Pontifícia Universidade Católica
do Rio de Janeiro

Pedro Abdalla Teixeira

Non-Asymptotic Random Matrix Theory and The Small Ball Method

Dissertação de Mestrado

Thesis presented to the Programa de Pós-graduação em Matemática da PUC-Rio in partial fulfillment of the requirements for the degree of Mestre em Matemática.

Advisor: Prof. Carlos Tomei

Rio de Janeiro
March 2020

Pontifícia Universidade Católica
do Rio de Janeiro

Pedro Abdalla Teixeira

Non-Asymptotic Random Matrix Theory and The Small Ball Method

Thesis presented to the Programa de Pós-graduação em Matemática da PUC-Rio in partial fulfillment of the requirements for the degree of Mestre em Matemática. Approved by the Examination Committee.

Prof. Carlos Tomei
Advisor
Departamento de Matemática – PUC-Rio

Prof. Carlos Hugo Jimenez
Departamento de Matemática – PUC-Rio

Prof. Edgard Pimentel
Departamento de Matemática – PUC-Rio

Prof. Hugo Carvalho
Universidade Federal do Rio de Janeiro – UFRJ

Prof. Paulo Orenstein
Instituto de Matemática Pura e Aplicada – IMPA

Prof. Roberto Oliveira
Instituto de Matemática Pura e Aplicada – IMPA

Rio de Janeiro, March 6th, 2020

Pedro Abdalla Teixeira

Majored in Mathematics and Electrical Engineering by PUC-Rio.

Bibliographic data

Teixeira,Pedro Abdalla

Non-Asymptotic Random Matrix Theory and The Small Ball Method / Pedro Abdalla Teixeira; advisor: Carlos Tomei. – Rio de janeiro: PUC-Rio , Departamento de Matemática, 2020.

v., 86 f: il. color. ; 30 cm

Dissertação (mestrado) - Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Matemática.

Inclui bibliografia

1. Matemática – Teses. 2. Matrizes Aleatórias;. 3. Valores Singulares;. 4. Recuperação Convexa;. 5. Método da Bola Pequena.. I. Tomei, Carlos. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Matemática. III. Título.

CDD: 004

To my family for their support.
To my friends, in special my Neide's friends.
To my mother, for me she will always be my number one.
To Raquel A. Amaro, you inspire me.

Acknowledgments

I would like to thank my advisor Professor Carlos Tomei for the stimulus and support during this work.

This study was financed in part by the coordenação de aperfeiçoamento de nível superior- Brasil (CAPES)- Finance Code 001.

I would like to thank you FAPERJ for the financial support.

I also would like to thank the Math Department of PUC-Rio: professors, colleagues and employees.

Abstract

Teixeira,Pedro Abdalla; Tomei, Carlos (Advisor). **Non-Asymptotic Random Matrix Theory and The Small Ball Method**. Rio de Janeiro, 2020. 86p. Dissertação de Mestrado – Departamento de Matemática, Pontifícia Universidade Católica do Rio de Janeiro.

Motivated by problems in the field of signal recovery by convex programming, the aim of this work is to provide a careful analysis of the celebrated small ball method and its connections with the non-asymptotic theory of random matrices. In particular, the study of the conic singular values of random matrices will play a key role to analyze such problems.

Keywords

Random Matrices; Singular Values; Convex Recovery; Small Ball Method.

Resumo

Teixeira,Pedro Abdalla; Tomei, Carlos. **Teoria não assintótica de matrizes aleatórias e o método da bola pequena** . Rio de Janeiro, 2020. 86p. Dissertação de Mestrado – Departamento de Matemática, Pontifícia Universidade Católica do Rio de Janeiro.

Motivado por problemas no campo da recuperação de sinais por programação convexa, o objetivo deste trabalho é fornecer uma análise precisa do método das bola pequena e suas conexões com a teoria não assintótica das matrizes aleatórias. Em particular, o estudo dos valores singulares cônicos de matrizes aleatórias desempenhará um papel fundamental na análise de tais problemas.

Palavras-chave

Matrizes Aleatórias; Valores Singulares; Recuperação Convexa; Método da Bola Pequena.

Table of contents

"If people do not believe that mathematics is simple, it is only because they do not realize how complicated life is"

John Von Neumann, .

1
Introduction

1.1
Asymptotic x Non-asymptotic Random Matrix Theory

Random matrices were introduced by Wishart in 1928 in the field of statistics [69]. Their popularity increased with the remarkable work of the physicist E.P.Wigner, who used random matrices to describe atomic levels in nuclear energy [68]. Since then, the asymptotic theory of random matrices has been connected to many different fields of mathematics such as analytic number theory [46], integrable systems [5], combinatorics [4], operator theory and orthogonal polynomials [17], mathematical physics and dynamical systems [1] among others. For an excellent introductory text we refer the reader to [40].

Asymptotic random matrix theory considers the asymptotic properties of random matrices in a parameter, frequently the increasing size of the matrix. Here is one example.

Theorem 1.1 *(Bai-Yin's law [3]) Let A be an $m \times n$ random matrix with entries given by independent copies of a distribution with zero mean, unit variance and finite fourth moment. Let $s_1(A) \geq s_n(A)$ be the largest and smallest singular values of A. As $n/m \to c \in [0,1]$,*

$$\frac{s_1(A)}{\sqrt{m}} \to 1+\sqrt{c}, \ \ \frac{s_n(A)}{\sqrt{m}} \to 1-\sqrt{c} \ \textit{almost surely.}$$

The theorem does not provide information about matrices with given dimensions $n, m < \infty$ or about the convergence rate. In contrast, consider the following non-asymptotic result.

Theorem 1.2 *(Gaussian deviation [16]). Let A be an $m \times n$ random matrix with independent entries satisfying the Gaussian $N(0,1)$ distribution. Then, for every $t \geq 0$, with probability at least $1-2e^{-t^2}$,*

$$\sqrt{m}-\sqrt{n}-t \leq s_n(A) \leq s_1(A) \leq \sqrt{m}+\sqrt{n}+t \ .$$

This is a quantitative information about the extrema of the set of singular values. On the other hand, we lost generality in the specification of the entries, as well as the the limit of such extrema, as in Theorem 1.1.

The non-asymptotic theory of random matrices relates to the geometry of Banach spaces and local operator theory [16], combinatorics [58], [15], numerical linear algebra [59], compressed sensing [22], mathematical statistics and data science [63], [65] and [12]. For excellent introductory texts in the so called high dimensional probability, see [63] and [9].

1.2 The Small Ball Method

The small ball method, developed by S. Mendelson and V. Koltchinskii, is a strategy to provide lower bounds of an empirical process under general assumptions [31]. The method originally was used on a problem related to s_n, the smallest singular value of a random matrix. Subsequent applications mainly by S. Mendelson and his co-authors consider a large variety of issues in statistical learning theory and sparse recovery of data [42], [43], [33], [34], compressed sensing [30] and the geometry of Banach spaces [41].

We point out three difficulties in non-asymptotic random matrix theory. Techniques associated to the phenomenon of concentration of measure are rarely valid for **heavy tail** distributions. The small ball method circumvents the use of concentration.

The second is related to **geometric constraints**. As we shall see in many sparse recovery problems, the error of an approximation can be quantified from lower bounds for the smallest conic singular value: given a cone $\mathcal{K}$, the smallest conic singular value of a matrix $A \in \mathbb{R}^{m\times n}$ is

$$s_n(A,\mathcal{K}) = \inf_{x\in\mathcal{K}\cap S^{n-1}} \|Ax\|_2 \ .$$

When $\mathcal{K}$ is the full space, $s_n(A,\mathcal{K}) = s_n(A)$. The small ball method is very useful in situations involving arbitrary cones, as are some techniques from stochastic process theory.

Moreover, the small ball method relies heavily on the **independence** of the underlying matrix rows.

1.3 Scope and content of the dissertation

We consider a general framework for recovery problems by means of convex optimization, following the use of the small ball method to the estimates for extremal singular values introduced by J.Tropp in [60].

Chapter 2 develops estimates for the usual (classical) singular values of random matrices with entries satisfying sub-gaussian distributions. Chapter 3 highlights unique features of Gaussian matrices. With such tools, in Chapter 4, we provide a rigorous treatment of a large class of recovery problems using the small ball method. Finally in Chapter 5 we apply the small ball method to the classical phase retrieval problem.

1.4 Basic notation

For a probability space $(\Omega, \Sigma, \mathbb{P})$ and a random variable $X : \Omega \to \mathbb{R}$, we denote the expectation (mean) by $\mathbb{E}X$ and the variance by $Var(X)$. All sets in this dissertation will be assumed to be measurable.

Let $L^p = L^p(\Omega, \Sigma, \mathbb{P})$ be the standard L^p spaces 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 $\mathbb{P}$.

Random vectors $\mathbf{X} = (X_1, \ldots, X_n) \in \mathbb{R}^n$ have mean $\mu = (\mathbb{E}X_1, \ldots, \mathbb{E}X_n)$. A standard Gaussian vector $\mathbf{g} = (g_1, \ldots, g_k) \in \mathbb{R}^k$ has entries given by independent standard normal variables $g_i \sim N(0, 1)$.

More generally, a Gaussian vector is a random vector $\mathbf{X} \in \mathbb{R}^n$ for which there exists a matrix $\Gamma \in \mathbb{R}^{n \times k}$ such that

$$\mathbf{X} = \Gamma \mathbf{g} + \mu,$$

where μ is the mean of $\mathbf{X}$ and $\mathbf{g}$ is a standard Gaussian vector.

We denote by $\|\mathbf{X}\|_2$ the standard Euclidean norm in $\mathbb{R}^n$ associated to the standard inner product $\langle ., . \rangle$.

Endow $\mathbb{R}^n$ and $\mathbb{R}^m$ with the ℓ_p and ℓ_q norms. Let $B_p^n \subset \mathbb{R}^n$ be the unit ball in $\mathbb{R}^n$ with the ℓ_p norm and set

$$\|A\|_{p \to q} = \sup_{x \in B_p^n} \|Ax\|_q.$$

The frequent case $\|A\|_{2 \to 2}$ will be denoted by $\|A\|$. We refer to this norm as operator norm.

Let f, g be functions assuming values in some common domain.

1. (Big $\mathcal{O}$): $f = \mathcal{O}(g)$ if exists $\alpha > 0$ such that $|f(x)| \leq \alpha |g(x)|$.

2. (Big Θ): $f = \Theta(g)$ if $f = \mathcal{O}(g)$ and $g = \mathcal{O}(f)$.

We say that f, g are equivalent if $f = \Theta(g)$.

Constants may change along a calculation: in this case, we replace C by other letters (as C_1). If the constant does not depend on any other parameter, we refer to it as an absolute constant. For a natural number n, denote $[n] = \{1, 2, \ldots, n\}$. For a set P, denote $\mathbb{I}_P(x)$ by its indicator function.

2
Concentration Inequalities in the Sub-gaussian Context

2.1
The Sub-gaussian Distributions

We start with a simple goal. We search for bounds that inform how a random variable X concentrates near to the mean, in other words, an expression of the form

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

where $\varepsilon(t)$ is a error function that converge fast, possibly exponentially, to zero. A systematic approach to this kind of concentration inequalities can be found in [9]. We consider the case when X is the sum of random variables X_i. A toy example is the following situation [63].

Question 2.1 *Toss a fair coin N times. What is the probability of obtaining at least $\frac{3}{4}N$ heads?*

Let S_N denote the number of heads. Clearly, S_N has mean $\frac{N}{2}$ and variance $\frac{N}{4}$. By the Chebyshev inequality (see the Appendix),

$$\mathbb{P}(S_N \geq \frac{3}{4}N) \ \leq \ \mathbb{P}(|S_N - \frac{N}{2}| \geq \frac{N}{4}) \ \leq \ \frac{4}{N}.$$

In this example, the decay in N is linear. If instead S_N satisfied a Gaussian distribution, the decay would be much faster.

Proposition 2.2 *(Tails for the Standard Gaussian Distribution). Let g be a standard Gaussian variable. Then, for every $t > 0$,*

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

Proof. By symmetry,

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

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

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

■

The standard Gaussian distribution concentrates heavily around the mean. An analogous result holds for Gaussian random variables. Since the sum of independent Gaussians is Gaussian, the sum concentrates equally well.

The celebrated central limit theorem is frequently used to obtain concentration bounds. Since we are interested in non-asymptotic results, we need a quantitative version of it.

Theorem 2.3 *(Berry-Esseen central limit Theorem [18]). Let X_1,X_2 ... be a sequence of i.i.d random variables with mean μ and variance σ^2. Consider S_N, the sum of the first N random variables X_i and the normalized sum*

$$Z_N = \frac{S_N - \mathbb{E}S_N}{\sqrt{Var(S_N)}}.$$

Then, for every N and for every $t \in \mathbb{R}$,

$$|\mathbb{P}(Z_N \geq t) - \mathbb{P}(g \geq t)| \leq \frac{\rho}{\sqrt{N}}.$$

Here $\rho = \mathbb{E}[|X_1 - \mu|^3]/\sigma^3$ and $g \sim N(0,1)$.

We show the sharpness of the theorem in an example. For even N, the probability of getting exactly $N/2$ heads is, by the standard Stirling approximation,

$$\mathbb{P}(S_N = \frac{N}{2}) \;=\; 2^{-N}\binom{N}{N/2} \;=\; \Theta(\frac{1}{\sqrt{N}}).$$

Hence $\mathbb{P}(Z_N = 0) = \Theta(\frac{1}{\sqrt{N}})$. On the other hand, $\mathbb{P}(g = 0) = 0$, so that the error must be of order $\Theta(\frac{1}{\sqrt{N}})$.

The bound $\mathcal{O}(1/\sqrt{N})$ is far from our desired exponential decay, so this theorem actually will be of little use. For the best known constant ρ, see [48].

Inspired by the strong concentration of the Gaussian variable, we restrict our attention to a special class of random variables that possess a fast decay. We list some properties of such variables: the proofs follow [63].

Proposition 2.4 *(Sub-gaussian Properties) 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

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

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. I $\Rightarrow$ II. Suppose property I holds. By homogeneity, we may assume $K_1 = 1$. Then

$$\begin{aligned}
\mathbb{E}|X|^p &= \int_0^\infty \mathbb{P}(|X|^p \geq u)du \text{ (integral layer cake identity, see the Appendix)} \\
&= \int_0^\infty \mathbb{P}(|X| \geq t)pt^{p-1}dt \text{ (change variables } u = t^p) \\
&\leq \int_0^\infty 2e^{-t^2}pt^{p-1}dt \text{ (property I)} \\
&= p\Gamma(p/2) \text{ (set } t^2 = s \text{ in the expression for the Gamma function)} \\
&\leq p(p/2)^{p/2} \text{ (as } \Gamma(x) \leq x^x).
\end{aligned}$$

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.

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

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

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

we thus obtain III for $K_3 = \sqrt{2e}$.

The implication III $\Rightarrow$ IV is straightforward, take λ sufficient small such that $e^{K_3^2 \lambda^2} \leq 2$. IV $\Rightarrow$ I. By homogeneity assume $K_4 = 1$. Then

$$\begin{aligned}
\mathbb{P}(|X| \geq t) &= \mathbb{P}(e^{X^2} \geq et^2) \\
&\leq e^{-t^2}\mathbb{E}[e^{X^2}] \text{ (by the Markov Inequality, see the Appendix)} \\
&\leq 2e^{-t^2} \text{ (by property IV).}
\end{aligned}$$

We obtain I with $K_1 = 1$.

We now prove III $\Rightarrow$ V and then V $\Rightarrow$ I.Assume III holds with $K_3 = 1$. Since $e^x \leq x + e^{x^2}$ for all real x,

$$\begin{aligned}
\mathbb{E}e^{\lambda X} &\leq \mathbb{E}[\lambda X + e^{\lambda^2 X^2}] \\
&= \mathbb{E}e^{\lambda^2 x^2} \text{ (since we assume } \mathbb{E}X = 0) \\
&\leq e^{\lambda^2} \text{ , (If } |\lambda| \leq 1).
\end{aligned}$$

We get the desired bound for $|\lambda| \leq 1$. 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 $K_5 = 1$.

For V $\Rightarrow$ I, we use a smart trick, the so called Laplace transform method.

Let $\lambda > 0$ be a parameter to be chosen later. Write

$$\begin{aligned}\mathbb{P}(X \geq t) &= \mathbb{P}(e^{\lambda X} \geq e^{\lambda t}) \\ &\leq e^{-\lambda t}\mathbb{E}e^{\lambda t} \text{ (by the Markov inequality)} \\ &\leq e^{-\lambda t}e^{\lambda^2} \text{ (by property V)} \\ &= e^{-\lambda t+\lambda^2} \\ &= e^{-t^2/4} \text{ (Minimizing in } \lambda\text{: } \lambda = \frac{t}{2}).\end{aligned}$$

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 $K_1 = 2$. ■

The argument in the proof of I $\Rightarrow$ II will be used repeatedly and will be referred to as the integral identity.

Definition 2.5 *(Sub-gaussian random variables, subgaussian norm) Any random variable that satisfies properties I-IV is called sub-gaussian random variables. The sub-gaussian norm* $\|X\|_{\psi_2}$ *is*

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

The reader may check that $\|X\|_{\psi_2}$ is indeed a norm. From property IV, X is sub-gaussian if and only if $\|X\|_{\psi_2}$ is finite. The sum of sub-gaussian random variables is also sub-gaussian.

Remark 2.6 The number 2 in the definitions can be replaced by another constant. It arises naturally in many scenarios when there is a common bound for X and $-X$, as in the proof of property V.

We rewrite Proposition 2.4 in the notation of sub-gaussian norms.

$$\mathbb{P}(|X| \geq t) \leq 2e^{-t^2/\|X\|_{\psi_2}^2}, \tag{2-3}$$

$$\|X\|_p \leq C\|X\|_{\psi_2}\sqrt{p} \quad \forall p \geq 1. \tag{2-4}$$

If X has zero mean,

$$\mathbb{E}e^{\lambda X} \leq e^{C\lambda^2\|X\|_{\psi_2}^2} \quad \forall \lambda \in \mathbb{R}, \tag{2-5}$$

where $C > 0$ is an absolute constant. The space of sub-gaussian random variables is a Banach space because of property 2-4.

Example 2.7 Some classical examples of sub-gaussian random variables:

1. - **(Gaussians)** If $X \sim N(0, \sigma^2)$, Proposition 2.2 and homogeneity implies that $\|X\|_{\psi_2} \leq C\sigma$, for an absolute constant C. By the triangular inequality, $X \sim N(\mu, \sigma^2)$, $\|X\|_{\psi_2} \leq C\sigma + |\mu|(1/\sqrt{\ln\ 2})$.

2. - **(Bounded variables)** Any bounded random variable X is sub-gaussian with $\|X\|_{\psi_2} \leq C\|X\|_\infty$, for $C = 1/\sqrt{\ln\ 2}$.

We end this section with an application of the Laplace transform method combined with the new ideas. We begin considering bounded random variables.

Proposition 2.8 *(Hoeffding's Inequality [29]). Let $X_1, \ldots, X_n$ be independent random variables. Assume that, for each $i \in [n]$, $X_i \in [m_i, M_i]$. Then*

$$\forall t > 0, \quad \mathbb{P}(|\sum_{i=1}^{n} X_i - \mathbb{E}X_i| \geq t) \leq 2e^{-2t^2/B}, \quad \textit{for } B = \sum_{i=1}^{n}(M_i - m_i)^2.$$

Proof. We suppose $\mathbb{E}X_i = 0$. After proving the bound in this case, apply the result for $X_i' = X_i - \mathbb{E}X_i$. For $\lambda > 0$ to be chosen later, consider the moment generating function $e^{\lambda\sum_{i=1}^n X_i}$ for the sum $\sum_{i=1}^n X_i$. Then

$$\begin{aligned}\mathbb{P}(\sum_{i=1}^{n} X_i \geq t) &= \mathbb{P}(e^{\lambda\sum_{i=1}^n X_i} \geq e^{\lambda t}) \\ &\leq e^{-\lambda t}\ \mathbb{E}e^{\lambda\sum_{i=1}^n X_i}\ \text{(Markov inequality)} \\ &\leq e^{-\lambda t}\ \prod_{i=1}^{n}\mathbb{E}e^{\lambda X_i}\ \text{(independence of } X_i)\ .\end{aligned} \tag{2-6}$$

We prove that

$$\mathbb{E}e^{\lambda X_i} \leq e^{\lambda^2(M_i - m_i)^2/8}. \tag{2-7}$$

Since $m_i \leq X_i \leq M_i$, $X_i = \alpha m_i + (1-\alpha)M_i$ where $\alpha = (X - m_i)/(M_i - m_i)$. By the convexity of the map $y \to e^{\lambda y}$, $e^{\lambda X} \leq \alpha e^{\lambda M_i} + (1-\alpha)e^{\lambda m_i}$.

Taking expectations and using $\mathbb{E}X_i = 0$,

$$\mathbb{E}e^{\lambda X} \leq \frac{-m_i}{M_i - m_i}e^{\lambda M_i} + \frac{M_i}{M_i - m_i}e^{\lambda m_i} = e^{g(u)},$$

where $u = \lambda(M_i - m_i)$, $g(u) = -\gamma u + \ln(1 - \gamma + \gamma e^u)$ and $\gamma = -m_i/(M_i - m_i)$.

We bound $g(u)$. Note that $g(0) = 0$, $g'(u) = -\gamma + (1 - \gamma + \gamma e^u)^{-1}\gamma e^u$, and $g'(0) = 0$. For the second derivative,

$$g''(u) = \frac{\gamma e^u(1 - \gamma + \gamma e^u) - \gamma^2 e^{2u}}{(1 - \gamma + \gamma e^u)^2} = \frac{\gamma e^u}{(1 - \gamma + \gamma e^u)}\Big(1 - \frac{\gamma e^u}{1 - \gamma + \gamma e^u}\Big)\ .$$

By the arithmetic-geometric mean inequality, the last term is (uniformly) bounded by $1/4$. Summarizing, $g(0) = g'(0) = 0$ and $g''(u) \leq 1/4$ for $u > 0$. By the Taylor expansion, there is $\theta \in (0, u)$ such that

$$g(u) = g(0) + ug'(0) + \frac{u^2}{2}g''(\theta) \leq \frac{u^2}{8} = \frac{\lambda^2(M_i - m_i)^2}{8}.$$

The proof of the bound 2-7 is now complete.

Substitute it in the right hand side of 2-6:

$$\begin{aligned}\mathbb{P}(\sum_{i=1}^{n} X_i \geq t) &\leq e^{-\lambda t} \prod_{i=1}^{n} \mathbb{E}e^{\lambda X_i} \\ &\leq e^{-\lambda t} \prod_{i=1}^{n} \mathbb{E}e^{\lambda^2 (M_i - m_i)^2/8} \text{ (by 2-7)} \\ &= e^{-\lambda t} e^{\lambda^2 B/8}, \text{ for } B = \sum_{i=1}^{n} (M_i - m_i)^2.\end{aligned}$$

and now the strictest bound is obtained for $\lambda = 4t/B$:

$$\mathbb{P}(\sum_{i=1}^{n} X_i \geq t) \leq e^{-2t^2/B}.$$

Finally, write $\mathbb{P}(|X| \geq t) \leq \mathbb{P}(X \geq t) + \mathbb{P}(X \leq -t)$ and observe that the proof for the other side bound $\mathbb{P}(\sum_{i=1}^{n} X_i \leq -t)$ is the same. We obtain $\mathbb{P}(|\sum_{i=1}^{n} X_i| \geq t) \leq 2e^{-2t^2/B}$. ■

We need an extension of this result, the so called bounded differences inequality [39]. We align the proof from [9] with the argument above.

Theorem 2.9 *(Bounded differences inequality). Let $X_1, \ldots, X_n$ be a sequence of independent random variables. Suppose that a function $g : \mathbb{R}^n \to \mathbb{R}$ obeys the bounded difference condition, i.e, for $i \in [n]$ there is $c_i > 0$ such that*

$$\sup_{x_1,\ldots,x_n,x_i'} |g(x_1, \ldots, x_i, \ldots, x_n) - g(x_1, \ldots, x_i', \ldots, x_n)| \leq c_i \ . \tag{2-8}$$

Then the following concentration holds:

$$\forall t > 0, \quad \mathbb{P}(|g(X_1, \ldots, X_n) - \mathbb{E}g(X_1, \ldots, X_n)| \geq t) \leq 2e^{-2t^2/A}, \quad \textit{for } A = \sum_{i=1}^{n} c_i^2 \ .$$

Proof. As in the end of the proof of 2.8, it is enough to prove one side bound. Let $V_i = \mathbb{E}[g(X_i|X_1, .., X_i)] - \mathbb{E}[g(X_i|X_1, .., X_{i-1})]$. Then $\sum_{i=1}^{n} V_i = g(X_1, \ldots, X_n) - \mathbb{E}g(X_1, \ldots, X_n)$. We apply the Laplace transform method

$$\begin{aligned}&\mathbb{P}(\sum_{i=1}^{n} V_i \geq t) \leq e^{-\lambda t}\ \mathbb{E}e^{t\sum_{i=1}^{n} V_i} \text{ (Same steps of the bound 2-6)} \\ &\leq e^{-\lambda t}\ \mathbb{E}[e^{t\sum_{i=1}^{n-1} V_i}\ \mathbb{E}[e^{tV_n}|X_1, \ldots, X_n]] \text{ (iterated expectation and independence)} \\ &\leq e^{-\lambda t + t^2 c_n^2/8}\ \mathbb{E}[e^{t\sum_{i=1}^{n-1} V_i}] \text{ (By 2.8).}\end{aligned}$$

We use iterated expectation combined with a bound 2.8 recursively, i.e, we apply this combination $n-1$ times and obtain

$$\mathbb{P}(\sum_{i=1}^{n} V_i \geq t) \leq e^{-\lambda t + t^2 A/8}, \quad A = \sum_{i=1}^{n} c_i^2.$$

We optimize for $\lambda = 4t/A$ as we did in the 2.8 and finish the proof. ■

Remark 2.10 The Azuma-Hoeffding inequality [2] is a (more general) martingale version of the inequality above. The proof is the same, using the vocabulary of martingales, which we do not use in this text.

The extension of Hoeffding's inequality to sub-gaussian random variables requires a lemma.

Lemma 2.11 *(Sum of sub-gaussian random variables) Let $X_1, \ldots, X_N$ 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. We consider again the moment generating function. For $\forall \lambda > 0$,

$$\begin{aligned} \mathbb{E} e^{\lambda \sum_{i=1}^{n} X_i} &= \prod_{i=1}^{n} \mathbb{E} e^{\lambda X_i} \text{ (independence)} \\ &\leq \prod_{i=1}^{n} \mathbb{E} e^{C\lambda^2 \| X_i \|_{\psi_2}^2} \text{ (property 2-5)} \\ &= e^{K\lambda^2} \text{ for } K = C \sum_{i=1}^{n} \| X_i \|_{\psi_2}^2. \end{aligned}$$

The sum $\sum_{i=1}^{n} X_i$ has zero mean by the linearity of the expected value. By property 2-5 and the equivalences in Proposition 2.4, $\| \sum_{i=1}^{n} X_i \|_{\psi_2}^2 \leq K$. ■

Proposition 2.12 *(Hoeffding's Inequality, sub-gaussian version [63]). Let $X_1, \ldots, X_n$ be independent, mean zero, sub-gaussian random variables. Then, for any $t \geq 0$,*

$$\mathbb{P}(|\sum_{i=1}^{n} X_i| \geq t) \leq 2e^{-t^2/A'}, \quad \textit{for } A' = \sum_{i=1}^{n} \| X_i \|_{\psi_2}^2 \ .$$

Proof. The proof is a consequence of Lemma 2.11 and 2-3. ■

2.2 Concentration of Norms

We consider the concentration of the Euclidean norm of random vectors. The material of this section is adapted from [63]. The following example motivates the definition of sub-exponential random variable.

Example 2.13 Let $\mathbf{g} = (g_1, \ldots, g_n)$ be a standard normal vector in $\mathbb{R}^n$. We should expect some concentration for the Euclidean norm $\|\mathbf{g}\|_2 = \sqrt{\sum_{i=1}^n g_i^2}$. Notice that g_i is Gaussian and consequently sub-gaussian, but g_i^2 is not:

$$\mathbb{P}(g_i^2 > t) = \mathbb{P}(|g_i| \geq \sqrt{t}) \sim e^{-t/2} \ .$$

Hoeffding inequalities are not applicable here: the tails are slightly heavier than the Gaussian tails. We can still mime the characterizations of sub-gaussian distributions as in the Proposition 2.4 to obtain concentration bounds in the sub-exponential case.

The following proposition is a counterpart of Proposition 2.4.

Proposition 2.14 *(Sub-exponential Properties). 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

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

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

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

Proof. We only prove that II is equivalent to V, the others properties can be proved using the same arguments in the proof of Proposition 2.4. 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 $p! \geq (p/e)^p$.

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

Now apply the bound above for $-X$ and conclude that it holds for all λ with absolute value less than $1/2e$ (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 $\lambda = -1$. Take the p-root and end the proof. ■

Definition 2.15 *(Sub-exponential random variable, sub-exponential norm). A random variable X is sub-exponential if it obeys I-IV. By Proposition 2.14, X is sub-exponential if and only if the following norm is finite*

$$\|X\|_{\psi_1} = \inf\{t > 0 : \mathbb{E}e^{|X|/t} \leq 2\}. \tag{2-9}$$

Proposition 2.16 *(Sub-exponentials are squares of sub-gaussians). A random variable X is sub-gaussian if and only if X^2 is sub-exponential. Moreover*

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

Proof. 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 $t_1 = t_2^2$. ■

We prove an analogous of the sub-gaussian Hoeffding inequality 2.12.

Proposition 2.17 *(Bernstein Inequality- First version) Let $X_1, \ldots, X_n$ be independent, mean zero, sub-exponential random variables. Then*

$$\forall t \geq 0, \quad \mathbb{P}(|\sum_{i=1}^{n} X_i| \geq t) \leq 2\max\{e^{-ct^2/\sigma^2},\ e^{-ct/K}\}$$

for an absolute constant $c > 0$, $\sigma^2 = \sum_{i=1}^{n} \|X_i\|_{\psi_1}^2$ and $K = \max_{i \leq n} \|X\|_{\psi_1}$.

Proof. We use the Laplace transform method. The proof follows closely Proposition 2.8 and some steps are omitted. For $S = \sum_{i=1}^{n} X_i$,

$$\mathbb{P}(S \geq t) \leq e^{\lambda t} \prod_{i=1}^{n} \mathbb{E}e^{\lambda X_i} \tag{2-10}$$

Take λ such that $|\lambda| \leq c/K$, and then, by property V of Proposition 2.14,

$$\mathbb{E}e^{\lambda X_i} \leq e^{C\lambda^2 K^2}.$$

Substituting the MGF bound into 2-10,

$$\mathbb{P}(S \geq t) \leq e^{-\lambda t + C\lambda^2\sigma^2} \ .$$

The best bound is attained for $\lambda = \min\{\frac{t}{2C\sigma^2}, \frac{c}{K}\}$:

$$\mathbb{P}(S \geq t) \leq \max\{e^{-ct^2/\sigma^2}, e^{-ct/K}\} \ ,$$

Again, we use the two sides bound to complete the proof. ■

For convenience we restate the Bernstein bound in a different form. We apply it with $a_i X_i$ instead of X_i where $a = (a_1, \ldots, a_n)$ is vector.

Theorem 2.18 *(Bernstein inequality - Usual form) Under the assumptions of Theorem 2.17, for $a = (a_1, \ldots, a_n) \in \mathbb{R}^n$,*

$$\forall t > 0, \quad \mathbb{P}(|\sum_{i=1}^{n} a_i X_i| \geq t) \leq 2\max\{e^{-ct^2/\|a\|_2^2\sigma^2},\ e^{-ct/K\|a\|_\infty}\}$$

Remark 2.19 The case $a_i = 1/n$ relates asymptotic and non-asymptotic results. Theorem 2.18 above is a quantitative law of large numbers.

Definition 2.20 *(Sub-gaussian random vectors). A random vector* $\mathbf{X}$ *in* $\mathbb{R}^n$ *is sub-gaussian if the one-dimensional marginals* $\langle \mathbf{X}, x\rangle$ *are sub-gaussian random variables for all* $x \in \mathbb{R}^n$*. The sub-gaussian norm of* $\mathbf{X}$ *is defined as the supremum over all possible directions,*

$$\|\mathbf{X}\|_{\psi_2} = \sup_{x\in S^{n-1}} \|\langle \mathbf{X}, x\rangle\|_{\psi_2} \ . \tag{2-11}$$

Theorem 2.21 *(Concentration of norm for vectors). Consider a sub-gaussian random vector* $\mathbf{X} = (X_1, \ldots, X_n) \in \mathbb{R}^n$ *of independent, sub-gaussian coordinates satisfying* $\mathbb{E}[X_i^2] = 1$*. Then*

$$\forall t > 0, \quad \mathbb{P}(|\ \|X\|_2 - \sqrt{n}\ | \geq t) \leq 2e^{-ct^2/K^4} \ ,$$

where $c > 0$ *is an absolute constant independent of* n *and* $K = \max_{i\in[n]} \|X_i\|_{\psi_2}$*.*

Proof. We may suppose $K \geq 1$. If not, multiply $\mathbf{X}$ by a constant larger than 1. We apply Theorem 2.18 on the expression

$$\frac{1}{n}\|\mathbf{X}\|_2^2 - 1 = \frac{1}{n}\sum_{i=1}^{n}(X_i^2 - 1) \ .$$

Since X_i is sub-gaussian, by Proposition 2.16, $X_i^2 - 1$ is sub-exponential with zero mean. The hypothesis of Theorem 2.18 with $a_i = 1/n$ holds: for $\forall i \in [n]$, we consider the sub-exponential norm,

$$\begin{aligned} \|X_i^2 - 1\|_{\psi_1} &\leq C\|X_i^2\|_{\psi_1} \text{ (by triangular inequality)} \\ &= C\|X_i\|_{\psi_2}^2 \text{ (by 2.16)} \\ &\leq CK^2 \end{aligned}$$

Then

$$\mathbb{P}(|\frac{1}{n}\|\mathbf{X}\|_2^2 - 1| \geq u) \leq 2\max\{e^{-cnu/K^4}, e^{-cnu^2/K^4}\} \text{ (Recall } K^4 \geq K^2) \ . \tag{2-12}$$

This is a concentration result for the squared norm. Now use

$$z \geq 0 \ , \ |z - 1| \geq \delta \Rightarrow |z^2 - 1| \geq \max\{\delta, \delta^2\} \tag{2-13}$$

so that, for every $\delta \geq 0$,

$$\begin{aligned} \mathbb{P}(|\frac{1}{\sqrt{n}}\|\mathbf{X}\|_2 - 1| \geq \delta) &\leq \mathbb{P}(|\frac{1}{n}\|\mathbf{X}\|_2^2 - 1| \geq \max\{\delta, \delta^2\}) \text{ (by 2-13)} \\ &\leq 2e^{-cn\delta^2/K^4} \quad \text{(by 2-12 with } u = \max\{\delta, \delta^2\}) \end{aligned}$$

We conclude the proof by setting $\delta = t/\sqrt{n}$. ■

Corollary 2.22 *Under the assumptions of the Theorem 2.21 we write*

$$\sqrt{n} - CK^2 \leq \mathbb{E}\|\mathbf{X}\|_2 \leq CK^2 + \sqrt{n}$$

Proof. Theorem 2.21 is equivalent to $\| \|\mathbf{X}\|_2 - \sqrt{n} \|_{\psi_2} \leq CK^2$, by Proposition 2.4 and the corollary follows. ■

2.3 Bounds for Singular Values via Net Arguments

In this section we use the material developed in the previous sections to present the ε-net arguments that provides bounds for norms of random matrices. The material covered in this section is from [63] and [62]. We start defining some geometric quantities.

Definition 2.23 *(ε-net). Let (T, d) be a metric space. Consider a compact subset $K \subset T$ 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 2.24 *(Covering numbers). The covering number $\mathcal{N}(K, d, \varepsilon)$ is the smallest possible cardinality of an ε-net $\mathcal{N}_\varepsilon$ of K.*

Equivalently, $\mathcal{N}(K, d, \varepsilon)$ is the smallest number of closed balls with center in K and radius ε whose union covers K.

Definition 2.25 *(Packing numbers). 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 $K \subset T$.*

Clearly covering numbers and packing numbers are finite. We omit the reference to d when we refer to the standard Euclidean distance and denote by $|.|$ the Lebesgue measure in $\mathbb{R}^n$. We now quantify the notions above.

Definition 2.26 *(Minkowski Sum). Let A and B subsets of $\mathbb{R}^n$. The Minkowski sum is $A + B = \{a + b :\ a \in A, b \in B\}$.*

Proposition 2.27 *(Volume bounds for Covering and Packing Numbers). Let K be a subset of $\mathbb{R}^n$ and $\varepsilon > 0$. Then*

$$\frac{|K|}{|\varepsilon B_2^n|} \leq \mathcal{N}(K, \varepsilon) \leq \mathcal{P}(K, \varepsilon) \leq \frac{|K + (\varepsilon/2)B_2^n|}{|(\varepsilon/2)B_2^n|}.$$

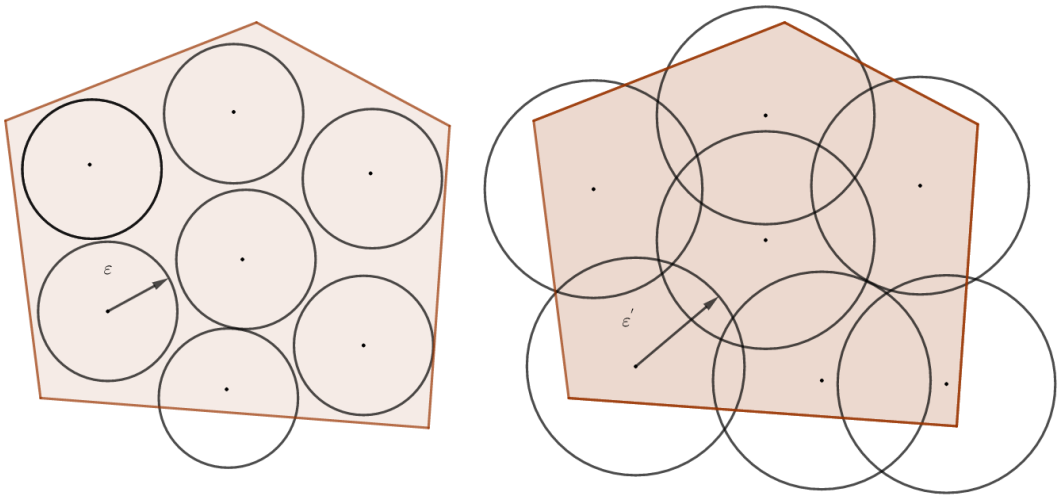

Figure 2.1: Left hand side: packing a pentagon K by seven ε-balls shows that $P(K,\varepsilon) \geq 7$. Right hand side: covering the same pentagon K by seven $\varepsilon^{'}$-balls shows that $N(K,\varepsilon^{'}) \leq 7$.

Proof. We begin with the lower bound. By comparing volumes,

$$|K| \leq \mathcal{N}(K,\varepsilon)\ |\varepsilon B_2^n|.$$

For the middle bound, observe that, if $\mathcal{P}$ is a ε-separated set of maximal cardinality, then $\mathcal{P}$ is a ε-net. Indeed, if for every $k \in K$ there is no $p \in \mathcal{P}$ with $d(p,k) \leq \epsilon$ then $\mathcal{P}' = \mathcal{P} \cup \{k\}$ is a ε-separated set larger than $\mathcal{P}$, a contradiction. The bound follows from the definition of ε-net.
Finally, for the upper bound, we consider the $\mathcal{P}(K,\varepsilon)$ disjoint balls centered in points in K and radius $\varepsilon/2$. The balls may not fit entirely in K, they lie in the larger set $K + (\varepsilon/2)B_2^n$. Now compare volumes,

$$\mathcal{P}(K,\varepsilon)\ |(\varepsilon/2)B_2^n| \leq |K + (\varepsilon/2)B_2^n|\ .$$

■

Corollary 2.28 *(Covering number for the Euclidean unit ball).*

$$\frac{1}{\varepsilon^n} \leq \mathcal{N}(B_2^n,\varepsilon) \leq (\frac{2}{\varepsilon}+1)^n.$$

The same upper bound holds for the sphere S^{n-1}.

Informally, nets are a discretization technique. For concreteness, we estimate the typical value of the norm of a random matrix $A \in \mathbb{R}^{m\times n}$,

$$\|A\| = \|A\|_{2\to 2} = \sup_{x\in B_2^n\ ,\ y\in B_2^m} \langle Ax, y\rangle$$

Consider ε-nets $\mathcal{N}_\varepsilon$ and $\mathcal{M}_\varepsilon$ of B_2^n and B_2^m respectively, with cardinalities $N = \mathcal{N}(B_2^n, \varepsilon)$ and $M = \mathcal{M}(B_2^m, \varepsilon)$. A discretization of the quadratic form is

$$\sup_{x \in \mathcal{N}_\varepsilon \ , \ y \in \mathcal{M}_\varepsilon} \langle Ax, y \rangle = \max_{x \in \mathcal{N}_\varepsilon \ , \ y \in \mathcal{M}_\varepsilon} \sum_{i=1}^{n} \sum_{j=1}^{m} A_{ij} x_i y_j \ .$$

We first evaluate the quality of this approximation.

Proposition 2.29 *(Norm and form over nets). Let $A \in \mathbb{R}^{m \times n}$ and $\varepsilon \in [0, 1)$. For an ε-net $\mathcal{N}_\varepsilon$ of the unit ball B_2^n,*

$$\sup_{x \in \mathcal{N}_\varepsilon} \|Ax\|_2 \leq \|A\| \leq \left(\frac{1}{1-\varepsilon}\right) \sup_{x \in \mathcal{N}_\varepsilon} \|Ax\|_2.$$

Similarly, let $\varepsilon \in [0, 1/2)$. Then, for ε-nets $\mathcal{N}_\varepsilon$ and $\mathcal{M}_\varepsilon$ of B_2^n and B_2^m,

$$\sup_{x \in \mathcal{N}_\varepsilon \ , \ y \in \mathcal{M}_\varepsilon} \langle Ax, y \rangle \leq \sup_{x \in B_2^n \ , \ y \in B_2^m} \langle Ax, y \rangle \leq \left(\frac{1}{1-2\varepsilon}\right) \sup_{x \in \mathcal{N}_\varepsilon \ , \ y \in \mathcal{M}_\varepsilon} \langle Ax, y \rangle.$$

Proof. We prove the slightly harder second statement. The lower bound is straightforward: every net is a subset of a ball. Now fix $x \in S^{n-1}$ and $y \in S^{m-1}$ such that $\langle Ax, y \rangle = \|A\|$ and choose $x_0 \in \mathcal{N}_\varepsilon$ with $\|x - x_0\| \leq \varepsilon$ and $y_0 \in \mathcal{M}_\varepsilon$ with $\|y - y_0\| \leq \varepsilon$. Then

$$|\langle Ax, y \rangle - \langle Ax_0, y_0 \rangle| = |\langle Ax, y - y_0 \rangle + \langle A(x - x_0), y \rangle| \leq 2\varepsilon \|A\|.$$

Also,

$$|\langle Ax_0, y_0 \rangle| \geq |\langle Ax, y \rangle| - |\langle Ax_0, y_0 \rangle| \geq \|A\| - 2\varepsilon\|A\| = (1 - 2\varepsilon)\|A\|$$

■

Theorem 2.30 *(Largest singular value of matrices with sub-gaussian entries). Let A be an $m \times n$ random matrix whose entries A_{ij} are independent, mean zero, sub-gaussian random variables. Then*

$$\forall t > 0, \quad \mathbb{P}(\|A\| > CK(\sqrt{m} + \sqrt{n} + t)) \leq 2e^{-t^2}.$$

Here $C > 0$ is an absolute constant and $K = \max_{i,j} \|A_{ij}\|_{\psi_2}$.

We use Lemma 2.11, as the vectors of interest belong to unit balls. The arguments envolving ε-nets can be divided in three steps: approximation, concentration, union bound.

Proof. **(Approximation)** Fix $\epsilon = 1/4$. From Corollary 2.28 there are ε-nets $\mathcal{N}_\varepsilon$ of the unit ball B_2^n and $\mathcal{M}_\varepsilon$ of the unit ball B_2^m with cardinalities

$$\mathcal{N}(B_2, \epsilon) \leq 9^n \text{ and } \mathcal{M}(B_2, \epsilon) \leq 9^m. \tag{2-14}$$

By Proposition 2.29,

$$\langle Ax, y\rangle \leq 2 \sup_{x\in\mathcal{N}_\varepsilon} \sup_{y\in\mathcal{M}_\varepsilon} \langle Ax, y\rangle \ . \tag{2-15}$$

(Concentration) For $x \in \mathcal{N}_\varepsilon$ and $y \in \mathcal{M}_\varepsilon$, by Lemma 2.11,

$$\|\langle Ax, y\rangle\|_{\psi_2}^2 \leq C \sum_{i=1}^n \sum_{j=1}^m \|A_{ij} x_i y_j\|_{\psi_2}^2 \leq CK^2 \sum_{i=1}^n \sum_{j=1}^m x_i^2 y_j^2 \leq CK^2,$$

since $\|x\| = \|y\| = 1$. By the property 2-3, for some $c > 0$,

$$\forall u \geq 0, \quad \mathbb{P}(\langle Ax, y\rangle \geq u) \leq 2e^{-cu^2/K^2} \ . \tag{2-16}$$

(Union Bound) For a fixed pair (x, y), we have the tail bound 2-16. We apply the crude union bound on nets by combining 2-16 and 2-14,

$$\mathbb{P}(\sup_{x\in\mathcal{N}_\varepsilon} \sup_{y\in\mathcal{M}_\varepsilon} \langle Ax, y\rangle \geq u) \leq \sum_{i=1}^n \sum_{j=1}^m \mathbb{P}(\langle Ax_i, y_j\rangle \geq u) \leq 9^{n+m} e^{-cu^2/K^2}. \tag{2-17}$$

Choose $u = CK(\sqrt{n} + \sqrt{m} + t)$. Then $u^2 \geq C^2K^2(n + m + t^2)$. Also choose C so that $cu^2/K^2 \geq 3(n + m) + t^2$. Combine 2-17 with 2-15,

$$\mathbb{P}(\|A\| \geq 2u) \leq \mathbb{P}(\sup_{x\in\mathcal{N}_\varepsilon} \sup_{y\in\mathcal{M}_\varepsilon} \langle Ax, y\rangle \geq u) \leq 9^{n+m}\ 2e^{-3(n+m)-t^2} \leq 2e^{-t^2}.$$

■

Corollary 2.31 *Under the same assumptions of Theorem 2.30, suppose that the entries A_{ij} of A has unit variance. Then,*

$$K'(\sqrt{m} + \sqrt{n}) \leq \mathbb{E}\|A\| \leq CK(\sqrt{m} + \sqrt{n}). \tag{2-18}$$

Here K' is a constant that depends on K and $C > 0$ is an absolute constant.

Proof. The upper bound follows directly from the integral identity (Appendix). Now, the operator norm of any matrix A is bounded below by the maximum between the operator norm of any row and the operator norm of any column. By Corollary 2.22 we obtain

$$\mathbb{E}\|A\| \geq K'(\sqrt{n} + \sqrt{m}).$$

■

We now consider concentration of extremal singular values for matrices whose entries are not independent: we assume independence of the rows.

Definition 2.32 *(Isotropy). A random vector* $\mathbf{X}$ *in* $\mathbb{R}^n$ *is isotropic if*

$$\forall x \in \mathbb{R}^n, \quad \mathbb{E}[\langle \mathbf{X}, x\rangle^2] = \|x\|_2^2.$$

Lemma 2.33 *(Characterization of Isotropic Random vectors). A random vector* $\mathbf{X}$ *is isotropic if and only if the covariance matrix* Σ *equals* I_n*, the identity matrix of size* n*.*

Proof. Clearly, two symmetric matrices M and N are equal if and only if $x^T M x = x^T N x$ for every vector x. Also

$$\mathbb{E}[\langle \mathbf{X}, x\rangle^2] = x^T(\mathbb{E}[\mathbf{X}\mathbf{X}^T])x = \|x\|_2^2 = x^T I_n x$$

■

Remark 2.34 *(Scaling to Isotropic Case).* Let $\mathbf{X}$ be a random vector with mean vector μ and covariance matrix Σ. For

$$Z = \Sigma^{-1/2}(\mathbf{X} - \mu),$$

we have $\mathbb{E}[ZZ^T] = I_n$. By Lemma 2.33, Z is an isotropic random vector.

We denote the singular values of A by $s_i(A)$ in a decreasing order:

$$s_1(A) = \max \frac{\|Ax\|_2}{\|x\|_2}, \quad s_n(A) = \min \frac{\|Ax\|_2}{\|x\|_2}.$$

Theorem 2.35 *(Sub-Gaussian deviation). Let* A *be an* $m \times n$ *matrix whose rows are independent, mean zero, isotropic sub-gaussian random vectors in* $\mathbb{R}^n$*. Set* $K = \max_{i\in[n]} \|A_i\|_{\psi_2}$*. Then, for every* $t \geq 0$*,*

$$\sqrt{m} - CK^2(\sqrt{n} + t) \leq s_n(A) \leq s_1(A) \leq \sqrt{m} + CK^2(\sqrt{n} + t)$$

with probability at least $1 - 2e^{-t^2}$*. Here* $C > 0$ *is an absolute constant.*

We need a standard result from linear algebra.

Lemma 2.36 *(Approximate Isometries). Let* A *be a* $m \times n$ *matrix for which*

$$\|\frac{1}{m}A^T A - I_n\| \leq \max\{\delta, \delta^2\}, \tag{2-19}$$

for some $\delta > 0$*. Then*

$$(1 - \delta) \leq s_n(A) \leq s_1(A) \leq (1 + \delta).$$

Proof. Suppose without loss of generality that $\|x\|_2 = 1$. By hypothesis,

$$\max\{\delta, \delta^2\} \geq |\langle \frac{1}{m} A^T A - I_n)x, x\rangle\| = |\ \|Ax\|_2^2 - 1\ |.$$

Use $\max\{|z-1|, |z-1|^2\} \leq |z^2 - 1|$ for $z = \|Ax\|_2$. ■

In the proof of Theorem 2.35, we consider nets in the sphere.

Proof. According to Lemma 2.36, it is enough to prove

$$\|\frac{1}{m} A^T A - I_n\| \leq K^2 \max\{\delta, \delta^2\},\ \delta = C(\sqrt{\frac{n}{m}} + \frac{t}{\sqrt{m}}) \tag{2-20}$$

with the required probability. We take the usual three steps.

(Approximation) Using 2.28, there is a 1/4-net $\mathcal{N}$ of the sphere S^{n-1} with cardinality bounded by 9^n. By Proposition 2.29,

$$\|\frac{1}{m} A^T A - I_n\| \leq 2 \max_{x \in \mathcal{N}} |\langle (\frac{1}{m}) A^T A - I_n)x, x\rangle| = 2 \max_{x \in \mathcal{N}} |\frac{1}{m}\|Ax\|_2^2 - 1|.$$

In order to prove 2.36, we show that

$$\max_{x \in \mathcal{N}} |\frac{1}{m}\|Ax\|_2^2 - 1| \leq \frac{\varepsilon}{2},\ \varepsilon = K^2 \max\{\delta.\delta^2\},$$

with the required probability.

(Concentration) For $x \in S^{n-1}$, $\|Ax\|_2^2 = \sum_{i=1}^n \langle A_i, x\rangle^2$. Moreover, $\langle A_i, x\rangle^2$ are independent sub-gaussian random variables with $\mathbb{E}\langle A_i, x\rangle^2 = 1$ (isotropy hypothesis) and $\|\langle A_i, x\rangle^2\|_{\psi_2} \leq K$. By Proposition 2.16, the random variables $\langle A_i, x\rangle^2 - 1$ are sub-exponential, mean zero and independent with $\|\langle A_i, x\rangle^2 - 1\|_{\psi_1} \leq CK^2$. From Theorem 2.18,

$$\begin{aligned}
\mathbb{P}(|\frac{1}{m}\|A_x\|_2^2 - 1| \geq \frac{\varepsilon}{2}) &\leq 2\max\{e^{-cm\varepsilon^2/K^4}, e^{-cm\varepsilon/K^2}\} \\
&= 2e^{-c\delta^2 m}\ (\text{since } \frac{\varepsilon}{K^2} = \max\{\delta, \delta^2\}) \\
&\leq 2e^{-cC^2(n+t^2)},
\end{aligned}$$

where δ was replaced by its value.

(Union Bound)

$$\mathbb{P}(\max_{x \in \mathcal{N}} |\frac{1}{m}\|Ax\|_2^2 - 1| \geq \frac{\varepsilon}{2}) \leq 9^n.2e^{-cC^2(n+t^2)} \leq 2e^{-t^2}.$$

The last step holds for a large $C > 0$. ■

Remark 2.37 Isotropy was used to obtain an absolute constant C.

The next result follows by the integral identity.

Corollary 2.38 *Under the same assumptions of Theorem 2.35,*

$$\sqrt{m} - CK^2\sqrt{n} \leq \mathbb{E}s_n(A) \leq \mathbb{E}s_1(A) \leq \sqrt{m} + CK^2\sqrt{n}.$$

2.4 Chaining

We obtained estimates for singular values of random matrices formed by sub-gaussian entries and sub-gaussian rows. We consider the analogous problem for conic singular values of random matrices in which ε-nets arguments does not work because such arguments relies on the simple geometry of the Euclidean ball. To circumvent the geometry of the cone, we develop tools from stochastic process theory, in particular estimates for the supremum of stochastic process using the method of chaining.

Definition 2.39 *(Stochastic Process). A stochastic process is a collection of random variables $(X_t)_{t\in T}$ defined on a common probability space. A stochastic process is centered if all random variables have zero mean.*

We assume that the index set T is countable, to avoid issues of measurability.

Definition 2.40 *(Lattice supremum). For a stochastic process $(X_t)_{t\in T}$, let $\mathcal{F}$ be the collection of all finite subsets of T. The lattice supremum of X_t is*

$$\mathbb{E}\sup_{t\in T} X_t = \sup_{F\in\mathcal{F}} \mathbb{E}\sup_{t\in F} X_t. \tag{2-21}$$

Definition 2.41 *(Gaussian process). A stochastic process $(X_t)_{t\in T}$ is a Gaussian process if for every finite subset $F = \{f_1, \ldots, f_n\}$ the random vector $\mathbf{X} = (X_{f_1}, \ldots, X_{f_n})$ is Gaussian.*

Example 2.42 Here are three important examples of stochastic processes.

1. The standard Gaussian process $(g_t)_{t\in T}$ formed by independent standard Gaussian random variables is a Gaussian process.

2. For $T \subset \mathbb{R}^n$, the stochastic process $X_t = \langle \mathbf{g}, t\rangle$ where $\mathbf{g} \in \mathbb{R}^n$ is a standard Gaussian vector.

3. Let A be a $m \times n$ random matrix. For $T = S^{n-1} \times S^{m-1}$,

$$s_1(A) = \|A\| = \max_{(u,v)\in T} \langle Au, v\rangle = \max_{(u,v)\in T} X_{(u,v)}.$$

$$s_n(A) = \min_u \max_v \langle Au, v\rangle = \min_u \max_v X_{(u,v)}.$$

The example above motivates the use of stochastic processes to estimate singular values of random matrices.

The definition below connects two notions of distance.

Definition 2.43 *(Stochastic process with sub-gaussian increments). Consider a metric space (T, d). A stochastic process $(X_t)_{t\in T}$ has sub-gaussian increments if, for some $K > 0$, for $t, s \in T$, the random variable $X_t - X_s$ is sub-gaussian with*

$$\|X_t - X_s\|_{\psi_2} \leq Kd(t, s).$$

Example 2.44 Let $(X_t)_{t\in T}$ be a Gaussian process and (T, d) a metric space with the canonical distance $d(t, s) = \|X_t - X_s\|_{L^2}$. The metric d has a probabilistic interpretation: points t and s are near from each other if $X_t - X_s$ has a small variance.

In a similar fashion, the inequality below, due to Dudley [19], relates probabilistic and geometric properties. The proof, based on [63] and [22], is a first example of a chaining argument. The estimate will be refined later for the Talagrand inequality, Theorem 2.50.

Theorem 2.45 *(Dudley integral inequality). Let $(X_t)_{t\in T}$ be a centered stochastic process indexed by a metric space (T, d), with sub-gaussian increments as in 2.43. Then, for an absolute constant $C > 0$,*

$$\mathbb{E}\sup_{t\in T} X_t \leq CK \int_0^\infty \sqrt{\ln(\mathcal{N}(T, d, \varepsilon))}d\varepsilon.$$

We need the following technical lemma.

Lemma 2.46 *(Maximum of a finite collection of sub-gaussian variables). Let $X_1, .., X_N$ be a sequence of sub-gaussian random variables (not necessarily independent). Then, for an absolute constant $C > 0$ and $L = \max_{i\in[N]} \|X_i\|_{\psi_2}$,*

$$\mathbb{E}\max_{i\in[N]} |X_i| \leq CL\sqrt{\ln(2N)}. \tag{2-22}$$

Proof. We use the Laplace transform method. For $\lambda > 0$ to be chosen later,

$$\begin{aligned}
\lambda \mathbb{E} \max_{i\in[N]} X_i &= \mathbb{E}[\ln \max_{i\in[N]}(e^{\lambda X_i})] \\
&\leq \mathbb{E}[\ln \sum_{i=1}^{N}(e^{\lambda X_i})] \text{ (union bound)} \\
&\leq \ln(\sum_{i=1}^{N} \mathbb{E}e^{\lambda X_i}) \text{ (Jensen inequality)} \\
&\leq \ln(Ne^{CL^2\lambda^2}) \text{ (by 2-5)} \\
&= \ln(N) + CL^2\lambda^2
\end{aligned}$$

We minimize the right hand side by choosing $\lambda = L^{-1}\sqrt{\ln(N)C^{-1}}$ to obtain

$$\mathbb{E} \max_{i\in[N]} X_i \leq CL\sqrt{\ln N}.$$

Apply it to $\sup_{i\in[N]} |X_i| = \sup_{i\in[N]}\{X_1, \ldots, X_N, -X_1, \ldots, -X_N\}$. ■

We now prove Theorem 2.45.

Proof. In analogy to the ε-net arguments, we divide the proof in three steps: chaining, controlling the increments and summing the increments.

(Chaining) We assume T finite. By homogeneity, assume $K = 1$. Consider a dyadic scale

$$\varepsilon_k = 2^{-k},\ k \in \mathbb{Z}\ .$$

We choose ε_k-nets T_k of T of cardinality $|T_k| = \mathcal{N}(T, d, \varepsilon_k)$. Since T is finite, there exist k_m and k_M such that $T_{k_m} = \{t_0\}$ for some $t_0 \in T$ and $T_{k_M} = T$. For a point $t \in T$, let $\pi_k(t)$ be a nearest point in T_k, and then

$$d(t, \pi_k(t)) \leq \varepsilon_k.$$

By assumption, the stochastic process is centered, therefore

$$\mathbb{E} \sup_{t\in T} X_t = \mathbb{E} \sup_{t\in T}(X_t - X_{t_0}).$$

Write $X_t - X_{t_0}$ as a telescoping sum

$$X_t - X_{t_0} = \sum_{k=k_m+1}^{k_M} X_{\pi_k(t)} - X_{\pi_{k-1}(t)}.$$

Then

$$\begin{aligned}\mathbb{E}\sup_{t\in T} X_t &= \mathbb{E}\sup_{t\in T}\sum_{k=k_m+1}^{k_M} X_{\pi_k(t)} - X_{\pi_{k-1}(t)} \\ &\leq \sum_{k=k_m+1}^{k_M} \mathbb{E}\sup_{t\in T}(X_{\pi_k(t)} - X_{\pi_{k-1}(t)}).\end{aligned} \tag{2-23}$$

(Controlling the increments) For a fixed t,

$$\begin{aligned}\|X_{\pi_k(t)} - X_{\pi_{k-1}(t)}\|_{\psi_2} &\leq d(\pi_k(t), \pi_{k-1}(t)) \text{ (by 2.43, as } K=1) \\ &\leq d(\pi_k(t), t)) + d(\pi_{k-1}(t), t) \text{ (triangle inequality)} \\ &\leq \varepsilon_k + \varepsilon_{k-1} \text{ (definition of } \pi_k(t)) \leq 2\varepsilon_{k-1}.\end{aligned}$$

Apply Lemma 2.46 for a constant C_1 and $L = 2\varepsilon_{k-1}$, where the supremum is taken for at most $|T_{k-1}||T_k| \leq |T_k|^2$ pairs $(\pi_{k-1}(t), \pi_k(t))$:

$$\mathbb{E}\sup_{t\in T} X_{\pi_k(t)} - X_{\pi_{k-1}(t)} \leq 2C_1\varepsilon_{k-1}\sqrt{\ln(|T_k|^2)} = C\varepsilon_k\sqrt{\ln(|T_k|)}. \tag{2-24}$$

(Summing the increments) Combine 2-23 and 2-24 ($|T_k| = \mathcal{N}(T, d, \varepsilon_k)$),

$$\mathbb{E}\sup_{t\in T} X_t \leq \sum_{k=k_m+1}^{k_M} \mathbb{E}\sup_{t\in T} X_{\pi_k(t)} - X_{\pi_{k-1}(t)} \leq C\sum_{k=k_m+1}^{k_M} \varepsilon_{k-1}\sqrt{\ln(\mathcal{N}(T, d, \varepsilon_k))}.$$

We interpret the right hand size as a Riemann summation.

$$\begin{aligned}C\sum_{k=k_m+1}^{k_M} \varepsilon_{k-1}\sqrt{\ln(\mathcal{N}(T, d, \varepsilon_k))} &= C\sum_{k=k_m+1}^{k_M} 2^{k-1}\sqrt{\ln(\mathcal{N}(T, d, 2^{k-1}))} \\ &\leq C\sum_{k\in\mathbb{Z}} 2\int_{2^{-k-1}}^{2^{-k}} \sqrt{\ln(\mathcal{N}(T, d, \varepsilon))}d\varepsilon \quad (\mathcal{N}(T, d, \varepsilon) \text{ is decreasing in } \varepsilon) \\ &= C\int_0^\infty \sqrt{\ln(\mathcal{N}(T, d, \varepsilon))}d\varepsilon.\end{aligned}$$

Finally, take the lattice supremum 2-21. ■

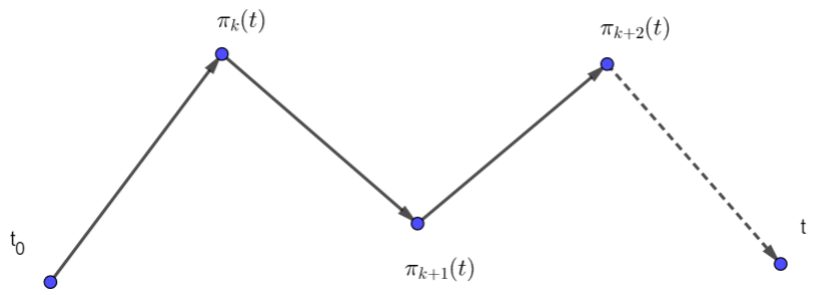

Figure 2.2: Illustration of the chain between t_0 and t.

We now prove Talagrand´s generic chaining bound. Its converse — the majorizing measure theorem, [56], [54] [55] — holds for Gaussian processes. d Now, we do not use ε-nets and specific T_k. Instead, we fix a largest cardinality for a sequence of subsets $(T_k)_{k\in\mathbb{N}}$ and choose appropriate scales.

A key concept is the well known Talagrand functional. To simplify notation, set $N_k = 2^{2^k},\ k \geq 1$ and $N_0 = 1$.

Definition 2.47 *(Admissible Sequences). For a set T, an admissible sequence is an increasing sequence of partitions $(\mathcal{A}_k)$ of T with cardinality $|\mathcal{A}_k| \leq N_k$.*

By increasing sequence of partitions we mean that every set of $\mathcal{A}_{k+1}$ is contained in a set of $\mathcal{A}_k$. Thus, the partition $\mathcal{A}_k$ may consist of a large number of subsets, and provides tight restrictions on the variables of interest.

For $A \subset T$, we denote by $\Delta(A)$ the diameter of A. Also, let $A_n(t) \in \mathcal{A}_n$ be the unique element containing t.

Definition 2.48 *(Talagrand Functional γ_2). The Talagrand functional of a metric space (T, d) is*

$$\gamma_2(T, d) = \inf_{\mathcal{A}_n} \sup_{t\in T} \sum_{k\geq 0} 2^{k/2}\Delta(A_k(t)).$$

Remark 2.49 An equivalent definition of the Talagrand functional is

$$\gamma_2^*(T, d) = \inf_{T_k\subset T} \sup_{t\in T} \sum_{k\geq 0} 2^{k/2} d(t, T_k)$$

for subsets T_k such that $\cup_k T_k = T$ and $|T_k| \leq N_k$. For a proof that $\gamma_2^*(T, d) = \Theta(\gamma_2(T, d))$, see [57].

The proof below is from [57]. We add some observations from [63].

Theorem 2.50 *(Generic Chaining Bound). Let $(X_t)_{t\in T}$ be a centered stochastic process with sub-gaussian increments as in 2.43. Then, for an absolute constant $C > 0$,*

$$\mathbb{E}\sup_{t\in T} X_t \leq CK\gamma_2(T, d).$$

Proof. **(Chaining)** As before, assume $K = 1$ and T finite. Let $(T_k)_{k\in\mathbb{N}}$ be a sequence of subsets of T obeying $|T_k| \leq N_k$ and whose union is T. Set $T_0 = \{t_0\}$ and choose $\pi_k(t) \in T_k$ such that $d(t, \pi_k(t)) = d(t, T_k)$. Suppose that $t \in T_{K_m}$, so that $\pi_{K_m}(t) = t$, and consider the chain with K_m elements

$$X_t - X_{t_0} = \sum_{k=1}^{K_m} X_{\pi_k(t)} - X_{\pi_{k-1}(t)}. \tag{2-25}$$

(Controlling the increments) We need to more careful than in Dudley inequality 2.45. We search for a bound of the type

$$\forall k \in \mathbb{N},\ \forall t \in T, \quad |X_{\pi_k(t)} - X_{\pi_{k-1}(t)}| \leq 2^{k/2} d(t, T_k)$$

with high probability.Fix k and t. By the sub-gaussian assumption,

$$\|X_{\pi_k(t)} - X_{\pi_{k-1}(t)}\|_{\psi_2} \leq d(\pi_k(t), \pi_{k-1}(t)).$$

By the sub-gaussian properties 2.4, for $u > 0$,

$$|X_{\pi_k(t)} - X_{\pi_{k-1}(t)}| \leq Cu 2^{k/2} d(\pi_k(t), \pi_{k-1}(t)), \tag{2-26}$$

with probability at least $1 - 2e^{-8u^2 2^k}$ (to get the constant 8, choose $C > 0$ large). Use the union bound, so that 2-26 holds for every k and t. There are $|T_{k-1}\|T_k| \leq N_k^2$ possible pairs $(\pi_k(t), \pi_{k-1}(t))$. The probability that 2-26 holds for every such pairs is

$$1 - \sum_{k=1}^{\infty} 2^{2^{k+1}} 2e^{-8u^2 2^k} \geq 1 - 2e^{-u^2},$$

for large u.

(Summing the increments) Since 2-26 holds for all k and t, we plug it into the chaining summation 2-25,

$$|X_t - X_{t_0}| \leq Cu \sum_{k=1}^{\infty} 2^{k/2} d(\pi_k(t), \pi_{k-1}(t)).$$

As $d(t, \pi_k(t)) = d(t, T_k)$, the triangle inequality $d(\pi_k(t), \pi_{k-1}(t)) \leq d(t, T_{k-1}) + d(t, T_k)$ gives

$$|X_t - X_{t_0}| \leq 2Cu \sum_{k=0}^{\infty} 2^{k/2} d(t, T_k).$$

Take the infimum over all sequences $(T_k)_{k\in\mathbb{N}}$ and then the supremum over t:

$$\sup_{t\in T} |X_t - X_{t_0}| \leq \inf_{(T_k)_{k\in\mathbb{N}}} \sup_{t\in T}\ 2Cu \sum_{k=0}^{\infty} 2^{k/2} d(t, T_k) = 2Cu\gamma_2^*(T, d),$$

with probability at least $1 - 2e^{-u^2}$. Therefore the supremum in the left hand side is sub-gaussian. By the equivalences 2-3 and 2-4,

$$\mathbb{E} \sup_{t\in T} X_t = \mathbb{E} \sup_{t\in T} X_t - X_{t_0} \leq \mathbb{E} \sup_{t\in T} |X_t - X_{t_0}| \leq \| \sup_{t\in T} |X_t - X_{t_0}| \|_{\psi_2} \leq C_1 \gamma_2^*(T, d).$$

Combine the above inequality with $\gamma_2^*(T,d) \leq \gamma_2(T,d)$ to complete the proof. For an admissible sequence $\mathcal{A}_k$, construct T_k by taking exactly one point of each set A of $\mathcal{A}_k$. Take $\pi_k(t) \in T_k \cap A_k(t)$, so that $d(t,\pi_k(t)) \leq \Delta(A_k(t))$. By Definition 2.48, the proof for finite sets T is complete. Now take the lattice supremum 2-21 to obtain the general result. ■

Remark 2.51 The generic chaining bound is actually sharper than Dudley inequality [57].

A deep result relating supremum of Gaussian processes and the Talagrand functional $\gamma_2(T,d)$ is

Theorem 2.52 *(Majorizing Measure Theorem). Let $(X_t)_{t\in T}$ be a centered Gaussian process for which the index set T is a metric space (T,d) equipped with the so called canonical distance $d(t,s) = \|X_t - X_s\|_{L^2}$. Then, for an absolute constant $C > 0$,*

$$C^{-1}\gamma_2(T,d) \leq E\sup_{t\in T} X_t \leq C\gamma_2(T,d).$$

The upper bound is a consequence of Theorem 2.50, as the constant K there is absolute for any Gaussian process indexed by (T,d). The lower bound, however, relies on specific features of the Gaussian process and does not extend to stochastic processes with sub-gaussian increments. The proof is very technical and can be found in [57].

Corollary 2.53 *(Talagrand Comparison Principle [63]) Let $(Y_t)_{t\in T}$ be a centered stochastic process with sub-gaussian increments and let $(X_t)_{t\in T}$ be a centered Gaussian process. Assume for all $t,s \in T$,*

$$\|Y_t - Y_s\|_{\psi_2} \leq K\|X_t - X_s\|_{L^2}.$$

Then, for an absolute constant $C > 0$,

$$\mathbb{E}\sup_{t\in T} Y_t \leq CK\mathbb{E}\sup_{t\in T} X_t.$$

Proof. Set $d = \|X_t - X_s\|_{L^2}$ and apply Theorems 2.50 and 2.52 to obtain

$$\mathbb{E}\sup_{t\in T} Y_t \leq CK\gamma_2(T,d) \leq CK\mathbb{E}\sup_{t\in T} X_t.$$

■

Corollary 2.53 motivates the development of tools to estimate the supremum of Gaussian process. Such estimates are studied in the next chapter.

3
Special Properties of Gaussian Processes

In this chapter we present two exclusive properties of Gaussian processes. The first one is the Gordon Comparison Lemma that provides estimates for the expectation of the supremum of such processes and the second one is the Gaussian concentration for Lipschitz functions to obtain concentration for this supremum. With such tools, we immediately obtain sharper results for singular values of Matrices formed by Gaussian entries

3.1
The Gordon Comparison Lemma

In this section we prove the following result due to Y.Gordon [25].

Theorem 3.1 *(Gordon Comparison Lemma). Consider two centered Gaussian process* $(X_{ut})_{(u,t)\in U\times T}$ *and* $(Y_{ut})_{(u,t)\in U\times T}$. *If*

$$\forall\ u,t,s,\quad \mathbb{E}[|X_{ut}-X_{us}|^2]\leq \mathbb{E}[|Y_{ut}-Y_{us}|^2], \tag{3-1}$$

$$\forall\ t,s,u\neq v,\quad \mathbb{E}[|X_{ut}-X_{vs}|^2]\geq \mathbb{E}[|Y_{ut}-Y_{vs}|^2]. \tag{3-2}$$

Then

$$\mathbb{E}\inf_{u\in U}\sup_{t\in T}X_{ut}\leq \mathbb{E}\inf_{u\in U}\sup_{t\in T}Y_{ut}.$$

We begin with analogous results in much less general situations.

Proposition 3.2 *(see [37]) Let* $(X_t)_{t\in T}$ *be a Gaussian process. Then for an absolute constant* $C>0$,

$$\mathbb{E}\sup_{t\in T}X_t\leq \mathbb{E}\sup_{t\in T}|X_t|\leq C\ \mathbb{E}\sup_{t\in T}X_t.$$

Proof. The first inequality is obvious. For the reverse inequality, fix an arbitrary point $t_0\in T$ and write

$$\mathbb{E}\sup_{t\in T}|\mathrm{X}_t|\leq \mathbb{E}\sup_{t\in T}|\mathrm{X}_t-\mathrm{X}_{t_0}|+\mathbb{E}|\mathrm{X}_{t_0}|\leq \mathbb{E}\sup_{t\in T,s\in T}|\mathrm{X}_t-\mathrm{X}_s|+C.$$

As $\mathbb{E}\sup_{t\in T,s\in T}\mathrm{X}_t-\mathrm{X}_s=\mathbb{E}\sup_{t\in T,s\in T}\mathrm{X}_s-\mathrm{X}_t$,

$$\mathbb{E}\sup_{t\in T,s\in T}|\mathrm{X}_t-\mathrm{X}_s|=\mathbb{E}\sup_{t\in T,s\in T}\mathrm{X}_t-\mathrm{X}_s=\mathbb{E}\sup_{t\in T,}\mathrm{X}_t+\mathbb{E}\sup_{s\in T,}\text{-}\mathrm{X}_s=2\ \mathbb{E}\sup_{t\in T,}\mathrm{X}_t.$$

The last step holds because X_s and $-X_s$ has the same distribution. ■

We omit the extension (with a similar proof) for generalized symmetric stochastic processes under mild conditions [37]. The next proposition is based on [37] and is a fundamental tool to bound Gaussian process.

Proposition 3.3 *Consider a standard Gaussian stochastic process $(g_t)_{t\in T}$ and T a finite set with cardinality N. Then for some absolute constant $C > 0$,*

$$C^{-1}\sqrt{\ln N} \leq \mathbb{E}\sup_{t\in T} g_t \leq C\sqrt{\ln N}.$$

Proof. By Proposition 3.2 it is enough to bound $\mathbb{E}\sup_{t\in T}|g_t|$. We begin with the upper bound. For $\delta > 0$, to be chosen later,

$$\mathbb{E}\sup_{t\in T}|g_t| = \int_0^\delta \mathbb{P}(\sup_{t\in T}|g_t| \geq u)du \; + \; \int_\delta^\infty \mathbb{P}(\sup_{t\in T}|g_t| \geq u)du.$$

The first integral is bounded by δ. Take the union bound,

$$\int_\delta^\infty \mathbb{P}(\sup_{t\in T}|g_t| \geq u)du \leq N\int_\delta^\infty \mathbb{P}(|g_t| \geq u)du \leq N\sqrt{\frac{\pi}{2}}e^{\frac{-\delta^2}{2}}$$

and, setting $\delta = \sqrt{2\ln N}$,

$$\mathbb{E}\sup_{t\in T}|g_t| \leq \sqrt{2\ln N} + \sqrt{\frac{\pi}{2}}.$$

For the lower bound, use the independence of the random variables,

$$\mathbb{E}\sup_{t\in T}|g_t| \geq \int_0^\delta \mathbb{P}(\sup_{t\in T}|g_t| \geq u)du = \int_0^\delta (1-(1-\mathbb{P}(|g_t| > u))^N)du.$$

By the monotonicity of the Gaussian tail,

$$\mathbb{E}\sup_{t\in T}|g_t| \geq \delta(1-(1-\mathbb{P}(|g_t| > \delta))^N).$$

and

$$\mathbb{P}(|g_t| > \delta) = \sqrt{\frac{2}{\pi}}\int_\delta^\infty e^{\frac{-u^2}{2}} \geq \sqrt{\frac{2}{\pi}}e^{\frac{-(\delta+1)^2}{2}}.$$

Choose $\delta = \sqrt{\ln N}$ and let N such that $\mathbb{P}(|g_t| > u) > \frac{1}{N}$,

$$\mathbb{E}\sup_{t\in T}|g_t| \geq \sqrt{\ln N}(1-(1-\frac{1}{N})^N) \geq \sqrt{\ln N}(1-\frac{1}{e}).$$

We proved that there exists N_0 such that for every $N \geq N_0$ the desired bound holds with absolute constant C_{N_0+1}. For $N < N_0$ we choose a constant C_N depending on N such that the desired bound holds. It is easy to see that

$C = \min\{C_2, .., C_{N_0+1}\}$ is also an absolute constant. ■

Remark 3.4 The bound $\sqrt{2 \ln N}$ is asymptotically sharp. This is Lemma 2.4.11 in [23]. The proof is essentially the same.

The proposition above does not hold for infinite sets T. The Slepian lemma obtains similar bounds for nonindependent Gaussian processes indexed by infinite countable sets T. The main argument in this section is Slepian's "comparison of stochastic processes" [49].

Theorem 3.5 *(Slepian Lemma) Let $(X_t)_{t\in T}$ and $(Y_t)_{t\in T}$ be two centered Gaussian process. Suppose, for all $t, s \in T$,*

$$\mathbb{E}[X_t^2] = \mathbb{E}[Y_t^2], \tag{3-3}$$

$$\mathbb{E}[(X_t - X_s)^2] \geq \mathbb{E}[(Y_t - Y_s)^2]. \tag{3-4}$$

Then for all $\tau \in \mathbb{R}$,

$$\mathbb{P}(\sup_{t\in T} X_t \geq \tau) \geq \mathbb{P}(\sup_{t\in T} Y_t \geq \tau).$$

Combining Slepian lemma and the integral identity we obtain the

Corollary 3.6 *Under the assumptions of Theorem 3.5,*

$$\mathbb{E} \sup_{t\in T} X_t \geq \mathbb{E} \sup_{t\in T} Y_t.$$

V.N.Sudakov and X.Fernique ([52], [53] and [21]) obtained this corollary without hypothesis 3-3. in In this case, Lemma 3.5 does not hold necessarily.

Theorem 3.7 *(Slepian-Sudakov-Fernique Lemma). Let $(X_t)_{t\in T}$ and $(Y_t)_{t\in T}$ be two centered Gaussian stochastic process. Suppose, for all $t, s \in T$,*

$$\mathbb{E}[(X_t - X_s)^2] \geq \mathbb{E}[(Y_t - Y_s)^2]. \tag{3-5}$$

Then $\mathbb{E} \sup_{t\in T} X_t \geq \mathbb{E} \sup_{t\in T} Y_t$.

We skip the proof of this result and consider instead a generalization, which also attains lower bounds that are necessary to estimate minimum singular values.

If U is a singleton, the theorem 3.1 reduces to Lemma 3.7. The proof below is from [22]. We split it in blocks: it is complete at the end of this subsection.

Consider finitely indexed families, i.e, $X_{i,j}$ and $Y_{i,j}$ with $i \in [n]$ and $j \in [m]$. We use a Gaussian interpolation technique. Set $U(t) = \sqrt{t}X + \sqrt{1-t}Y$, so that $U(0) = X$ and $U(1) = Y$. We then show that the function $\phi(t) = \mathbb{E}F(U(t))$ for $F(x) = \min_{i\in[n]} \max_{j\in[m]} \ x_{ij}$ is increasing. This is accomplished by studying the sign of $\phi'(t)$. The factor $\sqrt{t}$ is used to simplify calculations.

The next result is the classical Gaussian integration by parts, also referred as Stein's lemma.

Definition 3.8 *A function $F : \mathbb{R}^m \to \mathbb{R}$ has moderate growth if*

$$\forall \beta > 0, \quad \lim_{\|x\|_2 \to \infty} F(x)e^{-\beta\|x\|_2^2} = 0.$$

Lemma 3.9 *(Stein's lemma). Let $F : \mathbb{R}^m \to \mathbb{R}$ be a differentiable function whose partial derivatives satisfy the moderate growth condition. Then*
(1): For a mean zero Gaussian variable g and $m = 1$,

$$\mathbb{E}[gF(g)] = \mathbb{E}[g^2]\mathbb{E}F'(g) \tag{3-6}$$

(2): For a Gaussian vector $\mathbf{g} = (g_1, \ldots, g_m)$ and a Gaussian random variable g^* such that $(\mathbf{g}, g*)$ is a Gaussian random vector,

$$\mathbb{E}[F(g)g^*] = \sum_{j=1}^{m} \mathbb{E}g_j g^* \mathbb{E}\frac{\partial F}{\partial x_j \partial x_i}(g). \tag{3-7}$$

Proof. To prove (1), set $\sigma^2 = E[g^2]$ and integrate by parts,

$$\mathbb{E}[gF(g)] = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} te^{\frac{-t^2}{2\sigma^2}} F(t)dt = \frac{\sigma^2}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} e^{\frac{-t^2}{2\sigma^2}} F'(t)dt = \mathbb{E}[g^2]\mathbb{E}F'(g).$$

From the moderate growth assumption, $e^{-t^2}F(t)$ vanishes at infinity. We prove (2). Since g^* and $\mathbf{g}$ are not necessarily independent, set $g'_j = g_j - g^* \frac{\mathbb{E}g_j g^*}{\mathbb{E}g^{*2}}$ to obtain $\mathbb{E}[g'_j g^*] = 0$, which implies that $g^{'}$ and g^* are independent (uncorrelated Gaussian variables are independent). Now, condition on g' and apply (1) to the function $t \to F(g'_1 + t\mathbb{E}[g_1 g^*]/\mathbb{E}((g^*)^2), \ldots, g'_m + t\mathbb{E}[g_m g^*]/\mathbb{E}[(g^*)^2]$:

$$\begin{aligned}
\mathbb{E}[g^*F(\mathbf{g})] &= \mathbb{E}[g^*F(g'_1 + t\frac{\mathbb{E}[g_1 g^*]}{\mathbb{E}[(g^*)^2]}, \ldots, g'_m + t\frac{\mathbb{E}[g_m g^*]}{\mathbb{E}[(g^*)^2)]}] \\
&= \mathbb{E}[(g^*)^2] \sum_{j=1}^{m} \frac{\mathbb{E}[g_j g^*]}{\mathbb{E}[(g^*)^2]} \mathbb{E}\frac{\partial F}{\partial x_j}(g'_1 + t\frac{\mathbb{E}[g_1 g^*]}{\mathbb{E}[(g^*)^2]}, \ldots, g'_m + t\frac{\mathbb{E}[g_m g^*]}{\mathbb{E}([(g^*)^2]}) \\
&= \sum_{j=1}^{m} \mathbb{E}[g^* g_j]\mathbb{E}\frac{\partial F}{\partial x_j}(\mathbf{g}).
\end{aligned}$$

■

We need the following standard result in measure theory.

Lemma 3.10 *Let $J \subset \mathbb{R}$ be an open interval and $\psi : J \times \Omega \to \mathbb{R}$. Let X be a random variable in Ω such that $t \to \psi(t, X)$ is almost surely continuously differentiable in J. Assume that for each compact subinterval $I \subset J$,*

$$\mathbb{E} \sup_{t \in I} |\psi'(t, X)| < \infty. \tag{3-8}$$

Then the function $t \to \phi(t) = \mathbb{E}\psi(t, X)$ is continuously differentiable and

$$\phi'(t) = \mathbb{E}\psi'(t, X). \tag{3-9}$$

Proof. Take $t \in J$ and choose a compact subinterval $I \subset J$ containing t in its interior. For small, nonzero $h \in \mathbb{R}$ such that $t + h \in I$, set

$$\phi_h(t) = \frac{\phi(t+h) - \phi(t)}{h} \quad \text{and} \quad \psi_h(t, X) = \frac{\psi_h(t+h, X) - \psi_h(t, X)}{h}.$$

In order to take a derivative in ϕ, we check the hypothesis of the Lebesgue dominated convergence theorem. For positive (resp. negative) h, the mean value theorem gives $t_0 \in [t, t+h]$ (resp. $t_0 \in [t+h, t]$) such that $\psi'(t_0, X) = \psi_h(t, X)$. Therefore $|\psi_h(t, X)| \leq \sup_{t \in I} |\psi'(t, X)|$ and by 3-8, the quotient $\psi_h(t, X)$ has an integral majorant. From dominated convergence,

$$\phi'(t) = \lim_{h \to 0} \phi_h(t) = \lim_{h \to 0} \mathbb{E}\psi_h(t, X) = \mathbb{E}\psi'(t, X).$$

■

We now compute the derivative of the interpolation function ϕ.

Lemma 3.11 *Let $F : \mathbb{R}^m \to \mathbb{R}$ be a differentiable function with partial derivatives of moderate growth. Consider two independent mean zero Gaussian vectors $\mathbf{X} = (X_1, \ldots, X_m)$ and $\mathbf{Y} = (Y_1, .., Y_m)$. Define, for every $t \in [0.1]$, $\mathbf{U}(t) = (U_1(t), \ldots, U_m(t))$ where $U_i(t) = \sqrt{t}X_i + \sqrt{1-t}Y_i$ $i \in [m]$. Then the function $\phi(t) = \mathbb{E}F(\mathbf{U}(t))$ is differentiable and*

$$\phi'(t) = \sum_{i=1}^{m} \mathbb{E}[U_i'(t) \frac{\partial F}{\partial x_i}(\mathbf{U}(t))]. \tag{3-10}$$

Moreover, if F is twice differentiable with all partial derivatives of second order being of moderate growth,

$$\phi'(t) = \frac{1}{2} \sum_{i,j=1}^{m} (\mathbb{E}[X_i X_j] - \mathbb{E}[Y_i Y_j]) \frac{\partial F}{\partial x_i \partial x_j}(\mathbf{U}(t)). \tag{3-11}$$

Proof. First note that

$$\frac{d}{dt}F(\mathbf{U}(t)) = \sum_{i=1}^{m} U_i'(t)\frac{\partial F}{\partial x_i}(\mathbf{U}(t)),$$

with

$$U_i'(t) = \frac{1}{2\sqrt{t}}X_i - \frac{1}{2\sqrt{1-t}}Y_i.$$

Let $I = [a, b] \subset (0, 1)$. We verify condition 3-8.

$$\begin{aligned}\mathbb{E}\sup_{t\in I}[|U_i'(t)\frac{\partial F}{\partial x_i}(\mathbf{U}(t))|] &\leq \mathbb{E}\sup_{t\in I}|U_i'(t)|\sup_{t\in I}|\frac{\partial F}{\partial x_i}(\mathbf{U}(t))| \\ &\leq \sqrt{\mathbb{E}\sup_{t\in I}|U_i'(t)|^2]}\sqrt{\mathbb{E}[|\frac{\partial F}{\partial x_i}(\mathbf{U}(t))|^2]}\end{aligned}$$

The last inequality follows from Cauchy-Schwarz. We first bound the first term of the right hand side. By the triangle inequality and the fact every Gaussian variable has finite variance,

$$\sqrt{\mathbb{E}[\sup_{t\in I}|U_i'(t)|^2]} \leq \sqrt{\mathbb{E}\big[\frac{X_i^2}{4a}\big]} + \sqrt{\mathbb{E}\big[\frac{Y_i^2}{4(1-b)}\big]} < \infty$$

The second term is harder. By the growth condition assumption, there is $\beta > 0$ and $A > 0$ such that

$$\forall x \in \mathbb{R}^m, \quad |\frac{\partial F}{\partial x_i}(x)| \leq Ae^{\beta\|x\|_2}.$$

We bound $\|\mathbf{U}(t)\|_2$. Since $0 \leq t \leq 1$,

$$\|\mathbf{U}(t)\|_2 \leq \|\mathbf{X}\|_2 + \|\mathbf{Y}\|_2 \leq 2\max\{\|\mathbf{X}\|_2, \|\mathbf{Y}\|_2\}.$$

and then

$$\sup_{t\in I}|\frac{\partial F}{\partial x_i}(\mathbf{U}(t))| \leq A\max\{e^{4\beta\|\mathbf{X}\|_2^2}, e^{4\beta\|\mathbf{Y}\|_2^2}\}.$$

For the mean zero Gaussian vectors $\mathbf{X}$ and $\mathbf{Y}$, there are matrices Γ and Γ' such that $\mathbf{X} = \Gamma\mathbf{g}$ and $\mathbf{Y} = \Gamma'\mathbf{g}'$ for independent standard Gaussian vectors $\mathbf{g}$ and $\mathbf{g}'$. Then

$$\begin{aligned}\mathbb{E}[|\frac{\partial F}{\partial x_i}(\mathbf{U}(t))|^2] &\leq A^2\mathbb{E}[e^{8\beta\|\Gamma\|_{2\to2}\|\mathbf{g}\|_2 + 8\beta\|\Gamma'\|_{2\to2}\|\mathbf{g}'\|_2}] \\ &= A^2(\prod_{i=1}^{m}\mathbb{E}[e^{8\beta\|\Gamma\|_{2\to2}\mathbf{g}_i^2}]\prod_{j=1}^{m}\mathbb{E}[e^{8\beta\|\Gamma'\|_{2\to2}\mathbf{g}'^2_j}]).\end{aligned} \tag{3-12}$$

By Proposition 2.16, as $\mathbf{g}_i$ is a sub-gaussian random variable, $\mathbf{g}_i^2$ is sub-

exponential. From Proposition 2.14, there is $\beta > 0$ such that the moment generating function converges. Thus the last term of 3-12 is finite for every $i \in [m]$ and 3-12 is finite as well.
In order to prove 3-11. note that $\mathbb{E}[U_i'(t)U_i(t)] = \frac{1}{2}(\mathbb{E}[X_iX_j] - \mathbb{E}[Y_iY_j])$. Now apply Gaussian integration by parts, formula 3-7,

$$\mathbb{E}[U_i'(t)\frac{\partial F}{\partial x_i}(\mathbf{U}(t))] = \frac{1}{2}\sum_{j=1}^{m}(\mathbb{E}[X_iX_j] - \mathbb{E}[Y_iY_j])\frac{\partial F}{\partial x_i \partial x_j}(\mathbf{U}(t)).$$

■

The choice of $\sqrt{t}$ in the interpolation formula should now be clear: it provides a cancellation such that t does not appear in the right hand side of the expression $\mathbb{E}[U_i'(t)U_i(t)] = \frac{1}{2}(\mathbb{E}[X_iX_j] - \mathbb{E}[Y_iY_j])$.

We now have an expression for the derivative of ϕ, 3-11. We are interested however in a similar formula for a function F which is not twice differentiable, and we work with distributional derivatives (see the Appendix).

To simplify notation, set $r = x_{ij}$, $s = x_{kl}$. The function $F(\mathbf{x})$ is given by

$$F(\mathbf{x}) = \begin{cases} A(r,s) = \max\{\alpha(r), \beta(s)\} & \text{for } i = k \\ B(r,s) = \min\{\alpha(r), \beta(s)\} & \text{for } i \neq k. \end{cases} \tag{3-13}$$

The functions α and β in turn are truncated identities,

$$f(t) = \begin{cases} a, & \text{if } t < a \\ t, & \text{if } t \in [a,b] \\ b, & \text{if } t > b \end{cases} \tag{3-14}$$

We may take $a = -\infty$ or $b = \infty$.

For $i = k$, $F = A(r,s)$ and

$$A(r,s) = \frac{1}{2}(\alpha(r) + \beta(s) + |\alpha(r) - \beta(s)|).$$

The distributional partial derivative with respect to r is

$$\begin{aligned} \frac{\partial A}{\partial r}(r,s) &= \frac{1}{2}(\alpha'(r) + \beta'(r) + \alpha'(r)\text{sgn}(r - \beta(s))) \\ &= \begin{cases} 0, & \text{if } r \in [a,b] \\ \frac{1}{2} + \frac{1}{2}\text{sgn}(r - \beta(s)), & \text{if } r \in [a,b] \end{cases} \end{aligned}$$

The function $s \to sign(r - \beta(s))$ is non-increasing, so $\partial^2 A(r,s)/\partial r \partial s$ is nonpositive in the distributional sense.
Similarly, $\partial^2 B(r,s)/\partial r \partial s$ is nonnegative in the distributional sense. Therefore

$$\frac{\partial^2 F}{\partial x_{ij}\partial x_{kl}}(\mathbf{x}) \leq 0 \quad \text{if } i = k \tag{3-15}$$

and

$$\frac{\partial^2 F}{\partial x_{ij}\partial x_{kl}}(\mathbf{x}) \geq 0 \text{ if } i \neq k. \tag{3-16}$$

Combining 3-15 with assumption 3-1 and 3-16 with the assumption 3-2,

$$\forall i,j,k,l, \quad (\mathbb{E}[|X_{i,j} - X_{k,l}|^2] - \mathbb{E}[|Y_{i,j} - Y_{k,l}|^2])\frac{\partial^2 F}{\partial x_{ij}\partial x_{kl}}(\mathbf{x}) \leq 0. \tag{3-17}$$

We still did not get a formula like 3-11. Expand squares in formula 3-17 is not enough. But notice that the function $F(\mathbf{x}) = \min_{i\in[n]} \max_{j\in[m]} \mathbf{x}_{ij}$ satisfies $F(\mathbf{x} + c\mathbf{e}) = F(\mathbf{x}) + c$ for $\mathbf{e} = (1, 1, \ldots, 1)$. Indeed, if we sum $c\mathbf{e}$ to x the maximum and minimum are translated by the same c. The next lemma applied to F finishes the proof for finite sets U and T.

Lemma 3.12 *Let $F : \mathbb{R}^m \to \mathbb{R}$ be a Lipschitz function. Let $\mathbf{X} = (X_1, \ldots, X_m)$ and $\mathbf{Y} = (Y_1, \ldots, Y_m)$ be two mean zero Gaussian vectors. Assume that 3-17 holds in the sense of distributions and that $F(\mathbf{x} + c\mathbf{e}) = F(\mathbf{x}) + c$ for $\mathbf{e} = (1, 1, \ldots, 1)$. Then*

$$\mathbb{E}F(\mathbf{X}) \geq \mathbb{E}F(\mathbf{Y}).$$

Proof. Without loss of generality we can assume If $\mathbf{X}$ and $\mathbf{Y}$ are not independent, take $\mathbf{Y}'$ an independent copy of $\mathbf{Y}$ with the same expected value. We assume then that $\mathbf{X}$ and $\mathbf{Y}$ are independent.

Suppose first that F is twice differentiable. Since $F(\mathbf{x} + c\mathbf{e}) = F(\mathbf{x}) + c$,

$$\forall i \in [m], \mathbf{x} \in \mathbb{R}^m, \quad \sum_{j=1}^{m} \frac{\partial^2 F}{\partial x_i \partial x_j}(\mathbf{x}) = 0,$$

and

$$\sum_{j\neq i} \frac{\partial^2 F}{\partial x_i \partial x_j}(\mathbf{x}) = \frac{\partial^2 F}{\partial x_i^2}(\mathbf{x}). \tag{3-18}$$

We now relate equations 3-17 and 3-11 and control the sign of ϕ',

$$\begin{aligned}
\phi' &= \sum_{i,j=1}^{m} (\mathbb{E}[X_i X_j] - \mathbb{E}[Y_i Y_j])\frac{\partial^2 F}{\partial x_i \partial x_j} \\
&= -\sum_{i=1}^{m} (\mathbb{E}[X_i^2] - \mathbb{E}[Y_i^2]) \sum_{j\neq i} \frac{\partial^2 F}{\partial x_i \partial x_j} + \sum_{i\neq j} (\mathbb{E}[X_i X_j] - \mathbb{E}[Y_i Y_j])\frac{\partial^2 F}{\partial x_i \partial x_j} \\
&= -\frac{1}{2}\sum_{i\neq j} (\mathbb{E}[X_i^2] - \mathbb{E}[Y_i^2] + \mathbb{E}[X_j^2] - \mathbb{E}[Y_j^2] - 2(\mathbb{E}[X_i X_j] - \mathbb{E}[Y_i Y_j]))\frac{\partial^2 F}{\partial x_i \partial x_j} \\
&= -\frac{1}{2}\sum_{i,j=1}^{m} \frac{1}{2}\sum_{i,j=1}^{m} (\mathbb{E}[(X_i - X_j)^]] - \mathbb{E}[(Y_i Y_j)^2])\frac{\partial F}{\partial x_i \partial x_j} \geq 0
\end{aligned}$$

To prove the result in generality, we use compactly supported mollifiers ρ_ε (see

Appendix). For a smoothed version F_ε of F,

$$\begin{aligned}|F_\varepsilon(\mathbf{x}) - F(\mathbf{x})| &= |\int_{\mathbb{R}^m}(F(\mathbf{y}) - F(\mathbf{x}))\rho_\varepsilon(\mathbf{x}-\mathbf{y})| \\ &\leq \int_{\mathbb{B}(\mathbf{x},\varepsilon)}|F(\mathbf{y}) - F(\mathbf{x})|\rho_\varepsilon(\mathbf{x}-\mathbf{y}) \\ &= \mathcal{O}(\varepsilon) \ (F \text{ is Lipschitz}).\end{aligned}$$

In particular, F_ε converges uniformly to F when $\varepsilon \to 0$. Now we have to check that F_ε satisfies the same conditions of F. For a distribution f,

$$\begin{aligned}(F_\varepsilon, \frac{\partial^2 f}{\partial x_i \partial x_j}) &= \int_{\mathbb{R}^m}\int_{\mathbb{R}^m} F(\mathbf{y})\rho_\varepsilon(\mathbf{y}-\mathbf{x})d\mathbf{y}\frac{\partial^2 f}{\partial x_i \partial x_j}d\mathbf{x} \\ &= \int_{\mathbb{R}^m} F(\mathbf{y})\frac{\partial^2(f * \rho_\varepsilon)}{\partial x_i \partial x_j}(\mathbf{y})d\mathbf{y}.\end{aligned}$$

The last equality follows from Fubini theorem. We now check the additivity property of F_ε with respect to $\mathbf{e}$,

$$F_\varepsilon(\mathbf{x}-\mathbf{y}+\mathbf{ce}) = \int_{\mathbb{R}^n} F(\mathbf{x}-\mathbf{y}+\mathbf{ce})\rho_\varepsilon(\mathbf{y})d\mathbf{y} = F_\varepsilon(\mathbf{x})+\mathbf{ce}\int_{\mathbb{R}^m}\rho_\varepsilon(\mathbf{y})d\mathbf{y} = F_\varepsilon+\mathbf{ce}$$

The last equality follows from the fact that $\rho_\varepsilon \geq 0$ and $\|\rho_\varepsilon\|_{L^1} = 1$. Apply Lemma 3.12 to F_ε and use the uniform convergence to obtain

$$\mathbb{E}F(\mathbf{x}) = \lim_{\varepsilon\to 0}\mathbb{E}F_\varepsilon(\mathbf{x}) \geq \lim_{\varepsilon\to 0}\mathbb{E}F_\varepsilon(\mathbf{y}) = \mathbb{E}F(\mathbf{y}).$$

■

Finally, Theorem 3.1 follows by taking the lattice supremum 2-21.

3.2 Some Applications of the Gordon Comparison Lemma

Sudakov's theorem below [51] plays a key role in the proof of Talagrand majorizing measure theorem, Theorem 2.52. The proof is from [63], [37]. It is a generalization of the lower bound obtained in Proposition 3.3.

Theorem 3.13 *(Sudakov Minoration Inequality). Let $(X_t)_{t\in T}$ be a centered Gaussian process. Then there is an absolute constant $C > 0$ such that*

$$\forall \varepsilon > 0, \quad E\sup_{t\in T} X_t \geq C\varepsilon\sqrt{\ln(\mathcal{N}(T,d,\varepsilon))}.$$

Proof. For a ε-net $\mathcal{N}$ of T and a standard Gaussian process $(g_t)_{t\in T}$, set

$$Y_t = \frac{\varepsilon}{\sqrt{2}}g_t, \ t \in \mathcal{N}.$$

For two different points $t, s \in \mathcal{N}$,

$$E(X_t - X_s)^2 = d(t,s)^2 \geq \varepsilon^2,$$

$$E(Y_t - Y_s)^2 = \frac{\varepsilon^2}{2} E(g_t - g_s)^2 = \varepsilon^2.$$

The last equality follows from the fact that $g_t - g_s$ has distribution N(0,2). From Theorem 3.7 and Proposition 3.3,

$$E \sup_{t \in T} \mathrm{X}_t \geq E \sup_{t \in \mathcal{N}} \mathrm{X}_t \geq E \sup_{t \in \mathcal{N}} \mathrm{Y}_t \geq C\varepsilon \sqrt{\ln(\mathcal{N}(T,d,\varepsilon)}.$$

■

The next result is from [22]. Set $E_n = \sqrt{2}\frac{\Gamma((n+1)/2)}{\Gamma(n/2)}$.

Proposition 3.14 *(Euclidean norm of the standard Gaussian vector). Let* $\mathbf{g} \in \mathbb{R}^n$ *be a standard Gaussian vector. Then*

$$\frac{n}{\sqrt{n+1}} \leq \mathbb{E}\|\mathbf{g}\|_2 = E_n \leq \sqrt{n}.$$

Proof. Since the entries of $\mathbf{g}$ are independent, $\|\mathbf{g}\|_2^2$ follows a chi-square distribution with n degrees of freedom (see the Appendix). The expectation is

$$\begin{aligned}\mathbb{E}\|\mathbf{g}\|_2 &= \frac{1}{2^{n/2}\Gamma(n/2)} \int_0^\infty x^{1/2} x^{n/2-1} e^{-x/2} dx \\ &= \frac{2^{n/2}\sqrt{2}}{2^{n/2}\Gamma(n/2)} \int_0^\infty t^{(n-1)/2} e^{-t} dt \text{ (use } t = u/2) = E_n.\end{aligned}$$

By the functional equation of the gamma function,

$$E_{n+1}E_n = 2\frac{\Gamma(n/2+1)}{\Gamma(n/2)} = n$$

so that $E_n^2 \leq n$ and the upper bound for $\mathbb{E}\|\mathbf{g}\|_2$ follows. For the lower bound, note that $E_n = n/E_{n+1} \geq n/\sqrt{n+1}$. ■

The next theorem is an application of the theory of stochastic process to the study of random singular values. The proof is based on [63] and [22].

Theorem 3.15 *(Gordon Theorem for Gaussian Matrices [26]). Let* A *be* $m \times n$ *whose rows are independent standard Gaussian vectors. Then*

$$E_m - E_n \leq \mathbb{E}s_n(A) \leq \mathbb{E}s_1(A) \leq E_m + E_n.$$

Proof. For $(u, v) \in S^{n-1} \times S^{m-1}$, compare the stochastic processes

$$X_{uv} = \langle Au, v\rangle \text{ and } Y_{uv} = \langle \mathbf{g}, u\rangle + \langle \mathbf{h}, v\rangle,\ \mathbf{g} \sim N(0, I_n)\ , \mathbf{h} \sim N(0, I_m).$$

To use Theorem 3.1, we verify 3-5, 3-1 and 3-2. We begin with X_{uv}.

$$\begin{aligned}
\mathbb{E}[(X_{uv} - X_{u^*,v^*})^2] &= \mathbb{E}[|(\sum_{i=1}^{n}\sum_{j=1}^{m} A_{ij}(u_j v_i - u_j^* v_i^*)|^2] \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m} u_i^2 v_j^2 + (u_i^*)^2 (v_j^*)^2 - 2u_j u_j^* v_i v_i^* \\
&= \|u\|_2^2 \|v\|_2^2 + \|u^*\|_2^2 \|v^*\|_2^2 - 2\langle u, u^*\rangle\langle v, v^*\rangle \\
&= 2 - 2\langle u, u^*\rangle\langle v, v^*\rangle.
\end{aligned}$$

Similarly, for Y_{uv},

$$\begin{aligned}
\mathbb{E}[(Y_{uv} - Y_{u^*v^*})^2] &= \mathbb{E}[(\langle \mathbf{g}, u - u^*\rangle)^2] + \mathbb{E}[(\langle \mathbf{h}, v - v^*\rangle)^2] \\
&= \|u - u^*\|_2^2 + \|v - v^*\|_2^2 \\
&= 4 - 2\langle u, u^*\rangle - 2\langle v, v^*\rangle.
\end{aligned}$$

Combining the previous expressions,

$$\mathbb{E}[(Y_{uv} - Y_{u^*v^*})^2] - \mathbb{E}[(X_{uv} - X_{u^*,v^*})^2] = 2(1 - \langle u, u^*\rangle)(1 - \langle v, v^*\rangle) \geq 0.$$

The three conditions are thus satisfied. For ε-nets $\mathcal{N}_\varepsilon$ and $\mathcal{M}_\varepsilon$ of S^{n-1} and S^{m-1} respectively,

$$\begin{aligned}
\mathbb{E}\sup_{S^{n-1}\times S^{m-1}} \langle Au, v\rangle &\leq \frac{1}{1-2\varepsilon}\mathbb{E}\sup_{\mathcal{N}_\varepsilon \times \mathcal{M}_\varepsilon} \langle Au, v\rangle \ \text{(Proposition 2.29)} \\
&\leq \frac{1}{1-2\varepsilon}\mathbb{E}\sup_{\mathcal{N}_\varepsilon \times \mathcal{M}_\varepsilon} \langle \mathbf{g}, u\rangle + \mathbb{E}\langle \mathbf{h}, v\rangle \ \text{(Theorem 3.7)} \\
&\leq \frac{1}{1-2\varepsilon}\mathbb{E}\sup_{S^{n-1}\times S^{m-1}} \langle \mathbf{g}, u\rangle + \langle \mathbf{h}, v\rangle \\
&= \frac{1}{1-2\varepsilon}(\sqrt{2}\frac{\Gamma((n+1)/2)}{\Gamma(n/2)} + \sqrt{2}\frac{\Gamma((m+1)/2)}{\Gamma(m/2)}) \ \text{(Proposition 3.14).}
\end{aligned}$$

Take the limit $\varepsilon \to 0^+$ to obtain the upper bound on $\mathbb{E}s_1(A)$. We apply the same idea for $s_n(A)$ with a few modifications. Let u_0 be the vector that achieves the minimum of $\|A_u\|$ over the sphere and u_0^{net} the vector that achieves the minimum of $\|Au\|$ over the net. Moreover, consider u_1 the vector in the net $\mathcal{N}_\varepsilon$ such that $\|u_0 - u_1\| \leq \varepsilon$, then

$$s_n(A) = \|Au_0\| \leq \|A_{u_0^{\text{net}}}\| \leq \|Au_1\| \leq \|Au_0\| + \|A(u_1 - u_0)\| \leq s_n(A) + \varepsilon\|A\|.$$

As a consequence,

$$\inf_{S^{n-1}} \sup_{S^{m-1}} \langle Au, v\rangle \geq \inf_{\mathcal{N}_\varepsilon} \sup_{S^{m-1}} \langle Au, v\rangle - \varepsilon\|A\| \geq \inf_{\mathcal{N}_\varepsilon} \sup_{\mathcal{M}_\varepsilon} \langle Au, v\rangle - \varepsilon\|A\|$$

We bound the right hand side with Theorem 3.1:

$$\begin{aligned}
\mathbb{E}\inf_{\mathcal{N}_\varepsilon} \sup_{\mathcal{M}_\varepsilon} \langle Au, v\rangle - \varepsilon\|A\| &\geq \mathbb{E}\inf_{\mathcal{N}_\varepsilon} \sup_{\mathcal{M}_\varepsilon} \langle g, u\rangle + \langle h, v\rangle - \varepsilon\|A\| \text{ (by 3.1)} \\
&= \mathbb{E}\sup_{\mathcal{N}_\varepsilon} \langle g, -u\rangle + \mathbb{E}\sup_{\mathcal{M}_\varepsilon} \langle h, v\rangle - \varepsilon\|A\| \\
&\geq -\sqrt{2}\frac{\Gamma((n+1)/2)}{\Gamma(n/2)} + \sqrt{2}\frac{\Gamma((m+1)/2)}{\Gamma(m/2)}(1-\varepsilon) - \varepsilon\|A\| \text{ (by 2.29).}
\end{aligned}$$

Again, let $\varepsilon \to 0^+$ to finish the proof. ■

Remark 3.16 By Proposition 3.14, $\mathbb{E}s_1(A) \leq \sqrt{m} + \sqrt{n}$. Also, $\mathbb{E}s_n(A) \geq \sqrt{m} - \sqrt{n}$, from a long calculation related with the Gamma function (see Lemma C.4 in [22]).

3.3 The Gaussian Concentration Phenomenon for Lipschitz functions

We present a theorem proved independently by Sudakov and Borell in [50] and [8]. Our version follows [63]. For different versions, see [9], [22]. The importance of this theorem is to obtain tail bounds for singular values.

Theorem 3.17 *(Gaussian Concentration for Lipschitz functions). Let* $\mathbf{g}$ *be a standard Gaussian vector. Let* $f : \mathbb{R}^n \to \mathbb{R}$ *be a Lipschitz map with respect to the standard Euclidean norm and denote the Lipschitz constant by* $\|f\|_{Lip}$*. Then, for an absolute constant* $C > 0$*,*

$$\mathbb{P}(|f(\mathbf{g}) - \mathbb{E}f(\mathbf{g})| \geq t) \leq Ce^{-t^2/2\|f\|_{Lip}}.$$

Concentration of measure is an important concept in Analysis admitting a strong geometric interpretation. The proof below is based on isoperimetric inequalities. The first connection between the subject and concentration is due to P.Levy in his celebrated work on the isoperimetric problem for the sphere [38]. A similar connection has been established in many different scenarios (see [63], Chapter 5). We will be especially interested in concentration in Euclidean spaces with an appropriate Gaussian measure defined below. Gromov extended these ideas to Riemannian manifolds with positive Ricci curvature [27], but this is outside the scope of this text. For a broad view of the concentration phenomenon, see [35].

We denote by $d_g(.,.)$ by the geodesic distance on the sphere S^{n-1}, for which spherical caps are balls. An ε- neighborhood of a set $A \subset S^{n-1}$ is

$$A_\varepsilon = \{x \in S^{n-1} : \exists y \in A \text{ such that } d_g(x,y) \leq \varepsilon\}.$$

Equip the sphere S^{n-1} with the normalized Lebesgue measure σ_{n-1}, and let H be a hemisphere, so that $\sigma_{n-1}(S^{n-1}) = 1$ and $\sigma_{n-1}(H) = 1/2$. Levy [38] proved that, for a ε-neighborhood H_ε of H, $\sigma_{n-1}(H_\varepsilon) \geq 1 - 2e^{-c\varepsilon^2}$: H_ε essentially covers the whole sphere! Levy also proved the celebrated isoperimetric inequality for the sphere.

Theorem 3.18 *(Isoperimetric inequality for the sphere). Let $\mathcal{C}$ be a spherical cap and $A \subset S^{n-1}$ such that $\sigma_{n-1}(A) = \sigma_{n-1}(\mathcal{C})$. Then, for every $\varepsilon > 0$,*

$$\sigma_{n-1}(A_\varepsilon) \geq \sigma_{n-1}(\mathcal{C}_\varepsilon).$$

A proof is given in [20] and [23]. In contrast to the classic isoperimetric inequality for Euclidean space with Lebesgue measure, for which a short proofs are frequent (see, for example, [64]), the case of the sphere is hard.

The Gaussian context is analogous: instead of S^{n-1}, we consider $\mathbb{R}^n$ equipped with the canonical Gaussian measure.

Definition 3.19 *(Canonical Gaussian measure). Let $\mathbf{g}$ be a standard Gaussian vector and $A \subset \mathbb{R}^n$. The canonical Gaussian measure is*

$$\mu_n(A) = \mathbb{P}(\mathbf{g} \in A) = \frac{1}{(2\pi)^{n/2}} \int_A e^{-\|x\|_2^2/2}\, dx\ .$$

For $u \in \mathbb{R}^n$ and $a > 0$, consider the half-space $H = \{x \in \mathbb{R}^n : \langle x, u\rangle \leq a\}$. We prove an isoperimetric inequality for $\mathbb{R}^n$ with the Gaussian measure.

We now transfer the isoperimetric inequality obtained for the sphere to $\mathbb{R}^n$ equipped with Gaussian measure. The result was obtained independently by [50] and [8]. The proof below is from [36].

Theorem 3.20 *(Isoperimetric inequality in the Gaussian case). Let A be a measurable set in $\mathbb{R}^n$ and let H be a half-space such that $\mu_n(A) = \mu_n(H)$. Then, for every $\varepsilon > 0$,*

$$\mu_n(A_\varepsilon) \geq \mu_n(H_\varepsilon).$$

We proceed to the proof of Theorem 3.20. The following lemma is a standard argument in probability theory, incorrectly attributed to Poincaré (for historical notes see [37]). We follow [36]. Consider the sphere $\mathbb{S} = \sqrt{N}S^N \subset \mathbb{R}^{N+1}$.

Lemma 3.21 *(Poincaré Observation). Let $\sigma_N^{\sqrt{N}}$ be the normalized Lebesgue measure of the sphere $\mathbb{S}$ and let $P_{N+1,n}$ be the orthogonal projection from $\mathbb{R}^{N+1}$ onto $\mathbb{R}^n$. Then, for every set $A \subset \mathbb{R}^n$,*

$$\lim_{N\to\infty} \sigma_N^{\sqrt{N}}(P_{N+1,n}^{-1}(A) \cap \mathbb{S}) = \mu_n(A).$$

We use the inclusion $\mathbb{R}^n \subset \mathbb{R}^{N+1}$ given by $(x_1, \ldots, x_n) \mapsto (x_1, \ldots, x_n, 0, \ldots, 0)$.

Proof. Let $(g_i)_{i\geq 1}$ be a sequence of independent standard Gaussian variables. Set $R_k^2 = g_1^2 + \ldots + g_k^2$. Note that $(\sqrt{k}/R_{k+1})(g_1, .., g_{k+1})$ is equal in distribution to $\sigma_k^{\sqrt{k}}$. Thus $(\sqrt{k}/R_{k+1})(g_1, .., g_n)$ is equal in distribution to $P_{k+1,n}(\sigma_k^{\sqrt{k}})$ for $k \geq n$. Notice that $R_n^2, R_{N+1}^2 - R_n^2, (g_1, \ldots, g_n)/R_n$ are independent. Thus R_n^2/R_{N+1}^2 is independent of $(g_1, \ldots, g_n)/R_n$ and has β distribution with parameters $a = n/2$ and $b = (N + 1 - n)/2$ (see the Appendix). Then

$$\begin{aligned}\sigma_N^{\sqrt{N}}(P_{N+1,n}^{-1}(A) \cap \mathbb{S}) &= \mathbb{P}(\frac{\sqrt{N}}{R_{N+1}}(g_1, \ldots, g_n) \in A) \\ &= \mathbb{P}((N\frac{R_n^2}{R_{N+1}^2})^{1/2}\frac{1}{R_n}(g_1, \ldots, g_n) \in A).\end{aligned}$$

Use the probability density function of beta distribution in polar coordinates to rewrite the right hand side as

$$\beta(\frac{n}{2}, \frac{N+1-n}{2})^{-1} \int_{S^{n-1}} \int_0^1 \mathbb{I}_A(x\sqrt{Nt}) t^{n/2-1}(1-t)^{(N+1-n)/2} d\sigma_{n-1}(x)dt \ .$$

Change of variables, $r = \sqrt{Nt}$, so that

$$\beta(\frac{n}{2}, \frac{N+1-n}{2})^{-1} \frac{2}{N^{n/2}} \int_{S^{n-1}} \int_0^{\sqrt{N}} \mathbb{I}_A(xr) r^{n-1}(1 - \frac{u^2}{N})^{(N+1-n)/2} d\sigma_{n-1}(x)dr \ .$$

By the Lebesgue dominated convergence theorem, the limit for $N \to \infty$ is

$$\frac{2}{2^{n/2}\Gamma(n/2)} \int_{S^{n-1}} \int_0^{\infty} \mathbb{I}_A(rx) r^{n-1} e^{-r^2/2} d\sigma_{n-1}(x)dr \ ,$$

which is the Gaussian measure of A in polar coordinates. ■

We now prove Theorem 3.20. Let $\Phi(t)$ be the cumulative density function of a Gaussian variable,

$$\Phi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-x^2/2} dx \ .$$

For the half-space $H = \{x \in \mathbb{R}^n : \langle x, u\rangle \leq a\}$, $\mu_n(H) = \Phi(a)$ since the

Gaussian measure is a product of one dimensional Gaussian measures μ_1 and is invariant under rotations.

Proof. Theorem 3.20 is equivalent to $\mu(A_\varepsilon) \geq \Phi(a+\varepsilon)$. Assume that $a = \Phi^{-1}(\mu_n(A)) > -\infty$ and fix $b \in (-\infty, a)$. By hypothesis, $\mu_n(A) = \Phi(a) > \Phi(b) = \mu_1((-\infty, b])$. By Lemma 3.21, for large N,

$$\sigma_N^{\sqrt{N}}(P_{N+1,n}^{-1}(A_\varepsilon) \cap \mathbb{S}) > \sigma_N^{\sqrt{N}}(P_{N+1,1}^{-1}((-\infty, b]) \cap \mathbb{S}).$$

The set $P_{N+1,1}^{-1}((-\infty, b]) \cap \mathbb{S}$ is a spherical cap. Moreover, by the inclusion $(P_{N+1,n}^{-1}(A) \cap \mathbb{S})_\varepsilon \subset P_{N+1,n}^{-1}(A_\varepsilon) \cap \mathbb{S}$,

$$\begin{aligned}\sigma_N^{\sqrt{N}}(P_{N+1,n}^{-1}(A_\varepsilon) \cap \mathbb{S}) &\geq \sigma_N^{\sqrt{N}}((P_{N+1,n}^{-1}(A) \cap \mathbb{S})_\varepsilon) \\ &\geq \sigma_N^{\sqrt{N}}((P_{N+1,1}^{-1}((-\infty, b]) \cap \mathbb{S})_\varepsilon) \text{ (Theorem 3.18).}\end{aligned}$$

Set $\epsilon(N) = -b + \sqrt{N}\cos[\arccos(\frac{b}{\sqrt{N}}) - \frac{\varepsilon}{\sqrt{N}}]$ to obtain

$$(P_{N+1,1}^{-1}((-\infty, b]) \cap \mathbb{S})_\varepsilon = (P_{N+1,1}^{-1}((-\infty, \epsilon(N) + b]) \cap \mathbb{S}).$$

When $N \to \infty$, $\epsilon(N) \to \varepsilon$. By Lemma 3.21, $\mu_n(A_\varepsilon) \geq \Phi(b+\varepsilon)$. Since $b < a$, we may take the limit $b \to a^-$. ■

We now prove Theorem 3.17. The argument is based in [36] and [64].

Proof. By homogeneity assume $\|f\|_{\text{Lip}} = 1$. Let $M \in \mathbb{R}$ such that

$$\mu_n(f(x) \geq M) \geq \frac{1}{2} \quad \text{and} \quad \mu_n(f(x) \leq M) \geq \frac{1}{2}.$$

For $u \in \mathbb{R}^n$, consider the half-space $H = \{x \in \mathbb{R}^n : \langle x, u\rangle \leq 0\}$. Then

$$\mu_n(f(x) \geq M) \geq \frac{1}{2} = \Phi(0) = \mu_n(H).$$

Let $A = \{x \in \mathbb{R}^n : f(x) \leq M\}$. By Theorem 3.20,

$$\mu_n(A_\varepsilon) \geq \mu_n(H_\varepsilon) = \Phi(\varepsilon) \geq 1 - \frac{1}{2}e^{-\varepsilon^2/2}.$$

The last step follows from Proposition 2.2. We claim that

$$\mu_n(A_\varepsilon) \leq \mu_n(f(x) \leq M + \varepsilon).$$

Indeed, for $x \in A_\varepsilon$, there exists $y \in A$ such that $\|x - y\|_2 \leq \varepsilon$. Since f is Lipschitz, $f(x) \leq f(y) + \|x - y\|_2 \leq M + \varepsilon$. Repeat the steps for $-f$ to obtain

$$\mathbb{P}(|f(\mathbf{g}) - M| \geq \varepsilon) \leq e^{-\varepsilon^2/2}. \tag{3-19}$$

To finish the proof, replace the median by the expected value. It is enough to prove that $|M - \mathbb{E}f(\mathbf{g})| \leq C$ for an absolute constant $C > 0$. Then

$$\begin{aligned}|\mathbb{E}f(\mathbf{g}) - M| &= |\mathbb{E}[f(\mathbf{g}) - M]| \leq \mathbb{E}|f(\mathbf{g}) - M| \\ &= \int_0^\infty \mathbb{P}(|f(\mathbf{g}) - M| \geq t)dt \text{ (integral identity)} \\ &\leq \int_0^\infty e^{-t^2/2}dt = \sqrt{\frac{\pi}{2}} \text{ (by 3-19).}\end{aligned}$$

■

As an application of Theorem 3.17 we obtain a tail bound version of Theorem 3.15. The proof follows [63]. The Frobenius norm of a $m \times n$ matrix A with real valued entries is $||A||_F = \sqrt{\text{Tr}(A^t A)}$.

Corollary 3.22 *(Tail bounds for operator norm of Gaussian matrices). Under the assumptions of Theorem 3.15, for absolute constants $C, C_1 > 0$,*

$$\forall t > 0, \quad \mathbb{P}(\mid \|A\| - C_1(\sqrt{m} + \sqrt{n}) \mid \geq t) \leq Ce^{-t^2/2}.$$

Proof. Write the entries of the $m \times n$ Gaussian matrix A as a Gaussian vector $A^{'}$ with distribution $N(0, I_{nm})$. We apply Theorem 3.17. The map f given by $f(A^{'}) = \|A\|_2$ satisfies $f(A') = \|A\|_2 \leq \|A\|_F$. We prove that f is 1-Lipschitz

$$\mid f(A^{'}) - f(B^{'}) \mid \leq \|A - B\| = f(A - B) \leq \|A - B\|_F = \|A^{'} - B^{'}\|_2.$$

By Corollary 2.38 and Theorem 3.15, $\mathbb{E}\|A\| = \Theta(\sqrt{m} + \sqrt{n})$. Now, apply Theorem 3.17. ■

Remark 3.23 The argument above can be applied to prove Theorem 1.2.

4
Convex Recovery and the Small Ball Method

In this chapter, motivated by problems in signal recovery by convex optimization, we find accurate lower bounds for the conic singular value with the tools from stochastic process developed in the previous chapters. Let´s describe the scenario: the vector $\mathbf{x}^{\#} \in \mathbb{R}^n$ is an unknown signal to be recovered by observing $\mathbf{y} \in \mathbb{R}^m$,

$$\mathbf{y} = \Phi\mathbf{x}^{\#} + \mathbf{e}, \tag{4-1}$$

for a known sampling matrix $\Phi \in \mathbb{R}^{m\times n}$ and an error vector $\mathbf{e} \in \mathbb{R}^m$. Denote the extended real line $\mathbb{R} \cup \{\pm\infty\}$ by $\overline{\mathbb{R}}$. The presentation here is based on [60].

Definition 4.1 *(Proper Convex Functions). A convex function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is proper if, for every $x \in \mathbb{R}^n$, $f(x) > -\infty$ and there exists at least one point x_o such that $f(x_o) < +\infty$.*

For the rest of this chapter, $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is a proper convex function.

Given such $f : \mathbb{R}^n \to \overline{\mathbb{R}}$, we assume $\|\mathbf{e}\|_2 \leq \eta$ for some given $\eta > 0$. As a program for recovery, we take for an approximation of $\mathbf{x}^{\#}$ the solution of

$$\min_{\mathbf{x}} \quad f(\mathbf{x}) \tag{4-2a}$$

$$\text{subject to} \quad \|\Phi\mathbf{x} - \mathbf{y}\|_2 \leq \eta. \tag{4-2b}$$

We refer to this optimization problem as the convex recovery program and derive a deterministic error bound $\|\hat{x} - \mathbf{x}^{\#}\|_2$. Notice that *we assume the existence of a solution $\hat{x}$ to the optimization problem.*

How artificial is this hypothesis? The admissible set is not bounded, since the linear map Φ is highly non-injective. Still, a continuous function f for which $|f(x)| \to \infty$ as $|x| \to \infty$ necessarily realizes its minimum $\hat{x}$. Typical examples are norms (but not the ℓ_0 norm, which is not really a norm).

Example 4.2 (Recovery of Sparse Vectors): Consider an s-sparse vector $\mathbf{x}^{\#}$, i.e, a vector with at most s nonzero entries. It is well known that the ℓ_1 norm promotes sparsity ([12] or [22]). We set $f(x) = \|x\|_1$.

Matrices $\mathbf{X}^{\#} \in \mathbb{R}^{n_1\times n_2}$ may be interpreted as vectors in $\mathbb{R}^{n_1 n_2}$. A linear operator $\Lambda : \mathbb{R}^{n_1\times n_2} \to \mathbb{R}^m$ plays the role of the sampling matrix Φ, and we take $f : \mathbb{R}^{n_1\times n_2} \to \overline{\mathbb{R}}$. The Schatten 1- norm of a matrix $A \in \mathbb{R}^{m\times n}$ is $\|A\|_{S_1} = \sum_{i=1}^{\min\{m,n\}} |s_i(A)|$.

Example 4.3 (Recovery of Low Rank Matrices): A common approach to recover a low rank matrix $\mathbf{X}^{\#}$ from observations $y = \Lambda X + \mathbf{e}$ is to minimize the Schatten 1- norm $\|.\|_{S_1}$ ([12], [22]). We solve the problem below

$$\min_{\mathbf{X}} \quad \|X\|_{S_1} \tag{4-3a}$$

$$\text{subject to} \quad \|\Lambda(\mathbf{X}) - \mathbf{y}\|_2 \leq \eta. \tag{4-3b}$$

We first consider some motivation to derive the deterministic error bound $\|\hat{x} - \mathbf{x}^{\#}\|_2$. Suppose for simplicity that a linear system of equations $Ax = b$ with $A \in \mathbb{R}^{n\times n}$ invertible and $b \in \mathbb{R}^n$ is disturbed by an error vector $\mathbf{e}$ that satisfies $\|\mathbf{e}\|_2 \leq \eta$. Then the linear system $Ax = b$ becomes $Ax_e = b + \mathbf{e}$. Set $\delta = x - x_e$ and notice that $A\delta = \mathbf{e}$, then $\|\delta\|_2 = \|A^{-1}\mathbf{e}\|_2 \leq \frac{\eta}{s_n(A)}$.

As above, we derive an estimate for $\|\hat{x} - \mathbf{x}^{\#}\|_2$ in terms of the minimum singular value of Φ and the given constant η. However, we only consider directions in which the function f does not increases. Consequently the singular value is now restricted to a cone formed by these directions.

Definition 4.4 *(Cone) A measurable set $\mathcal{K} \subset \mathbb{R}^n$ is a cone if for every $x \in \mathcal{K}$ and $\tau > 0$, $\tau x \in \mathcal{K}$, i.e, a positive homogeneous measurable set.*

Definition 4.5 *(Descent Cone). The descent cone $\mathcal{D}(f, \mathbf{x})$ of f at $\mathbf{x}$ is*

$$\mathcal{D}(f, \mathbf{x}) = \bigcup_{\tau>0} \{\mathbf{u} \in \mathbb{R}^n : f(\mathbf{x} + \tau\mathbf{u}) \leq f(\mathbf{x})\}.$$

At a strict global minimum, the descent cone is empty. It is always convex, but not necessarily closed.

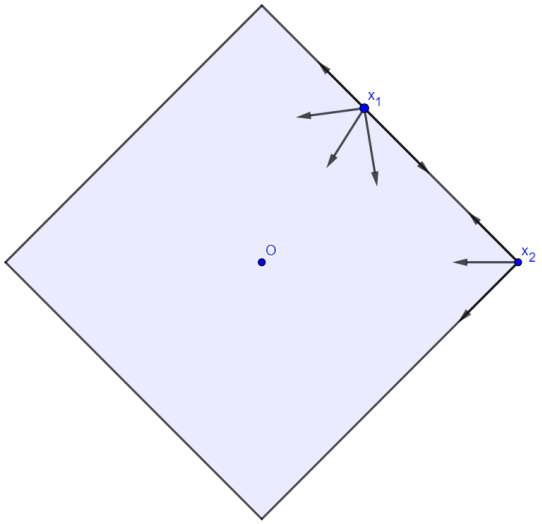

Figure 4.1: The descent cone in the ℓ_1 ball at the points x_1 and x_2. At the point x_1 the descent cone becomes a hyperplane.

Definition 4.6 *(Conic singular values). Let $\mathcal{K} \subset \mathbb{R}^n$ be a cone. The minimum conic singular value of matrix Φ with respect to $\mathcal{K}$ is*

$$s_n(\Phi, \mathcal{K}) = \inf_{x \in \mathcal{K} \cap S^{n-1}} \|\Phi x\|_2.$$

Theorem 4.7 *(A deterministic error bound). Any solution $\hat{x}$ of the optimization problem 4-2 satisfies*

$$\|\hat{x} - \mathbf{x}^\#\|_2 \leq \frac{2\eta}{s_n(\Phi, \mathcal{D}(f, \mathbf{x}^\#))}.$$

Proof. Let $\mathbf{u} = \hat{x} - \mathbf{x}^\#$. We rewrite 4-2 as

$$\min_{\mathbf{u}} f(\mathbf{x}^\# + \mathbf{u}) \quad \text{subject to} \quad \|\Phi\mathbf{u} - \mathbf{e}\|_2 \leq \eta.$$

Clearly $\mathbf{u} = 0$ is a feasible solution. Therefore any optimal point $\hat{\mathbf{u}}$ must satisfy $f(\mathbf{x}^\# + \hat{\mathbf{u}}) \leq f(\mathbf{x}^\#)$ and then $\hat{\mathbf{u}} \in \mathcal{D}(f, \mathbf{x}^\#)$. Then

$$s_n(\Phi, \mathcal{D}(f, \mathbf{x}^\#)) = \inf_{u \in \mathcal{D}(f, \mathbf{x}^\#)} \frac{\|\Phi\mathbf{u}\|_2}{\|\mathbf{u}\|_2} \leq \frac{\|\Phi\mathbf{u} - \mathbf{e}\|_2 + \|\mathbf{e}\|_2}{\|\mathbf{u}\|_2} \leq \frac{2\eta}{\|\mathbf{u}\|_2}.$$

■

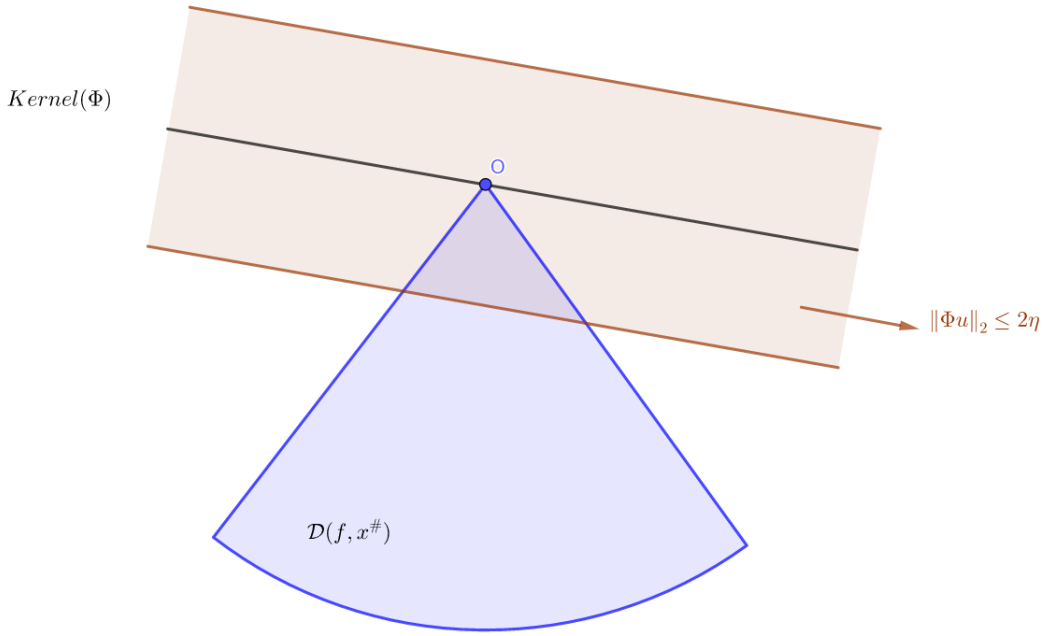

Figure 4.2: **Geometry of convex recovery**: The descent cone $\mathcal{D}(f, x^\#)$ contains the directions u in which f is decreasing at x. The assumption that $\|\mathbf{e}\|_2 \leq \eta$ added to the fact that $\|\Phi u - \mathbf{e}\|_2 \leq \eta$ creates the red tube. Each optimal point $\hat{u}$ lies in the intersection between the red and blue area.

Challenge: estimate the statistics of $s_n(\Phi, \mathcal{K})$ from below.

4.1 Gaussian Sampling Matrix

In this section we analyze, for a cone $\mathcal{K} \subset \mathbb{R}^n$, the conic singular value $s_n(\Phi, \mathcal{K})$ of the $m \times n$ Gaussian sampling matrix Φ, whose rows consist of independent standard Gaussian vectors. We mime the techniques developed in Chapter 3. We start with a definition.

Definition 4.8 *(Conic Gaussian Width). Let $\mathcal{K} \subset \mathbb{R}^n$ be a cone and $\mathbf{g}$ be a standard Gaussian vector. The conic Gaussian width of $\mathcal{K}$ is*

$$w(\mathcal{K}) = \mathbb{E} \sup_{u \in \mathcal{K} \cap S^{n-1}} \langle \mathbf{g}, u \rangle.$$

See [64], [63] for a geometric interpretation.

Theorem 4.9 *(Minimum Gaussian Conic singular value). Let $w(\mathcal{K})$ be the conic Gaussian width of a cone $\mathcal{K} \subset \mathbb{R}^n$. Then*

$$s_n(\Phi, \mathcal{K}) \geq \sqrt{m-1} - w(\mathcal{K}) - t,$$

with probability at least $1 - 2e^{-t^2/2}$.

Proof. For $(u, v) \in (\mathcal{K} \cap S^{n-1}) \times S^{m-1}$, compare the stochastic processes

$$X_{uv} = \langle \Phi u, v \rangle \text{ and } Y_{uv} = \langle \mathbf{g}, u \rangle + \langle \mathbf{h}, v \rangle, \ \mathbf{g} \sim N(0, I_m), \ \mathbf{h} \sim N(0, I_n).$$

Repeat the same steps of the proof of Theorem 3.15 for an ε-net $\mathcal{N}_\varepsilon$ of $\mathcal{K} \cap S^{n-1}$ and an ε-net $\mathcal{M}_\varepsilon$ of S^{m-1} to obtain

$$\begin{aligned}
\mathbb{E} s_n(\Phi, \mathcal{K}) &= \mathbb{E} \inf_{\mathcal{K} \cap S^{n-1}} \sup_{S^{m-1}} \langle \Phi u, v \rangle \\
&\geq \mathbb{E} \sup_{\mathcal{N}_\varepsilon} \langle g, -u \rangle + \mathbb{E} \sup_{\mathcal{M}_\varepsilon} \langle h, v \rangle - \varepsilon \|\Phi\| \\
&\geq -w(\mathcal{K}) + \sup_{S^{m-1}} \langle h, v \rangle (1 - \varepsilon) - \varepsilon \|\Phi\| \text{ (Proposition 2.29)} \\
&\geq -w(\mathcal{K}) + \sqrt{m-1}(1 - \varepsilon) - \varepsilon \|\Phi\| \text{ (Proposition 3.14).}
\end{aligned}$$

Taking the limit $\varepsilon \to 0^+$, $\mathbb{E} s_n(\Phi, \mathcal{K}) \geq \sqrt{m-1} - w(\mathcal{K})$. In order to apply Theorem 3.17, we check that the map $\psi(\Phi) = \inf_{\mathcal{K} \cap S^{n-1}} \|\Phi u\|_2$ is 1-Lipschitz with respect to the Frobenius norm. Indeed,

$$\begin{aligned}
\psi(\Phi) = \inf_{\mathcal{K} \cap S^{n-1}} \|\Phi u\|_2 &\leq \inf_{\mathcal{K} \cap S^{n-1}} (\|\Phi u\|_2 + \|(\Phi - \Phi_1) u\|_2) \\
&\leq \inf_{\mathcal{K} \cap S^{n-1}} (\|\Phi u\|_2 + \|\Phi - \Phi_1\|) \\
&\leq \psi(\Phi_1) + ||\Phi - \Phi_1||_F.
\end{aligned}$$

For the bound $\psi(\Phi_1) - \psi(\Phi)$, repeat the steps above interchanging Φ and Φ_1. Finally write Φ as a vector Φ' in $\mathbb{R}^{mn}$ and apply Theorem 3.17. ■

Denote $[a]_+ = \max\{0, a\}$. Combine Theorems 4.9 and 4.7 to conclude that any optimal solution $\hat{x}$ of the optimization problem 4-2 satisfies

$$\|\mathbf{x}^\# - \hat{x}\|_2 \leq \frac{2\eta}{[\sqrt{m-1} - w(\mathcal{D}(f, \mathbf{x}^\#)) - t]_+}$$

with probability at least $1 - 2e^{-t^2/2}$. The case in which the denominator above is larger than zero is called stable recovery. For stable recovery, we must have $m \geq C(w^2(\mathcal{D}(f, \mathbf{x}^\#)) + w(\mathcal{D}(f, \mathbf{x}^\#)))$, for an absolute constant $C > 0$. We use a duality argument to estimate the conic Gaussian width $w(\mathcal{D}(f, \mathbf{x}^\#))$.

Definition 4.10 *(Polar Cone). Let $\mathcal{K} \subset \mathbb{R}^n$ be a cone. The polar cone $\mathcal{K}^o$ is*

$$\mathcal{K}^o = \{v \in \mathbb{R}^n : \langle v, k \rangle \leq 0 \text{ for all } k \in \mathcal{K}\}.$$

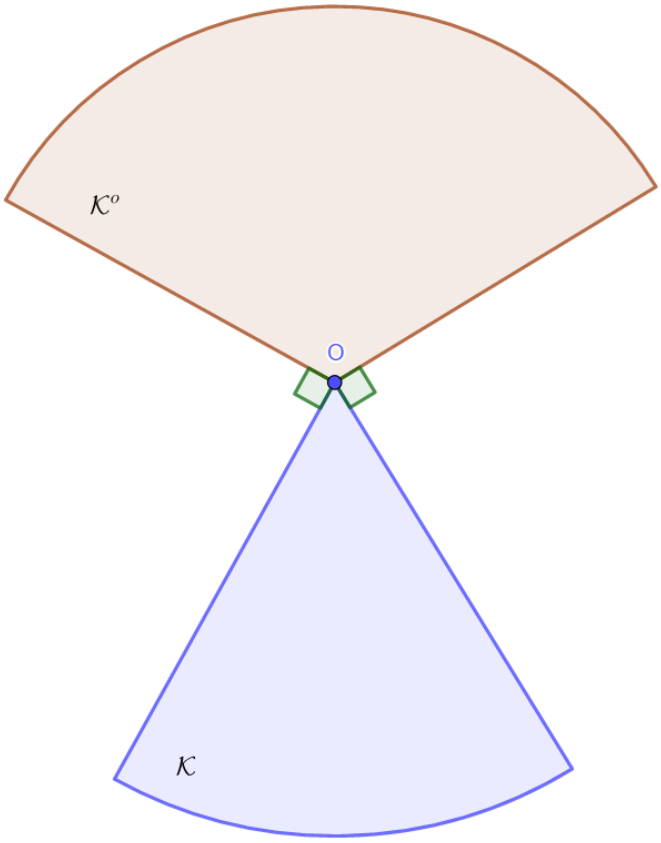

Figure 4.3: Illustration of the Polar Cone.

Denote by $d(\mathcal{K}, x)$ the Euclidean distance between the point $x \in \mathbb{R}^n$ and $\mathcal{K}$.

Proposition 4.11 *(Duality for Cones). For $x \in \mathbb{R}^n$,*

$$\sup_{u \in \mathcal{K} \cap S^{n-1}} \langle u, x \rangle \leq d(x, \mathcal{K}^o).$$

Proof. We express the Euclidean norm in terms of a functional. Write

$$d(\mathcal{K}^o, x) = \inf_{v \in \mathcal{K}^o} \|v - x\|_2 = \inf_{v \in \mathcal{K}^o} \sup_{u \in S^{n-1}} \langle x - v, u \rangle.$$

Apply the standard inf-sup inequality (see Appendix),

$$d(\mathcal{K}^o, x) \geq \sup_{u \in S^{n-1}} \inf_{v \in \mathcal{K}^o} \langle x - v, u \rangle = \sup_{u \in S^{n-1}} [\langle x, u \rangle - \sup_{v \in \mathcal{K}^o} \langle v, u \rangle].$$

By the definition of polarity, the supremum is infinite unless $u \in (\mathcal{K}^o)^o$. Assume that the left-hand side is finite (the infinite case is trivial). Therefore

$$d(\mathcal{K}^o, x) \geq \sup_{(\mathcal{K}^o)^o \cap S^{n-1}} \langle x, u \rangle \geq \sup_{\mathcal{K} \cap S^{n-1}} \langle x, u \rangle.$$

The last inequality follows from $\mathcal{K} \subset (\mathcal{K}^o)^o$. ■

We consider a standard notion in convex analysis [45] that is useful to analyze optimization problems with non differentiable objective function.

Definition 4.12 *(Subdifferential of a convex function). The subdifferential of* $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ *at* $x \in \mathbb{R}^n$ *is*

$$\partial f(x) = \{v \in \mathbb{R}^n : f(y) \geq f(x) + \langle v, y - x \rangle, \text{ for all } y \in \mathbb{R}^n\}.$$

Proposition 4.13 *(Width of a Descent Cone). Suppose that the subdifferential of* $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ *does not contain the origin and is nonempty. Then the conic Gaussian width of the the descent cone of* f *at a point* $x \in \mathbb{R}^n$ *satisfies*

$$w^2(\mathcal{D}(f, x)) \leq \mathbb{E} \inf_{\tau \geq 0} d^2(\mathbf{g}, \tau \partial f).$$

We use an auxiliary result from convex analysis [45].

Proposition 4.14 *(Polarity for Descent Cones). Assume that the subdifferential* $\partial f(x)$ *is nonempty and does not contain the origin. Then*

$$\mathcal{D}^o(f, x) = \overline{\bigcup_{\tau \geq 0} \tau \partial f(x)}.$$

Now we prove Proposition 4.13.

Proof. Proposition 4.11 implies that

$$w(\mathcal{D}(f, x)) = \mathbb{E} \sup_{u \in \mathcal{D}(f,x) \cap S^{n-1}} \langle \mathbf{g}, u \rangle \leq \mathbb{E}\ d(\mathbf{g}, \mathcal{D}^o(f, x)).$$

Apply Proposition 4.14 to the right hand side

$$w(\mathcal{D}(f, x)) \leq \mathbb{E}\ d(\mathbf{g}, \overline{\bigcup_{\tau \geq 0} \tau \partial f(x)}) = \mathbb{E} \inf_{\tau \geq 0} d(\mathbf{g}, \tau \partial f(x)).$$

Square both sides and apply Jensen inequality. ■

Corollary 4.15 *(Sparse Vectors).* Let $\mathbf{x}^{\#} \in \mathbb{R}^n$ be an s-sparse vector. A sufficient condition for stable recovery of the optimization problem 4-2 is that the $m \times n$ Gaussian matrix Φ has at least $2s\ln(\frac{n}{s})$ rows.
Proof. We claim that the conic Gaussian width $w(\mathcal{D}(\|.\|_1, \mathbf{x}^{\#}))$ satisfies

$$w^2(\mathcal{D}(\|.\|_1, \mathbf{x}^{\#})) \leq 2s\ln(\frac{n}{s}) + 2s.$$

We may assume that the entries of $\mathbf{x}^{\#}$ are non-increasing and non-negative, since the Gaussian distribution and the ℓ_1 norm are invariant under signed permutations. By Proposition 4.13, for every $\tau > 0$,

$$w^2(\mathcal{D}(\|.\|_1, \mathbf{x}^{\#})) \leq \mathbb{E}d^2(\mathbf{g}, \tau\partial\|\mathbf{x}^{\#}\|_1).$$

The subdifferential is of the form

$$\partial\|\mathbf{x}^{\#}\|_1 = \{[\mathbf{1}_s \ y]^t | \ \|y\|_\infty \leq 1\}, \quad \mathbf{1}_s = (1, \ldots, 1) \in \mathbb{R}^s.$$

Therefore

$$\mathbb{E}d^2(\mathbf{g}, \tau\partial\|\mathbf{x}^{\#}\|_1) = \sum_{j=1}^{s} \mathbb{E}(g_j - \tau)^2 + \sum_{j=s+1}^{n} \mathbb{E}[|g_j| - \tau]_+^2.$$

The first term in the right hand side is $1 + \tau^2$. From Proposition 2.2,

$$\mathbb{E}[|g_j| - \tau]_+^2 = \int_\tau^\infty (u - \tau)\mathbb{P}(|g_j| \geq u)du \leq \int_\tau^\infty u^2(\sqrt{\frac{2}{\pi}}u^{-1}e^{-u^2/2})du < e^{-\tau^2/2}.$$

so that

$$w^2(\mathcal{D}(\|.\|_1, \mathbf{x}^{\#})) \leq s(1 + \tau^2) + (n - s)e^{-\tau^2/2}.$$

Optimize on τ: $\tau_{\min}^2 = 2\ln(\frac{n}{s})$. Apply Theorems 4.9 and 4.7. ■

4.2 The Small Ball Method

In this section we describe Tropp's modification [60] of Mendelson's small ball method [31] for the estimation the conic singular value of non-Gaussian sampling matrices. Consider a random vector $\varphi \in \mathbb{R}^n$ and independent copies φ_i, $i \in [m]$ of φ. For the rest of this chapter, the sampling matrix is $\Phi = [\varphi_1 \ \ldots \ \varphi_m]^t$.

Definition 4.16 *(Empirical Process). Let $(S, \mathbf{S}, P)$ be a probability space and let X_i, $i \in \mathbb{N}$ be coordinate functions of the infinite product probability space*

$(\Omega, \Sigma, \mathbb{P}) = (S^{\mathbb{N}}, \mathbf{S}^{\mathbb{N}}, P^{\mathbb{N}})$, $X_i : S^{\mathbb{N}} \to S$. *Let* $\mathcal{F}$ *be a set of functions* $\mathbf{f} : S \to \mathbb{R}$. *The empirical process corresponding to* $X_1, \ldots, X_n$ *indexed by* $\mathcal{F}$ *is*

$$P_n\mathbf{f} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{f}(X_i).$$

For a cone $\mathcal{K} \in \mathbb{R}^n$, set $E = \mathcal{K} \cap S^{n-1}$. The minimum conic singular value $s_n(\Phi, \mathcal{K})$ is expressed in terms of a nonnegative empirical process

$$s_n(\Phi, \mathcal{K}) = \inf_{u \in E}(\sum_{i=1}^{m} |\langle \varphi_i, u\rangle|^2)^{1/2}. \tag{4-4}$$

We now describe the general strategy of the small ball method.

Definition 4.17 *(Marginal Tail function). The marginal tail function of the set* E *with respect a random vector* φ *is*

$$Q_\xi(E, \varphi) = \inf_{u \in E} \mathbb{P}(|\langle \varphi, u\rangle| \geq \xi).$$

The marginal tail function can be used to quantify the absolute continuity of the distribution of φ [31]. A Rademacher random variable X is the random variable such that $\mathbb{P}(X = 1) = \mathbb{P}(X = -1) = 1/2$.

Definition 4.18 *(Mean Empirical Width). Let* $\varepsilon_1, \ldots, \varepsilon_m$ *be independent Rademacher random variables. The mean empirical width of the set* E *with respect to the random vector* φ *is*

$$W_m(E, \varphi) = \mathbb{E}\sup_{u \in E}\langle h, u\rangle, \quad h = \frac{1}{\sqrt{m}}\sum_{i=1}^{m}\varepsilon_i\varphi_i.$$

Step 1 is to apply the following remarkable result.

Theorem 4.19 *(Lower Bound for Non-negative Empirical Process [31]). For every* $\xi > 0$ *and* $t > 0$

$$s_n(\Phi, \mathcal{K}) \geq \xi\sqrt{m}\; Q_{2\xi}(E, \varphi) - 2W_m(E, \varphi) - \xi t,$$

with probability at least $1 - e^{-t^2/2}$.

Step 2 and 3 handles with the terms in the right hand side. For $Q_{2\xi}(E, \varphi)$ use

Proposition 4.20 *(Paley-Zygmund inequality). Let* $Z \geq 0$ *be a random variable with finite variance. For* $0 \leq \theta \leq 1$,

$$\mathbb{P}(Z > \theta\; \mathbb{E}Z) \geq (1 - \theta)^2\frac{\mathbb{E}[Z]^2}{\mathbb{E}[Z^2]}.$$

The proof can be found in the Appendix. For $W_m(E,\varphi)$ use Theorem 2.53 when φ is sub-gaussian. In a more general case, apply proposition below

Proposition 4.21 *(Mean Empirical Width of a Descent Cone). Fix $x \in \mathbb{R}^n$. Assume that the subdifferential $\partial f(x)$ of the function f is nonempty and does not contain the origin. For independent copies $\varphi_1, \ldots, \varphi_m$ of a random vector φ and independent Rademacher random variables $\varepsilon_1, \ldots, \varepsilon_m$,*

$$W_m^2(\mathcal{D}(f,x) \cap S^{n-1}, \varphi) \leq \mathbb{E} \inf_{\tau \geq 0} d^2(h, \tau \partial f(x)), \quad \text{where } h = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} \varepsilon_i \varphi_{\mathbf{i}}.$$

Proposition 4.21 is proved as Proposition 4.13. Theorem 4.19 requires some preparation. A random variable X is symmetric if it has the same distribution of $-X$. We use the Gine-Zinn symmetrization [24].

Example 4.22 *Some classical examples of symmetric random variables:*

1. - ***(Gaussians)*** *If $X \sim N(0,\sigma^2)$ then $-X \sim N(0,\sigma^2)$ because the probability density function of X is an even function.*
2. - ***(Rademacher)*** *If X is a Rademacher random variable, then clearly $\mathbb{P}(-X=1) = \mathbb{P}(X=-1) = \frac{1}{2}$ and $\mathbb{P}(-X=-1) = \mathbb{P}(X=1) = \frac{1}{2}$. The random variable $-X$ is also Rademacher.*

Theorem 4.23 *(Symmetrization for Empirical Process). Let $P_n\mathbf{f}$ be an empirical process indexed by $\mathcal{F}$ and let $\varepsilon_1, \ldots, \varepsilon_n$ be independent Rademacher random variables. Then*

$$\mathbb{E} \sup_{\mathbf{f} \in \mathcal{F}} |\sum_{i=1}^{n} \mathbf{f}(X_i) - \mathbb{E}\mathbf{f}(X_i)| \leq 2\mathbb{E} \sup_{\mathbf{f} \in \mathcal{F}} |\sum_{i=1}^{n} \varepsilon_i \mathbf{f}(X_i)|.$$

The expectation in the third term is taken with respect to both the Rademacher random variables and X_i, $i \in [n]$.

Notice that $Z = \sum_{i=1}^{n} \varepsilon_i \mathbf{f}(X_i)$ is symmetric. The key idea of the proof is the fact that if X is a symmetric random variable and ε is an independent Rademacher random variable, then εX has the same distribution of X.

Proof. We use $\mathbb{E}_R$ to denote the expectation with respect to the random variable R. Now draw independent copies Y_i with the same distribution of X_i. Then

$$\begin{aligned}\mathbb{E}_X \sup_{\mathbf{f} \in \mathcal{F}} |\sum_{i=1}^{n} \mathbf{f}(X_i) - \mathbb{E}_Y \mathbf{f}(Y_i)| &\leq \mathbb{E}_{X,Y} \sup_{\mathbf{f} \in \mathcal{F}} |\sum_{i=1}^{n} \mathbf{f}(X_i) - \mathbf{f}(Y_i)| \\ &= \mathbb{E}_{X,Y,\varepsilon} \sup_{\mathbf{f} \in \mathcal{F}} |\sum_{i=1}^{n} \varepsilon_i (\mathbf{f}(X_i) - \mathbf{f}(Y_i))| := F_1\end{aligned}$$

for $X = (X_1, \dots, X_n), Y = (Y_1, \dots, Y_n), \varepsilon = (\varepsilon_1, \dots, \varepsilon_n)$ and thus

$$F_1 \leq \mathbb{E}_{X,\varepsilon} \sup_{\mathbf{f}\in\mathcal{F}} |\sum_{i=1}^{n} \varepsilon_i \mathbf{f}(X_i)| + \mathbb{E}_{Y,\varepsilon} |\sum_{i=1}^{n} \varepsilon_i \mathbf{f}(Y_i)| = 2\mathbb{E} \sup_{\mathbf{f}\in\mathcal{F}} |\sum_{i=1}^{n} \varepsilon_i \mathbf{f}(X_i)|.$$

For the last step, use that X_i and Y_i are i.i.d random variables. ■

Another important tool is the Talagrand contraction principle [37].

Theorem 4.24 *(Talagrand Contraction Principle). Consider a bounded subset $T \subset \mathbb{R}^n$, independent Rademacher random variables $\varepsilon_1, .., \varepsilon_n$ and 1-Lipschitz functions $\phi_i : \mathbb{R} \to \mathbb{R}$, for $i \in [n]$. Then*

$$\mathbb{E} \sup_{t\in T} \sum_{i=1}^{n} \varepsilon_i \phi_i(t_i) \leq \mathbb{E} \sup_{t\in T} \sum_{i=1}^{n} \varepsilon_i t_i.$$

Proof. Since each function ϕ_i is continuous, the suprema do not change if we assume that T is closed. The case $n = 1$ is straightforward. Suppose by induction that the theorem is proved for $n - 1$. Now, conditioning on $\varepsilon_1, \dots, \varepsilon_{n-1}$, we prove that, for $t_1^{'} = \sum_{i=1}^{n-1} \varepsilon_i \phi_i(t_i)$,

$$\mathbb{E} \sup_{t\in T} t_1^{'} + \varepsilon_n \phi_n(t_n) \leq \mathbb{E} \sup_{t\in T} t_n^{'} + \varepsilon_n(t_n).$$

As ε_n is a Rademacher random variable, the inequality reduces to

$$\frac{1}{2} \sup_{t\in T}(t_1^{'} - \phi_n(t_n)) + \frac{1}{2} \sup_{t\in T}(t_1^{'} + \phi_n(t_n)) \leq \frac{1}{2} \sup_{t\in T}(t_1^{'} + t_n) + \frac{1}{2} \sup_{t\in T}(t_1^{'} - t_n). \quad (4\text{-}5)$$

Since T is closed, take pairs $(t_{1a}^{'}, t_{na}^{'})$ and $(t_{1b}^{'}, t_{nb}^{'})$ that achieve the supremum of each summand in the left hand side. Then

$$\sup_{t\in T}(t_1^{'} - \phi_n(t_n)) + \sup_{t\in T}(t_1^{'} + \phi_n(t_n)) = t_{1a}^{'} - \phi_n(t_{na}) + t_{1b}^{'} + \phi_n(t_{nb})$$
$$\leq \sup_{t\in T}(t_1^{'} + t_n) + \sup_{t\in T}(t_1^{'} - t_n) \text{ (by the Lipschitz hypothesis of } \phi_n).$$

Finally, take the expectation with respect to $\varepsilon_1, \dots, \varepsilon_{n-1}$ in 4-5 and apply the induction hypothesis. ■

We now prove Theorem 4.19.

Proof. We introduce a directional version of the tail function: for $u \in E, \xi > 0$,

$$Q_\xi(u) = \mathbb{P}(|\langle \varphi, u \rangle| \geq \xi).$$

By the Cauchy-Schwartz and Markov inequalities,

$$(\frac{1}{m} \sum_{i=1}^{m} |\langle \varphi_i, u \rangle|^2)^{1/2} \geq \frac{1}{m} \sum_{i=1}^{m} |\langle \varphi_i, u \rangle| \geq \frac{\xi}{m} \sum_{i=1}^{m} \mathbb{1}(|\langle \varphi_i, u \rangle| \geq \xi).$$

so that

$$\begin{aligned}&\inf_{u\in E}(\frac{1}{m}\sum_{i=1}^{m}|\langle\varphi_i,u\rangle|^2)^{1/2}\\&\geq\xi\inf_{u\in E}Q_{2\xi}(u)-\frac{\xi}{m}\sup_{u\in E}\sum_{i=1}^{m}(Q_{2\xi}(u)-\mathbb{1}(|\langle\varphi_i,u\rangle|\geq\xi)).\end{aligned}\tag{4-6}$$

The summands in the right hand side are independent and bounded by one. By Theorem 2.9, with probability at least $1-e^{-t^2/2}$,

$$\begin{aligned}&\sup_{u\in E}\sum_{i=1}^{m}(Q_{2\xi}(u)-\mathbb{1}(|\langle\varphi_i,u\rangle|\geq\xi))\\&\leq\mathbb{E}\sup_{u\in E}\sum_{i=1}^{m}(Q_{2\xi}(u)-\mathbb{1}(|\langle\varphi_i,u\rangle|\geq\xi))+t\sqrt{m},\end{aligned}\tag{4-7}$$

For $\xi>0$, we introduce the soft indicator function $\psi_\xi:\mathbb{R}\to[0,1]$,

$$\psi_\xi(s)=\begin{cases}0,|s|\leq\xi\\(|s|-\xi)/\xi,\xi<|s|\leq2\xi\\1,|s|>2\xi\end{cases}$$

There are two desired properties of the soft indicator function. Clearly, for all $s\in\mathbb{R}$ $\mathbb{1}(|s|\geq2\xi)\leq\psi_\xi(s)\leq\mathbb{1}(|s|\geq\xi)$. Moreover, $\xi\psi_\xi$ is 1-Lipschitz. The first term in the right hand of 4-7 is

$$\begin{aligned}&\mathbb{E}\sup_{u\in E}\sum_{i=1}^{m}(Q_{2\xi}(u)-\mathbb{1}(|\langle\varphi_i,u\rangle|\geq\xi))\\&=\mathbb{E}\sup_{u\in E}\sum_{i=1}^{m}\mathbb{E}\mathbb{1}(|\langle\varphi_i,u\rangle|\geq2\xi)-\mathbb{1}(|\langle\varphi_i,u\rangle|\geq\xi)\\&\leq\mathbb{E}\sup_{u\in E}\sum_{i=1}^{m}\mathbb{E}\psi_\xi(\langle\varphi,u\rangle)-\psi_\xi(\langle\varphi_i,u\rangle)\\&\leq2\mathbb{E}\sup_{u\in E}\sum_{i=1}^{m}\varepsilon_i\psi_\xi(\langle\varphi_i,u\rangle)\text{ (by Theorem 4.23)}\\&\leq\frac{2}{\xi}\mathbb{E}\sup_{u\in E}\sum_{i=1}^{m}\varepsilon_i\langle\varphi_i,u\rangle\text{ (by Theorem 4.24).}\end{aligned}$$

Combine the last inequality with 4-6 and 4-7 to obtain

$$\inf_{u\in E}(\frac{1}{m}\sum_{i=1}^{m}|\langle\varphi_i,u\rangle|^2)^{1/2}\geq\inf_{u\in E}Q_{2\xi}(u)-\frac{\xi}{m}(\frac{2}{\xi}\mathbb{E}\sup_{u\in E}\sum_{i=1}^{m}\varepsilon_i\langle\varphi_i,u\rangle+t\sqrt{m}).$$

■

We now proceed to apply the small ball method to estimate the minimum conic singular value with a mean zero sub-gaussian random vector φ. Notice

that we need some small ball assumption to guarantee that $Q_{2\xi}(E)$ is bounded away from the origin, otherwise the estimate in step 1 becomes trivial. Assume:

1. $\exists\ \alpha > 0$, $\forall u \in S^{n-1}$, $\ \alpha \leq \mathbb{E}[|\langle \varphi, u\rangle|]$ (the *small ball assumption*).
2. $\exists K > 0$, $\forall u \in S^{n-1}$, $\|\langle \varphi, u\rangle\|_{\psi_2} \leq K$.

Consider a sub-gaussian sampling matrix Φ whose rows are independent copies of φ. Recall $[a]_+ = \max\{0, a\}$.

Theorem 4.25 *(Signal Recovery with Sub-gaussian Sampling Matrix). Let φ and Φ as above. Then, for $\rho = K/\alpha$ and absolute constants $c, C > 0$, any optimal solution $\hat{x}$ to the optimization problem 4-2 satisfies*

$$\|\hat{x} - \mathbf{x}^{\#}\| \leq \frac{2\eta}{[c\alpha\rho^{-2}\sqrt{m} - CKw(\mathcal{D}(f, \mathbf{x}^{\#})) - \alpha t)]_+}.$$

The small ball method usually splits in three steps.

Proof. **(Minimum Conic Singular Value).** Apply Theorem 4.19 to obtain

$$s_n(\Phi, \mathcal{K}) \geq \xi\sqrt{m}Q_{2\xi}(E) - 2W_m(E) - \xi t, \tag{4-8}$$

with probability at least $1 - 2e^{-t^2/2}$.

(Marginal Tail Function). By Proposition 4.20,

$$\mathbb{P}(|\langle \varphi, u\rangle| \geq 2\xi) \geq \frac{\mathbb{E}[|\langle \varphi, u\rangle| - 2\xi]_+^2}{\mathbb{E}[|\langle \varphi, u\rangle|^2]}.$$

Since φ is sub-gaussian

$$\mathbb{E}[|\langle \varphi, u\rangle|^2] = \int_0^\infty 2s\mathbb{P}(|\langle \varphi, u\rangle| \geq s)ds \leq 4K^2.$$

By the small ball assumption, $\mathbb{E}[|\langle \varphi, u\rangle| - 2\xi]_+^2 \geq (\alpha - 2\xi)^2$. For any $\xi < \alpha/2$,

$$Q_{2\xi}(E) \geq \frac{(\alpha - 2\xi)^2}{4K^2}. \tag{4-9}$$

(Mean Empirical Width). By Theorem 2.53,

$$W_m(E) = \mathbb{E}\sup_{u \in E}\langle h, u\rangle \leq \mathbf{g}, u\rangle = C_1 Kw(\mathcal{K}). \tag{4-10}$$

Set $\xi = \alpha/6$. Combine 4-9 and 4-10 with inequality 4-8 to obtain

$$s_n(\Phi, \mathcal{K}) \geq C_3\frac{\alpha^3}{K^2}\sqrt{m} - C_4 Kw(\mathcal{K}) - \frac{\alpha}{6}t,$$

with probability at least $1 - 2e^{-t^2/2}$. Apply Theorem 4.7. ■

For a stable recovery from sub-gaussian Φ, the number of rows of Φ need to be at least $m \geq C\rho^6 w^2(\mathcal{D}(f, x^{\#}))$. It has the same order of the estimate for the Gaussian sampling matrix.

A natural question arises here: how restrictive is the small ball assumption? Typical examples that satisfies the small ball assumption are absolute continuous distributions with respect to the Lebesgue measure and random vector formed by independent symmetric bounded entries [60]. A general characterization still unsolved.

The extension of Theorem 4.25 for a general non sub-gaussian random vector φ is a major open problem [7]. An interesting problem in which the small ball method can be applied to non sub-gaussian φ was considered by J. Tropp [60]. It is the topic of the next chapter.

5
Phase Retrieval

In this chapter we apply the small ball method to the well known phase retrieval problem first considered in [28] [44]. The problem occurs when detectors in X-ray crystallography can only record the squared modulus of the Fresnel or Fraunhofer diffraction pattern of the radiation that is scattered from an object. The phase of the optical wave reaching the detector cannot be measure and much information about the scattered object or the optical field is lost. Similar situation occurs in the field of image processing in which the magnitude of the Fourier transforms can be directly measured but its phases are lost [47]. For a survey about the applications of the phase retrieval problem, see [13]. We follow [60]

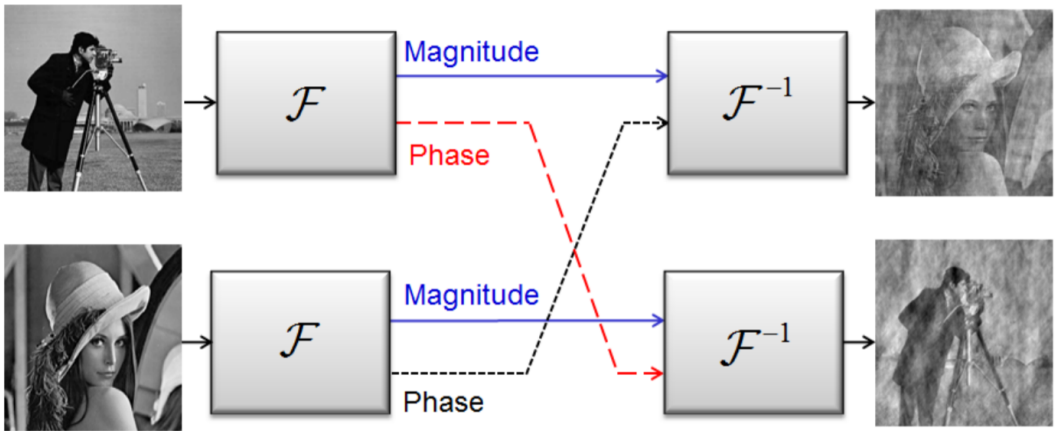

Figure 5.1: The importance of the Fourier phase in image processing [47]. Two classical images, Cameraman and Lena, are Fourier transformed, their phases are swapped and then they are inverse Fourier transformed.

5.1
Phase Retrieval Formulation via Convex Optimization

Fix a sampling ensemble $\psi_1, \ldots, \psi_m \in \mathbb{R}^n$. We want to recover a vector $\mathbf{x}^\# \in \mathbb{R}^n$ from observations

$$y_i = \langle \psi_i, \mathbf{x}^\# \rangle^2, \quad i \in [m].$$

A change of variable (the Lifting method [6]) linearizes the problem. Write

$$|\langle \psi, \mathbf{x} \rangle|^2 = \psi^t \mathbf{x}\mathbf{x}^t \psi = \mathrm{Tr}(\mathbf{x}\mathbf{x}^t \psi \psi^t),$$

and define

$$\mathbf{X}^{\#} = (\mathbf{x}^{\#})(\mathbf{x}^{\#})^t \in \mathbb{R}^{n\times n}\ ,\ \Psi_i = \psi_i\psi_i^t \in \mathbb{R}^{n\times n}, \quad i \in [m].$$

Clearly $\mathbf{x}$ and $\mathbf{y}$ induce the same matrix $\mathbf{X}^{\#}$ if and only if $\mathbf{x} = \pm\mathbf{y}$, in accordance to the fact that the observations y_i associated to both vectors are the same. Write y_i as linear functions of $\mathbf{X}^{\#}$,

$$y_i = \mathrm{Tr}(\mathbf{X}^{\#}\Psi_i), \quad i \in [m].$$

The standard approach to recover $\mathbf{x}^{\#}$ (up to sign) is to obtain the matrix $\mathbf{X}^{\#}$ as the (hopefully unique) minimizer of the optimization problem [14]

$$\min_{\mathbf{X}\in\mathbb{R}^{n\times n}} \quad \mathrm{Tr}(\mathbf{X}) \tag{5-1a}$$

$$\text{subject to} \quad \mathbf{X} \text{ positive semidefinite} \quad \text{and} \quad y_i = \mathrm{Tr}(\mathbf{X}\Psi_i). \tag{5-1b}$$

5.2 Phase Retrieval with The Gaussian Sampling Ensemble

Consider each sampling vector $\psi_i \in \mathbb{R}^n, i \in [m]$, chosen independently from a standard Gaussian distribution. The matrices $\Psi_i = \psi_i\psi_i^t$, $i \in [m]$, follows a Wishart distribution [65] and its rows are not sub-gaussian, therefore it is necessary to use Proposition 4.21 in the third step of the small ball method because Theorem 2.53 does not apply here.

Theorem 5.1 *(Phase Retrieval with Gaussian Sampling Ensemble). For absolute constants $C, c > 0$, if $m \geq Cn$, the matrix $\mathbf{X}^{\#}$ is recovered uniquely from the optimization problem 5-1 with probability at least $1 - e^{-cm}$.*

The remarkable fact about this theorem is that, with high probability, the optimal point $\mathbf{X}^{\#}$ is a matrix of rank one [14].

We need some preparation. The next lemma is adapted from [63] and [65]. Recall that matrix norms $\|A\|$ with no indices are L^2 norms, $\|A\| = \|A\|_{2\to 2}$.

Lemma 5.2 *(Covariance estimation). Let $\mathbf{X} \in \mathbb{R}^n$ be a mean zero Gaussian vector with covariance matrix Σ. Let $\psi_i \in \mathbb{R}^n$, $i \in [m]$ be independent copies of $\mathbf{X}$. Set $\Sigma_m = \frac{1}{m}\sum_{i=1}^m \psi_i\psi_i^t$. Then, for an absolute constant $C > 0$ and for every $u > 0$,*

$$\|\Sigma_m - \Sigma\| \leq C\left(\sqrt{\frac{n+u}{m}} + \frac{n+u}{m}\right)\|\Sigma\|,$$

with probability at least $1 - 2e^{-u}$.

Proof. Change variables to convert the vectors X and ψ_i into isotropic vectors,

$$Z = \Sigma^{-1/2}\mathbf{X}, \quad Z_i = \Sigma^{-1/2}\psi_i.$$

Now write

$$\|\Sigma_m - \Sigma\| = \|\Sigma^{1/2} R_m \Sigma^{1/2}\| \leq \|R_m\|\|\Sigma\|, \quad \text{where } R_m = \frac{1}{m}\sum_{i=1}^{m} Z_i Z_i^t - I_n.$$

Consider the $m \times n$ random matrix A whose rows are Z_i^t. Then

$$\frac{1}{m} A^t A - I_n = R_m.$$

Apply Theorem 2.35 to A. ■

Theorem 4.23 extends for random matrices [61] with the same proof.

Theorem 5.3 *(Symmetrization for Matrices). Let $A_1, \ldots, A_n$ be independent random matrices of the same size and let $\varepsilon_1, \ldots, \varepsilon_n$ be independent Rademacher random variables. Then, for every $p \geq 1$,*

$$\frac{1}{2}(\mathbb{E}\|\sum_{i=1}^{n} \varepsilon_i A_i\|^p)^{1/p} \leq (\mathbb{E}\|\sum_{i=1}^{n} A_i - \mathbb{E}A_i\|^p)^{1/p} \leq 2(\mathbb{E}\|\sum_{i=1}^{n} \varepsilon_i A_i\|^p)^{1/p}.$$

We use some results from Chapter 4 with a systematic alteration: vectors and matrices will be replaced by matrices and linear maps between spaces of matrices. Thus, new statements about a matrix $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2}$ is derived by interpreting it as a vector $\mathbf{X}' \in \mathbb{R}^{n_1 n_2}$.

Similarly, a convex proper function $f : \mathbb{R}^{n_1 \times n_2} \to \overline{\mathbb{R}}$ and observations $y = \Lambda X + \mathbf{e}$ with $e \in \mathbb{R}^m$ and $\Lambda : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^m$ for $\|\mathbf{e}\|_2 \leq \eta$, induce an optimization problem analogous to the problem 4-2,

$$\min_{\mathbf{X}} \quad f(\mathbf{X}) \tag{5-2a}$$

$$\text{subject to} \quad \|\Lambda\mathbf{X} - \mathbf{y}\|_2 \leq \eta. \tag{5-2b}$$

Again, *we assume the existence of a solution $\mathbf{X}^{\#} \in \mathbb{R}^{n_1 n_2}$ to the problem.*

Definition 5.4 *(Descent Cone for Matrices). The descent cone $\mathcal{D}(f, \mathbf{X})$ of f at a matrix $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2}$ is*

$$\mathcal{D}(f, \mathbf{X}) = \bigcup_{\tau > 0} \{\mathbf{U} \in \mathbb{R}^{n_1 \times n_2} : f(\mathbf{X} + \tau\mathbf{U}) \leq f(\mathbf{X})\}.$$

Recall that $\|.\|_F$ is the Frobenius norm for matrices, induced by the inner product $\langle A, B\rangle = \mathrm{Tr}(B^t A)$, for $A, B \in \mathbb{R}^{n_1 \times n_2}$. The unit sphere is

$$S_F^{n_1 \times n_2 - 1} = \{U \in \mathbb{R}^{n_1 \times n_2} : \|U\|_F = 1\} .$$

Definition 5.5 *(Conic singular values). Let $\mathcal{K} \in \mathbb{R}^{n_1 \times n_2}$ be a cone. The minimum conic singular value of the linear map $\Lambda : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^m$ is*

$$s_n(\Lambda, \mathcal{K}) = \inf_{U \in \mathcal{K} \cap S_F^{n_1 \times n_2 - 1}} \|\Lambda U\|_F.$$

Theorem 5.6 *(A deterministic error bound). Any optimal solution $\hat{X}$ of the optimization problem 5-2 satisfies*

$$\|\hat{X} - \mathbf{X}^{\#}\|_F \leq \frac{2\eta}{s_n(\Lambda, \mathcal{D}(f, \mathbf{X}^{\#}))}.$$

Consider a cone $\mathcal{K} \in \mathbb{R}^{n_1 \times n_2}$ and set $E = S_F^{n_1 \times n_2 - 1} \cap \mathcal{K}$.

Definition 5.7 *(Marginal Tail function for Matrices). The marginal tail function of the set E with respect a random matrix Ψ is*

$$Q_\xi(E, \Psi) = \inf_{U \in E} \mathbb{P}(|\langle \Psi, U\rangle| \geq \xi).$$

Definition 5.8 *(Mean Empirical Width for Matrices). Let $\varepsilon_1, \ldots, \varepsilon_m$ be independent Rademacher random variables, the mean empirical width of the set E with respect to the random matrix Ψ is*

$$W_m(E, \Psi) = \mathbb{E} \sup_{U \in E} \langle H, U\rangle, \quad \textit{where } H = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} \varepsilon_i \Psi_i.$$

Consider independent copies $\Psi_1, \ldots, \Psi_m$ of the matrix Ψ and a random linear map of observations $\Lambda : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^m$ with $[\Lambda(X)]_i = \langle \Psi_i, X\rangle$.

Theorem 5.9 *(Theorem 4.19 for Matrices). For every $\xi > 0$ and $t > 0$,*

$$s_n(\Lambda, \mathcal{K}) \geq \xi\sqrt{m} Q_{2\xi}(E, \Psi) - 2W_m(E, \Psi) - \xi t,$$

with probability at least $1 - e^{-t^2/2}$.

Here is another definition converted from vectors to matrix functions.

Definition 5.10 *(Subdifferential for Matrix Functions). The subdifferential of the proper convex function $f : \mathbb{R}^{n_1 \times n_2} \to \overline{\mathbb{R}}$ at $X \in \mathbb{R}^{n_1 \times n_2}$ is*

$$\partial f(X) = \{V \in \mathbb{R}^{n_1 \times n_2} : f(Y) \geq f(X) + \langle V, Y - X\rangle, \textit{ for all } Y \in \mathbb{R}^{n_1 \times n_2}\}.$$

Denote by $d_F(\mathcal{K}, X)$ the Frobenius distance between the cone $\mathcal{K}$ and X.

Proposition 5.11 *(Mean empirical width of a Descent Cone for Matrices). Fix $X \in \mathbb{R}^{n_1 \times n_2}$. Assume that the subdifferential $\partial f(X)$ is nonempty and does not contain the origin. For independent copies $\Psi_1, \ldots, \Psi_m$ of a random matrix Ψ and independent Rademacher random variables $\varepsilon_1, \ldots, \varepsilon_m$,*

$$W_m^2(\mathcal{D}(f, \mathbf{U}^{\#}) \cap S_F^{n_1 \times n_2 - 1}, \mathbf{\Psi}) \leq \mathbb{E} \inf_{\tau \geq 0} d_F^2(H, \tau \partial f(x)), \quad \textit{where } H = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} \varepsilon_i \mathbf{\Psi_i}.$$

We prove Theorem 5.1. Equip the space of symmetric matrices $\mathbb{R}_{sym}^{n \times n}$ with the Frobenius norm. Let $\lambda_{\max}(Z)$ be the maximum eigenvalue $Z \in \mathbb{R}_{sym}^{n \times n}$. Write δ_{ij} for the matrix whose only nonzero entry, equal to 1, has indices (i, j).

Proof. We rewrite the convex optimization program 5-1. Let

$$\Phi : \mathbb{R}_{sym}^{n \times n} :\to \mathbb{R}^m, \quad [\Phi(\mathbf{X})]_i = \operatorname{Tr}(\Psi_i \mathbf{X}), \quad i \in [m].$$

Let $i : \mathbb{R}_{sym}^{n \times n} \to \overline{\mathbb{R}}$ be the penalty indicator function which is zero at positive semidefinite matrices and infinity otherwise. Consider the convex proper function $f : \mathbb{R}_{sym}^{n \times n} \to \mathbb{R}$,

$$f(\mathbf{X}) = i(\mathbf{X}) + \operatorname{Tr}(\mathbf{X}).$$

The optimization problem 5-1 becomes

$$\min_{\mathbf{X}} \quad f(\mathbf{X}) \tag{5-3a}$$

$$\text{subject to} \quad y = \Phi(\mathbf{X}). \tag{5-3b}$$

This formulation matches the optimization problem 5-2. By Theorem 5.6, there exists an unique solution to 5-1 if the minimum conic singular value $s_n(\Phi, \mathcal{D}(f, \mathbf{X}^{\#})) > 0$ (notice there is no error vector, $\eta = 0$). We now argue as in the small ball method.

(Minimum conic singular value). Let

$$E = \{U \in \mathcal{D}(f, \mathbf{X}^{\#}) | \ \|U\|_F = 1\}.$$

From Theorem 5.9,

$$s_n(\Phi, \mathcal{D}(f, \mathbf{X}^{\#})) \geq \xi \sqrt{m} Q_{2\xi}(E) - 2W_m(E) - \xi t, \tag{5-4}$$

with probability at least $1 - e^{-t^2/2}$.

(Marginal Tail function). By Proposition 4.20,

$$\mathbb{P}(|\langle \Psi_1, U\rangle|^2) \geq \frac{1}{2}\mathbb{E}(|\langle \Psi_1, U\rangle|^2) \geq \frac{1}{4}\frac{\mathbb{E}[|\langle \Psi_1, U\rangle|^2]^2}{\mathbb{E}[|\langle \Psi_1, U\rangle|^4]}.$$

Apply the Gaussian hypercontractivity inequality (see Appendix) with $q = 4$,

$$\mathbb{E}(|\langle \Psi_1, U\rangle|^4)^{1/4} \leq 3\mathbb{E}(|\langle \Psi_1, U\rangle|^2)^{1/2}.$$

Since U is symmetric and $\|U\|_F = 1$,

$$\mathbb{E}(|\langle \Psi_1, U\rangle|^2) = 3\sum_{i=1}^{m}|u_{ii}|^2 + 2\sum_{i,j=1}^{m}|u_{ij}|^2 + |\sum_{i=1}^{m}u_{ii}|^2 \geq 2$$

Hence, there exists an absolute constant $c > 0$ such that, for every $U \in E$,

$$\mathbb{P}(|\langle \Psi_1, U\rangle|^2 \geq 1) \geq c. \tag{5-5}$$

(Mean Empirical Width). Recall that $\mathbf{X}^{\#} \in \mathbb{R}^{n\times n}_{sym}$ is a rank one, positive semidefinite matrix. Select a basis in $\mathbb{R}^n$ in which $\mathbf{X}^{\#}$ is

$$\mathbf{X}^{\#} = \begin{vmatrix} a & \mathbf{0}^t \\ \mathbf{0} & \mathbf{0} \end{vmatrix},$$

with $a > 0$. In the same basis, H is

$$H = \begin{vmatrix} h_{11} & h_{21}^t \\ h_{21} & H_{22} \end{vmatrix}.$$

By Proposition 5.11, for every $\tau \geq 0$,

$$W_m(E) \leq \mathbb{E}\,[d_F^2(H, \tau\partial f(\mathbf{X}^{\#}))]^{1/2}.$$

We claim that the subdifferential $\partial f(\mathbf{X}^{\#})$ is

$$\partial f(\mathbf{X}^{\#}) = \left\{ \begin{vmatrix} 1 & 0^t \\ 0 & Y \end{vmatrix} \in \mathbb{R}^{n\times n}_{sym} |\ \lambda_{max}(Y) \leq 1 \right\}. \tag{5-6}$$

It is enough to prove that any matrix $V \in \mathbb{R}^{n\times n}_{sym}$ satisfying

$$\mathrm{Tr}(Z) \geq \mathrm{Tr}(\mathbf{X}^{\#}) + \langle V, Z - \mathbf{X}^{\#}\rangle, \tag{5-7}$$

for every positive semidefinite $Z \in \mathbb{R}^{n\times n}_{sym}$, assumes the form described in 5-6. Denote the entries of Z by z_{ij}. Use $Z = z_{11}\delta_{11}$, $z_{11} > 0$, in 5-7 to obtain

$$z_{11}(1 - v_{11}) \geq a(1 - v_{11}).$$

For $z_{11} = 2a$ and $z_{11} = \frac{a}{2}$, we obtain $(1-v_{11}) \geq 0$ and $(1-v_{11}) \leq 0$ respectively, so that $v_{11} = 1$. The proof of $v_{1j} = 0$ for all $j \geq 2$ and $v_{i1} = 0$ for all $i \geq 2$ is similar. Now let Z be of the form

$$Z = \begin{vmatrix} 0 & \mathbf{0}^t \\ \mathbf{0} & Z_{22} \end{vmatrix}.$$

Let $V_{22} \in \mathbb{R}^{(n-1)\times(n-1)}_{sym}$ be the sub-matrix of V obtained by deleting the first row and the first column of V. From inequality 5-7,

$$\mathrm{Tr}(Z_{22}) \geq \mathrm{Tr}(Z_{22}V_{22}).$$

To check that $\lambda_{\max}(V_{22}) \leq 1$, consider the decomposition associated with the spectral theorem, $V_{22} = QDQ^t$, for an orthogonal matrix Q and a real diagonal matrix D. Set $Z_{22} = Q^t\delta_{ii}Q$ with $i \in [n]$ to obtain that $D_{ii} \leq 1$ for all $i \in [n]$ and then $\lambda_{\max}(V_{22}) \leq 1$. For $\tau = \lambda_{max}(H_{22})$:

$$\mathbb{E}d_F^2(H, \tau\partial f(\mathbf{X}^{\#})) = \mathbb{E}(h_{11}-\tau)^2 + 2\mathbb{E}\|h_{21}\|^2 + \mathbb{E}\inf_{\lambda_{\max}(S)\leq 1}\|H_{22}-\tau Y\|_F^2. \quad (5\text{-}8)$$

The third term in the right hand side is zero by construction. By a direct calculation, the second is $2(n-1)$. For the first,

$$\begin{aligned}\tau = \sup_{y\in S^{n-2}}\langle H_{22}y, y\rangle = \sup_{y\in S^{n-2}}[0\ y]H[0\ y]^t &\leq \sup_{x\in S^{n-1}}\langle Hx, x\rangle = \lambda_{\max}(H) \\ &= \frac{1}{\sqrt{m}}\lambda_{max}(\sum_{i=1}^{m}\varepsilon_i\psi_i\psi_i^t).\end{aligned}$$

For $m \geq Cn$, Lemma 5.2 gives

$$\mathbb{P}(\frac{1}{\sqrt{m}}\lambda_{max}(\sum_{i=1}^{m}\psi_i\psi_i^t)) > C_1\sqrt{n}) \leq e^{-c_1 n}.$$

By Theorem 5.3 and integral identity,

$$\mathbb{E}[\tau^2] \leq C_2\ \mathbb{E}[\|\frac{1}{\sqrt{m}}\sum_{i=1}^{m}\psi_i\psi_i^t\|^2] \leq C_3 n.$$

Combine the estimates for the three terms in 5-8 to obtain

$$W_m(E) \leq C_4\sqrt{n}. \quad (5\text{-}9)$$

Substitute 5-5 and 5-9 into 5-4 to obtain

$$s_n(\Phi, \mathcal{D}(f, \mathbf{X}^{\#})) \geq C_5\sqrt{m} - C_6\sqrt{n} - \frac{t}{2},$$

with probability at least $1 - e^{t^2/2}$. Set $t = C_7\sqrt{m}$ and choose m such that the minimum conic singular value $s_n(\Phi, \mathcal{D}(f, \mathbf{X}^{\#}))$ is positive with the required probability. ■

It can be shown that this result is optimal up to absolute constants [14].

6
Conclusion

In this dissertation we explored the powerful ideas of the small ball method and some of its connections to non-asymptotic random matrix theory. As an application, we analyzed the performance of convex optimization programs to reconstruct signals, in particular the phase retrieval problem with Gaussian sampling ensemble.

7
Appendix

7.1
Basic Probability Theory

Theorem 7.1 *(Markov Inequality [63]). 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 7.2 *(Chebyshev Inequality [63]). 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 7.3 *(Integral identity/Layer cake representation [63]). 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 7.4 *(Jensen Inequality [63]). Let X be a random variable and $\phi : \mathbb{R} \to \mathbb{R}$ be a convex function. Then*

$$\phi(\mathbb{E}X) \leq \mathbb{E}\phi(X).$$

Definition 7.5 *(Chi-Square Distribution [67]). A random variable X follows a Chi-Square distribution of degree n if its probability density function is*

$$u \geq 0, \quad \phi_n(u) = \frac{1}{2^{n/2}\Gamma(n/2)} u^{(n/2)-1} e^{-u/2}.$$

Theorem 7.6 *(Characterization of the Chi-Square Distribution [67]). Let $g_1, .., g_n$ be a sequence of standard Gaussian variables. Then the random variable $X = \sum_{i=1}^n g_i^2$ has the probability density function of a chi-square distribution of degree n.*

Definition 7.7 *(Beta Distribution [67]). A random variable X follows a Beta distribution with parameters a, b if it has the probability density function*

$$x \in [0, 1], \quad p_{a,b}(x) = \beta(a, b)^{-1} x^{a-1} (1 - x)^{b-1}.$$

Here the constant $\beta(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ is the Beta number.

Theorem 7.8 *(Beta Distribution as ratio of chi-squares [66]). Let X be a random variable following a chi-square distribution of degree m and let Y be a random variable following a chi-square distribution of degree n independent from X. Then the ration $\frac{X}{X+Y}$ follows a Beta distribution with parameters $a = \frac{m}{2}$ and $b = \frac{n}{2}$.*

7.2
Proof of the Paley-Zygmund Inequality

Here we prove Proposition 4.20.

Proof. Write

$$\mathbb{E}Z = \mathbb{E}[Z\mathbb{1}(Z \leq \theta\ \mathbb{E}Z)] + \mathbb{E}[Z\mathbb{1}(Z > \theta\ \mathbb{E}Z)].$$

The first summand is at most $\theta\ \mathbb{E}Z$. By the Cauchy-Schwartz inequality $\mathbb{E}[Z\mathbb{1}(Z > \theta\mathbb{E}Z)] \leq \mathbb{E}[Z^2]^{1/2}\ \mathbb{P}(Z > \theta\mathbb{E}Z)^{1/2}$. ■

7.3
Proof of the Gaussian Hypercontractivity Inequality

Here we prove the following result.

Theorem 7.9 *(Gaussian Hypercontractivity Inequality [9]). Let $f(t) = \sum_{i=0}^{k} a_i t^k$ be a polynomial of degree k and let $\mathbf{g}$ be a standard Gaussian variable. Then, for $q > 2$,*

$$\|f(\mathbf{g})\|_q \leq (q-1)^{k/2}\|f(\mathbf{g})\|_2.$$

We need some preparation. The functions $f : \{-1, +1\}^n \to \mathbb{R}$, with the obvious operation, give rise to a real vector space V of dimension 2^n, in which we define the inner product

$$\langle f, g\rangle = 2^{-n} \sum_{x\in\{-1,+1\}^n} f(x)g(x).$$

An orthonormal basis is defined by the monomials $u_S(x) = \prod_{i\in S} x_i$ (as usual, $\prod_{i\in\emptyset} x_i = 1$). The orthonormal expansion of f in this basis becomes

$$f(x) = \sum_{S\subset[n]} \alpha_S u_S(x), \quad \alpha_S = \langle f, u_S\rangle,$$

where the summation runs over all subsets S of $[n]$. Denote the cardinality of S by $|S|$. For $1 \leq p < \infty$, define the norm $\|f\|_{L^p} = (2^{-n} \sum_{x\in\{-1,+1\}^n} f(x)^p)^{1/p}$. For $\gamma > 0$, consider the linear transformation $T_\gamma : L^q \to L^p$ that maps

$f(x) = \sum_{S\subset[n]} \alpha_S u_S(x)$ to

$$(T_\gamma f)(x) = \sum_{S\subset[n]} \gamma^{|S|}\alpha_S u_S(x).$$

Theorem 7.10 *(Bonami-Beckner Inequality). Let $2 < q < \infty$. Then*

$$\|f\|_{L^q} \leq \|T_\gamma f\|_{L^2}, \quad \textit{for } \gamma = \sqrt{q-1}.$$

For the proof and the application below, see [9]. For a random variable X recall, the norm $\|X\|_p = (\mathbb{E}|X|^p)^{1/p}$. We now prove Theorem 7.9.
Proof. Let the entries of a random vector $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$ be independent Rademacher variables. Use the central limit theorem to approximate a Gaussian by a sum of ε_i: it suffices to show that, for $n \geq 1$,

$$\|f(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i)\|_q \leq (q-1)^{k/2}\|f(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i)\|_2.$$

Set $\tilde{f}(\varepsilon_1, \ldots, \varepsilon_n) = f(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i)$. Use Theorem 7.10 for $\tilde{f}$, $\gamma = \sqrt{q-1} > 1$:

$$\|\tilde{f}\|_{L^q}^2 \leq \|T_\gamma \tilde{f}\|_{L^2}^2 = \sum_{S\subset[n]:|S|\leq k} \alpha_S^2\gamma^{2|S|} \leq \gamma^{2k}\sum_{S\subset[n]:|S|\leq k}\alpha_S^2 = \gamma^{2k}\|\tilde{f}\|_{L^2}^2.$$

Now use that $\|\tilde{f}\|_{L^p} = \|\tilde{f}\|_p$. ■

7.4 Functional Analysis

Denote $C_o^\infty(\mathbb{R}^n)$ by the space of all real valued smooth functions with compact support in $\mathbb{R}^n$ and $\text{supp}(\rho_n)$ by the support set of the function ρ_n.

Definition 7.11 *(Mollifiers [11]). A sequence of compactly supported mollifiers in $\mathbb{R}^n$ is any sequence of functions $\rho_n : \mathbb{R}^n \to \mathbb{R}$ satisfying*

$$\rho_n \in C_o^\infty(\mathbb{R}^n), \quad supp(\rho_n) \subset \overline{(B(0, \frac{1}{n}))}, \quad \rho_n \geq 0, \quad ||\rho_n||_{L^1} = 1.$$

A sequence of functions $\{u_k\}_{k\in\mathbb{N}} \in C_o^\infty(\mathbb{R}^n)$ converges sequentially to a function $u \in C_o^\infty(\mathbb{R}^n)$ if the support of every function u_k is contained in the same compact set K and all partial derivatives of u_k converges uniformly to the respective partial derivatives of u. Denote this convergence by $u_k \to u$.

Definition 7.12 *(Distributions [32]). A distribution l is a linear functional of the space $C_o^\infty(\mathbb{R}^n)$ which is continuous with respect to the sequential convergence: $u_k \to u$ implies that $l(u_k) \to l(u)$.*

Definition 7.13 *(Distributional Derivative [22]). The distributional derivative of a distribution l with respect to the coordinate x_i is the continuous linear functional*

$$\frac{\partial l}{\partial x_i}(\phi) = -l(\frac{\partial \phi}{\partial x_i}), \quad \textit{for all } \phi \in C_o^\infty(\mathbb{R}^n).$$

7.5 Convex Analysis

Theorem 7.14 *(Inf-Sup Inequality [10]). For a function $f : A \times B \to \mathbb{R}$,*

$$\sup_{a\in A} \inf_{b\in B} f(a,b) \leq \inf_{b\in B} \sup_{a\in A} f(a,b).$$

8
References

[1] AIZENMAN, M.; WARZEL, S.. **Random Operators**, volumen 168. American Mathematical Society, 2015. 1.1

[2] AZUMA, K.. **Weighted sums of certain dependent random variables**. Tohoku Mathematical Journal, Second Series, 19:357–367, 1967. 2.10

[3] BAI, Z. D.; YIN, Y. Q.. **Limit of the smallest eigenvalue of a large dimensional sample covariance matrix**. The Annals of Probability, 21:1275–1294, 1993. 1.1

[4] BAIK, J.; DEIFT, P. ; SUIDAN, T.. **Combinatorics and Random Matrix Theory**, volumen 172. American Mathematical Society, 2016. 1.1

[5] BAIK, J.; KRIECHERBAUER, T.; LI, L.-C.; MCLAUGHLIN, K. D.-R. ; TOMEI, C.. **Integrable Systems and Random Matrices: In honor of Percy Deift**, volumen 458. American Mathematical Society, 2008. 1.1

[6] BALAN, R.; BODMANN, B. G.; CASAZZA, P. G. ; EDIDIN, D.. **Painless reconstruction from magnitudes of frame coefficients**. Journal of Fourier Analysis and Applications, 15:488–501, 2009. 5.1

[7] BANERJEE, A.; CHEN, S. ; SIVAKUMAR, V.. **Open problem: Restricted eigenvalue condition for heavy tailed designs**. In: PROCEEDINGS OF THE 28TH CONFERENCE ON LEARNING THEORY, volumen 40 de **Proceedings of Machine Learning Research**, p. 1752–1755, 2015. 4.2

[8] BORELL, C.. **The Brunn-Minkowski inequality in the Gauss space**. Inventiones mathematicae, 30(2):207–216, 1975. 3.3, 3.3

[9] BOUCHERON, S.; LUGOSI, G. ; MASSART, P.. **Concentration inequalities: A Nonasymptotic Theory of Independence**. Oxford university press, 2013. 1.1, 2.1, 2.1, 3.3, 7.9, 7.3

[10] BOYD, S.; VANDENBERGHE, L.. **Convex Optimization**. Cambridge university press, 2004. 7.14

[11] BREZIS, H.. **Functional analysis, Sobolev spaces and Partial Differential Equations**. Springer Science & Business Media, 2010. 7.11

[12] BÜHLMANN, P.; VAN DE GEER, S.. **Statistics for High-Dimensional Data: Methods, Theory and Applications**. Springer Science & Business Media, 2011. 1.1, 4.2, 4.3

[13] CANDÈS, E. J.; ELDAR, Y. C.; STROHMER, T. ; VORONINSKI, V.. **Phase retrieval via matrix completion**. SIAM Journal on Imaging Sciences, 6:199–225, 2013. 5

[14] CANDÈS, E. J.; LI, X.. **Solving quadratic equations via Phaselift when there are about as many equations as unknowns**. Foundations of Computational Mathematics, 14:1017–1026, 2014. 5.1, 5.2, 5.2

[15] CHUNG, F. R.; GRAHAM, F. C.. **Spectral Graph Theory**, volumen 92. American Mathematical Society, 1997. 1.1

[16] DAVIDSON, K. R.; SZAREK, S. J.. **Local operator theory, random matrices and Banach spaces**. In: HANDBOOK OF THE GEOMETRY OF BANACH SPACES, p. 67–101. Springer, 2015. 1.2, 1.1

[17] DEIFT, P.. **Orthogonal Polynomials and Random Matrices: A Riemann-Hilbert Approach**, volumen 3. American Mathematical Society, 1999. 1.1

[18] DURRETT, R.. **Probability: Theory and Examples**, volumen 49. Cambridge university press, 2019. 2.3

[19] DUDLEY, R.. **The sizes of compact subsets of Hilbert space and continuity of Gaussian processes**. Journal of Functional Analysis, 1:290–330, 1967. 2.4

[20] FIGIEL, T.; LINDENSTRAUSS, J. ; MILMAN, V. D.. **The dimension of almost spherical sections of convex bodies**. Acta Mathematica., 139:53–94, 1977. 3.3

[21] FERNIQUE, X.. **Regularité des trajectoires des fonctions aléatoires Gaussiennes**. In: ECOLE D'ETÉ DE PROBABILITÉS DE SAINT-FLOUR IV, p. 1–96. Springer, 1975. 3.1

[22] FOUCART, S.; RAUHUT, H.. **A Mathematical Introduction to Compressive Sensing**. Applied and Numerical Harmonic Analysis. Birkhäuser/Springer, New York, 2013. 1.1, 2.4, 3.1, 3.2, 3.2, 3.16, 3.3, 4.2, 4.3, 7.13

[23] GINÉ, E.; NICKL, R.. **Mathematical Foundations of Infinite-Dimensional Statistical Models**, volumen 40. Cambridge University Press, 2016. 3.4, 3.3

[24] GINÉ, E.; ZINN, J.. **Some limit theorems for empirical processes**. The Annals of Probability, 12:929–989, 1984. 4.2

[25] GORDON, Y.. **Some inequalities for gaussian processes and applications**. Israel Journal of Mathematics, 50:265–289, 1985. 3.1

[26] GORDON, Y.. **On Milman's inequality and random subspaces which escape through a mesh in $\mathbb{R}^n$**. Geometric Aspects of Functional Analysis, p. 84–106, 1988. 3.15

[27] GROMOV, M.. **Paul levy's isoperimetric inequality**. preprint IHES, 1980. 3.3

[28] HARRISON, R. W.. **Phase problem in crystallography**. Journal of the Optical Society of America, 10:1046–1055, 1993. 5

[29] HOEFFDING, W.. **Probability inequalities for sums of bounded random variables**. Journal of the American Statistical Association, 58:13–30, 1963. 2.8

[30] KABANAVA, M.; RAUHUT, H. ; TERSTIEGE, U.. **Analysis of low rank matrix recovery via mendelson's small ball method**. In: 2015 INTERNATIONAL CONFERENCE ON SAMPLING THEORY AND APPLICATIONS, p. 387–391. IEEE, 2015. 1.2

[31] KOLTCHINSKII, V.; MENDELSON, S.. **Bounding the smallest singular value of a random matrix without concentration**. International Mathematics Research Notices, 2015:12991–13008, 2015. 1.2, 4.2, 4.2, 4.19

[32] LAX, P. D.. **Functional Analysis**. Pure and Applied Mathematics. Wiley-Interscience, 2002. 7.12

[33] LECUÉ, G.; MENDELSON, S.. **Regularization and the small-ball method I: sparse recovery**. The Annals of Statistics, 46:611–641, 2018. 1.2

[34] LECUÉ, G.; MENDELSON, S.. **Regularization and the small-ball method II: complexity dependent error rates**. The Journal of Machine Learning Research, 18:5356–5403, 2017. 1.2

[35] LEDOUX, M.. **The Concentration of Measure Phenomenon**. American Mathematical Society, 2001. 3.3

[36] LEDOUX, M.. **Isoperimetry and Gaussian analysis**. In: LECTURES ON PROBABILITY THEORY AND STATISTICS, p. 165–294. Springer, 1996. 3.3, 3.3, 3.3

[37] LEDOUX, M.; TALAGRAND, M.. **Probability in Banach Spaces: Isoperimetry and Processes**. Springer Science & Business Media, 2013. 3.2, 3.1, 3.2, 3.3, 4.2

[38] LÉVY, P.. **Leçons d'Analyse Fonctionnelle**. BiblioBazaar, 2009. 3.3

[39] MCDIARMID, C.. **Concentration, probabilistic methods for algorithmic discrete mathematics**. Algorithms and Combinatorics, 16:195–248, 1998. 2.1

[40] MEHTA, M. L.. **Random Matrices**, volumen 142. Elsevier, 2004. 1.1

[41] MENDELSON, S.. **A remark on the diameter of random sections of convex bodies**. In: GEOMETRIC ASPECTS OF FUNCTIONAL ANALYSIS, p. 395–404. Springer, 2014. 1.2

[42] MENDELSON, S.. **Learning without concentration**. Journal of the Association for Computing Machinery, 62:21:1–21:25, 2015. 1.2

[43] MENDELSON, S.. **Learning without concentration for general loss functions**. Probability Theory and Related Fields, 171:459–502, 2018. 1.2

[44] MILLANE, R. P.. **Phase retrieval in crystallography and optics**. Journal of the Optical Society of America, 7:394–411, 1990. 5

[45] ROCKAFELLAR, R. T.. **Convex Analysis**, volumen 28. Princeton university press, 1970. 4.1, 4.1

[46] RUDNICK, Z.; SARNAK, P.. **Zeros of principal L-functions and random matrix theory**. Duke Mathematical Journal, 81:269–322, 1996. 1.1

[47] SHECHTMAN, Y.; ELDAR, Y. C.; COHEN, O.; CHAPMAN, H. N.; MIAO, J. ; SEGEV, M.. **Phase retrieval with application to optical imaging: A contemporary overview**. IEEE Signal Processing Magazine, 32:87–109, 2015. 5, 5.1

[48] SHEVTSOVA, I.. **On the absolute constants in the Berry-Esseen-type inequalities**. Doklady Mathematics, 89:378–381, 2014. 2.1

[49] SLEPIAN, D.. **The one-sided barrier problem for Gaussian noise**. Bell System Technical Journal, 41:463–501, 1962. 3.1

[50] SUDAKOV, V. N.. **Extremal properties of half-spaces for spherically invariant measures**. Journal of Soviet Mathematics, 41:14–24, 1978. 3.3, 3.3

[51] SUDAKOV, V. N.. **Gaussian measures, Cauchy measures and epsilon-entropy**. Doklady Mathematics, 10:310–13, 1969. 3.2

[52] SUDAKOV, V. N.. **Gaussian random processes and measures of solid angles in Hilbert space**. Doklady Akademii Nauk SSSR, 197:43–45, 1971. 3.1

[53] SUDAKOV, V. N.. **Geometric Problems in the Theory of Infinite-Dimensional Probability distributions**, volumen 141. American Mathematical Society, 1979. 3.1

[54] TALAGRAND, M.. **Majorizing measures: the generic chaining**. The Annals of Probability, 24:1049–1103, 1996. 2.4

[55] TALAGRAND, M.. **Majorizing measures without measures**. The Annals of Probability, 29:411–417, 2001. 2.4

[56] TALAGRAND, M.. **Regularity of Gaussian processes**. Acta Mathematica., 159:99–149, 1987. 2.4

[57] TALAGRAND, M.. **Upper and Lower bounds for Stochastic Processes: Modern Methods and Classical Problems**, volumen 60. Springer Science & Business Media, 2014. 2.49, 2.4, 2.51, 2.4

[58] TAO, T.; VU, V. H.. **Additive Combinatorics**, volumen 105. Cambridge University Press, 2006. 1.1

[59] TROPP, J. A.. **An Introduction to Matrix Concentration Inequalities**, volumen 8. Now Publishers, 2015. 1.1

[60] TROPP, J. A.. **Convex recovery of a structured signal from independent random linear measurements**. In: SAMPLING THEORY, A RENAISSANCE, p. 67–101. Springer, 2015. 1.3, 4, 4.2, 4.2, 5

[61] TROPP, J. A.. **The expected norm of a sum of independent random matrices: an elementary approach**. In: HIGH DIMENSIONAL PROBABILITY VII, p. 173–202. Springer, 2016. 5.2

[62] VERSHYNIN, R.. **Introduction to the non-asymptotic analysis of random matrices**, p. 210–268. Cambridge University Press, 2012. 2.3

[63] VERSHYNIN, R.. **High-Dimensional Probability: An Introduction with Applications in Data Science**. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018. 1.1, 2.1, 2.1, 2.12, 2.2, 2.3, 2.4, 2.4, 2.53, 3.2, 3.2, 3.3, 3.3, 3.3, 4.1, 5.2, 7.1, 7.2, 7.3, 7.4

[64] VERSHYNIN, R.. **Lectures in geometric functional analysis**. Unpublished manuscript. Available at http://www-personal. umich. edu/romanv/papers/GFA-book/GFA-book. pdf, 2011. 3.3, 3.3, 4.1

[65] WAINWRIGHT, M. J.. **High-Dimensional Statistics: A Non-asymptotic Viewpoint**, volumen 48. Cambridge University Press, 2019. 1.1, 5.2, 5.2

[66] WALCK, C.. **Handbook on statistical distributions for experimentalists**. In:report, 1996. 7.8

[67] WASSERMAN, L.. **All of Statistics: A Concise Course in Statistical Inference**. Springer Science & Business Media, 2013. 7.5, 7.6, 7.7

[68] WIGNER, E. P.. **On the statistical distribution of the widths and spacings of nuclear resonance levels**. Mathematical Proceedings of the Cambridge Philosophical Society, 47:790–798, 1951. 1.1

[69] WISHART, J.. **The generalised product moment distribution in samples from a normal multivariate population**. Biometrika, 20:32–52, 1928. 1.1