\mathbb{E}\langle X_i, t \rangle^2}{\delta^2} && \text{by Markov's inequality} \\ &= \frac{1}{100} && \text{by the isotropic assumption and choosing } \delta = 10 \end{aligned}$$

Therefore, if μ_X is the log-concave distribution from which $X_1, \dots, X_n$ are sampled, one can say that $\mu_X(\{t : \langle X_i, t \rangle \leq \delta\}) > 0.9 = \theta > \frac{1}{2}$. Therefore, we can finally apply Borell's Lemma and obtain:

$$\mathbb{P}(\langle X_i, t \rangle > \delta) \leq \theta \cdot \left(\frac{1-\theta}{\theta}\right)^{\frac{1+t}{2}} \leq e^{-t} \quad \forall t > 1$$

It follows that $\|\langle X_i, t \rangle\|_{\psi_1} \leq \delta$.

The second claim is a consequence of Paouris' inequality, Theorem 7.16.

2. From Paouris' inequality, Theorem 7.16, we know that it holds:

$$\mathbb{P}(\|X_i\|_2 \geq C_1 t \sqrt{d}) \leq e^{-t\sqrt{d}} \quad \forall 1 \leq i \leq n \text{ and } \forall t \geq 1$$

Then, by union bound and recalling that we assumed $n \leq e^{\sqrt{d}}$ the result follows, adjusting absolute constants.

We prove, now, the third claim.

3. Observe that, if $(\varepsilon_i)_{i=1}^n$ are independent Bernoulli random variables, then:

$$\mathbb{E} \langle \sum_{i=1}^n \varepsilon_i x_i, \sum_{j=1}^n (1 - \varepsilon_j) x_j \rangle = \frac{1}{4} \sum_{i \neq j} \langle x_i, x_j \rangle$$

Therefore, the result follows.

The last of the claims, instead, follows from a standard volume estimate. $\square$

Remark 7.10 *In the following of this Chapter, for any set E that appears in the argument, we will be fixing an ε -net of $B_2^n \cap \alpha B_\infty^n \cap \mathbb{R}^E$ that satisfies the fourth result from Claim 7.9.*

Moreover, we will denote it by $\mathcal{N}(E, \varepsilon, \alpha)$.

To proceed with the proof of Theorem 7.7, we also need the two following lemmas.

Lemma 7.11 ([2]) *Let $X_1, \dots, X_n$ be i.i.d. isotropic log-concave random vectors in $\mathbb{R}^d$. Then, if $m \leq n$, $\forall \varepsilon, \alpha \in (0, 1]$ and $L \geq 2m \log \frac{24en}{m\varepsilon}$, we have:*

$$\mathbb{P} \left(\sup_{F, E, z} \sum_{i \in E} \left| \langle z_i X_i, \sum_{j \in F \setminus E} z_j X_j \rangle \right| > \rho \alpha L \mathbf{X}_m \right) \leq e^{-L/2},$$

where $\sup_{F, E, z}$ denotes the triple supremum taken over all $F \subset \{1, \dots, n\}$ with $|F| \leq m$, all $E \subset F$ and all $z \in \mathcal{N}(F, \varepsilon, \alpha)$. Moreover, the constant ρ is the one such that $\|X_i, t\|_{\psi_1} \leq \delta \quad \forall 1 \leq i \leq n$.

Proof (of Lemma 7.11) Denote the underlying probability space by Ω . For $F \subset \{1, \dots, n\}$ with $|F| \leq m$, $E \subset F$, and $z \in \mathcal{N}(F, \varepsilon, \alpha)$, define the subset $\Omega(F, E, z)$ of Ω by:

$$\Omega(F, E, z) = \left\{ \sum_{i \in E} \left| \langle z_i X_i, \sum_{j \in F \setminus E} z_j X_j \rangle \right| > \rho \alpha L \mathbf{X}_m \right\}$$

Fix (F, E, z) and set $Y = \sum_{j \in F \setminus E} z_j X_j$. Since the X_i 's are independent by assumption, then Y is independent from X_i , for $i \in E$. Moreover, by definition of $\mathbf{X}_m$, we have $\|Y\|_2 \leq \mathbf{X}_m$.

Note that, if $\omega \in \Omega(F, E, z)$ then $Y(\omega) \neq 0$. Indeed, if this was not the case, we would have that $\langle z_i X_i, Y \rangle = 0$ for all $i \in E$ and the condition in the definition of $\Omega(F, E, z)$ would not be satisfied.

Thus, using also the fact that $\|z\|_\infty \leq \alpha$, which follows from the definition of $\mathcal{N}(F, \varepsilon, \alpha)$, we obtain:

$$\begin{aligned}
 \sum_{i \in E} |\langle z_i X_i, Y \rangle| &\leq \sum_{i \in E} |z_i| \cdot |\langle X_i, Y \rangle| \leq \alpha \sum_{i \in E} |\langle X_i, Y \rangle| && \text{since } \|z\|_\infty \leq \alpha \\
 &= \alpha \|Y\|_2 \sum_{i \in E} \left| \left\langle X_i, \frac{Y}{\|Y\|_2} \right\rangle \right| \\
 &\leq \alpha \mathbf{X}_m \sum_{i \in E} \left| \left\langle X_i, \frac{Y}{\|Y\|_2} \right\rangle \right| && \text{since } \|Y\|_2 \leq \mathbf{X}_m
 \end{aligned}$$

Since $\mathbf{X}_m \geq \|Y\|_2 > 0$, we have:

$$\begin{aligned}
 \mathbb{P}(\Omega(F, E, z)) &= \mathbb{P} \left(\left\{ \sum_{i \in E} |\langle z_i X_i, Y \rangle| > \rho \alpha L \mathbf{X}_m \right\} \right) && \text{from the definition of } \Omega(F, E, z) \\
 &\leq \mathbb{P} \left(\left\{ \alpha \mathbf{X}_m \sum_{i \in E} \left| \left\langle X_i, \frac{Y}{\|Y\|_2} \right\rangle \right| > \rho \alpha L \mathbf{X}_m \right\} \right) && \text{from the inequality above} \\
 &\leq \mathbb{P} \left(\sum_{i \in E} \left| \left\langle X_i, \frac{Y}{\|Y\|_2} \right\rangle \right| > \rho L \right)
 \end{aligned}$$

Using Markov's inequality and since by the first of the four assertions in Claim 7.9 we have that $\|\langle X_i, t \rangle\|_{\psi_1} \leq \rho \quad \forall 1 \leq i \leq n$ and $t \in \mathcal{S}^{d-1}$:

$$\begin{aligned}
 \mathbb{P} \left(\sum_{i \in E} \left| \left\langle X_i, \frac{Y}{\|Y\|_2} \right\rangle \right| > \rho L \right) &= \mathbb{P} \left(\exp \left(\frac{1}{\rho} \sum_{i \in E} \left| \left\langle X_i, \frac{Y}{\|Y\|_2} \right\rangle \right| \right) > e^L \right) \\
 &\leq \frac{\mathbb{E} \exp \left(\frac{1}{\rho} \sum_{i \in E} \left| \left\langle X_i, \frac{Y}{\|Y\|_2} \right\rangle \right| \right)}{e^L} && \text{by Markov's inequality} \\
 &\leq \frac{2^{|E|}}{e^L} \leq \frac{2^m}{e^L}
 \end{aligned}$$

Finally, by taking the union bound over all (F, E, z) and taking into account the forth statement from Claim 7.9, we get:

$$\begin{aligned}
 &\mathbb{P} \left(\sup_{F, E, z} \sum_{i \in E} \left| \left\langle z_i X_i, \sum_{j \in F \setminus E} z_j X_j \right\rangle \right| > \rho \alpha L A_m \right) \\
 &\leq \left[\left(\sum_{k=1}^m \binom{n}{k} 2^k \right) \left(\frac{3}{\varepsilon} \right)^m \right] \sup_{F, E, z} \mathbb{P}(\Omega(F, E, z)) && \text{by union bound and } |\mathcal{N}| \leq \left(\frac{3}{\varepsilon} \right)^m \\
 &\leq \sum_{k=1}^m \binom{n}{k} 2^k \left(\frac{3}{\varepsilon} \right)^m \frac{2^m}{e^L} \leq \left(\sum_{k=1}^m \left(\frac{2en}{k} \right)^k \right) \left(\frac{3}{\varepsilon} \right)^m \frac{2^m}{e^L} && \text{by the standard inequality of Theorem A.1} \\
 &\leq \left(\frac{4en}{m} \right)^m \left(\frac{6}{\varepsilon} \right)^m e^{-L} = \left(\frac{24en}{m\varepsilon} \right)^m e^{-L} = \exp \left(m \log \frac{24en}{m\varepsilon} - L \right),
 \end{aligned}$$

where in the last row we used that $(\frac{2en}{k})^k$ is an increasing function of k , which implies $\sum_{k=1}^m (\frac{2en}{k})^k \leq m \cdot (\frac{2en}{m})^m \leq (\frac{4en}{m})^m$. Since we assumed $L \geq 2m \log \frac{24en}{m\epsilon}$ then $m \log \frac{24en}{m\epsilon} \leq \frac{L}{2}$, which implies the result. $\square$

The second lemma we need is of the same type.

Lemma 7.12 ([2]) *Let $X_1, \dots, X_n$ be i.i.d. isotropic log-concave random vectors in $\mathbb{R}^d$. Let $1 \leq k, m \leq n, \epsilon$ and let $\epsilon, \alpha \in (0, 1]$, $\beta > 0$, and $L > 0$. We write $B(m, \beta)$ for the set of all vectors $x \in \beta B_2^n$ with $|\text{supp } x| \leq m$. We also consider a subset $\mathcal{B}$ of $B(m, \beta)$ with cardinality equal to M . Then:*

$$\mathbb{P} \left(\sup_{F, x, z} \sum_{i \in F} \left| \left\langle z_i X_i, \sum_{j \notin F} x_j X_j \right\rangle \right| > \rho \alpha \beta L \mathbf{X}_m \right) \leq M \left(\frac{12en}{k\epsilon} \right)^k e^{-L},$$

where $\sup_{F, x, z}$ denotes the triple supremum taken over all $F \subset \{1, \dots, n\}$ with $|F| \leq k$, all $x \in B(m, \beta)$ and all $z \in \mathcal{N}(F, \epsilon, \alpha)$. Moreover, the constant ρ is the one such that $\|\langle X_i, t \rangle\|_{\psi_1} \leq \delta \quad \forall 1 \leq i \leq n$.

Proof (of Lemma 7.12) For $F \subset \{1, \dots, n\}$ with $|F| \leq k$ and every $x \in B(m, \beta)$, and $z \in \mathcal{N}(F, \epsilon, \alpha)$, we define:

$$\Omega(F, x, z) = \left\{ \omega \in \Omega : \sum_{i \in F} \left| \left\langle z_i X_i, \sum_{j \notin F} x_j X_j \right\rangle \right| > \rho \alpha \beta L \mathbf{X}_m \right\}$$

We fix (F, x, z) and we set $Y = \sum_{j \notin F} x_j X_j$. Since the X_i 's are independent, Y is independent of all the vectors X_i , for $i \in F$.

Moreover, $\|Y\|_2 \leq \beta \mathbf{X}_m$. Note that, as in the proof of Lemma 7.11, we have that if $\omega \in \Omega(F, x, z)$ then $Y(\omega) \neq 0$. Then, using the fact that $\|z\|_\infty \leq \alpha$:

$$\sum_{i \in F} |\langle z_i X_i, Y \rangle| \leq \alpha \beta \mathbf{X}_m \sum_{i \in F} \left| \left\langle X_i, \frac{Y}{\|Y\|_2} \right\rangle \right|$$

Now, since $\mathbf{X}_m \geq \frac{\|Y\|_2}{\beta}$, we have:

$$\begin{aligned} \mathbb{P}(\Omega(F, x, z)) &= \mathbb{P} \left(\sup_{F, x, z} \sum_{i \in F} |\langle z_i X_i, Y \rangle| > \rho \alpha \beta L \mathbf{X}_m \right) \\ &\leq \mathbb{P} \left(\sum_{i \in F} \left| \left\langle X_i, \frac{Y}{\|Y\|_2} \right\rangle \right| > \rho L \right) && \text{by the previous inequality and } \mathbf{X}_m \geq \frac{\|Y\|_2}{\beta} \\ &\leq \frac{\mathbb{E} \exp \left(\frac{1}{\rho} \sum_{i \in F} \left| \left\langle X_i, \frac{Y}{\|Y\|_2} \right\rangle \right| \right)}{e^L} && \text{by Markov's inequality} \\ &\leq 2^{|F|} e^{-L} \leq 2^k e^{-L}. \end{aligned}$$

Taking the union bound, proceeding in a similar way as in Lemma 7.11:

$$\begin{aligned}
 & \mathbb{P} \left(\sup_{F, x, z} \sum_{i \in F} \left| \left\langle z_i X_i, \sum_{j \notin F} x_j X_j \right\rangle \right| > \rho \alpha \beta L \mathbf{X}_m \right) \\
 & \leq \left(\sum_{l=1}^k \binom{n}{l} \right) \cdot M \left(\frac{3}{\varepsilon} \right)^k \cdot \sup_{F, x, z} \mathbb{P}(\Omega(F, x, z)) \\
 & \leq M \left(\frac{3}{\varepsilon} \right)^k 2^k e^{-L} \left(\sum_{l=1}^k \left(\frac{2en}{l} \right)^l \frac{1}{2^l} \right) \leq M \left(\frac{3}{\varepsilon} \right)^k 2^k e^{-L} \left(\frac{2en}{k} \right)^j \\
 & = M e^{-L} \left(\frac{12en}{\varepsilon k} \right)^k,
 \end{aligned}$$

which proves the result. $\square$

We can finally prove Theorem 7.7, the main technical step we will need for the proof of the main result Theorem 7.5.

Proof (of Theorem 7.7) Since we have the assumption $n \leq e^{\sqrt{d}}$, the union bound shows that, to prove the main result, it is enough to check that, for t large enough, adjusting the absolute constants and for any fixed $m \leq n$, we have:

$$\mathbb{P} \left(\mathbf{X}_m \geq Ct \left(\sqrt{d} + \sqrt{m} \log \frac{2n}{m} \right) \right) \leq e^{-ct\sqrt{d}} \quad (7.1)$$

Therefore, we will proceed with this scope in mind.

We define a finite set $\mathcal{M} \subseteq \mathbb{R}^d$, with $|\text{support}(\mathcal{M})| \leq m$. This set will serve the following purposes:

- we will estimate $\sup_{x \in \mathcal{M}} \|\mathbf{X}x\|_2$ with high probability
- using $\mathcal{M}$, we will approximate any point from B_2^n with at most m non-zero coordinates

Let's start with the first item on the list.

Note that if for some $x \in \mathbb{S}^{n-1}$ we know that $|\text{support}(x)| \sim s$ and $\|x\|_\infty \leq \frac{1}{\sqrt{s}}$ for some $s \geq 1$, then we can estimate $\|\mathbf{X}x\|_2$ with probability close to one. To see this fact, we need to construct the set $\mathcal{M}$ such that it will consist of sums of vectors with disjoint support that roughly satisfy these two conditions. Therefore, to make this more precise, we define $\mathcal{M}$, distinguishing two cases:

1. **First case** Assume that $m \log \frac{96en}{m} \leq \sqrt{d}$. Then, we set:

$$\mathcal{M} = \cup_{\substack{E \subset \{1, \dots, n\} \\ |E| = m}} \mathcal{N} \left(E, \frac{1}{4}, 1 \right)$$

2. **Second case** Assume $m \log \frac{96en}{m} > \sqrt{d}$ and consider the smallest integer ℓ for which: $\frac{m}{2^\ell} \log \frac{96 \cdot 2^\ell en}{m} \leq \sqrt{d}$.
Fix some positive integers $a_0, a_1, \dots, a_\ell$ so that the next three conditions are satisfied:

- (i) $\sum_{k=0}^{\ell} a_k = m$
- (ii) $a_k \leq m \cdot 2^{-k+1} \quad \forall 1 \leq k \leq \ell$
- (iii) $a_0 \leq m \cdot 2^{-\ell}$

Then, we set $\mathcal{M} = \mathcal{M}_0 \cap 2B_2^n$, where $\mathcal{M}_0$ is the set of all vectors of the form $x = \sum_{k=0}^{\ell} x_k$, where the x_i 's have disjoint supports and:

$$x_0 \in \bigcup_{\substack{E \subset \{1, \dots, n\} \\ |E| \leq a_0}} \mathcal{N}\left(E, \frac{1}{4}, 1\right)$$

and

$$x_k \in \bigcup_{\substack{E \subset \{1, \dots, n\} \\ |E| \leq a_k}} \mathcal{N}\left(E, 2^{-k}, \sqrt{\frac{2^k}{m}}\right) \quad \text{for } 1 \leq k \leq \ell$$

Observe that, if $x \in \mathcal{M}$, then $|\text{support}(x)| \leq \sum_{k=0}^{\ell} a_k = m$, and, since $\mathcal{M} = \mathcal{M}_0 \cap 2B_2^n$, $\|x\|_2 \leq 2$.

We will prove that in both cases we are able to achieve such purposes.

1. Let's begin with the proof of the first case. In this case, observe that $\mathcal{M}$ is the union of several ε -nets, therefore we can affirm that $\mathcal{M}$ is an ε -net of $B_2^n \cap \alpha B_\infty^n \cap (\cup \mathbb{R}^E)$, with $\varepsilon = \frac{1}{4}$ and $\alpha = 1$. Now, we fix $x \in \mathcal{M}$ and denote the coordinates of x by $x(i)$, for $1 \leq i \leq n$. Then, we have:

$$\begin{aligned} \|Xx\|_2^2 &= \left\langle \sum_{i=1}^n x(i) X_i, \sum_{j=1}^n x(j) X_j \right\rangle \\ &= \sum_{i=1}^n x(i)^2 \|X_i\|_2^2 + \sum_{i \neq j} \langle x(i) X_i, x(j) X_j \rangle \\ &\leq \sum_{i=1}^n x(i)^2 \cdot \left(\max_i \|X_i\|_2^2 \right) + \sum_{i \neq j} \langle x(i) X_i, x(j) X_j \rangle \\ &\leq 1 \cdot \max_i \|X_i\|_2^2 + D_x \leq 2 \max \left\{ \max_i \|X_i\|_2^2, D_x \right\}, \end{aligned} \tag{7.2}$$

where $D_x = \sum_{i \neq j} \langle x(i) X_i, x(j) X_j \rangle$.

Observe that, from the second assertion of Claim 7.9, we know that $\max_i \|X_i\|_2 \leq C_0 t \sqrt{d}$ with probability greater than $1 - e^{-t\sqrt{d}}$. The goal is to find an equivalent estimate also for the second term in the maximum above, D_x .

By the third statement of Claim 7.9, there exists a set $E \subset \{1, \dots, n\}$ such that:

$$\begin{aligned}
 D_x &= \sum_{i \neq j} \langle x(i)X_i, x(j)X_j \rangle \leq 4 \sum_{i \in E} \sum_{j \in E^c} \langle x(i)X_i, x(j)X_j \rangle \\
 &\leq 4 \max_{\substack{F \subset \{1, \dots, n\} \\ |F|=m}} \max_{E \subset F} \max_{v \in \mathcal{N}(F, \frac{1}{4}, 1)} \sum_{i \in E} \left| \left\langle v_i X_i, \sum_{j \in F \setminus E} v_j X_j \right\rangle \right|
 \end{aligned}$$

Therefore, we can apply Lemma 7.11, with $L = 2t\sqrt{d}$, $\varepsilon = \frac{1}{4}$ and $\alpha = 1$ to obtain:

$$\mathbb{P} \left(\max_{x \in \mathcal{M}} D_x > 8t\sqrt{d}\rho\mathbf{X}_m \right) \leq e^{-t\sqrt{d}} \quad (7.3)$$

2. For the second case, we fix $x \in \mathcal{M}$ and write $x = \sum_{k=0}^{\ell} x_k$. We denote the support of x_k by F_k , and the coordinates of x by $x(i)$, for $1 \leq i \leq n$. Then, we have:

$$\begin{aligned}
 \|\mathbf{X}x\|_2^2 &= \left\langle \sum_{i=1}^n x(i)X_i, \sum_{j=1}^n x(j)X_j \right\rangle \\
 &= \sum_{i=1}^n x(i)^2 \|X_i\|_2^2 + \sum_{i \neq j} \langle x(i)X_i, x(j)X_j \rangle \\
 &\leq \sum_{i=1}^n x(i)^2 \cdot \left(\max_i \|X_i\|_2^2 \right) + \sum_{i \neq j} \langle x(i)X_i, x(j)X_j \rangle \\
 &\leq 2 \max_i \|X_i\|_2^2 + D_x \leq 2 \max \left\{ 4 \max_i \|X_i\|_2^2, D_x \right\},
 \end{aligned} \quad (7.4)$$

where $D_x = \sum_{i \neq j} \langle x(i)X_i, x(j)X_j \rangle$.

From the second assertion of Claim 7.9, we know that $\max_i \|X_i\|_2 \leq C_0 t\sqrt{d}$ with probability greater than $1 - e^{-t\sqrt{d}}$.

Now, we would like to have a similar estimate for D_x . Therefore, we split D_x into two parts, corresponding to the construction of x :

$$\begin{aligned}
 \bullet \quad D'_x &= \sum_{k=0}^{\ell} \sum_{\substack{i, j \in F_k \\ i \neq j}} \langle x(i)X_i, x(j)X_j \rangle \\
 D''_x &= \sum_{k=0}^{\ell} \sum_{\substack{i \in F_k \\ j \in F_k^c}} \langle x(i)X_i, x(j)X_j \rangle \\
 \bullet \quad &= 2 \sum_{k=1}^{\ell} \sum_{i \in F_k} \sum_{r \in G_k} \left\langle x(i)X_i, \sum_{j \in F_r} x(j)X_j \right\rangle
 \end{aligned}$$

where $G_k = \{0, k+1, k+2, \dots, \ell\}$.

Following this split, observe that: $D_x = D'_x + D''_x$. The problem has turned into the estimation of D'_x and D''_x . To accomplish this task, we will use the third statement of Claim 7.9 and Lemma 7.11.

We start with the estimation of D'_x . For every k there exists a subset F'_k of F_k , the support of x_k , such that:

$$\begin{aligned}
 D'_x &\leq 4 \sum_{k=0}^l \sum_{\substack{i \in F'_k \\ j \in F_k \setminus F'_k}} \langle x(i) X_i, x(j) X_j \rangle && \text{from the third statement of Claim 7.9} \\
 &\leq 4 \max_{\substack{F \subset \{1, \dots, n\} \\ |F| \leq \frac{m}{2^l}}} \max_{E \subset F} \max_{v \in \mathcal{N}(F, \frac{1}{4}, 1)} \sum_{i \in E} \left| \left\langle v_i X_i, \sum_{j \in F \setminus E} v_j X_j \right\rangle \right| && \text{recalling that } x = \sum_{i=0}^l x_i = x_0 + \sum_{k=1}^l x_k \\
 &\quad + 4 \sum_{k=1}^l \max_{\substack{F \subset \{1, \dots, N\} \\ |F| \leq \frac{2m}{2^k}}} \max_{E \subset F} \max_{v \in \mathcal{N}(F, 2^{-k}, \sqrt{\frac{2^k}{m}})} \sum_{i \in E} \left| \left\langle v_i X_i, \sum_{j \in F \setminus E} v_j X_j \right\rangle \right|
 \end{aligned}$$

Now, we will use Lemma 7.11 to bound the two terms in the sum above.

- a) To bound the first term, by choosing $L = 2t\sqrt{d}$, $\varepsilon = \frac{1}{4}$, $\alpha = 1$, we obtain:

$$\mathbb{P} \left(\sup_{F, E, \mathcal{N}} \sum_{i \in E} \left| \langle v_i X_i, \sum_{j \in F \setminus E} v_j X_j \rangle \right| > \rho \alpha L \mathbf{X}_m \right) \leq e^{-L/2} = e^{-t\sqrt{d}}$$

Then, it follows that:

$$\mathbb{P} \left(4 \sup_{F, E, v} \sum_{i \in E} \left| \langle v_i X_i, \sum_{j \in F \setminus E} v_j X_j \rangle \right| > 8t\sqrt{d} \rho \mathbf{X}_m \right) \leq e^{-t\sqrt{d}}$$

- b) To bound the second term, by choosing $L = \frac{4mt}{2^k} \log \frac{24en \cdot 4^k}{m}$, $\varepsilon = 2^{-k}$, $\alpha = \sqrt{\frac{2^k}{m}}$, we get:

$$\mathbb{P} \left(\sup_{F, E, v} \sum_{i \in E} \left| \langle v_i X_i, \sum_{j \in F \setminus E} v_j X_j \rangle \right| > \rho \alpha L \mathbf{X}_m \right) \leq e^{-L/2}$$

Then, it follows that:

$$\begin{aligned}
 &\mathbb{P} \left(4 \sum_{k=1}^l \sup_{F, E, v} \sum_{i \in E} \left| \langle v_i X_i, \sum_{j \in F \setminus E} v_j X_j \rangle \right| > 2\rho t \mathbf{X}_m \sum_{k=1}^l \sqrt{\frac{2^k}{m}} \frac{8m}{2^k} \log \left(\frac{24 \cdot 4^k en}{m} \right) \right) \\
 &\leq \sum_{k=1}^l \exp \left(-t \frac{2m}{2^k} \log \left(\frac{24en 4^k}{m} \right) \right) \\
 &\leq \sum_{k=1}^l \max_k \left[\exp \left(-t \frac{2m}{2^k} \log \left(\frac{24en 4^k}{m} \right) \right) \right] \\
 &\leq \ell \cdot \exp \left(-t \frac{2m}{2^\ell} \log \frac{24 \cdot 4^\ell en}{m} \right)
 \end{aligned}$$

Then, by union bound and putting together the bounds we just found for the two terms above, we finally obtain the bound for D'_x :

$$\begin{aligned} \mathbb{P} \left(\max_{x \in \mathcal{M}} D'_x > 8\rho t \mathbf{X}_m \sqrt{d} + 2\rho t \mathbf{X}_m \sum_{k=1}^{\ell} \sqrt{\frac{2^k}{m}} \frac{8m}{2^k} \log \frac{24 \cdot 4^k e n}{m} \right) \\ \leq \exp(-t\sqrt{d}) + \ell \exp \left(-t \frac{2m}{2^\ell} \log \frac{24 \cdot 4^\ell e n}{m} \right) \end{aligned}$$

Therefore, by the choice of ℓ we made in the Second case, and the fact that $\ell \leq 2\sqrt{d}$, since $m \leq n \leq e^{\sqrt{d}}$, we see that there exists an absolute constant $C > 0$ such that:

$$\begin{aligned} \mathbb{P} \left(\max_{x \in \mathcal{M}} D'_x > t \mathbf{X}_m \left(8\rho \sqrt{d} + C\rho \sqrt{m} \log \frac{2n}{m} \right) \right) \\ \leq \exp(-t\sqrt{d}) + \ell \exp(-t\sqrt{d}) \leq (2\sqrt{d} + 1) \exp(-t\sqrt{d}). \end{aligned}$$

Now that we have found an estimate for D'_x , we can find an estimate for D''_x . This process will be carried out in a similar way.

For every $1 \leq k \leq \ell$, consider the set $\mathcal{M}_k = \mathcal{M}'_k \cap 2B_2^n$, where $\mathcal{M}'_k$ is the set of all vectors of the form $x = x_0 + \sum_{s=k+1}^{\ell} x_s$, where the x_i 's have disjoint supports for $s \geq k+1$, and are such that:

$$x_0 \in \bigcup_{\substack{E \subset \{1, \dots, n\} \\ |E| \leq a_0}} \mathcal{N}(E, \frac{1}{4}, 1)$$

and

$$x_s \in \bigcup_{\substack{E \subset \{1, \dots, n\} \\ |E| \leq a_s}} \mathcal{N}(E, 2^{-s}, \sqrt{\frac{2^s}{m}}), \quad \forall s \geq k+1$$

Now, observe that from the definition of $\mathcal{M}_k$, we have that $\mathcal{M}_k \subset 2B_2^n$ and recall, from the forth statement in Claim 7.9 that $|\mathcal{N}| \leq \left(\frac{3}{\varepsilon}\right)^m$. Therefore, we can find the following bound on the size of $\mathcal{M}_k$:

$$\begin{aligned} |\mathcal{M}_k| &\leq \left[\left(\frac{3}{1/4} \right)^{a_0} \binom{n}{a_0} \right] \prod_{s=k+1}^{\ell} \left(\frac{3}{2^{-s}} \right)^{a_s} \binom{n}{a_s} && \text{recalling that } x = x_0 + \sum_{s=k+1}^{\ell} x_s \\ &\leq \left(\frac{12en}{a_0} \right)^{a_0} \prod_{s=k+1}^{\ell} \left(\frac{3 \cdot 2^s en}{a_s} \right)^{a_s} && \text{from the basic inequality of Theorem A.1} \\ &\leq \exp \left(\frac{m}{2^\ell} \log \frac{12 \cdot 2^\ell en}{m} + \sum_{s=k+1}^{\ell} \frac{2m}{2^s} \log \frac{3 \cdot 4^s en}{2m} \right) && \text{from the bounds on } |a_0| \text{ and } |a_k| \\ &\leq \exp \left(\sum_{s=k+1}^{\ell+1} \frac{2m}{2^s} \log \frac{3 \cdot 4^s en}{2m} \right) \\ &\leq \exp \left(\frac{m}{2^k} \left(\log \frac{6 \cdot 4^k e n}{m} \sum_{s=0}^{\ell-k} \frac{1}{2^s} + \log 4 \sum_{s=1}^{\ell-k} \frac{s}{2^s} \right) \right) \\ &\leq \exp \left(\frac{4m}{2^k} \log \frac{6 \cdot 4^k en}{m} \right) \end{aligned}$$

Moreover, observe that:

$$\begin{aligned} D_x'' &= 2 \sum_{k=1}^{\ell} \sum_{i \in F_k} \left\langle x(i) X_i, \sum_{r \in G_k} \sum_{j \in F_r} x(j) X_j \right\rangle \\ &\leq 2 \sum_{k=1}^{\ell} \max_{\substack{F \subset \{1, \dots, N\} \\ |F| \leq \frac{2m}{2^k}}} \max_{u \in \mathcal{N}\left(F, 2^{-k}, \sqrt{\frac{2^k}{m}}\right)} \max_{v \in \mathcal{M}_k} \sum_{i \in F} \left| \left\langle u_i X_i, \sum_{j \in F^c} v_j X_j \right\rangle \right| \end{aligned}$$

Then, notice that we can apply Lemma 7.12 to every term of the sum $\sum_{k=1}^{\ell}$, using $L_k = \frac{12m}{2^k} t \log \frac{12 \cdot 4^k en}{m}$, $\varepsilon_k = 2^{-k}$, $\alpha_k = \sqrt{\frac{2^k}{m}}$, $\beta = 2$, $\mathcal{B}_k = \mathcal{M}_k$. We obtain:

$$\begin{aligned} \mathbb{P} \left(D_x'' > 48\rho t \mathbf{X}_m \sum_{k=1}^{\ell} \sqrt{\frac{2^k}{m}} \log \frac{12 \cdot 4^k en}{m} \right) \\ \leq \sum_{k=1}^{\ell} \exp \left(\frac{4m}{2^k} \log \frac{6 \cdot 4^k en}{m} + \frac{2m}{2^k} \log \frac{6 \cdot 4^k en}{m} - t \frac{12m}{2^k} \log \frac{12 \cdot 4^k en}{m} \right) \\ \leq \sum_{k=1}^{\ell} \exp \left(-t \frac{6m}{2^k} \log \frac{12 \cdot 4^k en}{m} \right) \leq \ell \exp \left(-t \frac{6m}{2^{\ell}} \log \frac{12 \cdot 4^{\ell} en}{m} \right) \end{aligned}$$

As in the case of D'_x , one can check that:

$$\mathbb{P} \left(\max_{x \in \mathcal{M}} D_x'' > 3C\rho \mathbf{X}_m t \sqrt{m} \log \frac{2n}{m} \right) \leq 2\sqrt{d} \exp(-t\sqrt{d}),$$

where $C > 0$ is an absolute constant.

Now that we have found appropriate bounds for D'_x and D_x'' , since $D_x = D'_x + D_x''$, we have that:

$$\begin{aligned} \mathbb{P} \left(\max_{x \in \mathcal{M}} D_x > \mathbf{X}_m t \left(8\rho\sqrt{d} + 4C\rho\sqrt{m} \log \frac{2n}{m} \right) \right) \\ \leq (4\sqrt{d} + 1) \exp(-t\sqrt{d}) \end{aligned} \quad (7.5)$$

Now that we have found, for both the First case and the Second case, a bound on D_x , we need to finish proving that the way we've constructed the set $\mathcal{M}$ allows us to approximate any vector from the set B_2^n , whose support has a cardinality of m or less, using a vector from $\mathcal{M}$.

We consider $z \in \mathbb{S}^{n-1}$ with $|\text{supp}(z)| \leq m$ and denote its coordinates by z_i , $1 \leq i \leq n$. We can find $n_1, \dots, n_m$ so that $|z_{n_1}| \geq |z_{n_2}| \geq \dots \geq |z_{n_m}|$, and $z_{n_i} = 0$ if $i > m$.

If $m \log \frac{96en}{M} \leq \sqrt{d}$, then we can consider only the support of z , which we denote by E_0 .

Otherwise, we set:

- $E_0 = \{n_i\}_{1 \leq i \leq \frac{m}{2^t}}$

- $E_1 = \{n_i\}_{\frac{m}{2} < i \leq m}$
- $E_2 = \{n_i\}_{\frac{m}{4} < i \leq \frac{m}{2}}$
- $\vdots$
- $E_\ell = \{n_i\}_{\frac{m}{2^\ell} < i \leq \frac{m}{2^{\ell-1}}}$

where ℓ is the least integer satisfying: $\frac{m}{2^\ell} \log \frac{96 \cdot 2^\ell e n}{m} \leq \sqrt{d}$.
Then, it is clear that:

- $a_0 := |E_0| \leq \frac{m}{2^\ell}$
- $a_k := |E_k| \leq \frac{m}{2^k} + 1 \leq \frac{m}{2^{k-1}}, \quad \text{for } 1 \leq k \leq \ell$
- and $\sum_{i=0}^\ell a_i = m$

Observe that the quantities a_k 's do not depend on $z \in \mathbb{S}^{n-1}$, although the sets E_k 's do.

We denote by $P_{E_k} z$ the projection of the vector z onto the set E_k . Then, observe that for every $k \geq 1$ we have:

$$\|P_{E_k} z\|_\infty \leq |z_{n_s}| \leq \sqrt{\frac{2^k}{m}},$$

where $s = \lfloor \frac{m}{2^k} \rfloor$.

Thus, for every $k \geq 1$ the vector $P_{E_k} z$ can be approximated by a vector in the net $\mathcal{N}\left(E_k, 2^{-k}, \sqrt{\frac{2^k}{m}}\right)$ and the vector $P_{E_0} z$ can be approximated by a vector in the net $\mathcal{N}\left(E_0, \frac{1}{4}, 1\right)$.

Then, we can find $x \in \mathcal{M}$ of the form $x = \sum_{k=0}^\ell x_k$, such that:

$$\|z - x\|_2^2 \leq \sum_{k=0}^\ell \|P_{E_k} z - x_k\|_2^2 \leq \left(\frac{1}{4}\right)^2 + \sum_{k=1}^\ell (2^{-k})^2 < \frac{4}{10}$$

Moreover, x is chosen to have the same support as z , i.e. $\text{supp}(x) = \text{supp}(z)$, therefore the vector $w = z - x$ satisfies $|\text{supp}(w)| \leq m$.

Therefore, we can approximate any vector z with $x \in \mathcal{M}$.

Considering all $z \in \mathbb{S}^{n-1}$ with $|\text{supp}(z)| \leq m$, it follows:

$$\begin{aligned} \mathbf{X}_m &= \max_{\substack{z \in \mathbb{S}^{n-1} \\ |\text{supp}(z)| \leq m}} \|\mathbf{X}z\|_2 \leq \max_{x \in \mathcal{M}} \|\mathbf{X}x\|_2 + \sqrt{\frac{4}{10}} \cdot \max_{\substack{w \in \mathbb{S}^{n-1} \\ |\text{supp}(w)| \leq m}} \|\mathbf{X}w\|_2 \\ &= \max_{x \in \mathcal{M}} \|\mathbf{X}x\|_2 + \sqrt{\frac{4}{10}} \mathbf{X}_m, \end{aligned}$$

Thus, it follows that:

$$\mathbf{X}_m \leq \frac{1}{1-\sqrt{\frac{4}{10}}} \max_{x \in \mathcal{M}} \|\mathbf{X}x\|_2 \leq 3 \max_{x \in \mathcal{M}} \|\mathbf{X}x\|_2$$

And from the inequalities (7.2), (7.4), we get:

$$\mathbf{X}_m^2 \leq 9 \max_{x \in \mathcal{M}} \|\mathbf{X}x\|_2^2 \leq 9 \max \left\{ 8 \max_i \|X_i\|_2^2, 2 \max_{x \in \mathcal{M}} D_x \right\}$$

Finally, using what we just found, the second assertion of Claim 7.9, and the bound on D_x of inequalities (7.3), (7.5), we find:

$$\begin{aligned} \mathbb{P} \left(\mathbf{X}_m \leq Ct\rho \left(\sqrt{d} + \sqrt{m} \log \frac{2n}{m} \right) \right) &\geq 1 - c' \sqrt{d} e^{-c_1 t \sqrt{d}} \\ &\geq 1 - \exp(-ct\sqrt{d}), \end{aligned}$$

where $C, c > 0$ are absolute constants. This proves Theorem 7.7. $\square$

7.4 Proof of the Main Result

Finally, we are almost ready to prove the Main Result of this Chapter, Theorem 7.5. Let $X_1, \dots, X_n$ be *i.i.d.* isotropic log-concave random vectors in $\mathbb{R}^d$. Our aim is to estimate:

$$\max_{\theta \in \mathbb{S}^{d-1}} \left| \frac{1}{n} \sum_{i=1}^n \left(\langle X_i, \theta \rangle^2 - 1 \right) \right|$$

The next lemma will allow us to work with a finite set of points θ in $\mathbb{S}^{n-1}$.

Lemma 7.13 ([2]) *Let $x_1, \dots, x_n \in \mathbb{R}^d$. Let $\gamma, \varepsilon \in (0, 1)$ and let $\mathcal{N}$ be a $\gamma\varepsilon$ -net for $\mathbb{S}^{n-1}$. Then:*

$$\max_{\theta \in \mathcal{N}} \left| \frac{1}{n} \sum_{i=1}^n \left(\langle x_i, \theta \rangle^2 - 1 \right) \right| \leq \varepsilon$$

implies

$$\max_{\theta \in \mathbb{S}^{d-1}} \left| \frac{1}{n} \sum_{i=1}^n \left(\langle x_i, \theta \rangle^2 - 1 \right) \right| \leq c_1(\gamma)\varepsilon,$$

where $c_1(\gamma) > 0$ is a constant depending only on γ .

Proof (of Lemma 7.13) We define a semi-norm $\|\cdot\|_x$ on $\mathbb{R}^d$ by

$$\|y\|_x = \left(\frac{1}{n} \sum_{i=1}^n \langle x_i, y \rangle^2 \right)^{1/2}$$

Then, from the assumption of the Lemma, we have

$$1 - \varepsilon \leq \sqrt{1 - \varepsilon} \leq \max_{\theta \in \mathcal{N}} \|\theta\|_x \leq \sqrt{1 + \varepsilon} \leq 1 + \varepsilon,$$

Since $\mathcal{N}$ is a $\gamma\varepsilon$ -net, we can use Lemma 2.17 and, together with the triangle inequality and the homogeneity of $\|\cdot\|_x$, we get:

$$\max_{\theta \in \mathbb{S}^{d-1}} \|\theta\|_x \leq \frac{1+\varepsilon}{1-2\gamma\varepsilon} = \frac{1-2\gamma\varepsilon+2\gamma\varepsilon+\varepsilon}{1-2\gamma\varepsilon} \leq 1 + \frac{2\gamma+1}{1-2\gamma}\varepsilon$$

For the lower bound, we consider any $\theta \in \mathbb{S}^{d-1}$ and find $\theta_1 \in \mathcal{N}$ such that $\|\theta - \theta_1\|_2 \leq \gamma\varepsilon$. Then,

$$\|\theta\|_x \geq \|\theta_1\|_x - \|\theta - \theta_1\|_x \geq (1 - \varepsilon) - \gamma\varepsilon \left(1 + \frac{2\gamma+1}{1-2\gamma}\varepsilon\right) \geq 1 - \frac{2\gamma+3}{1-2\gamma}\varepsilon$$

This shows that $|\|\theta\|_x - 1| \leq c_1(\gamma)\varepsilon$, where $c_1(\gamma) = \frac{2\gamma+3}{1-2\gamma}$. Observe that, since $t \mapsto t^2$ is Lipschitz with constant $2(1 + c_1(\gamma))$ on the interval $[0, 1 + c_1(\gamma)]$, we get:

$$\max_{\theta \in \mathbb{S}^{d-1}} \left| \frac{1}{n} \sum_{i=1}^n \left(\langle x_i, \theta \rangle^2 - 1 \right) \right| \leq c_2(\gamma)\varepsilon,$$

where $c_2(\gamma) = 2(1 + c_1(\gamma))$. □

We also need the following Bernstein-type inequality for the proof of Theorem 7.5.

Lemma 7.14 *Let $Z_1, \dots, Z_n$ be i.i.d. random variables with $\mathbb{E}(Z_i) = 0$, $\|Z_i\|_\infty \leq a$ and $\text{Var}(Z_i) = \mathbb{E}(Z_i^2) = \sigma^2$. Then, for all $s > 0$:*

$$\mathbb{P}(Z_1 + \dots + Z_n \geq s \cdot n) \leq \exp\left(-\frac{3n}{8} \min\left\{\frac{s^2}{\sigma^2}, \frac{s}{a}\right\}\right)$$

Proof (of Lemma 7.14) Let $s > 0$. Then:

$$\begin{aligned} \mathbb{P}\left(\sum_{j=1}^n Z_j \geq sn\right) &= \mathbb{P}\left(\exp\left(\frac{\lambda}{n} \sum_{j=1}^n Z_j\right) \geq e^{\lambda s}\right) && \text{for } \lambda \geq 0 \\ &\leq \frac{\mathbb{E} \exp\left(\frac{\lambda}{n} \sum_{j=1}^n Z_j\right)}{e^{\lambda s}} && \text{by Markov's inequality} \\ &= e^{-\lambda s} \prod_{j=1}^n \mathbb{E} \exp\left(\frac{\lambda Z_j}{n}\right) && \text{by independence} \end{aligned}$$

Next, observe that:

$$\begin{aligned}\mathbb{E} \exp \left(\frac{\lambda Z_j}{n} \right) &= 1 + \sum_{k=2}^{\infty} \frac{\lambda^k \mathbb{E} \left(Z_j^k \right)}{n^k k!} \leq 1 + \mathbb{E} \left(Z_j^2 \right) \sum_{k=2}^{\infty} \frac{\lambda^k a^{k-2}}{n^k k!} \\ &= 1 + \frac{\sigma^2}{a^2} \left(e^{\frac{\lambda a}{n}} - \frac{\lambda a}{n} - 1 \right)\end{aligned}$$

Also, since $1 + u \leq e^u$ for all $u \in \mathbb{R}$, we get:

$$\prod_{j=1}^n \mathbb{E} \exp \left(\frac{\lambda Z_j}{n} \right) \leq \exp \left(\frac{\sigma^2 n}{a^2} \left(e^{\frac{\lambda a}{n}} - \frac{\lambda a}{n} - 1 \right) \right)$$

Therefore, using this result in the inequality obtained with Markov's inequality:

$$\mathbb{P} \left(\sum_{j=1}^n Z_j \geq sn \right) \leq \exp \left(\frac{\sigma^2 n}{a^2} \left(e^{\frac{\lambda a}{n}} - \frac{\lambda a}{n} - 1 \right) - \lambda s \right),$$

for all $\lambda > 0$.

We choose λ so that $\exp(\frac{\lambda a}{n}) = 1 + \frac{s}{\sigma^2}$. This leads to the bound:

$$\mathbb{P} \left(\sum_{j=1}^n Z_j \geq sn \right) \leq \exp \left(-\frac{\sigma^2 n}{a^2} F \left(\frac{as}{\sigma^2} \right) \right),$$

where $F(r) = (1+r) \log(1+r) - r$, for $r > 0$.

Now, it remains to check that $F(r) \geq \frac{r^2}{2+\frac{2}{3}}$, and this implies the assertion of the lemma, because $F(r) \geq \frac{3r}{8}$ if $r \geq 1$ and $F(r) \geq \frac{3r^2}{8}$ if $0 < r < 1$. $\square$

With these two lemmas and the main technical step (Theorem 7.7), we can finally prove Theorem 7.5.

Proof of Theorem 7.5

We fix $\gamma, \eta \in (0, 1)$ and a $\gamma\eta$ -net $\mathcal{N}$ of $\mathbb{S}^{d-1}$ with cardinality $|\mathcal{N}| \leq \left(\frac{3}{\gamma\eta} \right)^d$.

Thanks to Lemma 7.12, our goal is now to estimate:

$$\max_{\theta \in \mathcal{N}} \left| \frac{1}{n} \sum_{i=1}^n \left(\langle X_i, \theta \rangle^2 - \mathbb{E} \left(\langle X_i, \theta \rangle^2 \right) \right) \right|$$

Fix $B > 1$ that will be specified later and write:

$$\begin{aligned}
 & \max_{\theta \in \mathcal{N}} \left| \frac{1}{n} \sum_{i=1}^n (\langle X_i, \theta \rangle^2 - \mathbb{E} [\langle X_i, \theta \rangle^2]) \right| \\
 & \leq \frac{1}{n} \max_{\theta \in \mathcal{N}} \left| \sum_{i=1}^n \left(\min(|\langle X_i, \theta \rangle|, B)^2 - \mathbb{E} [\min(|\langle X_i, \theta \rangle|, B)^2] \right) \right| \quad (\text{A}) \\
 & \quad + \max_{\theta \in \mathcal{N}} \frac{1}{n} \sum_{i=1}^n (|\langle X_i, \theta \rangle|^2 - B^2) \mathbb{1}_{\{|\langle X_i, \theta \rangle| \geq B\}} \quad (\text{B}) \\
 & \quad + \max_{\theta \in \mathcal{N}} \frac{1}{n} \mathbb{E} \left[\sum_{i=1}^n (|\langle X_i, \theta \rangle|^2 - B^2) \mathbb{1}_{\{|\langle X_i, \theta \rangle| \geq B\}} \right], \quad (\text{C})
 \end{aligned}$$

We want to estimate each term (A), (B), and (C) separately.

- **(A)** In order to control the first term, for fixed $\theta \in \mathbb{S}^{d-1}$ we apply Bernstein-type inequality from Lemma 7.14 for the random variables

$$Z_i = \min(|\langle X_i, \theta \rangle|, B)^2 - \mathbb{E} [\min(|\langle X_i, \theta \rangle|, B)^2]$$

Note that $\|Z_i\|_\infty \leq B^2$ and that, using Theorem 7.17:

$$\text{Var}(Z_i) \leq \mathbb{E} (\langle X_i, \theta \rangle^4) \leq C_1 \mathbb{E} (\langle X_i, \theta \rangle^2) = C_1,$$

for some absolute constant $C_1 > 0$.

If we choose $s = tB\sqrt{\frac{d}{n}}$, from Lemma 7.14 we get:

$$\begin{aligned}
 & \mathbb{P} \left(\left| \frac{1}{n} \sum_{i=1}^n \left(\min(|\langle X_i, \theta \rangle|, B)^2 - \mathbb{E} [\min(|\langle X_i, \theta \rangle|, B)^2] \right) \right| \geq tB\sqrt{\frac{d}{n}} \right) \\
 & \leq \exp \left(-c_1 \min \left(t^2 B^2 d, \frac{t\sqrt{nd}}{B} \right) \right),
 \end{aligned}$$

where $c_1 > 0$ is an absolute constant. It follows that, with probability greater than $1 - \exp \left(d \log \left(\frac{3}{\gamma\eta} \right) - c_1 \min \left(t^2 B^2 d, \frac{t\sqrt{nd}}{B} \right) \right)$:

$$\begin{aligned}
 & \max_{\theta \in \mathcal{N}} \left| \frac{1}{n} \sum_{i=1}^n \left(\min(|\langle X_i, \theta \rangle|, B)^2 - \mathbb{E} [\min(|\langle X_i, \theta \rangle|, B)^2] \right) \right| \\
 & \leq tB\sqrt{\frac{d}{n}}
 \end{aligned}$$

- **(B)** To control the second term, first observe that the term $\sqrt{d} + \sqrt{m} \log \frac{2n}{m}$ can be replaced by $\sqrt{d} + \sqrt{m} \log \frac{2n}{\max(d, m)}$. Indeed, if $m < d$:

$$\sqrt{d} + \sqrt{m} \log \frac{2n}{m} = \sqrt{d} + \sqrt{m} \log \frac{d}{m} + \sqrt{m} \log \frac{2n}{d}$$

Then, taking into account this observation, Theorem 7.7, with inequality (7.1) affirms that, if $t \geq 1$, with probability greater than

$1 - \exp(-ct\sqrt{d})$ we have, for every $m \leq n$ and for every $z \in \mathbb{S}^{n-1}$ with $|\text{supp}(z)| = m$:

$$\|\sum_{i=1}^n z_i X_i\|_2 \leq C_2 t \left(\sqrt{d} + \sqrt{m} \log \frac{2n}{d} \right)$$

Then, for every $E \subset \{1, \dots, n\}$ we have:

$$\begin{aligned} \max_{\theta \in \mathbb{S}^{d-1}} \left(\sum_{i \in E} |\langle X_i, \theta \rangle|^2 \right)^{1/2} &\leq \max \left\{ \left\| \sum_{i=1}^n z_i X_i \right\|_2 \mid z \in \mathbb{S}^{n-1} \cap \mathbb{R}^E \right\} \\ &\leq C_2 t \left(\sqrt{d} + \sqrt{|E|} \log \frac{2n}{d} \right) \end{aligned} \quad (7.6)$$

Let $\theta \in \mathcal{N}$. We write:

$$\sum_{i=1}^n |\langle X_i, \theta \rangle|^2 \mathbf{1}_{\{|\langle X_i, \theta \rangle| \geq B\}} = \sum_{i \in E_B(\theta)} \langle X_i, \theta \rangle^2,$$

where $E_B(\theta) = \{1 \leq i \leq n : |\langle X_i, \theta \rangle| \geq B\}$. Using inequality (7.6), we see that:

$$B |E_B(\theta)|^{1/2} \leq \left(\sum_{i \in E_B(\theta)} \langle X_i, \theta \rangle^2 \right)^{1/2} \leq C_2 t \left(\sqrt{d} + \sqrt{|E_B(\theta)|} \log \frac{2n}{d} \right)$$

If we choose $B = 2C_2 t \log \frac{2n}{d}$, then $|E_B(\theta)| \leq \frac{4C_2^2 t^2 d}{B^2} = \frac{d}{\log^2(\frac{2n}{d})}$.

For this choice of B , we get that $\forall \theta \in \mathcal{N}$, with probability greater than $1 - \exp(-ct\sqrt{d})$:

$$\max_{\theta \in \mathcal{N}} \left(\sum_{i=1}^n |\langle X_i, \theta \rangle|^2 \mathbf{1}_{\{|\langle X_i, \theta \rangle| \geq B\}} \right) \leq 4C_2^2 t^2 n$$

- (C) In order to control the third term, observe:

$$\begin{aligned} \mathbb{E} \left(\langle X_i, \theta \rangle^2 \mathbf{1}_{\{|\langle X_i, \theta \rangle| \geq B\}} \right) &\leq \left(\mathbb{E} |\langle X_i, \theta \rangle|^4 \right)^{1/2} (\mathbb{P}(|\langle X_i, \theta \rangle| \geq B))^{1/2} && \text{by Cauchy-Schwarz's inequality} \\ &\leq C_3 \exp \left(-\frac{B}{2\rho} \right) && \text{by the first statement of Claim 7.9} \end{aligned}$$

Combining the bounds for (A), (B) and (C) we get:

$$\begin{aligned} \max_{\theta \in \mathcal{N}} \left| \frac{1}{n} \sum_{i=1}^n \left(\langle X_i, \theta \rangle^2 - \mathbb{E} \left(\langle X_i, \theta \rangle^2 \right) \right) \right| \\ \leq 2C_2 t^2 \log \left(\frac{2n}{d} \right) \sqrt{\frac{d}{n}} + 4C_2^2 t^2 \frac{d}{n} + C_3 \left(\frac{d}{2n} \right)^{\frac{C_3 t}{\rho}}, \end{aligned}$$

with probability greater than:

$$1 - \exp(-ct\sqrt{d}) - \exp \left(d \log \left(\frac{3}{\gamma\eta} \right) - c_1 \min \left(t^4 d \log^2 \left(\frac{2n}{d} \right), \frac{\sqrt{nd}}{\log(\frac{2n}{d})} \right) \right)$$

Given $\varepsilon \in (0, 1)$ we choose the parameters γ and η as follows: $\gamma = \frac{1}{2}$ and $\eta = \frac{\varepsilon}{40}$. Also, set $C(\varepsilon, t) := C_0 t^4 \varepsilon^{-2} \log^2(2t^2 \varepsilon^{-2})$, where $C > 0$ is a large enough absolute constant.

If $n \geq C(\varepsilon, t)d$ then we see that the lower bound for the probability in (7.4) is greater than $1 - \exp(-ct\sqrt{n})$ and that, with the same probability, we have:

$$\max_{\theta \in \mathcal{N}} \left| \frac{1}{n} \sum_{i=1}^n \left(\langle X_i, \theta \rangle^2 - \mathbb{E} \left(\langle X_i, \theta \rangle^2 \right) \right) \right| \leq \frac{\varepsilon}{40}$$

To complete the proof of Theorem 7.5, we use Lemma 7.13 in order to pass from $\mathcal{N}$ to the sphere $\mathbb{S}^{d-1}$.

Remark 7.15 *The proof of Theorem 7.5 was mainly based on Theorem 7.7 and some other additional results which required $n \leq e^{\sqrt{d}}$. This requirement followed, in particular, from the use of Paouris' theorem, Theorem 7.17.*

However, this doesn't create a problem regarding the generality of Theorem 7.5. Indeed, we can use the following argument.

For now, we have proved Theorem 7.5 for $n \leq e^{\sqrt{d}}$. Now, let $n > e^{\sqrt{d}}$.

If $X_i = (X_{i1}, \dots, X_{id}) \quad \forall 1 \leq i \leq n$ are the random vectors in the theorem, consider the smallest integer m such that $n \leq e^{\sqrt{m}}$. Note that $m > d$. Now, we define a new random vector $Y_i = (Y_{i1}, \dots, Y_{im})$ in $\mathbb{R}^m$ by setting $Y_{ij} = X_{ij}$ if $j \leq d$ and $Y_{ij} = g_{ij}$ when $j > d$, where g_{ij} are independent standard Gaussian random variables. From this definition, it follows that $Y_1, \dots, Y_n$ are i.i.d. isotropic random vectors in $\mathbb{R}^m$ and, since $n \leq e^{\sqrt{m}}$ we can apply the result as before, when $n \leq e^{\sqrt{d}}$. We have, if we identify $\theta = (\theta_1, \dots, \theta_d) \in \mathbb{S}^{d-1}$ with $Z = (Z_1, \dots, Z_m) \in \mathbb{S}^{m-1}$, where $Z_j = \theta_j$ if $j \leq d$ and $Z_j = 0$ if $j > d$, that:

$$\begin{aligned} \max_{\theta \in \mathbb{S}^{d-1}} \left| \frac{1}{n} \sum_{i=1}^n \left(\langle X_i, \theta \rangle^2 - \mathbb{E} \langle X_i, \theta \rangle^2 \right) \right| \\ \leq \max_{Z \in \mathbb{S}^{m-1}} \left| \frac{1}{m} \sum_{i=1}^m \left(\langle Y_i, \theta \rangle^2 - \mathbb{E} \langle Y_i, \theta \rangle^2 \right) \right| \leq \varepsilon \end{aligned}$$

Therefore, the assumption we made during the proof of Theorem 7.5 is justified, as with this extension we can recover the result for every value of n .

7.5 Appendix of this Chapter

In this Appendix, we will prove some useful results for log-concave distributions that come from the theory of convex bodies.

We start with the famous Paouris' inequality.

Paouris' inequality

Theorem 7.16 ([39]) *Let X be an isotropic log-concave vector on $\mathbb{R}^d$. Then:*

$$\mathbb{P} \left(\|X\|_2 \geq ct\sqrt{d} \right) \leq \exp(-t\sqrt{d}),$$

for every $t \geq 1$, where $c > 0$ is an absolute constant.

We will, now, prove this theorem. We first show that the proof can be reduced to the study of the moments of the function $x \mapsto \|x\|_2$. For every $q \geq 1$ we define:

$$I_q(X) = [\mathbb{E} (\|X\|_2^q)]^{1/q}$$

We will need the following estimate, which basically states that if X is sampled from an isotropic log-concave probability measure on $\mathbb{R}^d$, $I_q(X)$ remains essentially constant as long as $q \leq \sqrt{d}$. For the proof of it, we refer to [38].

Theorem 7.17 ([38]) *There exist absolute constants $c_1, c_2 > 0$ with the following property: if X is an isotropic log-concave random vector on $\mathbb{R}^s$, then:*

$$I_q(X) \leq c_2 I_2(X),$$

for every $q \leq c_1 \sqrt{d}$.

Finally, we are able to prove Theorem 7.16.

Proof (of Theorem 7.16) Assuming Theorem 7.17, we consider an isotropic log-concave vector X on $\mathbb{R}^d$. From Markov's inequality, for any $q \geq 2$ we have:

$$\mathbb{P} (\|X\|_2 \geq e^3 I_q(X)) \leq e^{-3q}$$

Now, consider Borell's lemma, in the form of (log-concave), and observe that the right-hand side $\theta \left(\frac{1-\theta}{\theta} \right)^{\frac{t+1}{2}}$ can be written in the form:

$$\frac{(1-\theta)^{\frac{t+1}{2}}}{\theta^{\frac{t-1}{2}}} < \frac{(1-\theta)^{\frac{t-1}{2}}}{\theta^{\frac{t-1}{2}}} = \left(\frac{1}{\theta} - 1 \right)^{\frac{t-1}{2}}$$

Then, from Borell's lemma and this observation, for every $s \geq 1$, we get:

$$\begin{aligned}\mathbb{P}(\|X\|_2 \geq e^3 I_q(X)s) &\leq (1 - e^{-3q}) \left(\frac{e^{-3q}}{1 - e^{-3q}} \right)^{(s+1)/2} \\ &\leq e^{-qs}\end{aligned}$$

Choosing $q = c_1 \sqrt{d}$, and using Theorem 7.17, for every $s \geq 1$, we get:

$$\mathbb{P}(\|X\|_2 \geq c_2 e^3 I_2(X)s) \leq \exp(-c_1 \sqrt{d}s)$$

Since X is isotropic, we have $I_2(X) = \sqrt{d}$. This proves the theorem. $\square$

Chapter 8

Case Study: Random Vectors with Sub-exponential marginals

In the previous Chapter, we analyzed the rate of convergence of the sample covariance estimator constructed with n samples from an isotropic log-concave distribution.

We saw that also in this case, if $n = \mathcal{O}(d)$, with overwhelming probability one has:

$$\|\Sigma_n - I\| \leq \varepsilon,$$

where $c > 0$ is an absolute constant.

Moreover, we saw that log-concave measures imply sub-exponential tails (Claim 7.9).

Therefore, we would like to extend the result to this framework.

8.1 Main Result

In this Chapter, we will consider the following class of vectors.

Definition 8.1 Let $X_1, \dots, X_n$ be a sequence of random vectors in $\mathbb{R}^d$, which are not even necessarily identically distributed.

We say that such a sequence is uniformly ψ_1 if, for some $\psi > 0$, we have:

$$\sup_{1 \leq i \leq n} \sup_{y \in \mathbb{S}^{d-1}} \|\langle X_i, y \rangle\|_{\psi_1} \leq \psi \quad (8.1)$$

We say that such a sequence satisfies the boundedness condition with constant K , where $K \geq 1$, if:

$$\mathbb{P} \left(\max_{1 \leq i \leq n} \frac{|X_i|}{\sqrt{d}} > K \max \left\{ 1, \left(\frac{n}{d} \right)^{1/4} \right\} \right) \leq \exp(-\sqrt{d}) \quad (8.2)$$

Observe that the conditions just stated are valid for many classes of distributions. For instance, the log-concave isotropic random vectors in $\mathbb{R}^d$ we studied in Chapter 7 satisfy (8.1) and (8.2) for appropriate absolute constants ψ and K . Indeed, we proved this in the first two assertions of Claim 7.9. Another example is random vectors uniformly distributed on the Euclidean ball of radius $K\sqrt{d}$, which clearly satisfy (8.2) and they also satisfy (8.1) with $\psi = CK$. The boundedness assumption is of the type stated in Remark 2.5. Observe that for this last class of vectors, we proved in Chapter 6 a version of Theorem 8.2 with a weaker probability estimate in the case of isotropic log-concave random vectors under an additional assumption of unconditionality. This was Theorem 6.10.

The main result of this Chapter is the following theorem.

Theorem 8.2 ([3]) *Let $n \geq d$ be positive integers and $\psi, K \geq 1$. Let $X_1, \dots, X_n$ be independent random vectors in $\mathbb{R}^d$ satisfying the conditions given in Definition 8.1. Then, with probability at least $1 - 2\exp(-c\sqrt{d})$, we have:*

$$\sup_{x \in \mathbb{S}^{d-1}} \left| \frac{1}{n} \sum_{i=1}^n \left(|\langle X_i, x \rangle|^2 - \mathbb{E} |\langle X_i, x \rangle|^2 \right) \right| \leq C(\psi + K)^2 \sqrt{\frac{d}{n}}$$

8.2 Proof of the Main Result

The proof of Theorem 8.2 is close to the arguments that have been used in Chapter 7 to prove Theorem 7.5.

The main technical tool will be the following theorem, a technical variant of Theorem 7.7.

Theorem 8.3 ([2]) *Let $X_1, \dots, X_n$ be independent random vectors in $\mathbb{R}^d$ such that the condition (8.1) is satisfied.*

Let $\mathbf{X}$ be a random $d \times n$ matrix whose columns are the X_i 's and let $\mathbf{X}_m$ defined as in Definition 7.6, for every $1 \leq m \leq n$. Then, for every $0 \leq \ell \leq \log m$ and every $K \geq 1$, one has:

$$\begin{aligned} \mathbb{P} \left(\mathbf{X}_m \geq C\psi K \left(\frac{m}{2^\ell} \log \frac{48en2^\ell}{m} + \sqrt{m} \log \frac{2n}{m} \right) + 6 \max_{1 \leq i \leq n} \|X_i\|_2 \right) \\ \leq (1 + 2^\ell) \exp \left(-2K \frac{m}{2^\ell} \log \frac{12en2^\ell}{m} \right), \end{aligned}$$

where C is an absolute constant.

In particular, choosing $0 \leq \ell \leq \log m$ to be the largest integer satisfying:

$$\frac{2m}{2^\ell} \log \frac{12en2^\ell}{m} \geq \sqrt{m} \log \frac{2n}{m}$$

we obtain that for every $K \geq 1$

$$\mathbb{P} \left(\mathbf{X}_m \geq C\psi K\sqrt{m} \log \frac{2n}{m} + 6 \max_{1 \leq i \leq n} \|X_i\|_2 \right) \leq (1 + 2 \log m) \exp \left(-K\sqrt{m} \log \frac{2n}{m} \right)$$

Proof (of Theorem 8.3) As in the proof of Theorem 7.7, we first construct a set $\mathcal{M}$, a set of vectors with $|\text{support}(\mathcal{M})| \leq m$, that will serve the following purposes:

- it will make us able to estimate $\sup_{x \in \mathcal{M}} \|\mathbf{X}x\|_2$ with high probability
- using $\mathcal{M}$, we will approximate any point from B_2^n with at most m non-zero coordinates

If $\ell = 0$, which means $m = 1$, we will construct $\mathcal{M}$ as in the first case of Theorem 7.7:

$$\mathcal{M} = \cup_{\substack{E \subset \{1, \dots, n\} \\ |E| = m}} \mathcal{N} \left(E, \frac{1}{4}, 1 \right)$$

If, instead, $\ell > 0$, we will construct $\mathcal{M}$ as in the second case of Theorem 7.7. Now, ℓ is a fixed integer and we fix some positive integers $a_0, \dots, a_\ell$ such that:

- (i) $\sum_{k=0}^{\ell} a_k = m$
- (ii) $a_k \leq m \cdot 2^{-k+1} \quad \forall 1 \leq k \leq \ell$
- (iii) $a_0 \leq m \cdot 2^{-\ell}$

Then, we set $\mathcal{M} = \mathcal{M}_0 \cap 2B_2^n$, where $\mathcal{M}_0$ is the set of all vectors of the form $x = \sum_{k=0}^{\ell} x_k$, where the x_i 's have disjoint supports and:

$$x_0 \in \cup_{\substack{E \subset \{1, \dots, n\} \\ |E| \leq a_0}} \mathcal{N} \left(E, \frac{1}{4}, 1 \right)$$

and

$$x_k \in \cup_{\substack{E \subset \{1, \dots, n\} \\ |E| \leq a_k}} \mathcal{N} \left(E, 2^{-k}, \sqrt{\frac{2^k}{m}} \right) \quad \text{for } 1 \leq k \leq \ell$$

Now, without repeating all of the steps we already did in Theorem 7.7, we upper bound $\|\mathbf{X}x\|_2^2 \leq C \max_i \|X_i\|_2^2 + D_x$, where $D_x = \sum_{i \neq j} \langle x(i)X_i, x(j)X_j \rangle$ and $x(i)$ denote the coordinates of x , a fixed vector in $\mathcal{M}$, for $1 \leq i \leq n$. Then, we need to estimate $D_x = D'_x + D''_x$.

Using Lemma 7.11 with $L = 4K \frac{m}{2^\ell} \log \frac{48en2^\ell}{m}$ we are able to estimate D'_x :

$$\begin{aligned} \mathbb{P} \left(\sup_{x \in \mathcal{M}} D'_x > 16\mathbf{X}_m K \psi \frac{m}{2^\ell} \log \frac{48en2^\ell}{m} + C\mathbf{X}_m K \psi \sqrt{m} \log \frac{2n}{m} \right) \\ \leq \exp \left(-2K \frac{m}{2^\ell} \log \frac{48en2^\ell}{m} \right) + \ell \exp \left(-2K \frac{m}{2^\ell} \log \frac{12en4^\ell}{m} \right) \end{aligned}$$

Also, by making the same choices that were done in the proof of Theorem 7.7 to estimate D''_x , we obtain:

$$\mathbb{P} \left(\sup_{x \in \mathcal{M}} D''_x > 3C\psi\mathbf{X}_m K \sqrt{m} \log \frac{2n}{m} \right) \leq \ell \exp \left(-K \frac{6m}{2^\ell} \log \frac{12e4^\ell n}{m} \right)$$

Now, it just remains to apply the approximation procedure. By (7.4) we have:

$$\|\mathbf{X}x\|_2^2 \leq 2 \max \{ 4 \max_i \|X_i\|_2^2, D_x \}$$

The, by approximation of a vector $z \in \mathbb{S}^{n-1}$ with a vector $x \in \mathcal{M}$ and the above inequality we get:

$$\mathbf{X}_m^2 \leq \max \{ 36 \max_i \|X_i\|_2^2, 18 \sup_{x \in \mathcal{M}} D_x \},$$

which implies the result, possibly by adjusting some constants.

The second claim, the "in particular" one, follows simply from the first one. $\square$

From this theorem, we can derive the following corollary.

Corollary 8.4 ([3]) *Let $X_1, \dots, X_n$ be independent random vectors in $\mathbb{R}^d$ such that the conditions given in Definition 8.1 are satisfied. Let $\mathbf{X}$ be a random $d \times n$ matrix whose columns are the X_i 's. Then for every $t \geq 1$ one has:*

$$\mathbb{P} \left(\exists m : \mathbf{X}_m \geq C\psi t \max \left\{ \sqrt{m} \ln \frac{2n}{m}, \sqrt{d} \right\} + 6 \max_{i \leq n} |X_i| \right) \leq \exp(-t\sqrt{d})$$

Proof (of Corollary 8.4) This is simply proved by using the "in particular" claim in Theorem 8.3, adjusting corresponding constants, and taking a union bound. $\square$

We are finally ready to prove Theorem 8.2.

Proof (of Theorem 8.2) For $x \in \mathbb{S}^{d-1}$ set:

$$S(x) = \left| \frac{1}{n} \sum_{i=1}^n \left(|\langle X_i, x \rangle|^2 - \mathbb{E} |\langle X_i, x \rangle|^2 \right) \right|$$

Let $B > 0$ be a parameter whose value will be specified later and observe that:

$$\sup_{x \in \mathbb{S}^{d-1}} S(x) \leq \sup_{x \in \mathbb{S}^{d-1}} \left(\left| \frac{1}{n} \sum_{i=1}^n \left(\min(|\langle X_i, x \rangle|, B)^2 - \mathbb{E} \left[\min(|\langle X_i, x \rangle|, B)^2 \right] \right) \right| \right) \quad (\text{A})$$

$$+ \frac{1}{n} \mathbb{E} \sum_{i=1}^n \left(|\langle X_i, x \rangle|^2 - B^2 \right) \mathbb{1}_{\{|\langle X_i, x \rangle| \geq B\}} \quad (\text{B})$$

$$+ \frac{1}{n} \sum_{i=1}^n \left(|\langle X_i, x \rangle|^2 - B^2 \right) \mathbb{1}_{\{|\langle X_i, x \rangle| \geq B\}} \quad (\text{C})$$

We will estimate each of the summands (A), (B), and (C) separately.

- **(A)** To estimate the first summand, given $x \in \mathbb{S}^{d-1}$ and $1 \leq i \leq n$, we let $Z_i = Z_i(x) = \min(|\langle X_i, x \rangle|, B)^2 - \mathbb{E} \left[\min(|\langle X_i, x \rangle|, B)^2 \right]$. Then $\|Z_i\|_\infty \leq B^2$. Moreover, using the second claim of Proposition 2.28:

$$\text{Var}(Z_i) \leq \mathbb{E}(|\langle X_i, x \rangle|^4) \leq C_1 \psi^4,$$

for some absolute constant $C_1 > 0$.

Observe that $\sigma^2 = \frac{1}{n} \sum_{i=1}^n \text{Var}(Z_i) \leq C_1 \psi^4$. Thus, by Bernstein's inequality (Proposition 2.31):

$$\mathbb{P}((A) \geq t) = \mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n Z_i \geq t\right) \leq \exp\left(-\frac{t^2 n}{2(C_1 \psi^4 + B^2 \frac{t}{3})}\right)$$

Now, let $\mathcal{N}$ be a $\frac{1}{3}$ -net of $\mathbb{S}^{d-1}$. From Lemma 2.15, we know that $|\mathcal{N}| \leq 7^d$. Then by the union bound we obtain that if

$$t^2 n > 8C_1 \psi^4 d \ln 7 \quad \text{and} \quad \theta n > \frac{8}{3} B^2 d \ln 7$$

then:

$$\begin{aligned} \mathbb{P}\left(\sup_{x \in \mathcal{N}} (A) \geq t\right) &\leq \exp\left(d \ln 7 - \frac{\theta^2 n}{2(C_1 \psi^4 + B^2 \frac{t}{3})}\right) \\ &\leq \exp\left(-\frac{\theta^2 n}{4(C_1 \psi^4 + B^2 \frac{t}{3})}\right) \end{aligned}$$

- **(B)** To bound the second term, observe:

$$\begin{aligned} \sup_{x \in \mathbb{S}^{d-1}} (B) &\leq \frac{1}{n} \sum_{i=1}^n \sup_{x \in \mathbb{S}^{d-1}} \|\langle X_i, x \rangle\|_4^2 \mathbb{P}(|\langle X_i, x \rangle| \geq B)^{1/2} && \text{By Hölder's inequality} \\ &\leq C_2 \psi^2 \exp\left(-\frac{B}{\psi}\right), && \text{from assumption (8.1)} \end{aligned}$$

for some absolute constant $C_2 \geq 1$.

- (C) To estimate the third term, we will use the following notation:

- $M = \max \left\{ \psi^2 d, \max_{i \leq n} |X_i|^2 \right\}$
- $E_B = E_B(x) = \{i \leq n : |\langle X_i, x \rangle| \geq B\}$
- $m = \sup_{x \in \mathbb{S}^{d-1}} |E_B(x)|$

By the definition of $\mathbf{X}_m$, we have for every $x \in \mathbb{S}^{d-1}$:

$$B^2 |E_B| \leq \sum_{i \in E_B} |\langle X_i, x \rangle|^2 \leq \sup_{|E| \leq m} \sum_{i \in E} |\langle X_i, x \rangle|^2 \leq \mathbf{X}_m^2$$

which yields $B^2 m \leq \mathbf{X}_m^2$ and $n(B) \leq \mathbf{X}_m^2$.

Corollary 8.4 implies that for some absolute constant $C \geq C_2$, with probability at least $1 - \exp(-\sqrt{d})$ one has

$$B^2 m \leq C \left(M + \psi^2 m \ln^2 \frac{2n}{m} \right)$$

and

$$\sup_{x \in \mathbb{S}^{d-1}} (C) \leq C \left(\frac{M}{n} + \psi^2 \frac{m}{n} \ln^2 \frac{2n}{m} \right)$$

Now that we have these estimates for (A), (B), and (C), choosing $B = 2\sqrt{2C}\psi \ln(5\frac{n}{d})$. Then, the estimate (B) gives:

$$(C) \leq C\psi^2 \frac{d}{n} \leq C \frac{M}{N} \text{ for all } x \in \mathbb{S}^{d-1}$$

and together with the estimate (C), it yields that with probability at least $1 - \exp(-\sqrt{d})$ one has:

$$\sup_{x \in \mathbb{S}^{d-1}} ((B) + (C)) \leq C \left((2\frac{M}{n}) + \psi^2 (\frac{m}{n}) \ln^2 (2\frac{n}{m}) \right)$$

Now, we prove the following claim.

Claim 8.5 *In the set where it holds that:*

$$B^2 m \leq C \left(M + \psi^2 m \ln^2 \frac{2n}{m} \right)$$

and

$$\sup_{x \in \mathbb{S}^{d-1}} (C) \leq C \left(\frac{M}{n} + \psi^2 \frac{m}{n} \ln^2 \frac{2n}{m} \right)$$

we have $M \geq \psi^2 m \left(\ln \frac{2n}{m} \right)^2$

Proof (of Claim 8.5) By contradiction, assume $M < \psi^2 m \left(\ln \frac{2n}{m} \right)^2$. Then by our assumption on the set, we observe that $B^2 \leq 2C\psi^2 \ln^2(\frac{2n}{m})$, which implies, solving for m :

$$m \leq 2n \exp\left(-\frac{B}{\psi\sqrt{2C}}\right) = \frac{2d^2}{25n}$$

By our hypothetical upper bound for M we are using for the contradiction and since $f(m) = m \ln^2(\frac{2n}{m})$ increases on $[1, \frac{2n}{e^2}]$, we get:

$$\psi^2 d \leq M \leq \psi^2 \left(\frac{8d^2}{25n}\right) \ln^2\left(\frac{5n}{d}\right)$$

which is impossible. $\square$

Therefore, following this claim, we can affirm that $M \geq \psi^2 m \left(\ln \frac{2n}{m}\right)^2$ and it implies that:

$$\mathbb{P}\left(\sup_{x \in \mathbb{S}^{d-1}} ((B) + (C)) \leq 3C \frac{M}{n}\right) \geq 1 - \exp(-\sqrt{d})$$

Combining this estimate with the one we obtained for the first term (A), we get:

$$\mathbb{P}\left(\sup_{x \in \mathcal{N}} S(x) \leq t + 3C \frac{M}{N}\right) \geq 1 - \exp(-\sqrt{d}) - \exp\left(-\frac{t^2 n}{4(C_1 \psi^4 + B^2 \frac{t}{3})}\right)$$

Now, set $t = C_3 \psi^2 \sqrt{\frac{d}{n}}$, where C_3 is a sufficiently large absolute positive constant so that we used in obtaining the estimate for (A) are satisfied. Then using the boundedness condition (8.2) with constant K we obtain:

$$\begin{aligned} \mathbb{P}\left(\sup_{x \in \mathcal{N}} S(x) \leq (C_3 \psi^2 + 3CK^2) \sqrt{\frac{d}{n}}\right) &\geq 1 - \exp(-\sqrt{d}) - \exp(-cd) \\ &\geq 1 - 2\exp(-c\sqrt{d}), \end{aligned}$$

where c is a sufficiently small positive constant.

Therefore, we just proved the desired estimate on the $\frac{1}{3}$ -net. Instead, we want to achieve it also on the whole sphere $\mathbb{S}^{d-1}$. To pass from $\mathcal{N}$ to the whole sphere note that $S(x)$ can be written as $|\langle Tx, x \rangle|$, where T is a self-adjoint operator on $\mathbb{R}^d$. Therefore, if we write each $x \in \mathbb{S}^{d-1}$ as $x = y + z$, where $y \in \mathcal{N}$ and $|z| \leq \frac{1}{3}$, we get:

$$\begin{aligned} \|T\| &= \sup_{x \in \mathbb{S}^{d-1}} |\langle Tx, x \rangle| \leq \sup_{y \in \mathcal{N}} |\langle Ty, y \rangle| + \frac{2}{3} \sup_{y \in \mathcal{N}} |Ty| + \sup_{|z| \leq \frac{1}{3}} |\langle Tz, z \rangle| \\ &\leq \sup_{y \in \mathcal{N}} S(y) + \frac{7}{9} \|T\| \end{aligned}$$

which implies the desired estimate on the whole sphere $\mathbb{S}^{d-1}$. $\square$

Chapter 9

Case Study: Random Vectors with bounded L_p marginal moments

In this Chapter, we will illustrate an approach that allows to bound the difference, in operator norm, between the empirical covariance matrix and the true covariance matrix of a distribution in $\mathbb{R}^d$ that we assume is such that:

- the distribution is supported in a relatively small ball
- the linear functionals are uniformly bounded in L^p for some $p > 8$.

As in the previous cases, we will consider the following setting:

- we take n independent samples $X_1, \dots, X_n$ from an isotropic probability distribution μ in $\mathbb{R}^d$
- constructing a $n \times d$ matrix $\mathbf{X}$ with rows X_i
- every linear functional has bounded p moments: $\sup_{t \in \mathbb{S}^{d-1}} \|\langle X, t \rangle\|_p \leq K_1$

Then, we aim to obtain the following result (w. h. p.):

$$\sup_{t \in B_2^d} \left| \frac{1}{n} \sum_{i=1}^n \langle X_i, t \rangle^2 - 1 \right| \leq C \sqrt{\frac{d}{n}}$$

Therefore, if $n \geq \mathcal{O}(d)$ we have that the sample covariance matrix Σ_n approximates the true covariance matrix with ε -accuracy.

9.1 Main Result

Let's state the main theorem of this chapter.

Theorem 9.1 ([34]) Let μ be an isotropic probability measure in $\mathbb{R}^d$. Let $n \geq d$ and assume that $\max_i \|X_i\|_2 \leq C_0 \cdot (nd)^{1/4}$. Let $K_1 \geq 1$ and let k_0 be the first integer which satisfies: $k_0 \log\left(\frac{en}{k_0}\right) \geq d$. If $p > 8$, $\sup_{t \in B_2^d} \|\langle t, \cdot \rangle\|_p \leq K_1$, and $1 \leq \beta \leq C_1 k_0$, then, with probability at least $1 - C_2 \left(\frac{1}{n^\beta} + \exp(-C_3 d)\right)$, one has:

$$\sup_{t \in S^{d-1}} \left| \frac{1}{n} \sum_{i=1}^n \langle X_i, t \rangle^2 - 1 \right| \leq C_4 \sqrt{\frac{d}{n}}$$

where C_1, C_2, C_3, C_4 depend only on β, p, C_0, K_1 .

Remark 9.2 Observe that the assumption $\max_i \|X_i\|_2 \leq C_0 \cdot (nd)^{1/4}$ is something like the one of Remark 2.5.

9.2 Proof of the Main Result

Let's now detail and establish all the essential elements that will guide us towards the proof of this theorem.

First, we state a result valid for non-increasing rearrangement of *i.i.d* random variables.

Lemma 9.3 ([34]) Let $Z_1, \dots, Z_n$ be *i.i.d* random variables, distributed according to Z . If $p > 4$, $C_0, \beta > 0$, then $\exists$ constants c_0, c_1, c_2, c_3 that depend only on p, C_0 and β such that:

if $\|Z\|_\infty \leq C_0 (nd)^{1/4}$ and $u \geq c_0$ then:

- $\sum_{i \leq u} (Z_i^*)^2 \leq c_1 \left(1 + u \|Z\|_p^2\right) (nd)^{1/2}$, with probability at least $1 - c_2 u^{-\beta}$
- $\sum_{i=uk_0+1}^n (Z_i^*)^4 \leq c_1 \|Z\|_p \cdot n$, with probability at least $1 - 2 \exp(-c_3 u d)$

Proof (of Lemma 9.3) To prove this result it will be fundamental the following claim.

Claim 9.4 Let $Z_1, \dots, Z_n$ be *i.i.d* random variables, distributed according to Z . Consider (Z_i^*) , the non-increasing rearrangement of $(|Z_i|)$. Then:

$$\mathbb{P}(Z_{2^s}^* \geq t) \leq \binom{n}{2^s} [\mathbb{P}(|Z| \geq t)]^{2^s}$$

Proof (of Claim 9.4) This claim is trivial and follows from union bound. Indeed, observe that $Z_{2^s}^*$ denote the 2^s -th largest element of the series $(|Z_i|)_{i=1}^n$, i.e. there are $2^s - 1$ elements that have a greater or equal value in the non-increasing rearrangement. Then, asking for $\mathbb{P}(Z_{2^s}^* \geq t)$ is equivalent to

asking that $\mathbb{P} \left(\min_{j \in I_{2^s}^*} |Z|_j \geq t \right)$, where $I_{2^s}^*$ denotes the set of indices corresponding to the first 2^s elements of the non-increasing rearrangement. Since we don't have access to $I_{2^s}^*$, using the fact that $|I_{2^s}^*| = 2^s$ and union bound, we obtain the result of the claim:

$$\mathbb{P} (Z_{2^s}^* \geq t) \leq \binom{n}{2^s} [\mathbb{P} (|Z| \geq t)]^{2^s} \quad \square$$

Continuing the proof of the lemma, we also observe that the assumption in Theorem 9.1 regarding the constant k_0 , i.e. the fact that k_0 is the first integer which satisfies: $k_0 \log \left(\frac{en}{k_0} \right) \geq d$, implies that:

- if $\log(en) \lesssim d$, then $k_0 \log \left(\frac{en}{k_0} \right) \sim d$
- if $\log(en) \gtrsim d$, $k_0 = 1$

We assume to be in the first of the two cases. If $k_0 = 1$, the modifications required are minimal and we will omit the proof in that case. Now, fix $\varepsilon = p/4 - 1$, let $\beta \geq 1$ and pick s_1 to be a constant that will be specified later and which depends only on β and p and. For 2^s in the interval $[2^{s_1}, k_0]$, define $\alpha_s = \frac{(en/2^s)^{(1+\varepsilon)/p}}{(nd)^{1/4}} = \frac{2^{s/4}}{d^{1/4}}$. Then, using Markov's inequality we get that: $\mathbb{P} (|Z| \geq \|Z\|_p \cdot t) \leq \frac{\mathbb{E}|Z|^p}{\|Z\|_p^p t^p} = \frac{\|Z\|_p^p}{\|Z\|_p^p t^p} \leq t^{-p}$. Now we can plug this estimate into Claim 9.4 and, by choosing $t = \alpha_s \cdot (nd)^{1/4}$, in the interval $[2^{s_1}, k_0]$, we obtain:

$$\mathbb{P} \left(Z_{2^s}^* \geq \|Z\|_p \alpha_s (nd)^{1/4} \right) \leq \left(\frac{en}{2^s} \right)^{-\varepsilon 2^s} \quad (9.1)$$

Moreover, observe that:

$$\begin{aligned} Z_1^* + \dots Z_n^* &= (Z_1^* + Z_2^*) + (Z_3^* + Z_4^*) + (Z_5^* + \dots + Z_8^*) + \dots (\dots + Z_n^*) \\ &\leq Z_1^* + Z_2^* + 2Z_4^* + 4Z_4^* + \dots \leq \sum_{2^s} 2^s Z_{2^s}^* \end{aligned}$$

It follows that, since we fixed previously $4(1+\varepsilon)/p = 1$ and, for a cautious choice of $s_1(\beta, p)$ to work out properly the computations, using the assumption $\|Z\|_\infty \leq C_0 (nd)^{1/4}$ and (9.1), with probability at least $1 - \left(\frac{en}{2^{s_1}} \right)^{-c\varepsilon 2^{s_1}} \geq 1 - c_0 n^{-\beta}$:

$$\begin{aligned}
 \sum_{i \leq u} (Z_i^*)^2 &\leq \sum_{2^s \leq uk_0} 2^s (Z_{2^s}^*)^2 \\
 &\leq \sum_{2^s \leq 2^{s_1}} \|Z\|_\infty^2 + \sum_{2^{s_1} \leq 2^s \leq uk_0} 2^s \|Z\|_p^2 \cdot \alpha_s^2 (nd)^{1/2} \\
 &\leq \|Z\|_\infty^2 2^{s_1} + \|Z\|_p^2 (nd)^{1/2} \sum_{2^{s_1} \leq 2^s \leq uk_0} 2^s \alpha_s^2 \\
 &\lesssim C_0 2^{s_1} (nd)^{1/2} + \|Z\|_p^2 n^{1/2} \sum_{2^{s_1} \leq 2^s \leq ud} 2^s \\
 &\lesssim_{\beta, p} \left(1 + u^{1/2} \|Z\|_p^2\right) (nd)^{1/2}
 \end{aligned}$$

We have, in this way, proved the first part of Lemma 9.3.

We turn, now, to the second part, which will be proved using again Claim 9.4.

For the second part of the statement, we choose $t = \|Z\|_p \left(\frac{en}{2^s}\right)^{\frac{1+\varepsilon}{p}} = \|Z\|_p \left(\frac{en}{2^s}\right)^{1/4}$. Also, choose u such that: $\max\{\frac{2}{\varepsilon}, 1\} < u \lesssim \left(\frac{n}{k_0}\right)^{1/2}$. Hence, by Claim 9.4 and Markov's inequality as before:

$$\begin{aligned}
 \sum_{i=uk_0+1} (Z_i^*)^4 &\leq \sum_{uk_0 \leq 2^s \leq n} 2^s (Z_{2^s}^*)^4 \leq \sum_{uk_0 \leq 2^s \leq n} 2^s \cdot t^4 \lesssim \|Z\|_p^4 \sum_{uk_0 \leq 2^s \leq n} 2^s \left(\frac{en}{2^s}\right)^{\frac{1.4}{4}} \\
 &\lesssim \|Z\|_p^4 n
 \end{aligned}$$

This is achieved with probability at least

$$\begin{aligned}
 1 - \sum_{uk_0 \leq 2^s \leq n} \exp\left(-\varepsilon 2^s \log\left(\frac{en}{2^s}\right)\right) &\geq 1 - \exp\left(-c_1 \varepsilon uk_0 \log\left(\frac{en}{2^s}\right)\right) \\
 &\geq 1 - \exp(-c_2 \varepsilon ud),
 \end{aligned}$$

as claimed in the lemma.

Now, recall that, at the beginning of the proof of the lemma we assumed $k_0 > 1$. Let's consider $k_0 = 1$ and see which modifications are needed in the proof in this case. For the first we can choose $\alpha_s = \frac{(en/1)^{(1+\varepsilon)/p}}{(nd)^{1/4}}$. Then, following the same steps, we get that:

$$\begin{aligned}
 \sum_{i \leq u} (Z_i^*)^2 &\leq \sum_{2^s \leq u} 2^s (Z_{2^s}^*)^2 \\
 &\leq \|Z\|_\infty^2 + \sum_{1 \leq 2^s \leq u} 2^s \|Z\|_p^2 \cdot \alpha_s^2 (nd)^{1/2} \\
 &\leq \|Z\|_\infty^2 + \|Z\|_p^2 (nd)^{1/2} \sum_{1 \leq 2^s \leq u} 2^s \alpha_s^2 \\
 &\lesssim C_0 (nd)^{1/2} + \|Z\|_p^2 n^{1/2} \sum_{1 \leq 2^s \leq ud} 2^{s/2} \\
 &\lesssim_{\beta, p} \left(1 + u^{1/2} \|Z\|_p^2\right) (nd)^{1/2}
 \end{aligned}$$

Also, for the second part of the statement, by choosing the same t as in the case $k_0 > 1$ and fixing, now $k_0 = 1$, we get the same bound with the same probability (the constants vary). $\square$

Let's continue with the proof of Theorem 9.1. For this, we will need also the following Corollary.

Corollary 9.5 ([34]) Fix $0 < \gamma < 1$ and k_3 , which will be specified later. Let $k_\ell = \gamma^\ell k_0$, with $k_0 > 1$, and let ℓ_0 be the first integer satisfying:

$$k_{\ell_0} \log \left(\frac{en}{k_{\ell_0}} \right) \leq k_3 (nd)^{1/4}$$

Then, $\exists$ a constant c_1 such that for every $0 < \gamma < 1$ there exists a constant $c_2 = c_2(\gamma)$ such that:

Let $p > 4$ and $\varepsilon = p/4 - 1$, let $\ell_1 > 0$ be any integer for which $k_{\ell_1} \geq 1$. Let $Z_1, \dots, Z_N$ be i.i.d. random variables, distributed according to Z with $\|Z\|_p < \infty$. Then, $\forall 0 \leq \ell < \ell_1$, with probability at least $1 - (en/k_{\ell+1})^{-\varepsilon k_{\ell+1}}$, we have:

$$\left(\sum_{j=k_{\ell+1}}^{k_\ell} (Z_j^*)^2 \right)^{1/2} \leq c_1 \|Z\|_p \eta_\ell,$$

where $\eta_\ell \sim (nk_\ell)^{1/4}$ and, in particular, $\sum_{\ell=0}^{\ell_1-1} \eta_\ell \leq c_2 (nd)^{1/4}$.

Proof (of Corollary 9.5) This corollary uses the same idea of the previous Lemma. Indeed, applying the result of the Claim 9.4 and Markov's inequality (as in the previous lemma), we get:

$$\mathbb{P} (Z_{2^s}^* \geq \|Z\|_p \cdot t) \leq \binom{n}{2^s} [t^{-p}]^{2^s}$$

Then, choosing $t = \left(\frac{en}{k_\ell} \right)^{(1+\varepsilon)/p} = \left(\frac{en}{k_\ell} \right)^{1/4}$, and also remembering that in the assumptions of the lemma, we chose $\eta_\ell \sim (nk_\ell)^{1/4}$:

$$\begin{aligned}
 \sum_{j=k_{\ell+1}}^{k_{\ell}} (Z_i^*)^2 &\leq \sum_{k_{\ell+1} \leq 2^s \leq k_{\ell}} 2^s (Z_{2^s}^*)^2 \leq \sum_{k_{\ell+1} \leq 2^s \leq k_{\ell}} 2^s \cdot \|Z\|_p^2 t^2 \\
 &\leq \|Z\|_p^2 \sum_{k_{\ell+1} \leq 2^s \leq k_{\ell}} 2^s \left(\frac{en}{k_{\ell}}\right)^{1/2} \leq \|Z\|_p^2 \sum_{k_{\ell+1} \leq 2^s \leq k_{\ell}} 2^{\frac{s}{2}} n^{1/2} \\
 &\lesssim \|Z\|_p^2 \eta_{\ell}^2
 \end{aligned}$$

By taking the square root and properly choosing the constant c_1 , the result follows. $\square$

To continue with the proof of Theorem 9.1, we also need to introduce:

- the sets $U_k = \{x \in \mathbb{S}^{n-1} : |\text{supp}(x)| \leq k\}$
- the sets $A_k = \sup_{a \in U_k} \left\| \sum_{i=1}^N a_i X_i \right\|$

Observe that, $\forall k \leq n$:

$$\begin{aligned}
 A_k &= \sup_{a \in U_k} \left\| \sum_{i=1}^N a_i X_i \right\| = \sup_{a \in U_k} \sup_{t \in B_2^d} \left\langle \sum_{i=1}^n a_i X_i, t \right\rangle \\
 &= \sup_{t \in B_2^d} \sup_{a \in U_k} \sum_{i=1}^n a_i \langle X_i, t \rangle = \sup_{t \in B_2^d} \left(\sum_{i=1}^k (\langle X_i, t \rangle^*)^2 \right)^{1/2}
 \end{aligned}$$

This motivates studying the quantity A_k .

For every k , let δ_k be determined later and let $\mathcal{N}_k$ be a subset of B_2^n such that for every $x \in \mathbb{R}^n$:

$$\sup_{y \in \mathcal{N}_k} \langle y, x \rangle \geq (1 - \delta_k) \sup_{z \in U_k} \langle z, x \rangle$$

It is standard to verify that there is a set $\mathcal{N}_k$ as above of cardinality at most $\exp\left(k \log\left(\frac{en}{k\delta_k}\right)\right)$, using the bound on the size of ε -net, Lemma 2.15 and Theorem A.1.

Making use of Corollary 9.5 we obtain the following lemma.

Lemma 9.6 ([34]) *For every $p > 8, C_0, K_1$ and $\beta > 0$ as in Theorem 1.2, $\exists$ constants c_1 and c_2 (depending only on p, C_0, K_1 and β) such that:*

If $I \subset \{1, \dots, n\}$, then probability at least $1 - \frac{c_1}{n^\beta}$,

$$\sup_{a \in U_{k_0}} \sup_{b \in U_{k_0}} \left\langle \sum_{i \in I} a_i X_i, \sum_{i \in I^c} b_i X_i \right\rangle \leq c_2 (nd)^{1/4} A_{k_0}$$

Proof (of Lemma 9.6) Again, we will restrict ourselves to the case in which $k_0 > 1$, since the modifications needed when $k_0 = 1$ are minor.

Consider the sets U_{k_ℓ} , defined as above and let:

$$B_{k_\ell} := \sup_{a \in U_{k_\ell}} \sup_{b \in U_{k_\ell}} \langle \sum_{i \in I} a_i X_i, \sum_{i \in I^c} b_i X_i \rangle$$

We claim the following.

Claim 9.7 $\forall 0 \leq \ell \leq \ell_1$:

$$\begin{aligned} (1 - \delta_{k_\ell}) (1 - \delta_{k_{\ell+1}}) B_{k_\ell} &\leq B_{k_{\ell+1}} + \sup_{b \in \mathcal{N}_{k_\ell}} \left[\sum_{i=k_{\ell+1}+1}^{k_\ell} \left(\left\langle \sum_{j \in I^c} b_j X_j, X_i \right\rangle^* \right)^2 \right]^{1/2} \\ &\quad + \sup_{a \in \mathcal{N}_{k_{\ell+1}}} \left[\sum_{j=k_{\ell+1}+1}^{k_\ell} \left(\left\langle \sum_{i \in I} a_i X_i, X_j \right\rangle^* \right)^2 \right]^{1/2}, \end{aligned}$$

where ℓ_1 will be defined later.

Proof (of Claim 9.7) Indeed, by the definition of $\mathcal{N}_{k_\ell}$:

$$B_{k_\ell} \leq \frac{1}{1 - \delta_{k_\ell}} \sup_{a \in U_{k_\ell}} \sup_{b \in \mathcal{N}_{k_\ell}} \left\langle \sum_{i \in I} a_i X_i, \sum_{i \in I^c} b_i X_i \right\rangle$$

Now, by exchanging the order in which we take the supremum, and defining $W_{b,i} = \langle X_i, \sum_{j \in I^c} b_j X_j \rangle$, we get:

$$\begin{aligned} B_{k_\ell} &\leq \frac{1}{1 - \delta_{k_\ell}} \sup_{a \in U_{k_\ell}} \sup_{b \in \mathcal{N}_{k_\ell}} \left\langle \sum_{i \in I} a_i X_i, \sum_{i \in I^c} b_i X_i \right\rangle \\ &= \frac{1}{1 - \delta_{k_\ell}} \sup_{b \in \mathcal{N}_{k_\ell}} \sup_{a \in U_{k_\ell}} \sum_{i \in I} a_i \left\langle X_i, \sum_{i \in I^c} b_i X_i \right\rangle = \frac{1}{1 - \delta_{k_\ell}} \sup_{b \in \mathcal{N}_{k_\ell}} \sup_{a \in U_{k_\ell}} \sum_{i \in I} a_i W_{b,i} \\ &\leq \frac{1}{1 - \delta_{k_\ell}} \sup_{b \in \mathcal{N}_{k_\ell}} \sup_{a \in U_{k_\ell}} \left| \sum_{i \in I} a_i W_{b,i} \right| = \frac{1}{1 - \delta_{k_\ell}} \sup_{b \in \mathcal{N}_{k_\ell}} \left[\sum_{i=1}^{k_\ell} (W_{b,i}^*)^2 \right]^{1/2} \\ &= \frac{1}{1 - \delta_{k_\ell}} \sup_{b \in \mathcal{N}_{k_\ell}} \left[\sum_{i=1}^{k_\ell} \left(\left\langle X_i, \sum_{j \in I^c} b_j X_j \right\rangle^* \right)^2 \right]^{1/2} \end{aligned}$$

By dividing the sum into two parts and using the standard inequality $(\sqrt{a} + \sqrt{b})^2 \geq \sqrt{a+b}$, $\forall a, b \in \mathbb{R}_+$, we get:

$$\begin{aligned}
 B_{k_\ell} &\leq \frac{1}{1 - \delta_{k_\ell}} \sup_{b \in \mathcal{N}_{k_\ell}} \left[\sum_{i=1}^{k_\ell} \left(\langle X_i, \sum_{j \in I^c} b_j X_j \rangle^* \right)^2 \right]^{1/2} \\
 &= \frac{1}{1 - \delta_{k_\ell}} \sup_{b \in \mathcal{N}_{k_\ell}} \left[\sum_{i=1}^{k_{\ell+1}} \left(\langle X_i, \sum_{j \in I^c} b_j X_j \rangle^* \right)^2 + \sum_{i=k_\ell+1}^{k_\ell} \left(\langle X_i, \sum_{j \in I^c} b_j X_j \rangle^* \right)^2 \right]^{1/2} \\
 &\leq \frac{1}{1 - \delta_{k_\ell}} \sup_{b \in \mathcal{N}_{k_\ell}} \left(\left[\sum_{i=1}^{k_{\ell+1}} \left(\langle X_i, \sum_{j \in I^c} b_j X_j \rangle^* \right)^2 \right]^{1/2} + \left[\sum_{i=k_\ell+1}^{k_\ell} \left(\langle X_i, \sum_{j \in I^c} b_j X_j \rangle^* \right)^2 \right]^{1/2} \right) \\
 &= \frac{1}{1 - \delta_{k_\ell}} \left(\sup_{a \in U_{k_{\ell+1}}} \sup_{b \in \mathcal{N}_{k_\ell}} \left\langle \sum_{i \in I} a_i X_i, \sum_{i \in I^c} b_i X_i \right\rangle + \sup_{b \in \mathcal{N}_{k_\ell}} \left[\sum_{i=k_{\ell+1}+1}^{k_\ell} (W_{b,i}^*)^2 \right]^{1/2} \right)
 \end{aligned}$$

Replacing $U_{k_{\ell+1}}$ by $\mathcal{N}_{k_{\ell+1}}$ and repeating the argument used above for the first term (while reversing the roles of a and b) proves the claim. $\square$

Coming back to the proof of Lemma 9.6, we notice that we can apply Corollary 9.5. Indeed, since $|\mathcal{N}_{k_\ell}| \leq \exp \left(k_\ell \log \left(\frac{en}{k_\ell \delta_{k_\ell}} \right) \right)$, and $(X_j)_{j \in I^c}$ and $(X_i)_{i \in I}$ are independent, with probability at least

$$1 - 2 \cdot \exp \left(- \left(\frac{p}{4} - 1 \right) k_{\ell+1} \log \left(\frac{en}{k_{\ell+1}} \right) + k_\ell \log \left(\frac{en}{k_\ell \delta_{k_\ell}} \right) \right), \quad (9.2)$$

for every $b \in \mathcal{N}_{k_\ell}$ and every $a \in \mathcal{N}_{k_{\ell+1}}$ we have:

$$\left(\sum_{i=k_{\ell+1}+1}^{k_\ell} \left(\left\langle \sum_{j \in I^c} b_j X_j, X_i \right\rangle^* \right)^2 \right)^{1/2} \leq (cnk_\ell)^{1/4} A_{k_0}$$

and

$$\left(\sum_{j=k_{\ell+1}+1}^{k_\ell} \left(\left\langle \sum_{i \in I} a_i X_i, X_j \right\rangle^* \right)^2 \right)^{1/2} \leq (cnk_\ell)^{1/4} A_{k_0}$$

Now, observe that, since $\frac{p}{4} > 2$, then $\exists \gamma < 1$ such that $(\frac{p}{4} - 1) \gamma > 1$. Thus, for $p > 8$ there are γ , c_1 and c_2 that depend only on p , and for which one may take $\delta_{k_\ell} = \left(\frac{k_\ell}{n} \right)^{c_1}$, in order to satisfy:

$$(9.2) \geq 1 - 2 \exp \left(-c_2 k_{\ell+1} \log \left(\frac{en}{k_{\ell+1}} \right) \right)$$

Now, let ℓ_1 be the largest integer ℓ for which both:

- $k_\ell - k_{\ell+1} > 1$
- $\sum_{j=0}^{\ell} \exp\left(-c_2 k_{j+1} \log\left(\frac{en}{k_{j+1}}\right)\right) \leq \frac{1}{n^\beta}$

Therefore, ℓ_1 is the first integer satisfying $(\frac{p}{4} - 2)k_\ell \log\left(\frac{en}{k_\ell}\right) \leq K_2 \beta \log n$ for an appropriate choice of K_2 .

Then, the strategy is to start from B_{k_ℓ} , that we know how bound from Claim 9.7, and applying iteratively this "dimension reduction" procedure up to $\ell = \ell_1$, observing that there is a constant c_3 that depends only on p for which $\frac{1}{\prod_{\ell=0}^{\ell_1} (1 - \delta_{k_\ell})^2} \leq \frac{1}{c_3}$.

To finish, we apply the "large coordinates" estimate from Lemma 9.3 for $B_{k_{\ell_1}}$ (while observing that $k_{\ell_1} \leq k_0$), thus concluding the proof. $\square$

Remark 9.8 Note that in the proof of the previous lemma 9.6 we used that $\frac{p}{4} - 1 > 1$. This is the only point in the proof of Theorem 9.1 where the critical assumption $p > 8$ is required.

Theorem 9.9 ([34]) Let μ be an isotropic probability measure in $\mathbb{R}^d$. Let $n \geq d$ and assume that $\max_i \|X_i\|_2 \leq C_0 \cdot (nd)^{1/4}$. Let $K_1 \geq 1$ and let k_0 be the first integer which satisfies: $k_0 \log\left(\frac{en}{k_0}\right) \geq d$. Assume also that $p > 8$, $\sup_{t \in B_2^d} \|\langle t, \cdot \rangle\|_p \leq K_1$, and $1 \leq \beta \leq C_1 k_0$. Then, there are constants c_1 and c_2 , depending only on β, p, C_0 and K_1 , for which, with probability at least $1 - c_1 n^{-\beta}$:

$$A_{k_0} \leq c_2 (nd)^{1/4}$$

Proof (of Theorem 9.9) Recall that we defined $A_{k_0} = \sup_{a \in U_{k_0}} \|\sum_{i=1}^n a_i X_i\|_2$. Our scope is to find a suitable upper bound for this quantity. First, observe that for every $a \in U_{k_0}$:

$$\|\sum_{i=1}^n a_i X_i\|_2^2 = \sum_{i \neq j} a_i a_j \langle X_i, X_j \rangle + \sum_{i=1}^n a_i^2 \|X_i\|_2^2$$

Since $\|a\|_2 \leq 1$ and since by assumption we have that $\max_i \|X_i\|_2 \leq C_0 \cdot (nd)^{1/4}$, the second term in the sum is at most $\max_{i \leq n} \|X_i\|_2^2 \leq C_0^2 (nd)^{1/2}$. To bound the first term in the sum, let $(\varepsilon_i)_{i=1}^n$ be independent Bernoulli random variables. Note that:

$$\mathbb{E}_\varepsilon \sum_{i \neq j} (1 + \varepsilon_i) (1 - \varepsilon_j) a_i a_j \langle X_i, X_j \rangle = \sum_{i \neq j} a_i a_j \langle X_i, X_j \rangle$$

Therefore, we can control the expectation.

$$\begin{aligned}
 & \sup_{a \in U_{k_0}} \mathbb{E}_\varepsilon \sum_{i \neq j} (1 + \varepsilon_i) (1 - \varepsilon_j) a_i a_j \langle X_i, X_j \rangle \\
 & \leq \mathbb{E}_\varepsilon \sup_{a \in U_{k_0}} \sum_{i \neq j} (1 + \varepsilon_i) (1 - \varepsilon_j) a_i a_j \langle X_i, X_j \rangle \\
 & = \mathbb{E}_\varepsilon H((\varepsilon_i)_{i=1}^n, (X_i)_{i=1}^n),
 \end{aligned}$$

where $H((\varepsilon_i)_{i=1}^n, (X_i)_{i=1}^n) = \sup_{a \in U_{k_0}} \sum_{i \neq j} (1 + \varepsilon_i) (1 - \varepsilon_j) a_i a_j \langle X_i, X_j \rangle$.

Observe that, if we define the set $I_\varepsilon = \{i : \varepsilon_i = 1\}$, then the function $H((\varepsilon_i)_{i=1}^n, (X_i)_{i=1}^n)$ takes on the following value for every realization of $(\varepsilon_i)_{i=1}^n$:

$$H((\varepsilon_i)_{i=1}^n, (X_i)_{i=1}^n) = 4 \cdot \sup_{a \in U_{k_0}} \left\langle \sum_{i \in I_\varepsilon} a_i X_i, \sum_{j \in I_\varepsilon^c} a_j X_j \right\rangle$$

Therefore, now fix a realization $(\varepsilon_i)_{i=1}^n$. Then

$$H((\varepsilon_i)_{i=1}^n, (X_i)_{i=1}^n) \lesssim \sup_{a \in U_{k_0}} \sup_{b \in U_{k_0}} \left\langle \sum_{i \in I_\varepsilon} a_i X_i, \sum_{j \in I_\varepsilon^c} b_j X_j \right\rangle$$

This inequality suggests to apply the result we obtained with Lemma 9.6. Indeed, if $p > 8$, we can affirm:

$$H((\varepsilon_i)_{i=1}^n, (X_i)_{i=1}^n) \lesssim_p (nd)^{1/4} A_{k_0}$$

with probability at least $1 - cn^{-\beta}$.

Therefore, there exists a set $\mathcal{B}$, in the probability space defined by the probability measure μ from which the vectors X_i are sampled, of probability at least $1 - c_1 n^{-\beta/2}$ on which, with probability at least $1 - c_2 n^{-\beta/2}$ with respect to the probability space defined by the sequence of Bernoulli random variables $(\varepsilon_i)_i$:

$$H((\varepsilon_i)_{i=1}^n, (X_i)_{i=1}^n) \lesssim_p (nd)^{1/4} A_{k_0}$$

Hence, for every $(X_i)_{i=1}^n \in \mathcal{B}$, we can affirm that:

$$H((\varepsilon_i)_{i=1}^n, (X_i)_{i=1}^n) \lesssim_p \begin{cases} (nd)^{1/4} A_{k_0} & \text{w. p. } \geq 1 - c_2 n^{-\beta/2} \\ \sup_{a \in U_{k_0}} \left| \sum_{i \neq j} a_i a_j \langle X_i, X_j \rangle \right| & \text{w. p. } \leq c_2 n^{-\frac{\beta}{2}} \end{cases}$$

It follows:

$$\begin{aligned}
 & \mathbb{E}_\varepsilon H((\varepsilon_i)_{i=1}^n, (X_i)_{i=1}^n) \\
 & \lesssim_p A_{k_0} (nd)^{1/4} \cdot (1 - n^{-\beta/2}) + n^{-\beta/2} \sup_{a \in U_{k_0}} \left| \sum_{i \neq j} a_i a_j \langle X_i, X_j \rangle \right| \\
 & \lesssim_{p, C_0} A_{k_0} (nd)^{1/4} + n^{-\beta/2} \left(A_{k_0}^2 + (Nn)^{1/2} \right) \quad \text{from the definition of } A_{k_0} \text{ and Cauchy-Schwarz}
 \end{aligned}$$

Therefore, on $\mathcal{B}$, if $\beta > 0$ and n is large enough, then:

$$A_{k_0}^2 \lesssim_{p,\beta,C_0} A_{k_0} (nd)^{1/4} + (nd)^{1/2}$$

and the claim follows. $\square$

We now have all the ingredients necessary to prove the main result, Theorem 9.1.

Proof of Theorem 9.1

First, we need a final observation, which follows from an application of Lemma 9.3 to the random variables $Y_t = \langle X, t \rangle$, for vectors t in a $\frac{1}{2}$ -net in B_2^d . Indeed, under the assumptions of Theorem 9.1 there exist constants c_1, c_2 and c_3 depending only on K_1 for which, if $\mathcal{N}$ is a $\frac{1}{2}$ -net in B_2^d , with probability at least $1 - 2 \exp(-c_1 d)$:

$$\sup_{t \in \mathcal{N}} \left(\sum_{i=c_3 k_0+1}^n (\langle X_i, t \rangle^*)^4 \right)^{1/2} \leq c_2 \sqrt{n} \quad (9.3)$$

This follows from the second of the two claim we previously proved in Lemma 9.3.

Now, let $\mathcal{C}$ be the intersection of the events from the observation (9.3) we just made and Theorem 9.9.

We prove the following claim.

Claim 9.10 *On $\mathcal{C}$, with probability at least $1 - 2 \exp(-c_1 d)$, we have:*

$$\sup_{t \in B_2^d} \left| \sum_{i=1}^d \varepsilon_i \langle X_i, t \rangle^2 \right| \lesssim C_0, p \sqrt{nd}$$

Proof (of Claim 9.10) Let c_3 be the constant we defined above.

Fix $t, t' \in \mathcal{N}$ and let J be the union of the sets of the largest $c_3 k_0$ coordinates of $(|\langle X_i, t \rangle|)_{i=1}^n$ and $(|\langle X_i, t' \rangle|)_{i=1}^n$.

Then, for every $v > 0$, with probability at least $1 - 2 \exp(-c_4 v^2)$:

9. CASE STUDY: RANDOM VECTORS WITH BOUNDED L_p MARGINAL MOMENTS

$$\begin{aligned}
\left| \sum_{i=1}^n \varepsilon_i \langle X_i, t \rangle \langle X_i, t' \rangle \right| &\leq \sum_{i \in J} |1 \cdot \langle X_i, t \rangle \langle X_i, t' \rangle| + \sum_{i \in J^c} |\varepsilon_i \langle X_i, t \rangle \langle X_i, t' \rangle| && \text{by the triangle inequality and since } \varepsilon_i \leq 1 \\
&\lesssim \sum_{i \in J} |\langle X_i, t \rangle \langle X_i, t' \rangle| + v \left(\sum_{i \in J^c} \langle X_i, t \rangle^2 \langle X_i, t' \rangle^2 \right)^{1/2} && \text{by Hoeffding's inequality:} \\
&\leq 2c_3 \left(\sum_{i=1}^{k_0} (\langle X_i, t \rangle^*)^2 \right)^{1/2} \left(\sum_{i=1}^{k_0} (\langle X_i, t' \rangle^*)^2 \right)^{1/2} && \forall v > 0: \mathbb{P} \left(\frac{\sum_i a_i \varepsilon_i}{(\sum_i a_i^2)^{1/2}} \geq v \right) \leq \exp(-c_4 v^2) \\
&\quad + v \left(\sum_{i=c_3 k_0+1}^n (\langle X_i, t \rangle^*)^4 \right)^{1/4} \left(\sum_{i=c_3 k_0+1}^n (\langle X_i, t' \rangle^*)^4 \right)^{1/4} && \text{from Cauchy-Schwarz inequality and definition of } J, J^c \\
&\lesssim A_{k_0}^2 + v\sqrt{n} && \text{from the definition of } A_{k_0} \text{ and by the observation (9.3)}
\end{aligned}$$

To continue, let $v \sim \sqrt{d}$. Then, from the previous computation and Theorem 9.9, we have:

$$\left| \sum_{i=1}^n \varepsilon_i \langle X_i, t \rangle \langle X_i, t' \rangle \right| \lesssim (nd)^{1/2} \tag{9.4}$$

Since, by Lemma 2.15 an $\frac{1}{2}$ -net has cardinality at most $|\mathcal{N}| \leq 5^d$, there is a set $\mathcal{D} \subset \{-1, 1\}^n$ of probability, with respect to the probability space defined by the Bernoulli's random variables, at least $1 - 2 \exp(-c_5 n)$ on which the inequality (9.4) holds for any pair t, t' taken from $\mathcal{N} \times \mathcal{N}$. Since each $t \in B_2^d$ can be written as $\sum_{i=1}^\infty \beta_i t_i$ with $0 \leq \beta_i \lesssim 2^{-i}$ and $t_i \in \mathcal{N}$, on $\mathcal{D}$ we have:

$$\sup_{t \in B_2^d} \left| \sum_{i=1}^n \varepsilon_i \langle X_i, t \rangle^2 \right| \lesssim (nd)^{1/2} \sum_{i,j=1}^\infty 2^{-i} 2^{-j} \lesssim (nd)^{1/2},$$

with constants that depend on k_0, C_0, p and β . □

Finally, to conclude the proof of Theorem 9.1, we make the following observation, using a symmetrization argument.

Observe that, for instance:

$$\begin{aligned}
 \mathbb{E}_X \sup_{t \in B_2^d} \left| \sum_{i=1}^n (\langle X_i, t \rangle^2 - 1) \right| &= \mathbb{E}_X \left\| \sum_{i=1}^n (X_i X_i^T - \mathbb{E} X_i X_i^T) \right\| \\
 &= \mathbb{E}_X \left\| \sum_{i=1}^n (X_i X_i^T - \mathbb{E}_Y Y_i Y_i^T) \right\| && \text{where } Y_i \text{ are} \\
 &&& \text{independent copies of } X_i \\
 &\leq \mathbb{E}_X \mathbb{E}_Y \left\| \sum_{i=1}^n (X_i X_i^T - Y_i Y_i^T) \right\| && \text{from Jensen's} \\
 &&& \text{inequality} \\
 &= \mathbb{E} \left\| \sum_{i=1}^n \varepsilon_i (X_i X_i^T - Y_i Y_i^T) \right\| && \text{where the } \varepsilon_i \text{ are independent} \\
 &&& \text{Bernoulli's random variables} \\
 &\leq 2 \mathbb{E} \left\| \sum_{i=1}^n \varepsilon_i X_i X_i^T \right\| = 2 \mathbb{E} \sup_{t \in B_2^d} \left| \sum_{i=1}^n \varepsilon_i \langle X_i, t \rangle^2 \right|
 \end{aligned}$$

Therefore, we can use the bound we found above with Claim 9.10 and dividing both sides of the inequality by n gives us the result of Theorem 9.1.

Conclusion

In this dissertation, we explored and examined non-asymptotic guarantees for sample covariance matrices. In most cases, the optimal sample size sufficient to obtain a ε -accuracy for approximating the true covariance matrix using the sample covariance matrix estimator was $n = \mathcal{O}(d)$. Sometimes, we have been even more precise and also showed the optimal rate of convergence of this approximation, which is $\sqrt{\frac{d}{n}}$, achieving a Bai–Yin type estimate for the largest singular value. We began our analysis with Gaussians and gradually reduced our assumptions on the underlying distribution. In the final chapter, our only assumption was the existence of $p > 8$ marginal moments.

There are other studies aligning with this perspective, with the most notable being K. Tikhomirov’s work, which assumes only the existence of $p > 4$ moments [43] to achieve ε -accuracy and also the optimal rate of convergence of the Bai–Yin type estimate mentioned above. In [43], ε -accuracy in the approximation of the covariance matrix is achieved also for $p > 2$ but not the Bai–Yin type estimate is not estimated. Nevertheless, it’s widely anticipated within the community that similar results might be achievable under even less stringent assumptions, such as the existence of $p > 2$ moments.

It’s also worth mentioning that there are numerous other research avenues in the field of covariance estimation that we chose not to address in this manuscript. As outlined by Chen, Gao, and Ren [17], the presence of even a single outlier in the sample can compromise the statistical performance of the sample covariance matrix. This has sparked considerable interest in robust covariance matrix estimators that can withstand adversarial contamination of the data. For examples of such studies, please refer to [19], [1], [36], and [37].

Beyond the sample covariance, several alternative estimators have been proposed for various applications. The sample covariance Σ_n is a natural esti-

mator for the population covariance Σ and it has a number of advantages: it is simple to construct, it's unbiased, and, under normality, it's the maximum likelihood estimator. However, it has notable limitations. Particularly in high-dimensional scenarios, where the number of variables d exceeds the number of observations n , Σ_n it is not invertible, even if the true underlying covariance matrix is.

Over time, new methods addressing these challenges have emerged, often tailored to specific fields like finance. Broadly, these techniques fall into four categories:

- **Shrinkage Method:** The shrinkage estimator is a linear combination of the sample estimator and another estimator (usually it is one implied by the Capital Asset Pricing Model theory). The challenge here is determining the optimal shrinkage. For a review of such method, you can refer to the guide from Ledoit and Wolf [31].
- **Factor Models:** This approach considers covariance matrices derived from large dimensional factor models, which can be observable or latent. Principal component and maximum likelihood methods are commonly used in latent models. For an introduction, refer to [23].
- **Bayesian and Empirical Bayes Estimators:** These estimators, related to the shrinkage estimator, offer alternative interpretations for the shrinkage method. For one example of such estimators, refer to [52].
- **Random Matrix Theory-Based Method:** This method attempts to mitigate the randomness of the sample covariance S by leveraging high-dimensional random matrix theory. For instance, look at [33].

In conclusion, the field of covariance is extremely large and in continuous evolution. There are several areas and open problems that still need to be explored.

Appendix A

Appendix

A.1 Basic Inequalities

Theorem A.1 *The binomial coefficient $\binom{n}{k}$ is the number of ways to choose k things from a set of n things and equals $\frac{n!}{k!(n-k)!}$. For positive integers n and k with $1 \leq k \leq n$, we have:*

$$\left(\frac{n}{k}\right)^k \leq \binom{n}{k} \leq \left(\frac{en}{k}\right)^k$$

Theorem A.2 *(Arithmetic-Geometric means inequality) For any real numbers $x_1, x_2, \dots, x_n \geq 0$:*

$$\frac{x_1 + x_2 + \dots + x_n}{n} \geq \sqrt[n]{x_1 x_2 \dots x_n},$$

with equality if and only if $x_1 = x_2 = \dots = x_n$.

A.2 Basic Probability Theory

Theorem A.3 *(Markov's Inequality [49]) Let $X \geq 0$ be a random variable. Then, for every $t > 0$:*

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}X}{t}$$

Corollary A.4 *(Chebyshev's Inequality [49]) Let $X \geq 0$ be a random variable. Then, for every $t > 0$:*

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X^p]}{t^p}$$

Theorem A.5 (*Integral identity/Layer cake representation [49]*) Let X be a random variable. Then:

$$\mathbb{E}X = \int_0^\infty \mathbb{P}(X > t)dt + \int_{-\infty}^0 \mathbb{P}(X < t)dt$$

Theorem A.6 (*Jensen Inequality [49]*). Let X be a random variable and $\varphi : \mathbb{R} \rightarrow \mathbb{R}$ be a convex function. Then:

$$\varphi(\mathbb{E}X) \leq \mathbb{E}\varphi(X)$$

Theorem A.7 (*Hölder's inequality [49]*) Let $p, q > 1$ such that they satisfy $\frac{1}{p} + \frac{1}{q} = 1$. Let X and Y be two random variables. Then:

$$|\mathbb{E}[XY]| \leq [\mathbb{E}|X|^p]^{1/p} [\mathbb{E}|Y|^q]^{1/q}$$

Theorem A.8 (*Cauchy-Schwarz Inequality [49]*) For any two random variables X and Y , we have

$$|\mathbb{E}XY| \leq \sqrt{\mathbb{E}[X^2] \mathbb{E}[Y^2]},$$

where equality holds if and only if $X = \alpha Y$, for some constant $\alpha \in \mathbb{R}$.

Bibliography

- [1] Pedro Abdalla and Nikita Zhivotovskiy. Covariance estimation: Optimal dimension-free guarantees for adversarial corruption and heavy tails. *ArXiv*, abs/2205.08494, 2022.
- [2] Radosław Adamczak, Alexander E. Litvak, Alain Pajor, and Nicole Tomczak-Jaegermann. Quantitative estimates of the convergence of the empirical covariance matrix in log-concave ensembles. *Journal of the American Mathematical Society*, 23:535–561, 2009.
- [3] Radosław Adamczak, Alexander E. Litvak, Alain Pajor, and Nicole Tomczak-Jaegermann. Sharp bounds on the rate of convergence of the empirical covariance matrix. *arXiv: Probability*, 2010.
- [4] Greg W. Anderson, Alice Guionnet, and Ofer Zeitouni. An introduction to random matrices. 2009.
- [5] Guillaume Aubrun. Sampling convex bodies: a random matrix approach. 2007.
- [6] Zhidong Bai and Jack W. Silverstein. Spectral analysis of large dimensional random matrices. 2009.
- [7] Zhidong Bai and Y. Q. Yin. Limit of the smallest eigenvalue of a large dimensional sample covariance matrix. *Annals of Probability*, 21:1275–1294, 1993.
- [8] Jinho Baik, Percy Deift, and Toufic M. Suidan. Combinatorics and random matrix theory. 2016.
- [9] Keith Ball and I. Perissinaki. The subindependence of coordinate slabs in ℓ_p balls. *Israel Journal of Mathematics*, 107:289–299, 1998.

- [10] Afonso Bandeira, Amit Singer, and Thomas Strohmer. Mathematics of data science. 2020.
- [11] Peter J. Bickel and Elizaveta Levina. Covariance regularization by thresholding. *arXiv: Statistics Theory*, 2009.
- [12] Sergey G. Bobkov and Fedor Nazarov. Large deviations of typical linear functionals on a convex body with unconditional basis. 2003.
- [13] Sergey G. Bobkov and Fedor Nazarov. On convex bodies and log-concave probability measures with unconditional basis. 2003.
- [14] Christer Borell. The brunn-minkowski inequality in gauss space. *Inventiones mathematicae*, 30:207–216, 1975.
- [15] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. Concentration inequalities - a nonasymptotic theory of independence. In *Concentration Inequalities*, 2013.
- [16] Silouanos Brazitikos. Geometry of isotropic convex bodies. 2014.
- [17] Mengjie Chen, Chao Gao, and Zhao Ren. Robust covariance and scatter matrix estimation under huber’s contamination model. *The Annals of Statistics*, 2015.
- [18] Richard Y. Chen, Alex Gittens, and Joel A. Tropp. The masked sample covariance estimator: An analysis via matrix concentration inequalities. *Information and Inference: A Journal of the IMA*, 1:2–20, 2011.
- [19] Yu Cheng, Ilias Diakonikolas, Rong Ge, and David P. Woodruff. Faster algorithms for high-dimensional robust covariance estimation. In *Annual Conference Computational Learning Theory*, 2019.
- [20] Kenneth R. Davidson and Stanislaw J. Szarek. Chapter 8 - local operator theory, random matrices and banach spaces. 2001.
- [21] Percy Deift. Orthogonal polynomials and random matrices: A riemann-hilbert approach. 2000.
- [22] Percy Deift and Dimitri Gioev. Random matrix theory: Invariant ensembles and universality. 2009.
- [23] Jianqing Fan, Yuan Liao, and Martina Mincheva. High dimensional covariance matrix estimation in approximate factor models. *Capital Markets: Asset Pricing & Valuation eJournal*, 2011.

-
- [24] S. Yu. Favorov. A generalized kahane-khinchin inequality. *Studia Mathematica*, 130:101–107, 1998.
- [25] Matthieu Fradelizi. Sections of convex bodies through their centroid. *Archiv der Mathematik*, 69:515–522, 1997.
- [26] David A. Freedman. Statistical models: Theory and practice. *Technometrics*, 48:315 – 315, 2005.
- [27] Ian T. Jolliffe and Jorge Cadima. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374, 2016.
- [28] Ravi Kannan, László Miklós Lovász, and Miklós Simonovits. Random walks and an $O(n^5)$ volume algorithm for convex bodies. *Random Structures and Algorithms*, 11:1–50, 1997.
- [29] Vladimir Koltchinskii and Karim Lounici. Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli*, 23:110–133, 2014.
- [30] Rafal Latała. Estimation of moments of sums of independent real random variables. *Annals of Probability*, 25:1502–1513, 1997.
- [31] Olivier Ledoit and Michael Wolf. The power of (non-)linear shrinking: A review and guide to covariance matrix estimation. *Capital Markets: Asset Pricing & Valuation eJournal*, 2020.
- [32] Elliott H. Lieb. Convex trace functions and the wigner-yanase-dyson conjecture. *Advances in Mathematics*, 11:267–288, 1973.
- [33] Tongshu Ma and Ravi Jagannathan. Risk reduction in large portfolios: Why imposing the wrong constraints helps. *Capital Markets: Asset Pricing & Valuation*, 2002.
- [34] Shahar Mendelson and Grigoris Paouris. On the singular values of random matrices. *Journal of the European Mathematical Society*, 16:823–834, 2014.
- [35] V. D. Milman and Alain Pajor. Isotropic position and inertia ellipsoids and zonoids of the unit ball of a normed n -dimensional space. 1989.
- [36] Arshak Minasyan and Nikita Zhivotovskiy. Statistically optimal robust mean and covariance estimation for anisotropic gaussians. *ArXiv*, abs/2301.09024, 2023.

- [37] Stanislav Minsker and Lan Wang. Robust estimation of covariance matrices: Adversarial contamination and beyond. *Statistica Sinica*, 2022.
- [38] Grigoris Paouris. Concentration of mass and central limit properties of isotropic convex bodies. 2004.
- [39] Grigoris Paouris. Concentration of mass on convex bodies. *Geometric & Functional Analysis GAFA*, 16:1021–1049, 2006.
- [40] Zeév Rudnick and Peter Sarnak. Zeros of principal ζ -functions and random matrix theory. *Duke Mathematical Journal*, 81:269–322, 1996.
- [41] Rolf Schneider. Convex bodies: The brunn-minkowski theory. 1993.
- [42] Mikhail A. Stephanov, Jacobus J. M. Verbaarschot, and Tilo Wettig. Random matrices. 2005.
- [43] Konstantin E. Tikhomirov. Sample covariance matrices of heavy-tailed distributions. *arXiv: Probability*, 2016.
- [44] Joel Tropp. Probability in high dimensions. 2021.
- [45] Joel A. Tropp. An introduction to matrix concentration inequalities. *ArXiv*, abs/1501.01571, 2015.
- [46] Ramon van Handel. Probability in high dimension. 2014.
- [47] Ramon van Handel. Structured random matrices. *arXiv: Probability*, pages 107–156, 2016.
- [48] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing*, 2010.
- [49] Roman Vershynin. High-dimensional probability. 2018.
- [50] Eugene Paul Wigner and Paul Adrien Maurice Dirac. On the statistical distribution of the widths and spacings of nuclear resonance levels. *Mathematical Proceedings of the Cambridge Philosophical Society*, 47:790 – 798, 1951.
- [51] John Wishart. The generalised product moment distribution in samples from a normal multivariate population. *Biometrika*, pages 32–52, 1928.
- [52] Ruoyong Yang and James O. Berger. Estimation of a covariance matrix using the reference prior. *Annals of Statistics*, 22:1195–1211, 1994.



Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Declaration of originality

The signed declaration of originality is a component of every semester paper, Bachelor's thesis, Master's thesis and any other degree paper undertaken during the course of studies, including the respective electronic versions.

Lecturers may also require a declaration of originality for other written papers compiled for their courses.

I hereby confirm that I am the sole author of the written work here enclosed and that I have compiled it in my own words. Parts excepted are corrections of form and content by the supervisor.

Title of work (in block letters):

Authored by (in block letters):

For papers written by groups the names of all authors are required.

Name(s):

First name(s):

With my signature I confirm that

- I have committed none of the forms of plagiarism described in the '[Citation etiquette](#)' information sheet.
- I have documented all methods, data and processes truthfully.
- I have not manipulated any data.
- I have mentioned all persons who were significant facilitators of the work.

I am aware that the work may be screened electronically for plagiarism.

Place, date

Signature(s)

For papers written by groups the names of all authors are required. Their signatures collectively guarantee the entire content of the written paper.