

MA3K0 - High-Dimensional Probability

Lecture Notes

Stefan Adams

2022, update 16.10.2022 (new: Definition 1.12; 1.14; 1.15 and Example 1.13),
 update 31.10.2022: typos/errors and Section 3.2 on the geometry of
 high-dimensional spaces
 update 6.11.2022: Section 3.2, typos and errors.
 update 20.11.2022: typos and errors.
 update 04.12.2022: typos and errors

Notes will be updated during the term

Contents

1 Preliminaries on Probability Theory	1
1.1 Random variables	1
1.2 Classical Inequalities	4
1.3 L^p -spaces	5
1.4 Limit Theorems	7
2 Concentration inequalities for independent random variables	10
2.1 Why concentration inequalities	10
2.2 Hoeffding's Inequality	12
2.3 Chernoff's Inequality	15
2.4 Sub-Gaussian random variables	17
2.5 Sub-Exponential random variables	26
3 Random vectors in High Dimensions	31
3.1 Concentration of the Euclidean norm	31
3.2 The geometry of high dimensions	35
3.3 Covariance matrices and Principal Component Analysis (PCA)	38
3.4 Examples of High-Dimensional distributions	41
3.5 Sub-Gaussian random variables in higher dimensions	43
3.6 Application: Grothendieck's inequality	45
4 Random Matrices	45
4.1 Geometrics concepts	45
4.2 Concentration of the operator norm of random matrices	48
4.3 Application: Community Detection in Networks	51
4.4 Application: Covariance Estimation and Clustering	52
5 Concentration of measure - general case	53
5.1 Concentration by entropic techniques	53
5.2 Concentration via Isoperimetric Inequalities	63
5.3 Some matrix calculus and covariance estimation	66
5.4 Application - Johnson-Lindenstrauss Lemma	70
6 Basic tools in high-dimensional probability	74
6.1 Decoupling	74
6.2 Concentration for Anisotropic random vectors	79
6.3 Symmetrisation	80

7	Random Processes	81
7.1	Basic concepts and examples	81
7.2	Slepian's inequality and Gaussian interpolation	83
7.3	The supremum of a process	87
7.4	Uniform law of large numbers	93
8	Application: Statistical Learning theory	97

Preface

Introduction

We discuss an elegant argument that showcases the usefulness of probabilistic reasoning in geometry. First recall that a *convex combination* of points $z_1, \dots, z_m \in \mathbb{R}^n$ is a linear combination with coefficients that are non-negative and sum to 1, i.e., it is a sum of the form

$$\sum_{i=1}^m \lambda_i z_i \quad \text{where} \quad \lambda_i \geq 0 \quad \text{and} \quad \sum_{i=1}^m \lambda_i = 1. \quad (0.1)$$

Given a set $M \subset \mathbb{R}^n$, the *convex hull* of M is the set of all convex combinations of all finite collections of points in M , defined as

$$\text{conv}(M) := \{\text{convex combinations of } z_1, \dots, z_m \in M \text{ for } m \in \mathbb{N}\}.$$

The number m of elements defining a convex combination in \mathbb{R}^n is not restricted a priori. The classical theorem of Caratheodory states that one always take $m \leq n + 1$. For the convenience of the reader we briefly state that classical theorem.

Theorem 0.1 (Caratheodory's theorem) *Every point in the convex hull of a set $M \subset \mathbb{R}^n$ can be expressed as a convex combination of at most $n + 1$ points from M .*

Unfortunately the bound $n + 1$ cannot be improved (it is clearly attained for a simplex M)¹ This is some bad news. However, in most applications we only want to *approximate* a point $x \in \text{conv}(M)$ rather than to represent it exactly as a convex combination.

Can we do this with fewer than $n + 1$ points?

We now show that it is possible, and actually the number of required points does not need to depend on the dimension n at all! This is certainly brilliant news for any applications in mind - in particular for those where the dimension of the data set is extremely high (data science and machine learning and high-dimensional geometry and statistical mechanics models).

¹A simplex is a generalisation of the notion of a triangle to arbitrary dimensions. Specifically, a k -simplex S is the convex hull of its $k + 1$ vertices: Suppose $u_0, \dots, u_k \in \mathbb{R}^k$ are affinely independent, which means that $u_1 - u_0, \dots, u_k - u_0$ are linearly independent. Then, the simplex determined by these vertices is the set of points

$$S = \left\{ \lambda_0 u_0 + \dots + \lambda_k u_k : \sum_{i=0}^k \lambda_i = 1, \lambda_i \geq 0 \text{ for } i = 0, \dots, k \right\}.$$

Theorem 0.2 (Approximate form Caratheodory's theorem) Consider a set $M \subset \mathbb{R}^n$ whose diameter $\text{diam}(M) := \sup\{\|x - y\| : x, y \in M\}$ is bounded by 1. Then, for every point $x \in \text{conv}(M)$ and every positive integer k , one can find points $x_1, \dots, x_k \in M$ such that

$$\left\| x - \frac{1}{k} \sum_{j=1}^k x_j \right\| \leq \frac{1}{\sqrt{k}}.$$

Here $\|x\| = \sqrt{x_1^2 + \dots + x_n^2}$, $x = (x_1, \dots, x_n) \in \mathbb{R}^n$, denotes the Euclidean norm on \mathbb{R}^n .

Remark 0.3 We have assumed $\text{diam}(M) \leq 1$ for simplicity. For a general set M , the bound in the theorem changes to $\text{diam}(M)/\sqrt{k}$.

Why is this result surprising?

First, the number of points k in convex combinations does not depend on the dimension n . Second, the coefficients of convex combinations can be made all equal.

Proof. The argument upon which our proof is based is known as the *empirical method* of B. Maurey. W.l.o.g., we may assume that not only the diameter but also the *radius* of M is bounded by 1, i.e.,

$$\|w\| \leq 1 \quad \text{for all } w \in M.$$

We pick a point $x \in \text{conv}(M)$ and express it as a convex combination of some vectors $z_1, \dots, z_m \in M$ as in (0.1). Now we consider the numbers λ_i in that convex combination as probabilities that a random vector Z takes the values $z_i, i = 1, \dots, m$, respectively. That is, we define

$$\mathbb{P}(Z = z_i) = \lambda_i, \quad i = 1, \dots, m.$$

This is possible by the fact that the weights $\lambda_i \in [0, 1]$ and sum to 1. Consider now a sequence $(Z_j)_{j \in \mathbb{N}}$ of copies of Z . This sequence is an independent identically distributed sequence of \mathbb{R}^n -valued random variables. By the strong law of large numbers,

$$\frac{1}{k} \sum_{j=1}^k Z_j \rightarrow x \quad \text{almost surely as } k \rightarrow \infty.$$

We shall now get a quantitative version of this limiting statement, that is, we wish to obtain an error bound. For this we shall compute the variance of $\frac{1}{k} \sum_{j=1}^k Z_j$. We obtain

$$\begin{aligned} \mathbb{E} \left[\left\| x - \frac{1}{k} \sum_{j=1}^k Z_j \right\|^2 \right] &= \frac{1}{k^2} \mathbb{E} \left[\left\| \sum_{j=1}^k (Z_j - x) \right\|^2 \right] \quad (\text{since } \mathbb{E}[Z_i - x] = 0) \\ &= \frac{1}{k^2} \sum_{j=1}^k \mathbb{E}[\|Z_j - x\|^2]. \end{aligned}$$

The last identity is just a higher-dimensional version of the basic fact that the variance of a sum of independent random variables equals the sum of the variances. To bound the variances of the single terms we compute using that Z_j is copy of Z and that $\|Z\| \leq 1$ as $Z \in M$,

$$\mathbb{E}[\|Z_j - x\|^2] = \mathbb{E}[\|Z - \mathbb{E}[Z]\|^2] = \mathbb{E}[\|Z\|^2] - \|\mathbb{E}[Z]\|^2 \leq \mathbb{E}[\|Z\|^2] \leq 1,$$

where the second equality follows from the well-known property of the variance, namely, for $n = 1$,

$$\mathbb{E}[\|Z - \mathbb{E}[Z]\|^2] = \mathbb{E}[(Z - \mathbb{E}[Z])^2] = \mathbb{E}[Z^2 - 2Z\mathbb{E}[Z] + \mathbb{E}[Z]^2] = \mathbb{E}[Z^2] - \mathbb{E}[Z]^2,$$

and the cases for $n > 1$ follow similarly. We have thus shown that

$$\mathbb{E}\left[\left\|x - \frac{1}{k} \sum_{j=1}^k Z_j\right\|^2\right] \leq \frac{1}{k}.$$

Therefore, there exists a realisation of the random variables Z_1, \dots, Z_k such that

$$\left\|x - \frac{1}{k} \sum_{j=1}^k Z_j\right\|^2 \leq \frac{1}{k}.$$

Since by construction each Z_j takes values in M , the proof is complete. \square

We shall give one application of Theorem 0.2 in computational geometry. Suppose that we are given a subset $P \subset \mathbb{R}^n$ (say a polygon²) and asked to cover it by balls of a given radius $\varepsilon > 0$. What is the smallest number of balls needed, and how should we place them?

Corollary 0.4 (Covering polytopes by balls) *Let P be a polygon in \mathbb{R}^n with N vertices and whose diameter is bounded by 1. Then P can be covered by at most $N^{\lceil 1/\varepsilon^2 \rceil}$ Euclidean balls of radii $\varepsilon > 0$.*

Proof. We shall define the centres of the balls as follows. Let $k := \lceil 1/\varepsilon^2 \rceil$ and consider the set

$$\mathcal{N} := \left\{ \frac{1}{k} \sum_{j=1}^k x_j : x_j \text{ are vertices of } P \right\}.$$

The polytope P is the convex hull of the set of its vertices, which we denote by M . We then apply Theorem 0.2 to any point $x \in P = \text{conv}(M)$ and deduce that x is within a distance $1/\sqrt{k} \leq \varepsilon$ from some point in \mathcal{N} . This shows that the ε -balls centred at \mathcal{N} do indeed cover P . \square

In this lecture we will learn several other approaches to the covering problem in relation to packing, entropy and coding, and random processes.

²In geometry, a polytope is a geometric object with 'flat' sides. It is a generalisation of the three-dimensional polyhedron which is a solid with flat polygonal faces, straight edges and sharp corners/vertices. Flat sides mean that the sides of a $(k + 1)$ -polytope consist of k -polytopes.

1 Preliminaries on Probability Theory

In this chapter we recall some basic concepts and results of probability theory. The reader should be familiar with most of this material some of which is taught in elementary probability courses in the first year. To make these lectures self-contained we review the material mostly without proof and refer the reader to basic chapters of common undergraduate textbooks in probability theory, e.g. [Dur19] and [Geo12]. In Section 1.1 we present basic definitions for probability space and probability measure as well as random variables along with expectation, variance and moments. Vital for the lecture will be the review of all classical inequalities in Section 1.2. Finally, in Section 1.4 we review well-know limit theorems.

1.1 Random variables

A probability space (Ω, \mathcal{F}, P) is a triple consisting of a set Ω , a σ -algebra \mathcal{F} and a probability measure P . We write $\mathcal{P}(\Omega)$ for the power set of Ω which is the set of all subsets of Ω .

Definition 1.1 (σ -algebra) Suppose $\Omega \neq \emptyset$. A system $\mathcal{F} \subset \mathcal{P}(\Omega)$ satisfying

- (a) $\Omega \in \mathcal{F}$
- (b) $A \in \mathcal{F} \Rightarrow A^c := \Omega \setminus A \in \mathcal{F}$
- (c) $A_1, A_2, \dots \in \mathcal{F} \Rightarrow \bigcup_{i \geq 1} A_i \in \mathcal{F}$.

is called σ -algebra (or σ -field) on Ω . The pair (Ω, \mathcal{F}) is then called an event space or measurable space.

Example 1.2 (Borel σ -algebra) Let $\Omega = \mathbb{R}^n$, $n \in \mathbb{N}$ and

$$\mathcal{G} = \left\{ \prod_{i=1}^n [a_i, b_i] : a_i < b_i, a_i, b_i \in \mathbb{Q} \right\}$$

be the system consisting of all compact rectangular boxes in \mathbb{R}^n with rational vertices and edges parallel to the axes. In honour of Émile Borel (1871–1956), the system $\mathcal{B}^n = \sigma(\mathcal{G})$ is called the Borel σ -algebra on \mathbb{R}^n , and every $A \in \mathcal{B}^n$ a Borel set. Here, $\sigma(\mathcal{G})$ denotes the smallest σ -algebra generated by the system \mathcal{G} . Note that the \mathcal{B}^n can also be generated by the system of open or half-open rectangular boxes, see [Dur19, Geo12].



The decisive point in the process of building a stochastic model is the next step: For each $A \in \mathcal{F}$ we need to define a value $P(A) \in [0, 1]$ that indicates the probability of A . Sensibly, this should be done so that the following holds.

(N) *Normalisation:* $P(\Omega) = 1$.

(A) *σ -Additivity:* For pairwise disjoint events $A_1, A_2, \dots \in \mathcal{F}$ one has

$$P\left(\bigcup_{i \geq 1} A_i\right) = \sum_{i \geq 1} P(A_i).$$

Definition 1.3 (Probability measure) Let (Ω, \mathcal{F}) be a measurable space. A function $P: \mathcal{F} \rightarrow [0, 1]$ satisfying the properties (N) and (A) is called a probability measure or a probability distribution, in short a distribution (or, a little old-fashioned, a probability law) on (Ω, \mathcal{F}) . Then the triple (Ω, \mathcal{F}, P) is called a probability space.

Theorem 1.4 (Construction of probability measures via densities) (a) *Discrete case: For countable Ω , the relations*

$$P(A) = \sum_{\omega \in A} \varrho(\omega) \text{ for } A \in \mathcal{P}(\Omega), \quad \varrho(\omega) = P(\{\omega\}) \text{ for } \omega \in \Omega$$

establish a one-to-one correspondence between the set of all probability measures P on $(\Omega, \mathcal{P}(\Omega))$ and the set of all sequences $\varrho = (\varrho(\omega))_{\omega \in \Omega}$ in $[0, 1]$ such that $\sum_{\omega \in \Omega} \varrho(\omega) = 1$.

(b) *Continuous case: If $\Omega \subset \mathbb{R}^n$ is Borel, then every function $\varrho: \Omega \rightarrow [0, \infty)$ satisfying the properties*

$$(i) \{x \in \Omega: \varrho(x) \leq c\} \in \mathcal{B}_\Omega^n \text{ for all } c > 0,$$

$$(ii) \int_\Omega \varrho(x) \, dx = 1$$

determines a unique probability measure on $(\Omega, \mathcal{B}_\Omega^n)$ via

$$P(A) = \int_A \varrho(x) \, dx \text{ for } A \in \mathcal{B}_\Omega^n$$

(but not every probability measure on $(\Omega, \mathcal{B}_\Omega^n)$ is of this form).

Proof. See [Dur19, Geo12]. □

Definition 1.5 A sequence or function ϱ as in Theorem 1.4 above is called a density (of P) or, more explicitly (to emphasise normalisation), a probability density (function), often abbreviated as *pdf*. If a distinction between the discrete and continuous case is required, a sequence $\varrho = (\varrho(\omega))_{\omega \in \Omega}$ as in case (a) is called a discrete density, and a function ϱ in case (b) a Lebesgue density.

In probability theory one often considers the transition from a measurable space (event space) (Ω, \mathcal{F}) to a coarser measurable (event) space (Ω', \mathcal{F}') . In general such a mapping should satisfy the requirement

$$A' \in \mathcal{F}' \Rightarrow X^{-1}A' := \{\omega \in \Omega: X(\omega) \in A'\} \in \mathcal{F}. \quad (1.1)$$

Definition 1.6 Let (Ω, \mathcal{F}) and (Ω', \mathcal{F}') be two measurable (event) spaces. Then every mapping $X: \Omega \rightarrow \Omega'$ satisfying property (1.1) is called a *random variable* from (Ω, \mathcal{F}) to (Ω', \mathcal{F}') , or a random element of Ω' , or a Ω' -valued random variable. Alternatively (in the terminology of measure theory), X is said to be measurable relative to \mathcal{F} and \mathcal{F}' .

In probability theory it is common to write $\{X \in A'\} := X^{-1}A'$.

Theorem 1.7 (Distribution of a random variable) *If X is a random variable from a probability space (Ω, \mathcal{F}, P) to a measurable space (Ω', \mathcal{F}') , then the prescription*

$$P'(A') := P(X^{-1}A') = P(\{X \in A'\}) \equiv P(X \in A') \quad \text{for any } A' \in \mathcal{F}'$$

defines a probability measure P' on (Ω', \mathcal{F}') .

Definition 1.8 (a) The probability measure P' in Theorem 1.7 is called the *distribution of X under P* , or the image of P under X , and is denoted by $P \circ X^{-1}$. (In the literature, one also finds the notations P_X or $\mathcal{L}(X; P)$. The letter \mathcal{L} stands for the more traditional term law, or loi in French.)

(b) Two random variables are said to be identically distributed if they have the same distribution.

We are considering real-valued or \mathbb{R}^n -valued random variables in the following and we just call them random variables for all these cases. In basic courses in probability theory, one learns about the two most important quantities associated with a random variable X , namely the expectation³ (also called the mean) and variance. They will be noted in this lecture by

$$\mathbb{E}[X] \quad \text{and} \quad \text{Var}(X) := \mathbb{E}[(X - \mathbb{E}(X))^2].$$

The distribution of a real-valued random variable X is determined by the *cumulative distribution function* (CDF) of X , defined as

$$F_X(t) = \mathbb{P}(X \leq t) = \mathbb{P}((-\infty, t]), \quad t \in \mathbb{R}. \quad (1.2)$$

It is often more convenient to work with the tails of random variables, namely with

$$\mathbb{P}(X > t) = 1 - F_X(t). \quad (1.3)$$

Here we write \mathbb{P} for the generic distribution of the random variable X which is given by the context.

For any real-valued random variable the *moment generating function* (MGF) (MGF) is defined

$$M_X(\lambda) := \mathbb{E}[e^{\lambda X}], \quad \lambda \in \mathbb{R}. \quad (1.4)$$

When M_X is finite for all λ in a neighbourhood of the origin, we can easily compute all moments by taking derivatives (interchanging differentiation and expectation (integration) in the usual way):

$$\mathbb{E}[X^k] = \left. \frac{d^k}{d\lambda^k} M_X(\lambda) \right|_{\lambda=0}, \quad k \in \mathbb{N}. \quad (1.5)$$

³In measure theory the expectation $\mathbb{E}[X]$ of a random variable on a probability space (Ω, \mathcal{F}, P) is the Lebesgue integral of the function $X: \Omega \rightarrow \mathbb{R}$. This makes theorems on Lebesgue integration applicable in probability theory for expectations of random variables

Lemma 1.9 (Integral Identity) *Let X be a real-valued non-negative random variable. Then*

$$\mathbb{E}[X] = \int_0^\infty \mathbb{P}(X > t) dt.$$

Proof. We can write any non-negative real number x via the following identity using indicator function ⁴:

$$x = \int_0^x 1 dt = \int_0^\infty \mathbb{1}_{\{t < x\}}(t) dt.$$

Substitute now the random variable X for x and take expectation (with respect to X) on both sides. This gives

$$\mathbb{E}[X] = \mathbb{E}\left[\int_0^\infty \mathbb{1}_{\{t < X\}}(t) dt\right] = \int_0^\infty \mathbb{E}[\mathbb{1}_{\{t < X\}}] dt = \int_0^\infty \mathbb{P}(t < X) dt.$$

To change the order of expectation and integration in the second inequality, we used the Fubini-Tonelli theorem. \square

Exercise 1.10 (Integral identity) Prove the extension of Lemma 1.9 to any real-valued random variable (not necessarily positive):

$$\mathbb{E}[X] = \int_0^\infty \mathbb{P}(X > t) dt - \int_{-\infty}^0 \mathbb{P}(X < t) dt.$$



1.2 Classical Inequalities

In this section fundamental classical inequalities are presented. Here, classical refers to typical estimates for analysing stochastic limits.

Proposition 1.11 (Jensen's inequality) *Suppose that $\Phi: I \rightarrow \mathbb{R}$, where $I \subset \mathbb{R}$ is an interval, is a convex function. Let X be a real-valued random variable. Then*

$$\Phi(\mathbb{E}[X]) \leq \mathbb{E}[\Phi(X)].$$

Proof. See [Dur19] or [Geo12] using either the existence of sub-derivatives for convex functions or the definition of convexity with the epi-graph of a function. The epi-graph of a function $f: I \rightarrow \mathbb{R}$, $I \subset \mathbb{R}$ some interval, is the set

$$\text{epi}(f) := \{(x, t) \in \mathbb{R}^2 : x \in I, f(x) \leq t\}.$$

A function $f: I \rightarrow \mathbb{R}$ is convex if and only if $\text{epi}(f)$ is a convex set in \mathbb{R}^2 . \square

⁴ $\mathbb{1}_A$ denotes the indicator function of the set A , that is, $\mathbb{1}_A(t) = 1$ if $t \in A$ and $\mathbb{1}_A(t) = 0$ if $t \notin A$.

1.3 L^p -spaces

In the following let X be a \mathbb{R} -valued random variable, i.e., there is a probability space (Ω, \mathcal{F}, P) such that $X: \Omega \rightarrow \mathbb{R}$ is a measurable function. By default, we equip the real line \mathbb{R} with its Borel- σ -algebra. We begin with the definition of the *essential supremum* of X .

Definition 1.12 (Essential supremum) Let X be \mathbb{R} -valued random variable. The *essential supremum* of X , written $\text{ess-sup}(X)$, is the smallest number $\alpha \in \mathbb{R}$ such that the set $\{x \in \Omega: X(x) > \alpha\}$ has measure zero, that is,

$$P(\{x \in \Omega: X(x) > \alpha\}) = 0.$$

If no such number exists we define $\text{ess-sup}(X) = \infty$.

To understand this definition better we shall check the following example.

Example 1.13 (Essential supremum being infinity) Suppose that $\Omega = (0, 1)$, $\mathcal{F} = \mathcal{B}((0, 1))$, and let P be the uniform measure on $(0, 1)$. This measure has constant probability density,

$$P(A) = \int_{\Omega} \mathbb{1}_A(t) dt = b - a, \quad \text{for any } A = (a, b) \text{ with } 0 \leq a < b \leq 1.$$

Define $X: (0, 1) \rightarrow \mathbb{R}, x \mapsto \frac{1}{x}$. Then X is continuous function and therefore measurable. Then $\text{ess sup}(X) = \infty$. To see this, pick any $\alpha \in \mathbb{R}_+$. Then

$$\{x \in (0, 1): \frac{1}{x} > \alpha\} = (0, \frac{1}{\alpha})$$

and

$$P((0, \frac{1}{\alpha})) = \frac{1}{\alpha} > 0.$$

As this holds for all $\alpha > 0$, we have that $\text{ess-sup}(X) = \infty$. ♣

Definition 1.14 Let (Ω, \mathcal{F}, P) be a probability space. Given two measurable functions $f, g: [0, \infty]$, we say that f is *equivalent to* g , written $f \sim g$, if

$$f(x) = g(x) \quad \text{for } P - a.e. x \in \Omega,$$

that is,

$$P(\{x \in \Omega: f(x) \neq g(x)\}) = 0.$$

We shall identify - with an abuse of notation - identify a measurable function f with its equivalence class $[f]$.

Definition 1.15 Let (Ω, \mathcal{F}, P) be a probability space and $1 \leq p < \infty$.

$$L^p \equiv L^p(\Omega, \mathcal{F}, P) := \{f: \Omega \rightarrow [-\infty, \infty]: f \text{ measurable and } \|f\|_{L^p} < \infty\},$$

where

$$\|f\|_{L^p} := \left(\int_{\Omega} |f|^p dP \right)^{\frac{1}{p}} = \left(\int_{\Omega} |f(x)|^p P(dx) \right)^{\frac{1}{p}}.$$

If $p = \infty$, then

$$L^\infty \equiv L^\infty(\Omega, \mathcal{F}, P) := \{f: \Omega \rightarrow [-\infty, \infty]: f \text{ measurable and } \|f\|_{L^\infty} < \infty\},$$

where


$$\|f\|_{L^\infty} := \text{ess-sup}(|f|),$$

and we write $\|f\|_\infty \equiv \|f\|_{L^\infty}$ occasionally.

A consequence of Jensen's inequality is that $\|X\|_{L^p}$ is an increasing function in the parameter p , i.e.,

$$\|X\|_{L^p} \leq \|X\|_{L^q} \quad 0 \leq p \leq q \leq \infty. \quad (1.6)$$

This follows from the convexity of $\Phi(x) = x^{\frac{q}{p}}$ when $q \geq p$.

Exercise 1.16 Show that (1.6) holds. 

Proposition 1.17 (Minkowski's inequality) For $p \in [1, \infty]$, let $X, Y \in L^p$, then

$$\|X + Y\|_{L^p} \leq \|X\|_{L^p} + \|Y\|_{L^p}.$$

Proposition 1.18 (Cauchy-Schwarz inequality) For $X, Y \in L^2$,

$$|\mathbb{E}[XY]| \leq \|X\|_{L^2} \|Y\|_{L^2}.$$

Proposition 1.19 (Hölder's inequality) For $p, q \in (1, \infty)$ with $1/p + 1/q = 1$ let $X \in L^p$ and $Y \in L^q$. Then

$$\mathbb{E}[XY] \leq \mathbb{E}[|XY|] \leq \|X\|_{L^p} \|Y\|_{L^q}.$$

Lemma 1.20 (Linear Markov's inequality) For non-negative random variables X and $t > 0$ the tail probability is bounded as

$$\mathbb{P}(X > t) \leq \frac{\mathbb{E}[X]}{t}.$$

Proof. Pick $t > 0$. Any positive number x can be written as

$$x = x \mathbb{1}_{\{X \geq t\}} + x \mathbb{1}_{\{X < t\}}.$$

As X is non-negative, we insert X into the above expression in place of x and take the expectation (integral) to obtain

$$\mathbb{E}[X] = \mathbb{E}[X \mathbb{1}_{\{X \geq t\}}] + \mathbb{E}[X \mathbb{1}_{\{X < t\}}] \geq \mathbb{E}[t \mathbb{1}_{\{X \geq t\}}] = t \mathbb{P}(X \geq t).$$

□

This is one particular version of the Markov inequality which provides linear decay in t . In the following proposition we obtain the general version which will be used frequently throughout the lecture.

Proposition 1.21 (Markov's inequality) *Let Y be a real-valued random variable and $f : [0, \infty) \rightarrow [0, \infty)$ an increasing function. Then, for all $\varepsilon > 0$ with $f(\varepsilon) > 0$,*

$$\mathbb{P}(|Y| \geq \varepsilon) \leq \frac{\mathbb{E}[f \circ |Y|]}{f(\varepsilon)}.$$

Proof. Clearly, the composition $f \circ |Y|$ is a positive random variable such that

$$f(\varepsilon)\mathbb{1}_{\{|Y| \geq \varepsilon\}} \leq f \circ |Y|.$$

Taking the expectation on both sides of that inequality gives

$$f(\varepsilon)\mathbb{P}(|Y| \geq \varepsilon) = \mathbb{E}[f(\varepsilon)\mathbb{1}_{\{|Y| \geq \varepsilon\}}] \leq \mathbb{E}[f \circ |Y|].$$

□

The following version of the Markov inequality is often called Chebyshev's inequality.

Corollary 1.22 (Chebyshev's inequality, 1867) *For all $Y \in L^2$ with $\mathbb{E}[Y] \in (-\infty, \infty)$ and $\varepsilon > 0$,*

$$\mathbb{P}(|Y - \mathbb{E}[Y]| \geq \varepsilon) \leq \frac{\text{Var}(Y)}{\varepsilon^2}.$$

1.4 Limit Theorems

Definition 1.23 (Variance and covariance) Let $X, Y \in L^2$ be real-valued random variables.

(a)

$$\text{Var}(X) := \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

is called the **variance**, and $\sqrt{\text{Var}(X)}$ the **standard deviation** of X with respect to \mathbb{P} .

(b)

$$\text{cov}(X, Y) := \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

is called the **covariance** of X and Y . It exists since $|XY| \leq X^2 + Y^2$.

(c) If $\text{cov}(X, Y) = 0$, then X and Y are called **uncorrelated**.

Recall that two independent random variables are uncorrelated, but two uncorrelated are not necessarily independent as the following example shows.

Example 1.24 Let $\Omega = \{1, 2, 3\}$ and let P the uniform distribution on Ω . Define two random variables by their images, that is,

$$(X(1), X(2), X(3)) = (1, 0, -1) \quad \text{and} \quad (Y(1), Y(2), Y(3)) = (0, 1, 0).$$

Then $XY = 0$ and $\mathbb{E}[XY] = 0$, and therefore $\text{cov}(X, Y) = 0$, but

$$\mathbb{P}(X = 1, Y = 1) = 0 \neq \frac{1}{9} = \mathbb{P}(X = 1)\mathbb{P}(Y = 1).$$

Hence X and Y are not independent. ♣

Theorem 1.25 (Weak law of large numbers, L^2 -version) *Let $(X_i)_{i \in \mathbb{N}}$ be a sequence of uncorrelated (e.g. independent) real-valued random variables in L^2 with bounded variance, in that $v := \sup_{i \in \mathbb{N}} \text{Var}(X_i) < \infty$. Then for all $\varepsilon > 0$*

$$\mathbb{P}\left(\left|\frac{1}{N} \sum_{i=1}^N (X_i - \mathbb{E}[X_i])\right| \geq \varepsilon\right) \leq \frac{v}{N\varepsilon^2} \xrightarrow{N \rightarrow \infty} 0,$$

and thus

$$\frac{1}{N} \sum_{i=1}^N (X_i - \mathbb{E}[X_i]) \xrightarrow[N \rightarrow \infty]{\text{P}} 0,$$

($\xrightarrow[N \rightarrow \infty]{\text{P}}$ means convergence in probability). In particular, if $\mathbb{E}[X_i] = \mathbb{E}[X_1]$ holds for all $i \in \mathbb{N}$, then

$$\frac{1}{N} \sum_{i=1}^N X_i \xrightarrow{\text{P}} \mathbb{E}[X_1].$$

We now present a second version of the weak law of large numbers, which does not require the existence of the variance. To compensate we must assume that the random variables, instead of being pairwise uncorrelated, are even pairwise independent and identically distributed.

Theorem 1.26 (Weak law of large numbers, L^1 -version) *Let $(X_i)_{i \in \mathbb{N}}$ be a sequence of pairwise independent, identically distributed real-valued random variables in L^1 . Then*

$$\frac{1}{N} \sum_{i=1}^N X_i \xrightarrow[N \rightarrow \infty]{\text{P}} \mathbb{E}[X_1].$$

Theorem 1.27 (Strong law of large numbers) *If $(X_i)_{i \in \mathbb{N}}$ is a sequence of pairwise uncorrelated real-valued random variables in L^2 with $v := \sup_{i \in \mathbb{N}} \text{Var}(X_i) < \infty$, then*

$$\frac{1}{N} \sum_{i=1}^N (X_i - \mathbb{E}[X_i]) \rightarrow 0 \text{ almost surely as } N \rightarrow \infty.$$

Theorem 1.28 (Central limit theorem) *A.M. Lyapunov 1901, J.W. Lindeberg 1922, P. Lévy 1922.*

Let $(X_i)_{i \in \mathbb{N}}$ be a sequence of independent, identically distributed real-valued random variables in L^2 with $\mathbb{E}[X_i] = m$ and $\text{Var}(X_i) = v > 0$. Then, as $N \rightarrow \infty$,

$$S_N^* := \frac{1}{\sqrt{N}} \sum_{i=1}^N \frac{X_i - m}{\sqrt{v}} \xrightarrow{d} \mathbf{N}(0, 1).$$

The normal distribution is defined as follows.

A real-valued random variable X is **normally** distributed with mean μ and variance $\sigma^2 > 0$ if

$$\mathbb{P}(X > x) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_x^\infty e^{-\frac{(u-\mu)^2}{2\sigma^2}} du, \quad \text{for all } x \in \mathbb{R}.$$

We write $X \sim \mathbf{N}(\mu, \sigma^2)$. We say that X is standard normal distributed if $X \sim \mathbf{N}(0, 1)$.

A random vector $X = (X_1, \dots, X_n)$ is called a **Gaussian random vector** if there exists an $n \times m$ matrix A , and an n -dimensional vector $b \in \mathbb{R}^n$ such that $X^T = AY + b$, where Y is an m -dimensional vector with independent standard normal entries, i.e. $Y_i \sim \mathbf{N}(0, 1)$ for $i = 1, \dots, m$. Likewise, a random variable $Y = (Y_1, \dots, Y_m)$ with values in \mathbb{R}^m has the m -dimensional standard Gaussian distribution if the m coordinates are standard normally distributed and independent. The covariance matrix of $X = AY + b$ is then given by

$$\text{cov}(Y) = \mathbb{E}[(Y - \mathbb{E}[Y])(Y - \mathbb{E}[Y])^T] = AA^T.$$

Lemma 1.29 *If A is an orthogonal $n \times n$ matrix, i.e. $AA^T = \mathbb{1}$, and X is a n -dimensional standard Gaussian vector, then AX is also a n -dimensional standard Gaussian vector.*

Lemma 1.30 *Let X_1 and X_2 be independent and normally distributed with zero mean and variance $\sigma^2 > 0$. Then $X_1 + X_2$ and $X_1 - X_2$ are independent and normally distributed with mean 0 and variance $2\sigma^2$.*

Proposition 1.31 *If X and Y are n -dimensional Gaussian vectors with $\mathbb{E}[X] = \mathbb{E}[Y]$ and $\text{cov}(X) = \text{cov}(Y)$, then X and Y have the same distribution.*

Corollary 1.32 *A Gaussian random vector X has independent entries if and only if its covariance matrix is diagonal. In other words, the entries in a Gaussian vector are uncorrelated if and only if they are independent.*

Bernoulli: $p \in [0, 1]$, then $X \sim \text{Ber}(p)$ if $\mathbb{P}(X = 1) = p$ and $\mathbb{P}(X = 0) = 1 - p =: q$. If $X \sim \text{Ber}(p)$, then $\mathbb{E}[X] = p$ and $\text{Var}(X) = pq$. We call this random variable the *standard Bernoulli random variable*.

Binomial: $S_N = \sum_{i=1}^N X_i \sim \mathbf{B}(N, p)$ if $X_i \sim \text{Ber}(p)$ and $(X_i), i = 1, \dots, N$, independent family. $\mathbb{E}[S_N] = Np$ and $\text{Var}(S_N) = Npq$.

Exercise 1.33 (a) Let $X \sim \text{Ber}(p)$, $p \in [0, 1]$. Compute the expectation, the variance and the moment generating function M_X .

(b) Let $Z := X - 1$ with $X \sim \text{Ber}(p)$. Z is called *symmetric* Bernoulli variable. Compute the expectation, the variance and the moment generating function M_Z .



Poisson: $\lambda > 0$, then $X \sim \text{Poi}(\lambda)$ if

$$\mathbb{P}(X = k) = \frac{e^{-\lambda} \lambda^k}{k!} \quad k \in \mathbb{N}_0,$$

$\text{Poi}(\lambda) \in \mathcal{M}_1(\mathbb{N}_0, \mathcal{P}(\mathbb{N}_0))$. Here, $\mathcal{M}_1(\Omega)$ denotes the set of probability measures on Ω and $\mathcal{P}(\mathbb{N}_0)$ is the power set.

Exercise 1.34 Let $X \sim \text{Poi}(\lambda)$, $\lambda > 0$. Compute the expectation, the variance and the moment generating function M_X .



Exponential: A random variable X taking positive real values is exponentially distributed with parameter $\alpha > 0$ when the probability density function is

$$f_X(t) = \alpha e^{-\alpha t} \quad \text{for } t \geq 0.$$

We write $X \sim \text{Exp}(\alpha)$. If $X \sim \text{Exp}(\alpha)$, then $\mathbb{E}[X] = \frac{1}{\alpha}$ and $\text{Var}(X) = \frac{1}{\alpha^2}$.

Exercise 1.35 Let $X \sim \text{Exp}(\alpha)$, $\alpha > 0$. Compute the expectation, the variance and the moment generating function M_X .



Theorem 1.36 (Poisson limit theorem) Let $X_i^{(N)}$, $i = 1, \dots, N$, be independent Bernoulli random variables $X_i^{(N)} \sim \text{Ber}(p_i^{(N)})$, and denote $S_N = \sum_{i=1}^N X_i^{(N)}$ their sum. Assume that, as $N \rightarrow \infty$,

$$\max_{1 \leq i \leq N} \{p_i^{(N)}\} \xrightarrow{N \rightarrow \infty} 0 \quad \text{and} \quad \mathbb{E}[S_N] = \sum_{i=1}^N p_i^{(N)} \xrightarrow{N \rightarrow \infty} \lambda \in (0, \infty).$$

Then, as $N \rightarrow \infty$,

$$S_N \longrightarrow \text{Poi}(\lambda) \quad \text{in distribution.}$$

2 Concentration inequalities for independent random variables

2.1 Why concentration inequalities

Suppose a random variable X has mean μ , then, for any $t \geq 0$,

$$\mathbb{P}(|X - \mu| > t) \leq \text{something small}$$

is a concentration inequality. One is interested in cases where the bound on the right hand side decays with increasing parameter $t \geq 0$. We now discuss a very simple example to demonstrate that we need better concentration inequalities than the ones obtained for example from the central limit theorem (CLT). Toss a fair coin N times. What is the probability that we get at least $3N/4$ heads? Recall that $\mathbb{E}[S_N] = \frac{N}{2}$ and $\text{Var}(S_N) = \frac{N}{4}$ in conjunction with Corollary 1.22 gives the bound

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

This concentration bound vanishes linearly in N . One may wonder if we can do better using the CLT. The *De Moivre Laplace limit theorem* (variant of the CLT for Binomial distributions) states that the distribution of the normalised number of heads

$$Z_N = \frac{S_N - N/2}{\sqrt{N/4}}$$

converges to the standard normal distribution $N(0, 1)$. We therefore expect to get the following concentration bound

$$\mathbb{P}\left(S_N \geq \frac{3}{4}N\right) = \mathbb{P}\left(Z_N \geq \sqrt{N/4}\right) \approx \mathbb{P}(Y \geq \sqrt{N/4}), \quad (2.1)$$

where $Y \sim N(0, 1)$. To obtain explicit bounds we need estimates on the tail of the normal distribution.

Proposition 2.1 (Tails of the normal distribution) *Let $Y \sim N(0, 1)$. Then, for all $t > 0$, we have*

$$\left(\frac{1}{t} - \frac{1}{t^3}\right) \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \leq \mathbb{P}(Y \geq t) \leq \frac{1}{t} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}.$$

Proof. Denote $f(x) := \exp(-x^2/2)$. For the upper bound we use $x \geq t$ to get the estimate

$$\int_t^\infty f(x) dx \leq \int_t^\infty \frac{x}{t} e^{-x^2/2} dx = \frac{1}{t} e^{-t^2/2}.$$

For the lower bound we use integration by parts (IBP) and $f'(x) = -xe^{-x^2/2}$:

$$\begin{aligned} \int_t^\infty e^{-x^2/2} dx &= \int_t^\infty \frac{1}{x} x e^{-x^2/2} dx = \left[-\frac{1}{x} e^{-x^2/2}\right]_t^\infty - \int_t^\infty \frac{1}{x^2} e^{-x^2/2} dx \\ &= \frac{1}{t} e^{-t^2/2} - \int_t^\infty \frac{1}{x^3} x e^{-x^2/2} dx = \frac{1}{t} e^{-t^2/2} - \left[-\frac{1}{x^3} e^{-x^2/2}\right]_t^\infty \\ &\quad + \int_t^\infty \frac{3}{x^4} e^{-x^2/2} dx. \end{aligned}$$

Hence

$$\int_t^\infty e^{-x^2/2} dx = \left(\frac{1}{t} - \frac{1}{t^3}\right) e^{-t^2/2} + \int_t^\infty \frac{3}{x^4} e^{-x^2/2} dx,$$

and, as the integral on the right hand side is positive,

$$\mathbb{P}(Y \geq t) \geq \frac{1}{\sqrt{2\pi}}(1/t - 1/t^3)e^{-t^2/2}.$$

□

The lower bound in Proposition 2.1 is lower than the tail lower bound in Lemma C.5 in the appendix.

Lemma 2.2 (Lower tail bound for normal distribution) *Let $Y \sim N(0, 1)$. Then, for all $t \geq 0$,*

$$\mathbb{P}(Y \geq t) \geq \left(\frac{t}{t^2 + 1}\right) \frac{1}{\sqrt{2\pi}} e^{-t^2/2}.$$

Proof. Define

$$g(t) := \frac{1}{\sqrt{2\pi}} \left(\int_t^\infty e^{-x^2/2} dx - \frac{t}{t^2 + 1} e^{-t^2/2} \right).$$

Then $g(0) > 0$ and $g'(t) = \frac{-2e^{-t^2/2}}{(t^2+1)^2} < 0$. Thus g is strictly decreasing with $\lim_{t \rightarrow \infty} g(t) = 0$ implying that $g(t) > 0$ for all $t \in \mathbb{R}$. □

Using the tail estimates in Proposition 2.1 we expect to obtain an exponential bound for the right hand side of (2.1), namely that the probability of having at least $3N/4$ heads seems to be smaller than $\frac{1}{\sqrt{2\pi}} e^{-N/8}$. However, we have not taken into account the approximation errors of ' \approx ' in (2.1). Unfortunately, it turns out that the error decays too slowly, actually even more slowly than linearly in N . This can be seen from the following version of the CLT which we state without proof, see for example [Dur19].

Theorem 2.3 (Berry-Esseen CLT) *In the setting of Theorem 1.28, for every N and every $t \in \mathbb{R}$, we have*

$$|P(Z_N \geq t) - \mathbb{P}(Y \geq t)| \leq \frac{\varrho}{\sqrt{N}},$$

where $Y \sim N(0, 1)$ and $\varrho = \mathbb{E}[|X_1 - m|^3]/\sigma^3$.

Can we improve the approximation error ' \approx ' in (2.1) by simply computing the probabilities with the help of Stirling's formula? Suppose that N is even. Then

$$\mathbb{P}(S_N = N/2) = 2^{-N} \binom{N}{N/2} \sim \frac{1}{\sqrt{N}} \quad \text{and} \quad \mathbb{P}(Z_N = 0) \sim \frac{1}{\sqrt{N}},$$

but $\mathbb{P}(Y = 0) = 0$. We thus see that we shall get a better concentration bound. This is the content of the next section.

2.2 Hoeffding's Inequality

In this section we study sums of *symmetric Bernoulli random variables* defined as follows.

Definition 2.4 A random variable Z taking values in $\{-1, +1\}$ with

$$\mathbb{P}(Z = -1) = \mathbb{P}(Z = +1) = \frac{1}{2}$$

is called symmetric Bernoulli random variable.

Note that the 'standard' Bernoulli random variable X takes values in $\{0, 1\}$, and that one can easily switch between both via $Z = 2X - 1$.

Theorem 2.5 (Hoeffding's inequality) Suppose that X_1, \dots, X_N , are independent symmetric Bernoulli random variables and let $a = (a_1, \dots, a_N) \in \mathbb{R}^N$. Then, for any $t \geq 0$, we have

$$\mathbb{P}\left(\sum_{i=1}^N a_i X_i \geq t\right) \leq \exp\left(-\frac{t^2}{2\|a\|_2^2}\right).$$

Proof. Without loss of generality we can put $\|a\|_2 = 1$, namely, if $\|a\|_2 \neq 1$, then define $\tilde{a}_i = a_i/\|a\|_2, i = 1, \dots, N$, to obtain the bound for

$$\mathbb{P}\left(\sum_{i=1}^N \tilde{a}_i X_i \geq t\right) = \mathbb{P}\left(\sum_{i=1}^N a_i X_i \geq \|a\|_2 t\right) \leq \exp\left(-\frac{t^2}{2}\right).$$

Let $\lambda > 0$, then, using Markov's inequality, obtain

$$\mathbb{P}\left(\sum_{i=1}^N a_i X_i \geq t\right) = \mathbb{P}\left(\exp\left(\lambda \sum_{i=1}^N a_i X_i\right) \geq e^{\lambda t}\right) \leq e^{-\lambda t} \mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^N a_i X_i\right)\right].$$

To bound the right hand side use first that the random variables are independent,

$$\mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^N a_i X_i\right)\right] = \prod_{i=1}^N \mathbb{E}[e^{\lambda a_i X_i}]$$

and then compute for $i \in \{1, \dots, N\}$,

$$\mathbb{E}[e^{\lambda a_i X_i}] = \frac{e^{\lambda a_i} + e^{-\lambda a_i}}{2} = \cosh(\lambda a_i).$$

We are left to find a bound for the hyperbolic cosine. There are two ways to get the upper bound

$$\cosh(x) \leq e^{x^2/2}.$$

1.) Simply write down both Taylor series and compare term by term, that is,

$$\cosh(x) = 1 + \frac{x^2}{2!} + \frac{x^4}{4!} + \frac{x^6}{6!} + \dots$$

and

$$e^{x^2/2} = \sum_{k=0}^{\infty} \frac{(x^2/2)^k}{k!} = 1 + \frac{x^2}{2} + \frac{x^4}{2^2 2!} + \dots$$

2.) Use the product expansion (complex analysis – not needed for this course, just as background information)

$$\cosh(x) = \prod \left(1 + \frac{4x^2}{\pi^2(2k-1)^2} \right) \leq \exp \left(\sum_{k=1}^{\infty} \frac{4x^2}{\pi^2(2k-1)^2} \right) = e^{x^2/2},$$

where we used $1 + x \leq e^x$ for the inequality.

In any case, we obtain

$$\mathbb{E}[e^{\lambda a_i X_i}] \leq e^{\lambda^2 a_i^2 / 2},$$

and thus

$$\mathbb{P} \left(\sum_{i=1}^N a_i X_i \geq t \right) \leq \exp \left(-\lambda t + \frac{\lambda^2}{2} \|a\|_2^2 \right).$$

Using $\|a\|_2 = 1$, and optimising over the value of λ we get $\lambda = t$ ($0 = g'(\lambda) = -t + \lambda$, $g(\lambda) = -\lambda t + \lambda^2/2$) and thus the right hand is simply $\exp(-t^2/2)$. \square

With Hoeffding's inequality in Theorem 2.5 we are now in the position to answer our previous question on the probability for a fair coin to have at least $3/4N$ times with heads. The fair coin is a standard Bernoulli random variable Y with $\mathbb{P}(Y = 1) = \mathbb{P}(Y = 0) = \frac{1}{2}$, and the symmetric one is just $Z = 2Y - 1$. Thus we obtain the following bound,

$$\begin{aligned} \mathbb{P} \left(\sum_{i=1}^N Y_i \geq 3/4N \right) &= \mathbb{P} \left(\sum_{i=1}^N \frac{Z_i + 1}{2} \geq 3/4N \right) = \mathbb{P} \left(\sum_{i=1}^N \frac{Z_i}{2} \geq 1/4N \right) \\ &= \mathbb{P} \left(\sum_{i=1}^N \frac{Z_i}{\sqrt{N}} \geq \frac{1}{2} \sqrt{N} \right) \leq \exp \left(-\frac{1}{2 \cdot 4} N \right). \end{aligned}$$

This shows that Hoeffding's inequality is a good concentration estimate avoiding the approximating error using the CLT. We also get two-sided tail / concentration estimates for $S = \sum_{i=1}^n a_i X_i$ using Theorem 2.5 writing

$$\mathbb{P}(|S| \geq t) = \mathbb{P}(S \geq t) + \mathbb{P}(-S \geq t).$$

Theorem 2.6 (Two-sided Hoeffding's inequality) *Suppose that X_1, \dots, X_N , are independent symmetric Bernoulli random variables and let $a = (a_1, \dots, a_N) \in \mathbb{R}^N$. Then, for any $t \geq 0$, we have*

$$\mathbb{P} \left(\left| \sum_{i=1}^N a_i X_i \right| \geq t \right) \leq 2 \exp \left(-\frac{t^2}{2 \|a\|_2^2} \right).$$

Exercise 2.7 Prove Theorem 2.6. 

As the reader may have realised, the proof of Hoeffding's inequality, Theorem 2.5, is quite flexible as it is based on bounding the moment generating function. For example, the following version of Hoeffding's inequality is valid for general bounded random variables. The proof will be given as an exercise.

Theorem 2.8 (Hoeffding’s inequality for general bounded random variables) *Suppose that X_1, \dots, X_N are independent random variable with $X_i \in [m_i, M_i], m_i < M_i$, for $i = 1, \dots, N$. Then, for any $t \geq 0$, we have*

$$\mathbb{P}\left(\sum_{i=1}^N (X_i - \mathbb{E}[X_i]) \geq t\right) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^N (M_i - m_i)^2}\right).$$

Exercise 2.9 Prove Theorem 2.8.



Example 2.10 Let X_1, \dots, X_N non-negative independent random variables with continuous distribution (i.e., having a density with respect to the Lebesgue measure). Assume that the probability density functions f_i of X_i are uniformly bounded by 1. Then the following holds.

(a) For all $i = 1, \dots, N$,

$$\mathbb{E}[\exp(-tX_i)] = \int_0^\infty e^{-tx} f_i(x) dx \leq \int_0^\infty e^{-tx} dx = \frac{1}{t}.$$

(b) For any $\varepsilon > 0$, we have

$$\mathbb{P}\left(\sum_{i=1}^N X_i \leq \varepsilon N\right) \leq (e\varepsilon)^N. \tag{2.2}$$

To show (2.2), we use Markov’s inequality for some $\lambda > 0$, the independence, and part (a) to arrive at

$$\mathbb{P}\left(\sum_{i=1}^N \left(\frac{-X_i}{\varepsilon}\right) \geq -N\right) \leq e^{\lambda N} \prod_{i=1}^N \mathbb{E}[e^{-\lambda/\varepsilon X_i}] \leq e^{\lambda N} \left(\frac{\varepsilon}{\lambda}\right)^N = e^{\lambda N} e^{N \log(\varepsilon/\lambda)}.$$

Minimising over $\lambda > 0$ gives $\lambda = 1$ and thus the desired statement.



2.3 Chernoff’s Inequality

Hoeffding’s inequality is already good but it does not produce good results in case the success probabilities/parameter p_i are very small. In that case one knows that the sum S_N of N Bernoulli random variables has an approximately Poisson distribution. The Gaussian tails are not good enough to match Poisson tails as we will see later.

Theorem 2.11 (Chernoff’s inequality) *Let X_i be independent Bernoulli random variables with parameter $p_i \in [0, 1]$, i.e., $X_i \sim \text{Ber}(p_i)$, $i = 1, \dots, N$. Denote $S_N = \sum_{i=1}^N X_i$ their sum and $\mu = \mathbb{E}[S_N]$ its mean. Then, for any $t > \mu$, we have*

$$\mathbb{P}\left(S_N \geq t\right) \leq e^{-\mu} \left(\frac{e\mu}{t}\right)^t.$$

Proof. Let $\lambda > 0$, then Markov's inequality gives

$$\mathbb{P}(S_N \geq t) = \mathbb{P}(e^{\lambda S_N} \geq e^{\lambda t}) \leq e^{-\lambda t} \prod_{i=1}^N \mathbb{E}[\exp(\lambda X_i)].$$

We need to bound the MGF of each Bernoulli random variable X_i separately. Using $1 + x \leq e^x$, we get

$$\mathbb{E}[\exp(\lambda X_i)] = e^\lambda p_i + (1 - p_i) = 1 + (e^\lambda - 1)p_i \leq \exp((e^\lambda - 1)p_i).$$

Thus

$$\prod_{i=1}^N \mathbb{E}[\exp(\lambda X_i)] \leq \exp\left(\left(e^\lambda - 1\right) \sum_{i=1}^N p_i\right) = \exp\left(\left(e^\lambda - 1\right)\mu\right),$$

and therefore

$$\mathbb{P}(S_N \geq t) \leq e^{-\lambda t} \exp((e^\lambda - 1)\mu) \quad \text{for all } \lambda > 0.$$

Define $g(\lambda) := -\lambda t + (e^\lambda - 1)\mu$. Optimising g over $\lambda > 0$, we obtain $\lambda = \log(t/\mu)$ as $t > \mu$ for the minimal upper bound. \square

Exercise 2.12 In the setting of Theorem 2.11, prove that, for any $t < \mu$,

$$\mathbb{P}(S_N \leq t) \leq e^{-\mu} \left(\frac{e\mu}{t}\right)^t.$$



Solution. We get

$$\mathbb{P}(S_N \leq t) = \mathbb{P}(-S_N \geq -t) \leq \mathbb{E}[e^{-\lambda S_N}] e^{\lambda t}$$

and

$$\prod_{i=1}^N \mathbb{E}[e^{-\lambda X_i}] \leq \exp((e^{-\lambda} - 1)\mu).$$

Thus

$$\mathbb{P}(S_N \leq t) \leq e^{\lambda t} \exp((e^{-\lambda} - 1)\mu).$$

Minimising over $\lambda > 0$, gives $-\lambda = \log(t/\mu)$ which is valid as $t < \mu$. \odot

We shall now discuss some example on random graphs where concentration inequalities provide sufficient bounds.

Example 2.13 (Degrees of Random Graphs) We consider the Erdős-Rényi random graph model. This is the simplest stochastic model for large, real-world networks. The random graph $G(n, p)$ consists of n vertices, and each pair of distinct vertices is connected independently (from all other pairs) with probability $p \in [0, 1]$. The degree of a vertex is the

(random) number of edges incident to that vertex. The expected degree of every vertex i (we label the n vertices by numbers $1, \dots, n$), is

$$d(i) = \mathbb{E}[D(i)] = (n-1)p, \quad i = 1, \dots, n.$$

Note that $d(i) = d(j)$ for all $i \neq j, i, j = 1, \dots, n$, and thus we simply write d in the following. We call the random graph dense when $d \geq \log n$ holds, and the graph is denoted regular if the degree of all vertices approximately equal d .



Proposition 2.14 (Dense graphs are almost regular) *There is an absolute constant $C > 0$ such that the following holds: Suppose that $G(n, p)$ has expected degree $d \geq C \log n$. Then, with high probability, all vertices of $G(n, p)$ have degrees between $0.9d$ and $1.1d$.*

Proof. Pick a vertex i of the graph, the degree is simply a sum of Bernoulli random variables, i.e.,

$$D(i) = \sum_{k=1}^{n-1} X_k, \quad X_k \sim \text{Ber}(p).$$

We are using the two-sided version of the Chernoff bound in Theorem 2.11, see Exercise 2(a) of Example Sheet 1:

$$\mathbb{P}(|D(i) - d| \geq 0.1d) \leq 2e^{-cd} \quad \text{for some absolute constant } c > 0.$$

We can now 'unfix' i by taking the union bound over all vertices of the graph:

$$\mathbb{P}(\exists i \in \{1, \dots, n\}: |D(i) - d| \geq 0.1d) \leq \sum_{i=1}^n \mathbb{P}(|D(i) - d| \geq 0.1d) \leq 2ne^{-cd}.$$

If $d \geq C \log n$ for a sufficiently large absolute constant C , the probability on the left hand side is bounded as

$$\mathbb{P}(\exists i \in \{1, \dots, n\}: |D(i) - d| \geq 0.1d) \leq 2ne^{-cC \log n} = 2n^{1-cC} \leq 0.1.$$

This means that, with probability 0.9, the complementary event occurs and we have

$$\mathbb{P}(\forall i \in \{1, \dots, n\}: |D(i) - d| < 0.1d) \geq 0.9.$$



2.4 Sub-Gaussian random variables

In this section we introduce a family of random variables, the so-called *sub-Gaussian random variables*, who show exponential concentration bounds similar to the Normal distribution (Gaussian). Before we define this class via various properties, we discuss some facts on Chernoff bounds.

Suppose the real-valued random variable X has mean $\mu \in \mathbb{R}$ and there is some constant $b > 0$ such that the function

$$\Phi(\lambda) := \mathbb{E}[e^{\lambda(X - \mathbb{E}[X])}] \quad (2.3)$$

the so-called *centred moment generating function* exists for all $|\lambda| \leq b$. For $\lambda \in [0, b]$, we may apply Markov's Inequality in Theorem 1.21 to the random variable $Y = e^{\lambda(X - \mathbb{E}[X])}$, thereby obtaining the upper bound

$$\mathbb{P}((X - \mu) \geq t) = \mathbb{P}(e^{\lambda(X - \mathbb{E}[X])} \geq e^{\lambda t}) \leq \frac{\Phi(\lambda)}{e^{\lambda t}}.$$

Optimising our choice of λ so as to obtain the tightest results yields the so-called *Chernoff bound*, namely, the inequality

$$\log \mathbb{P}((X - \mu) \geq t) \leq \inf_{\lambda \in [0, b]} \{ \log \Phi(\lambda) - \lambda t \}. \quad (2.4)$$

Example 2.15 (Gaussian tail bounds) Let $X \sim N(\mu, \sigma^2)$, $\mu \in \mathbb{R}$, $\sigma^2 > 0$, and recall the probability density function (p.d.f) of X ,

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

The moment generating function (MGF) is easily computed as

$$\begin{aligned} M_X(\lambda) &= \mathbb{E}[e^{\lambda X}] = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{\mathbb{R}} e^{-\frac{(x-\mu)^2}{2\sigma^2} + \lambda x} dx = \frac{e^{\lambda\mu}}{\sqrt{2\pi\sigma^2}} \int_{\mathbb{R}} e^{-y^2/2\sigma^2 + \lambda y} dy \\ &= e^{\lambda\mu + \lambda^2\sigma^2/2} < \infty \end{aligned}$$

for all $\lambda \in \mathbb{R}$. We obtain the Chernoff bound by optimising over $\lambda \geq 0$ using $\Phi(\lambda) = e^{-\lambda\mu} M_X(\lambda) = e^{\lambda^2\sigma^2/2}$,

$$\inf_{\lambda \geq 0} \{ \log \Phi(\lambda) - \lambda t \} = \inf_{\lambda \geq 0} \left\{ \frac{\lambda^2\sigma^2}{2} - \lambda t \right\} = -\frac{t^2}{2\sigma^2}.$$

Thus any $N(\mu, \sigma^2)$ random variable X satisfies the *upper deviation inequality*

$$\mathbb{P}(X \geq \mu + t) \leq \exp\left(-\frac{t^2}{2\sigma^2}\right), \quad t \geq 0,$$

and the two-sided version

$$\mathbb{P}(|X - \mu| \geq t) \leq 2 \exp\left(-\frac{t^2}{2\sigma^2}\right), \quad t \geq 0.$$



Exercise 2.16 (Moments of the normal distribution) Let $p \geq 1$ and $X \sim N(0, 1)$. Then

$$\begin{aligned} \|X\|_{L^p} &= (\mathbb{E}[|X|^p])^{1/p} = \left(\frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} |x|^p e^{-x^2/2} dx \right)^{1/p} = \left(\frac{2}{\sqrt{2\pi}} \int_0^\infty x^p e^{-x^2/2} dx \right)^{1/p} \\ &\stackrel{(x^2/2=w)}{=} \left(\frac{2}{\sqrt{2\pi}} \int_0^\infty (\sqrt{2}\sqrt{w})^p e^{-w} \frac{\sqrt{2}}{\sqrt{w}} \frac{1}{2} dw \right)^{1/p} = \left(\frac{2}{\sqrt{2\pi}} \frac{\sqrt{2}^p \sqrt{2}}{2} \Gamma(p/2 + 1/2) \right)^{1/p} \\ &= \sqrt{2} \left(\frac{\Gamma(p/2 + 1/2)}{\Gamma(1/2)} \right)^{1/p}. \end{aligned}$$

Hence

$$\|X\|_{L^p} = O(\sqrt{p}) \quad \text{as } p \rightarrow \infty.$$



We also note that the MGF of $X \sim N(0, 1)$ is given as $M_X(\lambda) = e^{\lambda^2/2}$. In the following proposition we identify equivalent properties for a real-valued random variable which exhibits similar tail bounds, moment bounds and MGF estimates than a Gaussian random variable does.

Proposition 2.17 (Sub-Gaussian properties) Let X be a real-valued random variable. Then there are absolute constants $C_i > 0, i = 1, \dots, 5$, such that the following statements are equivalent:

(i) **Tails of X :**

$$\mathbb{P}(|X| \geq t) \leq 2 \exp(-t^2/C_1^2) \quad \text{for all } t \geq 0.$$

(ii) **Moments:**

$$\|X\|_{L^p} \leq C_2 \sqrt{p} \quad \text{for all } p \geq 1.$$

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

$$\mathbb{E}[\exp(\lambda^2 X^2)] \leq \exp(C_3^2 \lambda^2) \quad \text{for all } \lambda \text{ with } |\lambda| \leq \frac{1}{C_3}.$$

(iv) **MGF bound:**

$$\mathbb{E}[\exp(X^2/C_4^2)] \leq 2.$$

Moreover, if $\mathbb{E}[X] = 0$, then the statements (i)-(iv) are equivalent to the following statement

(v) **MGF bound of X :**

$$\mathbb{E}[\exp(\lambda X)] \leq \exp(C_5^2 \lambda^2) \quad \text{for all } \lambda \in \mathbb{R}.$$

Proof. (i) \Rightarrow (ii): Without loss of generality we can assume that $C_1 = 1$ for property (i). This can be seen by just multiplying the parameter t with C_1 which corresponds

to multiply X by $1/C_1$. The integral identity in Lemma 1.9 applied to the non-negative random variable $|X|^p$ gives

$$\begin{aligned}\mathbb{E}[|X|^p] &= \int_0^\infty \mathbb{P}(|X|^p > t) dt \stackrel{t=up}{=} \int_0^\infty \mathbb{P}(|X| \geq u) p u^{p-1} du \stackrel{(i)}{\leq} \int_0^\infty 2e^{-u^2} p u^{p-1} du \\ &\stackrel{u^2=s}{=} p \Gamma\left(\frac{p}{2}\right) \leq p \left(\frac{p}{2}\right)^{p/2},\end{aligned}$$

where we used the property $\Gamma(x) \leq x^x$ of the Gamma function.⁵ Taking the p th root gives $\sqrt[p]{p} \sqrt[p]{p} \sqrt{1/2}$ and $\sqrt[p]{p} \leq 2$ gives (ii) with some $C_2 \leq 2$. To see that $\sqrt[p]{p} \leq 2$ recall that $\lim_{n \rightarrow \infty} \sqrt[n]{n} = 1$. More precisely, in that proof one takes $\sqrt[n]{n} = 1 + \delta_n$ and shows with the binomial theorem that $0 < \delta_n < \sqrt{\frac{2}{n-1}}$.

(ii) \Rightarrow (iii): Again without loss of generality we can assume that $C_2 = 1$. Taylor expansion of the exponential function and linearity of the expectation gives

$$\mathbb{E}[\exp(\lambda^2 X^2)] = 1 + \sum_{k=1}^{\infty} \frac{\lambda^{2k} \mathbb{E}[X^{2k}]}{k!}.$$

Property (ii) implies that $\mathbb{E}[X^{2k}] \leq (2k)^k$ and Stirling's formula give⁶ we get that $k! \geq \left(\frac{k}{e}\right)^k$. Thus

$$\mathbb{E}[\exp(\lambda^2 X^2)] \leq 1 + \sum_{k=1}^{\infty} \frac{(2\lambda^2 k)^k}{(k/e)^k} = \sum_{k=0}^{\infty} (2e\lambda^2)^k = \frac{1}{1 - 2e\lambda^2},$$

provided that $2e\lambda^2 < 1$ (geometric series). To bound the right hand side, i.e., to bound $1/(1 - 2e\lambda^2)$, we use that

$$\frac{1}{1-x} \leq e^{2x} \quad \text{for all } x \in [0, 1/2].$$

For all $|\lambda| \leq \frac{1}{2\sqrt{e}}$ we have $2e\lambda^2 < \frac{1}{2}$, and thus

$$\mathbb{E}[\exp(\lambda^2 X^2)] \leq \exp(4e\lambda^2) \quad \text{for all } |\lambda| \leq \frac{1}{2\sqrt{e}},$$

⁵ $\Gamma(n) = (n-1)!, n \in \mathbb{N}$.

$$\begin{aligned}\Gamma(z) &:= \int_0^\infty x^{z-1} e^{-x} dx \quad \text{for all } z = x + iy \in \mathbb{C} \text{ with } x > 0, \\ \Gamma(z+1) &= z\Gamma(z), \\ \Gamma(1) &= 1, \\ \Gamma\left(\frac{1}{2}\right) &= \sqrt{\pi}.\end{aligned}$$

⁶Stirling's formula:

$$N! \sim \sqrt{2\pi N} e^{-N} N^N \quad \text{where } \sim \text{ means quotient goes to zero when } N \rightarrow \infty.$$

Alternatively,

$$N! \sim \sqrt{2\pi N} e^{-N} N^N \left(1 + \frac{1}{12N} + O\left(\frac{1}{N^2}\right)\right).$$

and (iii) follows with $C_3 = 2\sqrt{e}$.

(iii) \Rightarrow (iv): Trivial and left as an exercise.

(iv) \Rightarrow (i): Again, we assume without loss of generality that $C_4 = 1$. Using Markov's inequality 1.21 with $f = \exp \circ x^2$, we obtain

$$\mathbb{P}(|X| \geq t) = \mathbb{P}(e^{X^2} \geq e^{t^2}) \leq e^{-t^2} \mathbb{E}[e^{X^2}] \stackrel{(iv)}{\leq} 2e^{-t^2}.$$

Suppose now that $\mathbb{E}[X] = 0$. We show that **(iii) \Rightarrow (v)** and **(v) \rightarrow (i)**.

(iii) \Rightarrow (v): Again, we assume without loss of generality that $C_3 = 1$. We use the well-known inequality

$$e^x \leq x + e^{x^2}, \quad x \in \mathbb{R},$$

to obtain

$$\mathbb{E}[e^{\lambda X}] \leq \mathbb{E}[\lambda X + e^{\lambda^2 X^2}] = \mathbb{E}[e^{\lambda^2 X^2}] \stackrel{(iii)}{\leq} e^{\lambda^2} \quad \text{if } |\lambda| \leq 1,$$

and thus we have (v) for the range $|\lambda| \leq 1$. Now, assume $|\lambda| \geq 1$. This time, use the inequality $2\lambda x \leq \lambda^2 + x^2$, $\lambda, x \in \mathbb{R}$, to arrive at

$$\mathbb{E}[e^{\lambda X}] \leq e^{\lambda^2/2} \mathbb{E}[e^{X^2/2}] \leq e^{\lambda^2/2} \exp(1/2) \leq e^{\lambda^2} \quad \text{as } |\lambda| \geq 1.$$

(v) \Rightarrow (i): Again, assume that $C_5 = 1$. Let $\lambda \geq 0$.

$$\mathbb{P}(X \geq t) = \mathbb{P}(e^{\lambda X} \geq e^{\lambda t}) \leq e^{-\lambda t} \mathbb{E}[e^{\lambda X}] \leq e^{-\lambda t} e^{\lambda^2} = \exp(-\lambda t + \lambda^2).$$

Optimising over $\lambda \geq 0$, we obtain $\lambda = t/2$, and thus $\mathbb{P}(x \geq t) \leq e^{-t^2/4}$. Now we repeat the argument for $-X$ and obtain $\mathbb{P}(X \leq -t) \leq e^{-t^2/4}$, and thus

$$\mathbb{P}(|X| \geq t) \leq 2e^{-t^2/4},$$

and $C_1 = 2$ implies (i). □

Definition 2.18 (Sub-Gaussian random variable, first definition) A random real-valued variable X that satisfies one of the equivalent statements (i)-(iv) in Proposition 2.17 is called *sub-Gaussian random variable*. Define, for any sub-Gaussian random variable

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

Exercise 2.19 (Sub-Gaussian norm) Let X be a sub-Gaussian random variable and define

$$\|X\|_{\psi_2} := \inf \left\{ t > 0: \mathbb{E}[\exp(X^2/t^2)] \leq 2 \right\}.$$

Show that $\|\cdot\|_{\psi_2}$ is indeed a norm on the space of sub-Gaussian random variables.



Example 2.20 (Examples of Sub-Gaussian random variables) (a) $X \sim N(0, 1)$ is sub-Gaussian random variable: pick $t > 0$ with $1 > \frac{2}{t^2}$ and set $a(t) := 1 - \frac{2}{t^2}$. Then

$$\mathbb{E}[\exp(X^2/t^2)] = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{-\frac{a(t)x^2}{2}} dx = \sqrt{\frac{1}{a(t)}}$$

shows that there is an absolute constant $C > 0$ with $\|X\|_{\psi_2} \leq C$.

(b) Let X be a symmetric Bernoulli random variable, $\|X\| = 1$.

$$\mathbb{E}[e^{X^2/t^2}] = e^{1/t^2} \leq 2 \Leftrightarrow \frac{1}{\log 2} \leq t^2.$$

Thus $\|X\|_{\psi_2} = 1/\sqrt{\log 2}$.

(c) Let X be a real-valued, bounded random variable, that is, $|X| \leq b = \|X\|_{\infty}$. Thus


$$\mathbb{E}[e^{X^2/t^2}] \leq e^{\|X\|_{\infty}^2/t^2} \leq 2 \Leftrightarrow t \geq \|X\|_{\infty}/\sqrt{\log 2},$$

and $\|X\|_{\psi_2} = \|X\|_{\infty}/\sqrt{\log 2}$.



Exercise 2.21 (Exponential moments) Show that if $X \sim N(0, 1)$, the function

$$\mathbb{R} \ni \lambda \mapsto \mathbb{E}[\exp(\lambda^2 X^2)]$$

of X^2 is finite only in some bounded neighbourhood of zero. Determine this neighbourhood. 

Solution. Recall that when $X \sim N(0, 1)$, X has the probability density $\sqrt{2\pi}^{-1} e^{-x^2/2}$. Thus

$$\begin{aligned} \mathbb{E}[\exp(\lambda^2 X^2)] &= \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{\lambda^2 x^2} e^{-x^2/2} dx = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \exp\left(\left(\lambda^2 - \frac{1}{2}\right)x^2\right) dx < \infty \\ &\Leftrightarrow \left(\lambda^2 - \frac{1}{2}\right) < 0 \Leftrightarrow \lambda \in \left(-\sqrt{\frac{1}{2}}, \sqrt{\frac{1}{2}}\right) \end{aligned}$$

because $\lambda \in \left(-\sqrt{\frac{1}{2}}, \sqrt{\frac{1}{2}}\right)$ ensures that the integrand is a so-called *log-concave* density with finite integral, where a log-concave density is a probability density f which can be written as

$$f(x) = \exp(\varphi(x)), \quad \text{with } \varphi \text{ being concave, i.e., } \varphi''(x) \leq 0.$$



Recall that the sum of independent normal random variables $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$, $i = 1, \dots, N$, is normally distributed, that is,

$$S_N = \sum_{i=1}^N X_i \sim \mathcal{N}\left(\sum_{i=1}^N \mu_i, \sum_{i=1}^N \sigma_i^2\right).$$

For a proof see [Dur19] or [Geo12]. We then may wonder if the sum of independent sub-Gaussian random variables is sub-Gaussian as well. There are different statements on this depending whether $\mu_i = 0$ for all $i = 1, \dots, N$, as done in the book [Ver18], or not vanishing means. We shall follow the general case here. It proves useful to have the following equivalent definition of sub-Gaussian random variables.

Definition 2.22 (Sub-Gaussian random variables, second definition) A real-valued random variable X with mean $\mu = \mathbb{E}[X]$ is a sub-Gaussian random variable if there is a positive number σ such that

$$\mathbb{E}[e^{\lambda(X-\mu)}] \leq e^{\sigma^2 \lambda^2 / 2} \quad \text{for all } \lambda \in \mathbb{R}. \quad (2.5)$$

The constant σ satisfying (2.5) is referred to as the *sub-Gaussian parameter*; for instance, we say that X is sub-Gaussian with parameter σ when (2.5) holds.

Remark 2.23 (Sub-Gaussian definitions:) It is easy to see that our two definitions are equivalent when the random variable X has zero mean $\mu = 0$: use statement (v) of Proposition 2.17. If $\mu \neq 0$ and X satisfies (2.5), we obtain the following tail bound

$$\mathbb{P}(|X| \geq t) \leq 2 \exp\left(-\frac{1}{2\sigma^2}(t - \mu)^2\right) \quad \text{for all } t \geq 0. \quad (2.6)$$

This tail bound is not exactly the statement (i) of Proposition 2.17 as the parameter $t \geq 0$ is chosen with respect to the mean. In most cases, one is solely interested in tail estimates away from the mean. Thus the definitions are equivalent in case $\mu \neq 0$ if we limit the range for parameter t to $t \geq |\mu|$. In the literature and in applications Definition 2.22 is widely used and sometimes called sub-Gaussian for centred random variables. We use both definitions synonymously. \diamond

We now show that the sum of sub-Gaussian random variables is again sub-Gaussian, and we do so for each of the two definitions separately.

Proposition 2.24 (Sum of independent sub-Gaussian random variables)

(a) Let X_1, \dots, X_N be independent sub-Gaussian random variables with sub-Gaussian parameters σ_i , $i = 1, \dots, N$, respectively. Then $S_N = \sum_{i=1}^N X_i$ is a sub-Gaussian random variable with sub-Gaussian parameter $\sqrt{\sum_{i=1}^N \sigma_i^2}$.

(b) Let X_1, \dots, X_N be independent mean-zero sub-Gaussian random variables. Then $S_N = \sum_{i=1}^N X_i$ is a sub-Gaussian random variable, and

$$\left\| \sum_{i=1}^N X_i \right\|_{\psi_2}^2 \leq C \sum_{i=1}^N \|X_i\|_{\psi_2}^2, \quad (2.7)$$

where $C > 0$ is an absolute constant.

Proof. (a) Recall Definition 2.22, then

$$\mathbb{E}[\exp(\lambda \sum_{i=1}^N (X_i - \mu_i))] = \prod_{i=1}^N \mathbb{E}[\exp(\lambda(X_i - \mu_i))] \leq \exp(\lambda^2/2 \sum_{i=1}^N \sigma_i^2) \quad \text{for all } \lambda \in \mathbb{R}.$$

(b) We first compute the MGF of the sum S_N . Indeed, for $\lambda \in \mathbb{R}$,

$$\mathbb{E}[\exp(\lambda S_N)] = \prod_{i=1}^N \mathbb{E}[\exp(\lambda X_i)] \leq \prod_{i=1}^N \exp(C\lambda^2 \|X_i\|_{\psi_2}^2) = \exp(\lambda^2 K^2),$$

where $K^2 := C \sum_{i=1}^N \|X_i\|_{\psi_2}^2$. The inequality follows directly from Proposition 2.17, see Exercise 2.42. Recall the equivalence of (v) and (iv) in Proposition 2.17 to see that the sum S_N is sub-Gaussian and

$$\|S_N\|_{\psi_2} \leq C_1 K,$$

where $C_1 > 0$ is an absolute constant. \square

This allows us to obtain a Hoeffding bound for independent sum of sub-Gaussian random variables.

Proposition 2.25 (Hoeffding bounds for sums of independent sub-Gaussian random variables)

Let X_1, \dots, X_N be real-valued independent sub-Gaussian random variables with sub-Gaussian parameter σ_i and mean $\mu_i, i = 1, \dots, N$, respectively. Then, for all $t \geq 0$,

$$\mathbb{P}\left(\sum_{i=1}^N (X_i - \mu_i) \geq t\right) \leq \exp\left(-\frac{t^2}{2 \sum_{i=1}^N \sigma_i^2}\right).$$

Proof. Set $\tilde{S}_N := \sum_{i=1}^N (X_i - \mu_i)$. Using again the exponential version of the Markov inequality in Proposition 1.21, we obtain

$$\mathbb{P}(\tilde{S}_N \geq t) \leq \mathbb{E}[e^{\lambda \tilde{S}_N}] e^{-\lambda t} = e^{-\lambda t} \prod_{i=1}^N \mathbb{E}[e^{\lambda(X_i - \mu_i)}] \leq \exp\left(-\lambda t + \lambda^2/2 \sum_{i=1}^N \sigma_i^2\right).$$

Optimising over λ , we obtain $\lambda = t / \sum_{i=1}^N \sigma_i^2$, and conclude with the statement. \square

Example 2.26 (Bounded random variables) Let X be a real-valued random variable with mean-zero and (bounded) supported on $[a, b]$, $a < b$. Denote X' an identical independent copy of X , i.e., $X \sim X'$. For any $\lambda \in \mathbb{R}$, using Jensen's inequality 1.11,

$$\mathbb{E}_X[e^{\lambda X}] = \mathbb{E}_X[e^{\lambda(X - \mathbb{E}_{X'}[X'])}] \leq \mathbb{E}_{X, X'}[e^{\lambda(X - X')}],$$

where $\mathbb{E}_{X, X'}$ is expectation with respect to the two independent and identically distributed random variables X and X' . Suppose ε is an independent (from X and X' Rademacher function, that is ε is a symmetric Bernoulli random variable. Then $(X - X') \sim \varepsilon(X - X')$. For Rademacher functions/symmetric Bernoulli random variables ε we estimate the moment generating function as

$$\begin{aligned} \mathbb{E}[e^{\lambda \varepsilon}] &= \frac{1}{2}(e^\lambda + e^{-\lambda}) = \frac{1}{2} \left(\sum_{k=0}^{\infty} \frac{(\lambda)^k}{k!} + \sum_{k=0}^{\infty} \frac{(-\lambda)^k}{k!} \right) = \sum_{k=0}^{\infty} \frac{\lambda^{2k}}{(2k)!} \\ &\leq 1 + \sum_{k=1}^{\infty} \frac{\lambda^{2k}}{2^k k!} = e^{\lambda^2/2}. \end{aligned}$$

Thus

$$\begin{aligned} \mathbb{E}_{X, X'}[e^{\lambda(X - X')}] &= \mathbb{E}_{X, X'}[\mathbb{E}_\varepsilon[e^{\lambda \varepsilon(X - X')}]] \leq \mathbb{E}_{X, X'}[\exp(\lambda^2(X - X')^2/2)] \\ &\leq \mathbb{E}_{X, X'}[\exp(\lambda^2(b - a)^2/2)], \end{aligned}$$

as $|X - X'| \leq b - a$. Thus X sub-Gaussian with sub-Gaussian parameter $\sigma = b - a > 0$, see Definition 2.22. \clubsuit

As discussed above in Remark 2.23, in many situations and results we encounter later, we typically assume that the random variables have zero means. Then the two definitions are equivalent. In the next lemma we simply show that centering does not harm the sub-Gaussian property. This way we can see that we can actually use both our definitions for sub-Gaussian random variables.

Lemma 2.27 (Centering) *If X is sub-Gaussian random variable then $X - \mathbb{E}[X]$ is sub-Gaussian too with*

$$\|X - \mathbb{E}[X]\|_{\psi_2} \leq C\|X\|_{\psi_2}$$

for some absolute constant $C > 0$.

Proof. We use the fact that $\|\cdot\|_{\psi_2}$ is a norm. Triangle inequality gives

$$\|X - \mathbb{E}[X]\|_{\psi_2} \leq \|X\|_{\psi_2} + \|\mathbb{E}[X]\|_{\psi_2}.$$

For any $a \in \mathbb{R}$ one can find $t \geq \frac{|a|}{\log 2}$ such that $\mathbb{E}[\exp(a^2/t^2)] \leq 2$, thus

$$\|a\|_{\psi_2} = \frac{|a|}{\log 2} \leq c|a|$$

for some constant $c > 0$. Hence

$$\|\mathbb{E}[X]\|_{\psi_2} \leq c|\mathbb{E}[X]| \leq c\mathbb{E}[|X|] = c\|X\|_{L^1} \leq C\|X\|_{\psi_2}$$

using $\|X\|_{L^p} \leq C\|X\|_{\psi_2} \sqrt{p}$ for all $p \geq 1$.

□

2.5 Sub-Exponential random variables

Suppose $Y = (Y_1, \dots, Y_N) \in \mathbb{R}^N$ is a random vector with independent coordinates $Y_i \sim N(0, 1), i = 1, \dots, N$. We expect that the Euclidean norm $\|Y\|_2$ exhibits some form of concentration as the square of the norm $\|Y\|_2^2$ is the sum of independent random variables Y_i^2 . However, although the Y_i are sub-Gaussian random variables, the Y_i^2 are not. Recall our tail estimate for the Gaussian random variables and note that

$$\mathbb{P}(Y_i^2 > t) = \mathbb{P}(|Y_i| > \sqrt{t}) \leq C \exp(-\sqrt{t}^2/2) = C \exp(-t/2).$$

The tails are like those for the exponential distribution. To see that, suppose that $X \sim \text{Exp}(\lambda), \lambda > 0$, i.e., the probability density function (pdf) is

$$f_X(t) = \lambda e^{-\lambda t} \mathbb{1}\{t \geq 0\}.$$

Then, for all $t \geq 0$,

$$\mathbb{P}(X \geq t) = \int_t^\infty \lambda e^{-\lambda t} dt = e^{-\lambda t}.$$

We can therefore use our general Hoeffding bound. The following proposition summarise properties of a new class of random variables.

Proposition 2.28 (Sub-exponential properties) *Let X be a real-valued random variable. Then the following properties are equivalent for absolute constant $C_i > 0, i = 1, \dots, 5$:*

(i) **Tails:**

$$\mathbb{P}(|X| \geq t) \leq 2 \exp(-t/C_1) \quad \text{for all } t \geq 0.$$

(ii) **Moments:**

$$\|X\|_{L^p} = (\mathbb{E}[|X|^p])^{1/p} \leq C_2 p \quad \text{for all } p \geq 1.$$

(iii) **Moment generating function (MGF) of $|X|$:**

$$\mathbb{E}[\exp(\lambda|X|)] \leq \exp(C_3 \lambda) \quad \text{for all } \lambda \text{ such that } 0 \leq \lambda \leq 1/C_3.$$

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

$$\mathbb{E}[\exp(|X|/C_4)] \leq 2.$$

Moreover, if $\mathbb{E}[X] = 0$, the properties (i)-(iv) are also equivalent to (v):

(v) **MGF of X :**

$$\mathbb{E}[\exp(\lambda X)] \leq \exp(C_5^2 \lambda^2) \quad \text{for all } \lambda \text{ such that } |\lambda| \leq 1/C_5.$$

Definition 2.29 (Sub-exponential random variable, first definition) A real-valued random variable X is called *Sub-exponential* if it satisfies one of the equivalent properties (i)-(iv) of Proposition 2.28 (respectively properties (i)-(v) when $\mathbb{E}[X] = 0$). Define $\|X\|_{\psi_1}$ by

$$\|X\|_{\psi_1} := \inf\{t > 0: \mathbb{E}[\exp(|X|/t)] \leq 2\}. \quad (2.8)$$

Lemma 2.30 Equation (2.8) defines the sub-exponential norm $\|X\|_{\psi}$ on the set of all sub-exponential random variables X .

Proof. Supportclass. □

Lemma 2.31 A real-valued random variable X is sub-Gaussian if and only if X^2 is sub-exponential. Moreover,

$$\|X^2\|_{\psi_1} = \|X\|_{\psi_1}^2. \quad (2.9)$$

Proof. Suppose X is sub-Gaussian and $t \geq 0$. Then

$$\mathbb{P}(|X|^2 \geq t) = \mathbb{P}(|X| \geq \sqrt{t}) \leq 2 \exp(-\sqrt{t}^2/C_1^2),$$

and thus X^2 is sub-exponential according to Proposition 2.28 (i). If X^2 is sub-exponential we have

$$\mathbb{P}(|X| \geq t) = \mathbb{P}(|X|^2 \geq t^2) \leq 2 \exp(-t^2/C_1),$$

and thus X is sub-Gaussian. As for the norms, recall that $\|X^2\|_{\psi_1}$ is the infimum of all numbers $C > 0$ satisfying $\mathbb{E}[\exp(X^2/C)] \leq 2$, whereas $\|X\|_{\psi_2}$ is the infimum of all numbers $K > 0$ satisfying $\mathbb{E}[\exp(X^2/K^2)] \leq 2$. Putting $C = K^2$, one obtains $\|X^2\|_{\psi_1} = \|X\|_{\psi_2}^2$. □

Lemma 2.32 (Product of sub-Gaussians) Let X and Y real-valued sub-Gaussian random variables. Then XY is sub-exponential and

$$\|XY\|_{\psi_1} \leq \|X\|_{\psi_2} \|Y\|_{\psi_2}.$$

Proof. Suppose that $0 \neq \|X\|_{\psi_2}$ (if $\|X\|_{\psi_2} = 0$ and/or $\|Y\|_{\psi_2} = 0$ the statement trivially holds). Then $\tilde{X} = X/\|X\|_{\psi_2}$ is sub-Gaussian with $\|\tilde{X}\|_{\psi_2} = \frac{1}{\|X\|_{\psi_2}} \|X\|_{\psi_2} = 1$. Thus we assume without loss of generality that $\|X\|_{\psi_2} = \|Y\|_{\psi_2} = 1$. To prove the statement in the lemma we shall show that $\mathbb{E}[\exp(X^2)] \leq 2$ and $\mathbb{E}[\exp(Y^2)] \leq 2$ both imply that $\mathbb{E}[\exp(|XY|)] \leq 2$, where $\mathbb{E}[\exp(|XY|)] \leq 2$ implies that $\|XY\|_{\psi_1} \leq 1$. We are going to use *Young's inequality* :

$$ab \leq \frac{a^2}{2} + \frac{b^2}{2} \quad \text{for } a, b \in \mathbb{R}.$$

Thus

$$\begin{aligned} \mathbb{E}[\exp(|XY|)] &\leq \mathbb{E}\left[\exp\left(\frac{X^2}{2} + \frac{Y^2}{2}\right)\right] = \mathbb{E}\left[\exp\left(\frac{X^2}{2}\right) \exp\left(\frac{Y^2}{2}\right)\right] \quad \text{Young's inequality} \\ &\leq \frac{1}{2} \mathbb{E}\left[\exp(X^2) + \exp(Y^2)\right] = \frac{1}{2}(2 + 2) = 2 \quad \text{Young's inequality.} \end{aligned}$$

□

Example 2.33 (Exponential random variables) Suppose $X \sim \text{Exp}(\alpha)$, $\alpha > 0$. Then $\mathbb{E}[X] = \frac{1}{\alpha}$ and $\text{Var}(X) = \frac{1}{\alpha^2}$. We compute for every $t \neq \frac{1}{\alpha}$ to get

$$\mathbb{E}[\exp(|X|/t)] = \int_0^\infty \alpha e^{x/t} e^{-\alpha x} dx = \left[\frac{-\alpha}{(\alpha - 1/t)} e^{-x(\alpha - 1/t)} \right]_0^\infty = \frac{\alpha}{(\alpha - 1/t)} \leq 2$$

if and only if $t \geq 2/\alpha$. Hence $\|X\|_{\psi_1} = \frac{2}{\alpha}$. ♣

Example 2.34 (Sub-exponential but not sub-Gaussian) Suppose that $Z \sim N(0, 1)$, and define $X := Z^2$. Then $\mathbb{E}[X] = 1$. For $\lambda < \frac{1}{2}$ we have

$$\mathbb{E}[e^{\lambda(X-1)}] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty e^{\lambda(z^2-1)} e^{-z^2/2} dz = \frac{e^{-\lambda}}{\sqrt{1-2\lambda}},$$

whereas for $\lambda > \frac{1}{2}$ we have $\mathbb{E}[e^{\lambda(X-1)}] = +\infty$. Thus X is not sub-Gaussian. In fact one can show, after some computation, that

$$\frac{e^{-\lambda}}{\sqrt{1-2\lambda}} \leq e^{2\lambda^2} = e^{4\lambda^2/2} \quad \text{for all } |\lambda| < \frac{1}{4}.$$

This motivates the following alternative definition of sub-exponential random variables which corresponds to the second definition for sub-Gaussian random variables in Definition 2.22. ♣

Definition 2.35 (Sub-exponential random variables, second definition) A real-valued random variable X with mean $\mu \in \mathbb{R}$ is *sub-exponential* if there exist non-negative parameters (ν, α) such that

$$\mathbb{E}[\exp(\lambda(X - \mu))] \leq e^{\nu^2 \lambda^2 / 2} \quad \text{for all } |\lambda| < \frac{1}{\alpha}. \quad (2.10)$$

Remark 2.36 (a) The random variable X in Example 2.34 is sub-exponential with parameters $(\nu, \alpha) = (2, 4)$.

(b) It is easy to see that our two definitions are equivalent when the random variable X has zero mean $\mu = 0$: use statement (v) of Proposition 2.28. If $\mu \neq 0$ and X satisfies (2.10), we obtain the following tail bound

$$\mathbb{P}(|X| \geq t) \leq 2 \exp\left(-\frac{1}{2\nu^2}(t - \mu)^2\right) \quad \text{for all } t \geq 0. \quad (2.11)$$

This tail bound is not exactly the statement (i) of Proposition 2.28 as the parameter $t \geq 0$ is chosen with respect to the mean. In most cases, one is solely interested in tail estimates away from the mean. Thus the definitions are equivalent in case $\mu \neq 0$ if we limit the range for parameter t to $t \geq |\mu|$. In the literature and in applications Definition 2.35 is widely used and sometimes called sub-exponential for centred random variables. We use both definitions synonymously. ◇

Example 2.37 (Bounded random variable) Let X be a real-valued, mean-zero and bounded random variable support on the compact interval $[a, b]$, $a < b$. Furthermore, let X' be an independent copy of X and let ε be an independent (from both X and X') *Rademacher function*, that is, ε is a symmetric Bernoulli random variable with $\mathbb{P}(\varepsilon = -1) = \mathbb{P}(\varepsilon = 1) = \frac{1}{2}$. Using Jensen's inequality for X' we obtain

$$\mathbb{E}_X[e^{\lambda X}] = \mathbb{E}_X[e^{\lambda(X - \mathbb{E}_{X'}[X'])}] \leq \mathbb{E}_{X, X'}[e^{\lambda(X - X')}] = \mathbb{E}_{X, X'}[\mathbb{E}_\varepsilon[e^{\lambda\varepsilon(X - X')}]],$$

where we write $\mathbb{E}_{X, X'}$ for the expectation with respect to X and X' , \mathbb{E}_ε for the expectation with respect to ε , and where we used that $(X - X') \sim \varepsilon(X - X')$. Here, \sim means that both random variables have equal distribution. Hold $\alpha := (X - X')$ fixed and compute

$$\begin{aligned} \mathbb{E}_\varepsilon[e^{\lambda\varepsilon\alpha}] &= \frac{1}{2}[e^{-\lambda\alpha} + e^{\lambda\alpha}] = \frac{1}{2}\left(\sum_{k=0}^{\infty} \frac{(-\lambda\alpha)^k}{k!} + \sum_{k=0}^{\infty} \frac{(\lambda\alpha)^k}{k!}\right) = \sum_{k=0}^{\infty} \frac{(\lambda\alpha)^{2k}}{(2k)!} \\ &\leq 1 + \sum_{k=1}^{\infty} \frac{(\lambda\alpha)^{2k}}{2^k k!} = e^{\lambda^2\alpha^2/2}. \end{aligned}$$

Inserting this result in the previous one leads to

$$\mathbb{E}_X[e^{\lambda X}] \leq \mathbb{E}_{X, X'}\left[e^{\lambda^2(X - X')^2/2}\right] \leq e^{\lambda^2(b-a)^2/2}$$

as $|X - X'| \leq b - a$. ♣

Proposition 2.38 (Sub-exponential tail-bound) Let X be a real-valued sub-exponential random variable with parameters (ν, α) and mean $\mu = \mathbb{E}[X]$. Then, for every $t \geq 0$,

$$\mathbb{P}(X - \mu \geq t) \leq \begin{cases} e^{-\frac{t^2}{2\nu^2}} & \text{for } 0 \leq t < \nu^2/\alpha, \\ e^{-\frac{t}{2\alpha}} & \text{for } t \geq \nu^2/\alpha. \end{cases} \quad (2.12)$$

Proof. Recall Definition 2.35 and obtain

$$\mathbb{P}(X - \mu \geq t) \leq e^{-\lambda t} \mathbb{E}[e^{\lambda(X - \mu)}] \leq \exp\left(-\lambda t + \frac{\lambda^2\nu^2}{2}\right) \quad \text{for } \lambda \in [0, \alpha^{-1}).$$

Define $g(\lambda, t) := -\lambda t + \lambda^2\nu^2/2$. We need to determine $g^*(t) = \inf_{\lambda \in [0, \alpha^{-1})} \{g(\lambda, t)\}$. Suppose that t is fixed, then $\partial_\lambda g(\lambda, t) = -t + \lambda\nu^2 = 0$ if and only if $\lambda = \lambda^* = \frac{t}{\nu^2}$. If $0 \leq t < \nu^2/\alpha$, then the infimum equals the unconstrained one and $g^*(t) = -t^2/2\nu^2$ for $t \in [0, \nu^2/\alpha)$. Suppose now that $t \geq \nu^2/\alpha$. As $g(\cdot, t)$ is monotonically decreasing on $[0, \lambda^*)$ (derivative is not positive), the constrained infimum is achieved on the boundary $\lambda^* = \alpha^{-1}$, and hence $g^*(t) = -t/2\alpha$. \square

Definition 2.39 (Bernstein condition) A real-valued random variable X with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 \in (0, \infty)$ satisfies the *Bernstein condition* with parameter $b > 0$ if

$$|\mathbb{E}[(X - \mu)^k]| \leq \frac{1}{2}k!\sigma^2b^{k-2} \quad k = 2, 3, 4, \dots \quad (2.13)$$

Exercise 2.40 (a) Show that a bounded random variable X with $|X - \mu| \leq b$ with variance $\sigma^2 > 0$ satisfies the *Bernstein condition* (2.13).

(b) Show that the bounded random variable X in (a) is sub-exponential and derive a bound on the centred moment generating function

$$\mathbb{E}[\exp(\lambda(X - \mathbb{E}[X]))].$$



Solution. (a) From our assumption we have $\mathbb{E}[(X - \mu)^2] = \mathbb{E}[|X - \mu|^2] = \sigma^2$ and $\text{ess sup}|X - \mu|^{k-2} \leq b^{k-2} \leq b^{k-2} \frac{1}{2} k!$ for $k \in \mathbb{N}, k \geq 2$. Using Hölder's inequality we obtain

$$\mathbb{E}[|X - \mu|^{k-2} |X - \mu|^2] \leq \mathbb{E}[|X - \mu|^2] \text{ess sup}|X - \mu|^{k-2} \leq \sigma^2 b^{k-2} \frac{1}{2} k!$$

for all $k \in \mathbb{N}, k \geq 2$.

(b) By power series expansion we have (using the Bernstein bound from (a)),

$$\begin{aligned} \mathbb{E}[e^{\lambda(X-\mu)}] &= 1 + \frac{\lambda^2 \sigma^2}{2} + \sum_{k=3}^{\infty} \lambda^k \frac{\mathbb{E}[(X - \mu)^k]}{k!} \\ &\leq 1 + \frac{\lambda^2 \sigma^2}{2} + \frac{\lambda^2 \sigma^2}{2} \sum_{k=3}^{\infty} (|\lambda| b)^{k-2}, \end{aligned}$$

and for $|\lambda| < 1/b$ we can sum the geometric series to obtain

$$\mathbb{E}[e^{\lambda(X-\mu)}] \leq 1 + \frac{\lambda^2 \sigma^2 / 2}{1 - b|\lambda|} \leq \exp\left(\frac{\lambda^2 \sigma^2 / 2}{1 - b|\lambda|}\right)$$

by using $1 + t \leq e^t$. Thus X is sub-exponential as we obtain

$$\mathbb{E}[e^{\lambda(X-\mu)}] \leq \exp(\lambda^2 (\sqrt{2}\sigma)^2 / 2)$$

for all $|\lambda| < 1/2b$.

□

Exercise 2.41 (general Hoeffding inequality) Let X_1, \dots, X_N independent mean-zero sub-Gaussian real-valued random variables, and let $a = (a_1, \dots, a_N) \in \mathbb{R}^N$. Then, for every $t \geq 0$, we have

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

where $K = \max_{1 \leq i \leq N} \{\|X_i\|_{\psi_2}\}$.



Hint: Use the fact that X_i sub-Gaussian, $i = 1, \dots, N$, implies that $\sum_{i=1}^N X_i$ is sub-Gaussian with

$$\left\|\sum_{i=1}^N X_i\right\|_{\psi_2}^2 \leq C \sum_{i=1}^N \|X_i\|_{\psi_2}^2,$$

see Proposition 2.24.

Exercise 2.42 Restate property (v) of Proposition 2.17 in terms of the sub-Gaussian norm, i.e., show that if $\mathbb{E}[X] = 0$ then

$$\mathbb{E}[\exp(\lambda X)] \leq \exp(C\lambda^2 \|X\|_{\psi_2}^2), \quad \text{for all } \lambda \in \mathbb{R}.$$



The following exercise explores different deviations from the mean.

Exercise 2.43 (Poisson distribution - various deviations) Let $X \sim \text{Poi}(\lambda)$, $\lambda > 0$. Then the following holds.

(a) For any $t > \lambda$, we have

$$\mathbb{P}(X \geq t) \leq e^{-\lambda} \left(\frac{e\lambda}{t}\right)^t.$$

(b) For $t \in (0, \lambda]$, we have

$$\mathbb{P}(|X - \lambda| \geq t) \leq 2 \exp\left(-\frac{ct^2}{\lambda}\right),$$

for some absolute $c > 0$.

Hint: Use the Poisson approximation in Theorem 1.36 in conjunction with the corresponding concentration bounds for Bernoulli random variables.



3 Random vectors in High Dimensions

We study random vectors $X = (X_1, \dots, X_n) \in \mathbb{R}^n$ and aim to obtain concentration properties of the Euclidean norm of random vectors X .

3.1 Concentration of the Euclidean norm

Suppose $X_i, i = 1, \dots, n$, are independent \mathbb{R} -valued random variables with $\mathbb{E}[X_i] = 0$ and $\text{Var}(X_i) = 1$. Then

$$\mathbb{E}[\|X\|_2^2] = \mathbb{E}\left[\sum_{i=1}^n X_i^2\right] = \sum_{i=1}^n \mathbb{E}[X_i^2] = n.$$

We thus expect that the expectation of the Euclidean norm is approximately \sqrt{n} . We will now see in a special case that the norm $\|X\|_2$ is indeed very close to \sqrt{n} with high probability.

Theorem 3.1 (Concentration of the norm) Let $X = (X_1, \dots, X_n) \in \mathbb{R}^n$ be a random vector with independent sub-Gaussian coordinates X_i that satisfy $\mathbb{E}[X_i^2] = 1$. Then

$$\left| \|X\|_2 - \sqrt{n} \right|_{\psi_2} \leq CK^2,$$

where $K := \max_{1 \leq i \leq n} \{ \|X_i\|_{\psi_2} \}$ and $C > 0$ an absolute constant.

The following two exercises are used in the proof of the theorem.

Exercise 3.2 (Centering for sub-exponential random variables) Show the centering lemma for sub-exponential random variables. This is an extension of Lemma 2.27 to sub-exponential random variables: Let X be a real-valued sub-exponential random variable. Then

$$\|X - \mathbb{E}[X]\|_{\psi_1} \leq C\|X\|_{\psi_1}.$$

☹

Exercise 3.3 (Bernstein inequality) Let X_1, \dots, X_N be independent mean-zero sub-exponential real-valued random variables. Then, for every $t \geq 0$, we have

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

for some absolute constant $c > 0$ and where $K := \max_{1 \leq i \leq N} \{\|X_i\|_{\psi_1}\}$.

☹☹

Solution. $S_N := \frac{1}{N} \sum_{i=1}^N X_i$. As usual we start with

$$\mathbb{P}(S_N \geq t) \leq e^{-\lambda t} \mathbb{E}[e^{\lambda S_N}] = e^{-\lambda t} \prod_{i=1}^N \mathbb{E}[e^{(\lambda/N)X_i}],$$

and (v) in Proposition 2.28 implies that, writing $\tilde{X}_i = X_i/N$, there are $\tilde{c} > 0$ and $C > 0$ such that

$$\mathbb{E}[e^{\lambda \tilde{X}_i}] \leq \exp(C\lambda^2 \|\tilde{X}_i\|_{\psi_1}^2) \quad \text{for } |\lambda| \leq \frac{\tilde{c}}{K},$$

and $\|\tilde{X}_i\|_{\psi_1}^2 = 1/N^2 \|X_i\|_{\psi_1}^2$, $\tilde{K} := \max_{1 \leq i \leq N} \{\|\tilde{X}_i\|_{\psi_1}\}$. With $\tilde{\sigma}^2 = \sum_{i=1}^N 1/N^2 \|X_i\|_{\psi_1}^2$ we thus get

$$\mathbb{P}(S_N \geq t) \leq \exp(-\lambda t + C\lambda^2 \tilde{\sigma}^2) \quad \text{for } |\lambda| \leq \frac{\tilde{c}}{\tilde{K}}.$$

Define $g(\lambda) := -\lambda t + C\lambda^2 \tilde{\sigma}^2$. Then $g'(\lambda) = -t + 2C\tilde{\sigma}^2 \lambda$. The zero of the derivative is at $\lambda = \frac{t}{2C\tilde{\sigma}^2}$. As long as $t \leq \frac{2C\tilde{\sigma}^2 \tilde{c}}{\tilde{K}}$, this $\lambda \leq \frac{\tilde{c}}{\tilde{K}}$ satisfies the constraint. For $t > \frac{2C\tilde{\sigma}^2 \tilde{c}}{\tilde{K}}$ we see that $g'(\lambda) \leq 0$, and thus the function is monotonically decreasing and the infimum will be attained at the upper bound for λ . Hence, optimising over λ one obtains

$$\lambda = \min\left\{\left(\frac{t}{2C\tilde{\sigma}^2}\right), \left(\frac{\tilde{c}}{\tilde{K}}\right)\right\}.$$

Inserting these values into the function g , we obtain

$$g(t/(2C\tilde{\sigma}^2)) = -t^2/(4C\tilde{\sigma}^2) \quad \text{for } t \leq \frac{2C\tilde{\sigma}^2 \tilde{c}}{\tilde{K}},$$

and for the other value we note that for $t > (2C\tilde{\sigma}^2 \tilde{c})/\tilde{K}$ we get

$$g\left(\frac{\tilde{c}}{\tilde{K}}\right) = -\frac{\tilde{c}}{\tilde{K}}t + \frac{C\tilde{c}^2 \tilde{\sigma}^2}{\tilde{K}^2} = -\frac{\tilde{c}t}{2\tilde{K}} - \frac{\tilde{c}t}{2\tilde{K}} + \frac{C\tilde{c}^2 \tilde{\sigma}^2}{\tilde{K}^2} \leq -\frac{\tilde{c}t}{2\tilde{K}},$$

which follows from

$$-\frac{\tilde{c}t}{2\tilde{K}} + \frac{C\tilde{c}^2\tilde{\sigma}^2}{\tilde{K}^2} \leq 0, \quad \text{for } t > (2C\tilde{\sigma}^2\tilde{c})/(\tilde{K}).$$

Thus

$$\mathbb{P}(S_N \geq t) \leq \exp\left(-\min\left\{\left(\frac{t^2}{4C\tilde{\sigma}^2}\right), \left(\frac{\tilde{c}t}{2\tilde{K}}\right)\right\}\right),$$

and finally, using $\tilde{\sigma}^2 \leq \frac{1}{N}K^2$ and $\tilde{K} \leq \frac{1}{N}K$, there is an absolute constant $c > 0$ such that

$$\mathbb{P}(S_N \geq t) \leq \exp\left(-c \min\left\{\left(\frac{t^2}{K^2}\right), \left(\frac{t}{K}\right)\right\}N\right).$$

To conclude the proof one needs to derive the complementary bound to derive the concentration for the absolute value. \odot

Remark 3.4 With high probability, e.g., with probability 0.99 (adjust the absolute constants $c > 0$ and $K > 0$ accordingly) X stays within a constant distance from the sphere of radius \sqrt{n} . $S_n := \|X\|_2^2$ has mean n and standard deviation $O(\sqrt{n})$:

$$\begin{aligned} \text{Var}\left(\|X\|_2^2\right) &= \mathbb{E}\left[\left(\|X\|_2^2 - n\right)^2\right] = \mathbb{E}\left[\sum_{i,j=1}^n (X_i X_j)^2\right] - 2n\|X\|_2^2 + n^2 \\ &= \mathbb{E}\left[\sum_{i,j=1}^n (X_i X_j)^2\right] - n^2 = \sum_{i=1}^n \mathbb{E}[X_i^4] + \sum_{i,j=1, i \neq j}^n \mathbb{E}[(X_i X_j)^2] - n^2. \end{aligned}$$

For any $i \neq j$ we have $\mathbb{E}[(X_i X_j)^2] = \mathbb{E}[X_i^2]\mathbb{E}[X_j^2] = 1$ (the X_i 's are independent). Furthermore, we get $\mathbb{E}[X_i^4] = O(1)$ because from (ii) in Proposition 2.17 we estimate

$$\mathbb{E}[X_i^4] \leq \|X_i\|_{L^4}^4 \leq C\sqrt{4}^4 = 16C = O(1).$$

Thus $\text{Var}(\|X\|_2^2) \leq 16nC + n(n-1) - n^2 = C'n = O(1)n$, and therefore $\sqrt{\text{Var}(\|X\|_2^2)} = O(\sqrt{n})$. Hence $\|X\|_2 = \sqrt{S_n}$ deviates by $O(1)$ around \sqrt{n} . Note that this follows from

$$\sqrt{n \pm O(\sqrt{n})} = \sqrt{n} \sqrt{1 \pm \frac{1}{n}O(\sqrt{n})} = \sqrt{n} \left(1 \pm O\left(\frac{1}{\sqrt{n}}\right)\right) \sqrt{n} \pm O(1).$$

\diamond

Proof of Theorem 3.1. We assume again without loss of generality that $K \geq 1$.

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

X_i sub-Gaussian implies that that $X_i^2 - 1$ is sub-exponential (Lemma 2.32). The centering property of Exercise 3.2 shows that there is an absolute constant $C > 0$ such that

$$\|X_i^2 - 1\|_{\psi_1} \leq C\|X_i^2\|_{\psi_1} \stackrel{\text{Lemma 2.31}}{=} C\|X_i\|_{\psi_2}^2 \leq CK^2.$$

We now apply the Bernstein inequality in Exercise 3.3 to obtain, for every $u \geq 0$,

$$\mathbb{P}\left(\left|\frac{1}{n}\|X\|_2^2 - 1\right| \geq u\right) \leq 2 \exp\left(-\frac{cn}{K^4} \min\{u^2, u\}\right), \quad (3.1)$$

where we used that $K \geq 1$ implies $K^4 \geq K^2$. Note that for $z \geq 0$ inequality $|z - 1| \geq \delta$ implies the inequality

$$|z^2 - 1| \geq \max\{\delta, \delta^2\}.$$

To see that, consider first $z \geq 1$ which implies that $z + 1 \geq z - 1 \geq \delta$ and thus $|z^2 - 1| = |z - 1||z + 1| \geq \delta^2$. For $0 \leq z < 1$ we have $z + 1 \geq 1$ and thus $|z^2 - 1| \geq \delta$. We apply this finding, (3.1) with $u = \max\{\delta, \delta^2\}$ to obtain

$$\mathbb{P}\left(\left|\frac{1}{\sqrt{n}}\|X\|_2 - 1\right| \geq \delta\right) \leq \mathbb{P}\left(\left|\frac{1}{n}\|X\|_2^2 - 1\right| \geq \max\{\delta, \delta^2\}\right) \leq 2 \exp\left(-\frac{cn}{K^4} \delta^2\right).$$

We used that $v = \min\{u, u^2\} = \delta^2$ when $u = \max\{\delta, \delta^2\}$. To see that, note that $\delta \geq \delta^2$ implies $\delta \leq 1$ and thus $u = \delta$ and $v = \delta^2$. If $\delta > 1$ we have $\delta^2 > \delta$ and thus $u = \delta^2$ and thus $v = \delta^2$. Setting $t = \delta\sqrt{n}$ we finally conclude with

$$\mathbb{P}(|\|X\|_2 - \sqrt{n}| \geq t) \leq 2 \exp\left(-\frac{ct^2}{K^4}\right),$$


which shows that $\|\|X\|_2 - \sqrt{n}\|$ is sub-Gaussian. \square

Exercise 3.5 (Small ball probabilities) Let $X = (X_1, \dots, X_n) \in \mathbb{R}^n$ be a random vector with independent coordinates X_i having continuous distribution with probability density functions $f_i: \mathbb{R} \rightarrow \mathbb{R}$ (Radon-Nikodym density with respect to the Lebesgue measure) satisfying

$$|f_i(x)| \leq 1, \quad i = 1, \dots, n, \text{ for all } x \in \mathbb{R}.$$

Show that, for any $\varepsilon > 0$, we have

$$\mathbb{P}(\|X\|_2 \leq \varepsilon\sqrt{n}) \leq (C\varepsilon)^n$$

for some absolute constant $C > 0$. 

Solution.

$$\mathbb{P}(\|X\|_2^2 \leq \varepsilon^2 n) = \mathbb{P}(-\|X\|_2^2 \geq -\varepsilon^2 n) \leq e^{\lambda\varepsilon^2 n} \mathbb{E}[\exp(-\lambda\|X\|_2^2)] = e^{\lambda\varepsilon^2 n} \prod_{i=1}^n \mathbb{E}[\exp(-\lambda X_i^2)],$$

and inserting

$$\mathbb{E}[\exp(-\lambda X_i^2)] \leq \int_{\mathbb{R}} e^{-\lambda x^2} |f_i(x)| dx \leq \int_{\mathbb{R}} e^{-\lambda x^2} dx = \sqrt{\frac{\pi}{\lambda}},$$

we obtain

$$\mathbb{P}(\|X\|_2^2 \leq \varepsilon^2 n) \leq \exp\left(\lambda\varepsilon^2 n - \frac{n}{2} \log(\lambda/\pi)\right) \leq (C\varepsilon)^n,$$

where the last inequality follows by optimising over λ and getting $\lambda = \frac{1}{2\varepsilon^2}$. \odot

3.2 The geometry of high dimensions

We collect a few facts about high-dimensional Euclidean vector spaces. We begin with the volume and the area of balls.

Let $R > 0$. Then

$$B_R^{(n)} := \{x \in \mathbb{R}^n : \|x\|_2 \leq R\}$$

is called the n -dimensional ball with radius R around the origin. If $a \in \mathbb{R}^n$, we denote $B_R^{(n)}(a)$ the ball with radius R around a , $B_R^{(n)}(a) = \{x \in \mathbb{R}^n : \|x - a\|_2 \leq R\}$. If $R = 1$, we write $B^{(n)}$ respectively $B^{(n)}(a)$.

$$S_R^{(n-1)} := \{x \in \mathbb{R}^n : \|x\|_2 = R\}$$

is called n -dimensional sphere with radius R around the origin, and $S_R^{(n-1)}(a) = \{x \in \mathbb{R}^n : \|x - a\|_2 = R\}$. If $R = 1$, we write $S^{(n-1)}$ respectively $S^{(n-1)}(a)$.

$$\begin{aligned} \mathbf{vol}(B_R^{(n)}) &= \frac{\pi^{n/2}}{\frac{n}{2}\Gamma(n/2)} R^n, \\ \mathbf{area}(S_R^{(n-1)}) &= \frac{2\pi^{n/2}}{\Gamma(n/2)} R^{n-1}. \end{aligned}$$

Example 3.6 $n = 3, R = 1$. Note that $\Gamma(3/2) = \frac{1}{2}\Gamma(1/2) = \frac{1}{2}\sqrt{\pi}$, and thus $\mathbf{area}(S^{(2)}) = 4 \in$ and $\mathbf{vol}(B^{(3)}) = \frac{4}{3}\pi$. ♣

In n -dimensional polar coordinates, the volume $\mathbf{vol}(B^{(n)})$ of the n -dimensional unit ball is given by

$$\mathbf{vol}(B^{(n)}) = \int_{S^{(n-1)}} \int_0^1 r^{n-1} dr d\sigma = \frac{1}{n} \int_{S^{(n-1)}} d\sigma = \frac{\mathbf{area}(S^{(n-1)})}{n}. \quad (3.2)$$

It remains to determine the surface area $\mathbf{area}(S^{(n-1)})$, that is, the surface integral in (3.2) for general $n \in \mathbb{N}$. In principle one can use the generalisation of the polar coordinates from $n = 3$ to any higher dimensions. This is slightly elaborate, and we therefore show a different and easier way to compute that area. For any $n \in \mathbb{N}$ we have

$$I(n) := \int_{\mathbb{R}} \dots \int_{\mathbb{R}} e^{-(x_1^2 + \dots + x_n^2)} dx_n \dots dx_1 = (\sqrt{\pi})^n = \pi^{n/2}. \quad (3.3)$$

Alternatively, we can compute $I(n)$ in (3.3) using polar coordinates with differential $r^{n-1} dr$ and change $t = r^2$ in the integral,

$$\begin{aligned} I(n) &= \int_{S^{(n-1)}} d\sigma \int_0^\infty e^{-r^2} r^{n-1} dr \\ &= \mathbf{area}(S^{(n-1)}) \int_0^\infty e^{-r^2} r^{n-1} dr = \mathbf{area}(S^{(n-1)}) \int_0^\infty e^{-t} t^{\frac{n-1}{2}} (1/2t^{-1/2}) dt \\ &= \mathbf{area}(S^{(n-1)}) \int_0^\infty \frac{1}{2} e^{-t} t^{\frac{n}{2}-1} dt = \mathbf{area}(S^{(n-1)}) \frac{1}{2} \Gamma\left(\frac{n}{2}\right). \end{aligned} \quad (3.4)$$

Thus

$$\mathbf{area}(S^{(n-1)}) = \frac{2\pi^{n/2}}{\Gamma(n/2)}. \quad (3.5)$$

Notation 3.7 (Landau symbols) Asymptotic analysis is concerned with the behaviour of function $f(n)$, $n \in \mathbb{N}$, as $n \rightarrow \infty$. Suppose $f, g: \mathbb{N} \rightarrow \mathbb{R}_+$ (or \mathbb{R}). We define the following Landau symbols, called big-O and little-o.

- $f(n)$ is $O(g(n))$ if there is a constant $C > 0$ such that $f(n) \leq Cg(n)$ for all $n \in \mathbb{N}$.

- $f(n)$ is $o(g(n))$ if

$$\lim_{n \rightarrow \infty} \frac{f(n)}{g(n)} = 0.$$

- $f(n) \sim g(n)$ if

$$\lim_{n \rightarrow \infty} \frac{f(n)}{g(n)} = 1.$$

We now discuss briefly the fact that most of the volume of high-dimensional objects (sets with non-vanishing volume) is near the surface of that object. Let $A \subset \mathbb{R}^n$ be a set with non-vanishing volume, i.e., $\mathbf{vol}(A) > 0$, and pick $\varepsilon > 0$ small. Now we shrink A by a small amount ε to produce

$$(1 - \varepsilon)A := \{(1 - \varepsilon)x : x \in A\}.$$

Then the following holds,

$$\mathbf{vol}((1 - \varepsilon)A) = (1 - \varepsilon)^n \mathbf{vol}(A). \quad (3.6)$$

To see (3.6), partition the set A into infinitesimal cubes (for a Riemann sum approximate of the volume integral). Then, $(1 - \varepsilon)A$ is the union of the set of cubes obtained by shrinking the cubes of the partition of A by a factor $(1 - \varepsilon)$. If we shrink each of the $2n$ sides of an n -dimensional cube Q by $(1 - \varepsilon)$, its volume $\mathbf{vol}((1 - \varepsilon)Q)$ shrinks by the factor $(1 - \varepsilon)^n$. Using that $1 - x \leq e^{-x}$, we get the following estimate of the ratio of the volumes:

$$\frac{\mathbf{vol}((1 - \varepsilon)A)}{\mathbf{vol}(A)} = (1 - \varepsilon)^n \leq e^{-n\varepsilon}. \quad (3.7)$$

Thus nearly all of the volume of A must be in the portion of A that does not belong to the region $(1 - \varepsilon)A$. For the unit ball $B^{(n)}$ we have at least a $(1 - e^{-\varepsilon n})$ fraction of the volume $\mathbf{vol}(B^{(n)})$ of the unit ball concentrated in $B^{(n)} \setminus (1 - \varepsilon)B^{(n)}$, namely in a small annulus of width ε at the boundary.

Proposition 3.8 (Volume near the equator) For $c \geq 1$ and $n \geq 3$ at least a $1 - \frac{2}{c}e^{-c^2/2}$ fraction of the volume $\mathbf{vol}(B^{(n)})$ of the unit ball has $|x_1| \leq \frac{c}{\sqrt{n-1}}$. Here, the coordinate x_1 points to the north pole.

Proof. By symmetry it suffices to prove that at most a $\frac{2}{c}e^{-c^2/2}$ fraction of the half of the ball with $x_1 \geq 0$ has $x_1 \geq \frac{c}{\sqrt{n-1}}$. Let denote $H = \{x \in B^{(n)} : x_1 \geq 0\}$ be the upper hemisphere (northern hemisphere) and $A = \{x \in B^{(n)} : x_1 \geq \frac{c}{\sqrt{n-1}}\}$. We need to show that the ratio of the volumes is bounded as

$$\frac{\mathbf{vol}(A)}{\mathbf{vol}(H)} \leq \frac{2}{c}e^{-c^2/2}. \quad (3.8)$$

We prove (3.8) by obtaining an upper bound for $\mathbf{vol}(A)$ and a lower bound for $\mathbf{vol}(H)$. To calculate the volume $\mathbf{vol}(A)$, integrate an incremental volume that is a disk of width dx_1 and whose face is a ball of dimension $n-1$ and radius $\sqrt{1-x_1^2}$. The surface area of the disk is $(1-x_1^2)^{(n-1)/2}\mathbf{vol}(B^{(n-1)})$ and the volume above the slice is

$$\mathbf{vol}(A) = \int_{c/\sqrt{n-1}}^1 (1-x_1^2)^{\frac{n-1}{2}} \mathbf{vol}(B^{(n-1)}) dx_1.$$

We obtain an upper bound by using $1-x \leq e^{-x}$, integrating up to infinity and by inserting $x_1\sqrt{n-1}/c \geq 1$ into the integral. Then

$$\begin{aligned} \mathbf{vol}(A) &\leq \mathbf{vol}(B^{(n-1)}) \frac{\sqrt{n-1}}{c} \int_{c/\sqrt{n-1}}^{\infty} x_1 e^{-\frac{(n-1)}{2}x_1^2} dx_1 \\ &= \mathbf{vol}(B^{(n-1)}) \frac{\sqrt{n-1}}{c} \left(\frac{1}{n-1}\right) e^{-c^2/2} = \frac{\mathbf{vol}(B^{(n-1)})}{c\sqrt{n-1}} e^{-c^2/2}. \end{aligned} \quad (3.9)$$

The volume of the hemisphere below the plane $x_1 = \frac{1}{\sqrt{n-1}}$ is a lower bound on the entire volume $\mathbf{vol}(H)$, and this volume is at least that of a cylinder of height $\frac{1}{\sqrt{n-1}}$ and radius $\sqrt{1-\frac{1}{n-1}}$. The volume of the cylinder is

$$\mathbf{vol}(B^{(n-1)}) \left(1 - \frac{1}{n-1}\right)^{\frac{n-1}{2}} \frac{1}{\sqrt{n-1}}.$$

Using the fact that $(1-x)^a \geq 1-ax$ for $a \geq 1$, the volume of the cylinder is at least $\frac{\mathbf{vol}(B^{(n-1)})}{2\sqrt{n-1}}$ for $n \geq 3$. Thus we obtain (3.8) from our bounds. \square

We consider the orthogonality of two random vectors. Draw two points at random from the unit ball $B^{(n)} \subset \mathbb{R}^n$. With high probability their vectors will be nearly orthogonal to each other. To understand that, recall from our previous considerations that most of the volume of the n -dimensional unit ball $B^{(n)}$ is contained in an annulus of width $O(1/n)$ near the boundary (surface), that is, we pick $\varepsilon = \frac{c}{n}$, $c \geq 1$, and have from (3.7) that

$$\frac{\mathbf{vol}((1-\varepsilon)B^{(n)})}{\mathbf{vol}(B^{(n)})} \leq e^{-\varepsilon n}.$$

Thus at least as $1 - e^{-\varepsilon n}$ fraction of $\mathbf{vol}(B^{(n)})$ is concentrated in the annulus of width ε at the boundary. Equivalently, using our result about the volume near the equator in

Proposition 3.8, if one vector points to the north pole, the other vector has projection in this direction of only $\pm O(1/\sqrt{n})$, and thus their dot/inner/scalar product will be of order $\pm O(1/\sqrt{n})$.

Proposition 3.9 *Suppose that we sample N points $X^{(1)}, \dots, X^{(N)}$ uniformly from the unit ball $B^{(n)}$. Then with probability $1 - O(1/N)$ the following holds:*

(a) $\|X^{(i)}\|_2 \geq 1 - \frac{2 \log N}{n}$ for $i = 1, \dots, N$;

(b)

$$\langle X^{(i)}, X^{(j)} \rangle \leq \frac{6 \log N}{\sqrt{n-1}} \quad \text{for all } i, j = 1, \dots, i \neq j.$$

Proof. (a) For any $i = 1, \dots, N$, the probability that $\|X^{(i)}\|_2 < 1 - \varepsilon$ is less than $e^{-\varepsilon n}$. Thus

$$\mathbb{P}\left(\|X^{(i)}\|_2 < 1 - \frac{2 \log N}{n}\right) \leq e^{-\left(\frac{2 \log N}{n}\right)n} = \frac{1}{N^2}.$$

By the union bound, the probability there exists an $i \in \{1, \dots, N\}$ such that $\|X^{(i)}\|_2 < 1 - \frac{2 \log N}{n}$ is at most $1/N$.

(b) From Proposition 3.8 we know that the component $X_1^{(i)}$ in direction of the north pole satisfies

$$\mathbb{P}\left(|X_1^{(i)}| > \frac{c}{\sqrt{n-1}}\right) \leq \frac{2}{c} e^{-c^2/2}.$$

There are $\binom{N}{2}$ pairs i and j , and for each pair we define $X^{(i)}$ as the direction of the north pole. Then the probability that the projection of the other pair vector $X^{(j)}$ onto the direction of the north pole is more than $\frac{\sqrt{6 \log N}}{\sqrt{n-1}}$ is at most $O(\exp(-6/2 \log N)) = O(1/N^3)$. Thus, the dot product condition is violated with probability at most

$$O\left(\binom{N}{2} N^{-3}\right) = O(1/N).$$

□

3.3 Covariance matrices and Principal Component Analysis (PCA)

Definition 3.10 Let $X = (X_1, \dots, X_n) \in \mathbb{R}^n$ be a random vector. Define the random matrix XX^T as the $(n \times n)$ matrix

$$XX^T := \begin{pmatrix} X_1 \\ \cdot \\ \cdot \\ \cdot \\ X_n \end{pmatrix} (X_1 \cdot \cdot \cdot X_n) = \begin{pmatrix} X_1^2 & X_1 X_2 & \cdot & \cdot & X_1 X_n \\ X_2 X_1 & X_2^2 & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ X_n X_1 & \cdot & \cdot & \cdot & X_n^2 \end{pmatrix}.$$

Let $X = (X_1, \dots, X_n) \in \mathbb{R}^n$ be a random vector with mean $\mu = \mathbb{E}[X]$ and matrix $\mu\mu^T = (\mu_i\mu_j)_{i,j=1,\dots,n}$. Then the *covariance matrix*, which is defined as

$$\text{cov}(X) := \mathbb{E}[(X - \mu)(X - \mu)^T] = \mathbb{E}[XX^T] - \mu\mu^T, \quad (3.10)$$

is a $(n \times n)$ symmetric positive-semidefinite matrix with entries

$$\text{cov}(X)_{i,j} = \mathbb{E}[(X_i - \mu_i)(X_j - \mu_j)], \quad i, j = 1, \dots, n. \quad (3.11)$$

The *second moment matrix* of a random vector $X \in \mathbb{R}^n$ is simply

$$\Sigma = \Sigma(X) = \mathbb{E}[XX^T] = (\mathbb{E}[X_i X_j])_{i,j=1,\dots,n}, \quad (3.12)$$

and $\Sigma(X)$ is symmetric and positive-semidefinite matrix which can be written as the spectral decomposition

$$\Sigma(X) = \sum_{i=1}^n s_i \mathbf{u}_i \mathbf{u}_i^T, \quad (3.13)$$


where $\mathbf{u}_i \in \mathbb{R}^n$ are the eigenvectors of $\Sigma(X)$ for the eigenvalues s_i . The second moment matrix allows the principal component analysis (PCA). We order the eigenvalues of $\Sigma(X)$ according to their size: $s_1 \geq s_2 \geq \dots \geq s_n$. For large values of the dimension n one aims to identify a few principal directions. These directions correspond to the eigenvectors with the largest eigenvalues. For example, suppose that the first m eigenvalues are significantly larger than the remaining $n - m$ ones. This allows to reduced the dimension of the given data to \mathbb{R}^m by neglecting all contributions from directions with eigenvalues significantly smaller than the chosen principal ones.

Definition 3.11 A random vector $X = (X_1, \dots, X_n) \in \mathbb{R}^n$ is called *isotropic* if

$$\Sigma(X) = \mathbb{E}[XX^T] = \mathbb{1}_n,$$

where $\mathbb{1}_n := \text{id}_n$ is the identity operator/matrix in \mathbb{R}^n .

Exercise 3.12 (a) Let Z be an isotropic mean-zero random vector in \mathbb{R}^n , $\mu \in \mathbb{R}^n$, and Σ be a $(n \times n)$ positive-semidefinite symmetric matrix. Show that then $X := \mu + \Sigma^{1/2}Z$ has mean μ and covariance matrix $\text{cov}(X) = \Sigma$.

(b) Let $X \in \mathbb{R}^n$ be a random vector with mean μ and invertible covariance matrix $\Sigma = \text{cov}(X)$. Show that then $Z := \Sigma^{-1/2}(X - \mu)$ is an isotropic mean-zero random vector. 

Lemma 3.13 (Isotropy) A random vector $X \in \mathbb{R}^n$ is isotropic if and only if

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

Proof. X is isotropic if and only if $(\mathbb{E}[X_i X_j])_{i,j=1,\dots,n} = (\delta_{i,j})_{i,j=1,\dots,n} = \mathbb{1}_n$. For every $x \in \mathbb{R}^n$ we have

$$\mathbb{E}[\langle X, x \rangle^2] = \sum_{i,j=1}^n \mathbb{E}[x_i X_i x_j X_j] = \sum_{i,j=1}^n x_i \mathbb{E}[X_i X_j] x_j = \langle x, \Sigma(X)x \rangle = \|x\|_2^2$$

if and only if $\Sigma(X) = \mathbb{1}_n$. □

It suffices to show $\mathbb{E}[\langle X, e_i \rangle^2] = 1$ for all basis vectors $e_i, i = 1, \dots, n$. Note that $\langle X, e_i \rangle$ is a one-dimensional marginal of the random vector X . Thus X is isotropic if and only if all one-dimensional marginals of X have unit variance. An isotropic distribution is evenly extended in all spatial directions.

Lemma 3.14 *Let X be an isotropic random vector in \mathbb{R}^n . Then*

$$\mathbb{E}[\|X\|_2^2] = n.$$

Moreover, if X and Y are two independent isotropic random vectors in \mathbb{R}^n , then

$$\mathbb{E}[\langle X, Y \rangle^2] = n.$$

Proof. For the first statement we view $X^T X$ as a 1×1 matrix and take advantage of the cyclic property of the trace operation on matrices:

$$\begin{aligned} \mathbb{E}[\|X\|_2^2] &= \mathbb{E}[X^T X] = \mathbb{E}[\text{Trace}(X^T X)] = \mathbb{E}[\text{Trace}(X X^T)] = \text{Trace}(\mathbb{E}[X X^T]) \\ &= \text{Trace}(\mathbb{1}_n) = n. \end{aligned}$$

We fix a realisation of Y , that is, we consider the conditional expectation of X with respect to Y which we denote \mathbb{E}_X . The law of total expectation says that

$$\mathbb{E}[\langle X, Y \rangle^2] = \mathbb{E}_Y[\mathbb{E}_X[\langle X, Y \rangle^2 | Y]],$$

where \mathbb{E}_Y denotes the expectation with respect to Y . To compute the innermost expectation we use Lemma 3.13 with $x = Y$ and obtain using the previous part that

$$\mathbb{E}[\langle X, Y \rangle^2] = \mathbb{E}_Y[\|Y\|_2^2] = n.$$

□

Suppose $X \perp Y$ are isotropic vectors in \mathbb{R}^n , and consider the normalised versions $\bar{X} := X/\|X\|_2$ and $\bar{Y} := Y/\|Y\|_2$. From the concentration results in this chapter we know that with high probability, $\|X\|_2 \sim \sqrt{n}$, $\|Y\|_2 \sim \sqrt{n}$, and $\langle X, Y \rangle \sim \sqrt{n}$. Thus, with high probability,

$$|\langle \bar{X}, \bar{Y} \rangle| \sim \frac{1}{\sqrt{n}}.$$

Thus, in high dimensions, independent and isotropic random vector are almost orthogonal.

3.4 Examples of High-Dimensional distributions

Definition 3.15 (Spherical distribution) A random vector X is called spherically distributed if it is uniformly distributed on the Euclidean sphere with radius \sqrt{n} and centre at the origin, i.e.,

$$X \sim \text{Unif}(\sqrt{n}S^{(n-1)}),$$

where

$$S^{(n-1)} = \left\{ x = (x_1, \dots, x_n) \in \mathbb{R}^n : \sum_{i=1}^n x_i^2 = 1 \right\} = \{x \in \mathbb{R}^n : \|x\|_2 = 1\}$$

is the *unit sphere* of radius 1 and centre at the origin.

Exercise 3.16 $X \sim \text{Unif}(\sqrt{n}S^{(n-1)})$ is isotropic but the coordinates $X_i, i = 1, \dots, n$, of X are not independent due to the condition $X_1^2 + \dots + X_n^2 = n$. ☹️☹️

Solution. We give a solution for $n = 2$. The solution for higher dimensions is similar and uses high-dimensional versions of the Polar coordinates. We use Polar coordinates to represent X as

$$X = \sqrt{n} \begin{pmatrix} \cos(\theta) \\ \sin(\theta) \end{pmatrix}, \quad \theta \in [0, 2\pi].$$

Let e_1 and e_2 the two unit basis vectors in \mathbb{R}^2 . It suffices to show $\mathbb{E}[\langle X, e_i \rangle^2] = 1, i = 1, 2$, as any vector can be written as a linear combination of the two basis vectors. Without loss of generality we pick e_1 (the proof for e_2 is just analogous):

$$\mathbb{E}[\langle X, e_1 \rangle^2] = \mathbb{E}[(\sqrt{n} \cos(\theta))^2] = \frac{n}{2\pi} \int_0^{2\pi} \cos^2(\theta) d\theta = \frac{2}{2\pi} \left[\frac{\theta}{2} + \frac{\sin(2\theta)}{4} \right]_0^{2\pi} = \frac{2\pi}{2\pi} = 1,$$

where we used that $\cos(2x) = 2 \cos^2(x) - 1$. ☺️

An example of a discrete isotropic distribution in \mathbb{R}^n is the *symmetric Bernoulli distribution in \mathbb{R}^n* . We say that a random vector $X = (X_1, \dots, X_n) \in \mathbb{R}^n$ is symmetric Bernoulli distributed if the coordinates X_i are independent, symmetric, Bernoulli random variables. A random variable ε is symmetric if $\mathbb{P}(\varepsilon = -1) = \mathbb{P}(\varepsilon = +1) = \frac{1}{2}$. Equivalently, we may say that X is uniformly distributed on the unit cube in \mathbb{R}^n :

$$X \sim \text{Unif}(\{-1, +1\}^n).$$

The symmetric Bernoulli distribution is isotropic. This can be easily seen again by checking $\mathbb{E}[\langle X, e_i \rangle^2] = 1$ for all $i = 1, \dots, n$, or for any $x \in \mathbb{R}^n$ by checking that

$$\mathbb{E}[\langle X, x \rangle^2] = \mathbb{E} \left[\sum_{i=1}^n X_i^2 x_i^2 \right] + \mathbb{E} \left[2 \sum_{1 \leq i < j \leq n} X_i X_j x_i x_j \right] = \|x\|_2^2,$$

as the second term vanishes because the X_i are independent mean-zero random variables .

Any random vector $X = (X_1, \dots, X_n) \in \mathbb{R}^n$ with independent mean-zero coordinates X_i with unit variance $\text{Var}(X_i) = 1$ is an isotropic random vector in \mathbb{R}^n .

For the following recall the definition of the normal distribution. See also the appendix sheets distributed at the begin of the lecture for useful Gaussian calculus formulae.

Definition 3.17 (Multivariate Normal / Gaussian distribution) We say a random vector $Y = (Y_1, \dots, Y_n) \in \mathbb{R}^n$ has standard normal distribution in \mathbb{R}^n , denoted

$$Y \sim \mathbf{N}(0, \mathbb{1}_n),$$

if the coordinates $Y_i, i = 1, \dots, n$, are independent, \mathbb{R} -valued standard normally distributed random variables, i.e., $Y_i \sim \mathbf{N}(0, 1)$. The probability density function (pdf) for Y is just the product

$$f_Y(x) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} e^{-x_i^2/2} = \frac{1}{(2\pi)^{n/2}} e^{-\|x\|_2^2/2}. \quad (3.14)$$

It is easy to check that Y is isotropic. Furthermore, as the density (3.14) only depends on the Euclidean norm, that is, the standard normal distribution in \mathbb{R}^n only depends on the length and not on the direction. In other words, the standard normal distribution in \mathbb{R}^n is rotation invariant. This reasoning is rigorously stated in the next proposition.

Proposition 3.18 Let $Y \sim \mathbf{N}(0, \mathbb{1}_n)$ and U be a $n \times n$ orthogonal matrix (i.e., $U^T U = U U^T = \mathbb{1}_n$, or equivalently, $U^{-1} = U^T$). Then

$$UY \sim \mathbf{N}(0, \mathbb{1}_n).$$

Proof. For $Z := UY$ we have

$$\|Z\|_2^2 = Z^T Z = Y^T U^T U Y = Y^T Y = \|Y\|_2^2.$$

Furthermore, $|\det(U)| = |\det(U^T)| = 1$, and thus for any vector $J \in \mathbb{C}^n$ (characteristic functions/Laplace transform), writing $z = Ux, x \in \mathbb{R}^n$,

$$\begin{aligned} \mathbb{E}_Z[e^{\langle J, Z \rangle}] &= \frac{1}{(2\pi)^{n/2}} \int_{\mathbb{R}^n} \exp\left(-\frac{1}{2}\langle z, z \rangle + \langle J, z \rangle\right) \prod_{i=1}^n dz_i \\ &= \frac{1}{(2\pi)^{n/2}} \int_{\mathbb{R}^n} \exp\left(-\frac{1}{2}\langle x, x \rangle + \langle U^T J, x \rangle\right) \prod_{i=1}^n dx_i \\ &= \frac{1}{(2\pi)^{n/2}} \exp\left(\frac{1}{2}\langle U^T J, U^T J \rangle\right) = \frac{1}{(2\pi)^{n/2}} \mathbb{E}[e^{\frac{1}{2}\langle J, J \rangle}] = \mathbb{E}_Y[e^{\langle J, Y \rangle}]. \end{aligned}$$

Thus we have shown that Z has the same characteristic function/Laplace transform then $Y \sim \mathbf{N}(0, \mathbb{1}_N)$ and therefore $Z = UY \sim \mathbf{N}(0, \mathbb{1}_n)$. \square

Let Σ be a symmetric positive-definite $n \times n$ matrix and $X \in \mathbb{R}^n$ random vector with mean $\mu = \mathbb{E}[X]$. Then

$$\begin{aligned} X \sim \mathbf{N}(\mu, \Sigma) &\Leftrightarrow Z = \Sigma^{-1/2}(X - \mu) \sim \mathbf{N}(0, \mathbb{1}_n) \\ &\Leftrightarrow f_X(x) \frac{1}{(2\pi)^{n/2} \det(\Sigma)^{1/2}} \exp\left(-\frac{1}{2}\langle X - \mu, \Sigma^{-1}(X - \mu) \rangle\right), \quad x \in \mathbb{R}^n. \end{aligned}$$

For large values of n the standard normal distribution $\mathbf{N}(0, \mathbb{1}_n)$ is not concentrated around the origin, instead it is concentrated in a thin spherical shell around the sphere of radius \sqrt{n} around the origin (shell with width of order $O(1)$). From Theorem 3.1 we obtain for $Y \sim \mathbf{N}(0, \mathbb{1}_n)$,

$$\mathbb{P}(\|Y\|_2 - \sqrt{n} \geq t) \leq 2 \exp(-Ct^2), \quad \text{for all } t \geq 0$$

and an absolute constant $C > 0$. Therefore with high probability $\|Y\|_2 \approx \sqrt{n}$, and thus with high probability,

$$Y \approx \sqrt{n}\Theta \sim \text{Unif}(\sqrt{n}\mathbb{S}^{(n-1)}),$$

with the unit direction vector $\Theta = Y/\|Y\|_2$. Henceforth, with high probability,

$$\mathbf{N}(0, \mathbb{1}_n) \approx \text{Unif}(\sqrt{n}\mathbb{S}^{(n-1)}).$$

3.5 Sub-Gaussian random variables in higher dimensions

Definition 3.19 (Sub-Gaussian random vectors) A random vector $X \in \mathbb{R}^n$ is sub-Gaussian if the one-dimensional marginals $\langle X, x \rangle$ are sub-Gaussian real-valued random variables for all $x \in \mathbb{R}^n$. Moreover,

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

Lemma 3.20 Let $X = (X_1, \dots, X_n) \in \mathbb{R}^n$ be a random vector with independent mean-zero sub-Gaussian coordinates X_i . Then X is sub-Gaussian and

$$\|X\|_{\psi_2} \leq C \max_{1 \leq i \leq n} \{\|X_i\|_{\psi_2}\},$$

for some absolute constant $C > 0$.

Proof. Let $x \in \mathbb{S}^{(n-1)}$. We are using Proposition 2.24 for the sum of independent sub-Gaussian random variables:

$$\|\langle X, x \rangle\|_{\psi_2}^2 = \left\| \sum_{i=1}^n x_i X_i \right\|_{\psi_2}^2 \stackrel{\text{Prop. 2.24}}{\leq} C \sum_{i=1}^n x_i^2 \|X_i\|_{\psi_2}^2 \leq C \max_{1 \leq i \leq n} \{\|X_i\|_{\psi_2}^2\},$$

where we used that $\sum_{i=1}^n x_i^2 = 1$, \square

Theorem 3.21 (Uniform distribution on the sphere) Let $X \in \mathbb{R}^n$ be a random vector uniformly distributed on the Euclidean sphere in \mathbb{R}^n with centre at the origin and radius \sqrt{n} , i.e.,

$$X \sim \text{Unif}(\sqrt{n}S^{(n-1)}).$$

Then X is sub-Gaussian, and $\|X\|_{\psi_2} \leq C$ for some absolute constant $C > 0$.

We will actually present two different proofs of this statement. The first uses concentration properties whereas the second employs a geometric approach.

Proof of Theorem 3.21 - Version I. See [Ver18] page 53-54. \square

Proof of Theorem 3.21 - Version II.

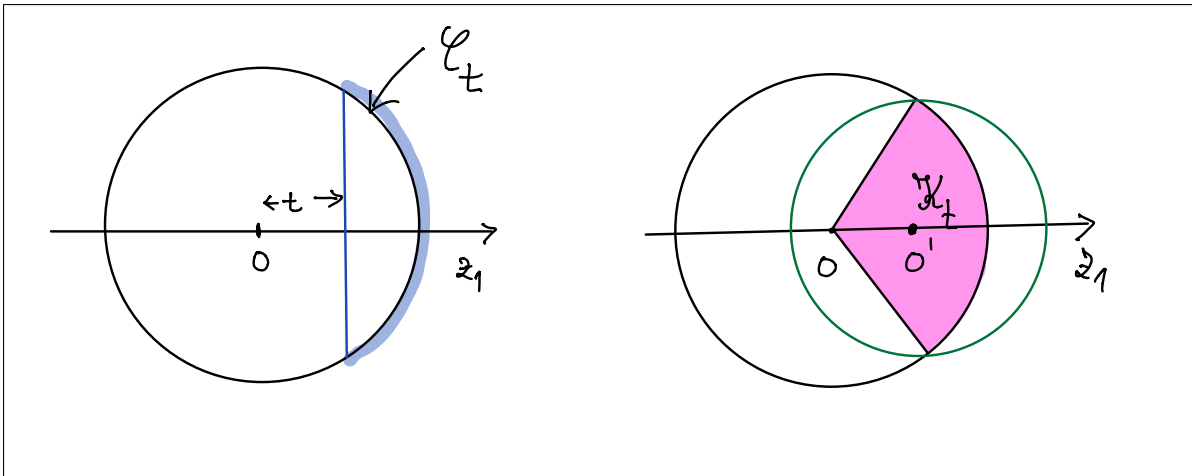


Figure 1:

For convenience we will work on the unit sphere, so let us rescale

$$Z := \frac{X}{\sqrt{n}} \sim \text{Unif}(S^{(n-1)}).$$

It suffices to show that $\|Z\|_{\psi_2} \leq C/\sqrt{n}$, which by definition means that $\|\langle Z, x \rangle\|_{\psi_2} \leq C$ for all unit vectors x . By rotation invariance, all marginals $\langle Z, x \rangle$ have the same distribution, and hence without loss of generality, we may prove our claim for $x = e_1 = (1, 0, \dots, 0) \in \mathbb{R}^n$. In other words, we shall show that

$$\mathbb{P}(|Z_1| \geq t) \leq 2 \exp(-ct^2n) \quad \text{for all } t \geq 0.$$

We use the fact that

$$P(Z_1 \geq t) = \mathbb{P}(Z \in \mathcal{C}_t) \quad \text{with the spherical cap } \mathcal{C}_t = \{z \in S^{(n-1)}: z_1 \geq t\}.$$

Denote by \mathcal{K}_t the "ice-cream" cone when we connect all points in \mathcal{C}_t to the origin, see Figure 1. The fraction of \mathcal{C}_t in the unit sphere (in terms of area) is the same as the fraction of \mathcal{K}_t in the unit ball $B(0, 1) = \{x \in \mathbb{R}^n: x_1^2 + \dots + x_n^2 \leq 1\}$. Thus

$$\mathbb{P}(Z \in \mathcal{C}_t) = \frac{\text{Vol}(\mathcal{K}_t)}{\text{Vol}(B(0, 1))}.$$

The set \mathcal{K}_t is contained in a ball $B(0', \sqrt{1-t^2})$ with radius $\sqrt{1-t^2}$ centred at $0' = (t, 0, \dots, 0)$. Using $1-x \leq e^{-x}$ for $0 \leq x < 1$, we get

$$\mathbb{P}(Z_1 \geq t) = (\sqrt{1-t^2})^n \leq \exp(-t^2 n/2) \quad \text{for } 0 \leq t \leq 1/\sqrt{2},$$

and we can easily extend this bound to all t by loosening the absolute constant (note that for $t \geq 1$ the probability is trivially zero). Indeed, in the range $1/\sqrt{2} \leq t \leq 1$,

$$\mathbb{P}(Z_1 \geq t) \leq \mathbb{P}(Z_1 \geq 1/\sqrt{2}) \leq \exp(-n/4) \leq \exp(-t^2 n/4)$$

We proved that $\mathbb{P}(Z_1 \geq t) \leq \exp(-t^2 n/4)$. By symmetry, the same inequality holds for $-Z_1$. Taking the union bound, we obtain the desired sub-Gaussian tail. \square

Remark 3.22 The so-called *Projective Central Limit Theorem* tells us that marginals of the uniform distribution on the sphere in \mathbb{R}^n become asymptotically normally distributed as the dimension n increases. Namely, if $X \sim \text{Unif}(\sqrt{n}S^{(n-1)})$ then for any fixed unit vector $u \in \mathbb{R}^n$ we have

$$\langle X, u \rangle \longrightarrow N(0, 1) \quad \text{in distribution as } n \rightarrow \infty.$$

\diamond

3.6 Application: Grothendieck's inequality

See Chapter 3.5 in [Ver18].

4 Random Matrices

4.1 Geometrics concepts

Definition 4.1 Let (T, d) be a metric space, $K \subset T$, and $\varepsilon > 0$.

(a) **ε -net:** A subset $\mathcal{N} \subset K$ is an ε -net of K if every point of K is within a distance ε of some point of \mathcal{N} ,

$$\forall x \in K \exists x_0 \in \mathcal{N} : d(x, x_0) \leq \varepsilon.$$

(b) **Covering number:** The smallest possible cardinality of an ε -net of K is the *covering number* of K and is denoted $\mathcal{N}(K, d, \varepsilon)$. Equivalently, $\mathcal{N}(K, d, \varepsilon)$ is the smallest number of closed balls with centres in K and radii ε whose union covers K .

(c) **ε -separated sets:** A subset $\mathcal{N} \subset T$ is ε -separated if $d(x, y) > \varepsilon$ for all distinct points $x, y \in \mathcal{N}$.

(d) **Packing numbers:** The largest possible cardinality of an ε -separated subset of K is called the *packing number* of K and denoted $\mathcal{P}(K, d, \varepsilon)$.

Lemma 4.2 *Let (T, d) be a metric space. Suppose that \mathcal{N} is a maximal ε separated subset of $K \subset T$. Here maximal means that adding any new point $x \in K$ to the set \mathcal{N} destroys the ε -separation property. Then \mathcal{N} is an ε -net.*

Proof. Let $x \in K$. If $x \in \mathcal{N} \subset K$, then choosing $x_0 = x$ we have $d(x_0, x_0) = 0 \leq \varepsilon$. Suppose $x \notin \mathcal{N}$. Then the maximality assumption implies that $\mathcal{N} \cup \{x\}$ is not ε -separated, thus $d(x, x_0) \leq \varepsilon$ for some $x_0 \in \mathcal{N}$. \square

Lemma 4.3 (Equivalence of packing and covering numbers) *Let (T, d) be a metric space. For any $K \subset T$ and any $\varepsilon > 0$, we have*

$$\mathcal{P}(K, d, 2\varepsilon) \leq \mathcal{N}(K, d, \varepsilon) \leq \mathcal{P}(K, d, \varepsilon). \quad (4.1)$$

Proof. Without loss of generality we consider Euclidean space $T = \mathbb{R}^n$ with $d = \|\cdot\|_2$. For the upper bound one can show that $\mathcal{P}(K, \|\cdot\|_2, \varepsilon)$ is the largest number of closed disjoint balls with centres in K and radii $\varepsilon/2$. Furthermore, $\mathcal{P}(K, \|\cdot\|_2, \varepsilon)$ is the largest cardinality of an ε -separated subset, any ε -separated set \mathcal{N} with $\#\mathcal{N} = \mathcal{P}(K, \|\cdot\|_2, \varepsilon)$ is maximal, and hence an ε -net according to Lemma 4.2. Thus

$$\mathcal{N}(K, \|\cdot\|_2, \varepsilon) \leq \#\mathcal{N}.$$

Pick an 2ε -separated subset $\mathcal{P} = \{x_i\}$ in K and an ε -net $\mathcal{N} = \{y_j\}$ of K . Each $x_i \in K$ belongs to some closed ball $\overline{B_\varepsilon(y_j)}$ with radius ε around some y_j . Any such ball $\overline{B_\varepsilon(y_j)}$ may contain at most one point of the x_i 's. Thus $|\mathcal{P}| = \#\mathcal{P} \leq |\mathcal{N}| = \#\mathcal{N}$. \square

In the following we return to the Euclidean space \mathbb{R}^n with its Euclidean norm, $d(x, y) = \|x - y\|_2$.

Definition 4.4 (Minkowski sum) $A, B \subset \mathbb{R}^n$.

$$A + B := \{a + b : a \in A, b \in B\}.$$

Proposition 4.5 (Covering numbers of the Euclidean ball) (a) *Let $K \subset \mathbb{R}^n$ and $\varepsilon > 0$. Denote $|K|$ the volume of K and denote $B^{(n)} = \{x \in \mathbb{R}^n : \|x\|_2 \leq 1\}$ the closed unit Euclidean ball. Then*

$$\frac{|K|}{|\varepsilon B^{(n)}|} \leq \mathcal{N}(K, \|\cdot\|_2, \varepsilon) \leq \mathcal{P}(K, \|\cdot\|_2, \varepsilon) \leq \frac{|(K + \varepsilon/2 B^{(n)})|}{|\varepsilon/2 B^{(n)}|}.$$

(b)

$$\left(\frac{1}{\varepsilon}\right)^n \leq \mathcal{N}(B^{(n)}, \|\cdot\|_2, \varepsilon) \leq \left(\frac{2}{\varepsilon} + 1\right)^n.$$

Proof.

(a) The centre inequality follows from Lemma 4.3.

Lower bound: Let $N := \mathcal{N}(K, \|\cdot\|_2, \varepsilon)$. Then we can cover K by N balls with radii ε . Then $|K| \leq N|\varepsilon B^{(n)}|$. Upper bound: Let $N := \mathcal{P}(K, \|\cdot\|_2, \varepsilon)$ and construct N closed

disjoint balls $B_{\frac{\varepsilon}{2}}(x_i)$ with centres $x_i \in K$ and radii $\varepsilon/2$. These balls might not fit entirely into the set K , but certainly into the extended set $K + \frac{\varepsilon}{2}B^{(n)}$. Thus

$$N|\frac{\varepsilon}{2}B^{(n)}| \leq |K + \frac{\varepsilon}{2}B^{(n)}|.$$

(b) The statement follows easily with part (a) and is left as an exercise. \square

Remark 4.6 To simplify the bound in Proposition 4.5, note that in the nontrivial range $\varepsilon \in (0, 1]$ we have

$$\left(\frac{1}{\varepsilon}\right)^n \leq \mathcal{N}(B^{(n)}, \|\cdot\|_2, \varepsilon) \leq \left(\frac{3}{\varepsilon}\right)^n.$$

\diamond

Example 4.7 (Euclidean balls - volume and surface area) Suppose $R > 0$ and denote $B_R^{(n)}$ the ball of radius R around the origin and $S_R^{(n-1)}$ its surface, i.e.,

$$B_R^{(n)} := \{x \in \mathbb{R}^n : \|x\|_2 \leq R\} \quad \text{and} \quad S_R^{(n-1)} := \{x \in \mathbb{R}^n : \|x\|_2 = R\}.$$

Then the volume and surface area is given as

$$\begin{aligned} \mathbf{vol}(B_R^{(n)}) &= |B_R^{(n)}| = \frac{\pi^{n/2}}{\Gamma(n/2 + 1)} R^n, \\ \mathbf{area}(S_R^{(n-1)}) &= |S_R^{(n-1)}| = \frac{2\pi^{n/2}}{\Gamma(n/2)} R^{n-1}, \end{aligned} \tag{4.2}$$

where Γ is the Gamma function. \clubsuit

Definition 4.8 (Hamming cube) The *Hamming cube* \mathcal{H} is the set of binary strings of length n , i.e.,

$$\mathcal{H} = \{0, 1\}^n.$$

Define the *Hamming distance* $d_{\mathcal{H}}$ between two binary strings as the number of bits where they disagree, i.e.,

$$d_{\mathcal{H}}(x, y) := \#\{i \in \{1, \dots, n\} : x(i) \neq y(i)\}, \quad x, y \in \{0, 1\}^n.$$

Exercise 4.9 Show that $d_{\mathcal{H}}$ is a metric on \mathcal{H} . \clubsuit

4.2 Concentration of the operator norm of random matrices

Definition 4.10 Let A be an $m \times n$ matrix with real entries. The matrix A represents a linear map $\mathbb{R}^n \rightarrow \mathbb{R}^m$.

(a) The *operator norm* or *simply the norm* of A is defined as

$$\|A\| := \max_{x \in \mathbb{R}^n \setminus \{0\}} \left\{ \frac{\|Ax\|_2}{\|x\|_2} \right\} = \max_{x \in \mathbb{S}^{(n-1)}} \{\|Ax\|_2\}.$$

Equivalently,

$$\|A\| = \max_{x \in \mathbb{S}^{(n-1)}, y \in \mathbb{S}^{(m-1)}} \{\langle Ax, y \rangle\}.$$

(b) The *singular values* $s_i = s_i(A)$ of the matrix A are the square roots of the eigenvalues of both AA^T and $A^T A$,

$$s_i(A) = \sqrt{\lambda_i(AA^T)} = \sqrt{\lambda_i(A^T A)},$$

and one orders them $s_1 \geq s_2 \geq \dots \geq s_n \geq 0$. If A is symmetric, then $s_i(A) = |\lambda_i(A)|$.

(c) Suppose $r = \text{rank}(A)$. The singular value decomposition of A is

$$A = \sum_{i=1}^r s_i u_i v_i^T,$$

where $s_i = s_i(A)$ are the singular values of A , the vectors $u_i \in \mathbb{R}^m$ are the left singular vectors, and the vectors $v_i \in \mathbb{R}^n$ are the right singular vectors of A .

Remark 4.11 (a) The extreme singular values $s_1(A)$ and $s_n(A)$ ($s_r(A)$) are respectively the smallest number M and the largest number m such that

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

Thus

$$s_n(A)\|x - y\|_2 \leq \|Ax - Ay\|_2 \leq s_1(A)\|x - y\|_2 \quad \text{for all } x, y \in \mathbb{R}^n.$$

(b) In terms of its spectrum, the operator norm of A equals the largest singular value of A ,

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

◇

Lemma 4.12 (Operator norm on a net) Let $\varepsilon \in [0, 1)$ and A be an $m \times n$ matrix. Then, for any ε -net \mathcal{N} of $\mathbb{S}^{(n-1)}$, we have

$$\sup_{x \in \mathcal{N}} \{\|Ax\|_2\} \leq \|A\| \leq \frac{1}{1 - \varepsilon} \sup_{x \in \mathcal{N}} \{\|Ax\|_2\}.$$

Proof. The lower bound is trivial as $\mathcal{N} \subset S^{(n-1)}$. For the upper bound pick an $x \in S^{(n-1)}$ for which $\|A\| = \|Ax\|_2$ and choose $x_0 \in \mathcal{N}$ for this x such that $\|x - x_0\|_2 \leq \varepsilon$. Then

$$\|Ax - Ax_0\|_2 \leq \|A\| \|x - x_0\|_2 \leq \varepsilon \|A\|.$$

The triangle inequality implies that

$$\|Ax_0\|_2 \geq \|Ax\|_2 - \|Ax - Ax_0\|_2 \geq (1 - \varepsilon) \|A\|,$$

and thus $\|A\| \leq \|Ax_0\|_2 / (1 - \varepsilon)$. \square

Exercise 4.13 Let \mathcal{N} be an ε -net of $S^{(n-1)}$ and \mathcal{M} be an ε -net of $S^{(m-1)}$. Show that for any $m \times n$ matrix A one has

$$\sup_{x \in \mathcal{N}, y \in \mathcal{M}} \{\langle Ax, y \rangle\} \leq \|A\| \leq \frac{1}{1 - 2\varepsilon} \sup_{x \in \mathcal{N}, y \in \mathcal{M}} \{\langle Ax, y \rangle\}.$$



Exercise 4.14 (Isometries) Let A be an $m \times n$ matrix with $m \geq n$. Prove the following equivalences:

$$\begin{aligned} A^T A = \mathbb{1}_n &\Leftrightarrow A \text{ isometry, i.e., } \|Ax\|_2 = \|x\|_2 \text{ for all } x \in \mathbb{R}^n \\ &\Leftrightarrow s_n(A) = s_1(A). \end{aligned}$$



Theorem 4.15 (Norm of sub-Gaussian random matrices) Let A be an $m \times n$ random matrix with independent mean-zero sub-Gaussian random entries $A_{ij}, i = 1, \dots, m, j = 1, \dots, n$. Then, for every $t > 0$, we have that

$$\|A\| \leq CK(\sqrt{m} + \sqrt{n} + t)$$

with probability at least $1 - 2 \exp(-t^2)$, where $K := \max_{\substack{1 \leq i \leq m, \\ 1 \leq j \leq n}} \{\|A_{ij}\|_{\psi_2}\}$ and where $C > 0$ is an absolute constant.

Proof. We use an ε -net argument for our proof.

Step 1: Using Proposition 4.5 (b) we can find for $\varepsilon = \frac{1}{4}$ an ε -net $\mathcal{N} \subset S^{(n-1)}$ and an ε -net $\mathcal{M} \subset S^{(m-1)}$ with

$$|\mathcal{N}| \leq 9^n \text{ and } |\mathcal{M}| \leq 9^m.$$

By Exercise 4.13, the operator norm of A can be bounded using our nets as follows

$$\|A\| \leq 2 \max_{x \in \mathcal{N}, y \in \mathcal{M}} \{\langle Ax, y \rangle\}.$$

Step 2: Concentration Pick $x \in \mathcal{N}$ and $y \in \mathcal{M}$. We compute (using the fact that the single matrix entries are sub-Gaussian random variables with their norm bounded by K)

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

for some absolute constant $C > 0$. We therefore obtain a tail bound for $\langle Ax, y \rangle$, i.e., for every $u \geq 0$,

$$\mathbb{P}(\langle Ax, y \rangle \geq u) \leq 2 \exp(-cu^2/K^2),$$

for some absolute constant $c > 0$.

Step 3: Union bound We unfix the choice of x and y in Step 2 by a union bound.

$$\mathbb{P}\left(\max_{x \in \mathcal{N}, y \in \mathcal{M}} \{\langle Ax, y \rangle\} \geq u\right) \leq \sum_{x \in \mathcal{N}, y \in \mathcal{M}} \mathbb{P}(\langle Ax, y \rangle \geq u) \leq 9^{n+m} 2 \exp(-cu^2/K^2).$$

We continue the estimate by choosing $u = CK(\sqrt{n} + \sqrt{m} + t)$ which leads to a lower bound $u^2 \geq C^2 K^2(n + m + t^2)$, and furthermore, adjust the constant $C > 0$ such that $cu^2/K^2 \geq 3(n + m) + t^2$. Inserting these choices we get

$$\mathbb{P}\left(\max_{x \in \mathcal{N}, y \in \mathcal{M}} \{\langle Ax, y \rangle\} \geq u\right) \leq 9^{n+m} 2 \exp(-3(n + m) - t^2) \leq 2 \exp(-t^2),$$

and thus

$$\mathbb{P}(\|A\| \geq 2u) \leq 2 \exp(-t^2).$$

□

Corollary 4.16 *Let A be an $n \times n$ random matrix whose entries on and above the diagonal are independent mean-zero sub-Gaussian random variables. Then, for every $t > 0$, we have*

$$\|A\| \leq CK(\sqrt{n} + t)$$

with probability at least $1 - 4 \exp(-t^2)$ and $K := \max_{1 \leq i, j \leq n} \{\|A_{ij}\|_{\psi_2}\}$.

Proof. Exercise. Decompose the matrix into an upper-triangular part and a lower-triangular part and use the proof of the previous theorem. □

We can actually improve the statement in Theorem 4.15 in two ways. First we obtain a two-sided bound, and, secondly, we can relax the independence assumption. As this is a refinement of our previous statement we only state the result and omit its proof. The statement is used in covariance estimation below.

Theorem 4.17 *Let A be an $m \times n$ matrix whose rows $A_i, i = 1, \dots, m$, are independent mean-zero sub-Gaussian isotropic random vectors in \mathbb{R}^n .*

(a) *Then, for every $t > 0$, we have*

$$\sqrt{m} - CK^2(\sqrt{n} + t) \leq s_n(A) \leq s_1(A) \leq \sqrt{m} + CK^2(\sqrt{n} + t), \quad (4.3)$$

with probability at least $1 - 2 \exp(-t^2)$ and $K := \max_{1 \leq i \leq m} \{\|A_i\|_{\psi_2}\}$. Furthermore, with probability at least $1 - 2 \exp(-t^2)$,

$$\left\| \frac{1}{m} A^T A - \mathbb{1}_n \right\| \leq K^2 \max\{\delta, \delta^2\} \quad \text{where } \delta = C \left(\sqrt{\frac{n}{m}} + \frac{t}{\sqrt{m}} \right). \quad (4.4)$$

(b) Property (4.4) implies that

$$\mathbb{E} \left[\left\| \frac{1}{m} A^T A - \mathbb{1}_n \right\| \right] \leq CK^2 \left(\sqrt{\frac{n}{m}} + \frac{n}{m} \right).$$

Proof. (a) The proof follows similarly to the proof of Theorem 4.15 and is thus left as an exercise. To show that (4.4) indeed implies (4.3) we use Lemma 4.18 below.

(b) To obtain the bound for the expected operator norm of the difference of $\frac{1}{m} A^T A$ to the identity, we use the integral identity (1.9) for the real valued random variable $\left\| \frac{1}{m} A^T A - \mathbb{1}_n \right\|$. This calculation is quite long and requires some computational work for the different cases and is omitted. □

Lemma 4.18 *Let A be an $(m \times n)$ -matrix and $\delta > 0$. Suppose that*

$$\|A^T A - \mathbb{1}_n\| \leq \max\{\delta, \delta^2\}.$$

Then

$$(1 - \delta)\|x\|_2 \leq \|Ax\|_2 \leq (1 + \delta)\|x\|_2 \quad \text{for all } x \in \mathbb{R}^n. \quad (4.5)$$

In particular, this means that all singular values of A lie between $1 - \delta$ and $1 + \delta$,

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

Proof. We first assume without loss of generality that $\|x\|_2 = 1$. Then our assumptions give

$$\max\{\delta, \delta^2\} \geq |\langle (A^T A - \mathbb{1}_n)x, x \rangle| = \left| \|Ax\|_2^2 - 1 \right|.$$

For every $z \geq 0$ the following elementary inequality holds:

$$\max\{|z - 1|, |z - 1|^2\} \leq |z^2 - 1|. \quad (4.6)$$

To show (4.6) use that for $z \geq 1$ we have $|z - 1|^2 = |z^2 - 2z + 1| \leq |z^2 - 2 + 1| = |z^2 - 1|$, and for $z \in [0, 1)$, use $|z - 1| \leq |z^2 - 1|$ as $z^2 \leq z$ for $z \in [0, 1)$. Then use (4.6) with $\|Ax\|_2$ to conclude that

$$\left| \|Ax\|_2 - 1 \right| \leq \delta.$$

This implies both statements of the lemma. □

4.3 Application: Community Detection in Networks

See Chapter 4.5 in [Ver18].

4.4 Application: Covariance Estimation and Clustering

Suppose that $X^{(1)}, \dots, X^{(N)}$ are empirical samples (random outcomes) of a random vector $X \in \mathbb{R}^n$. We do not have access to the full distribution, only to the empirical measure

$$L_N = \frac{1}{N} \sum_{i=1}^N \delta_{X^{(i)}},$$

which is a random (depending on the N random outcomes $X^{(i)}$) probability measure on \mathbb{R}^n . Here, the symbol δ_X is the Kronecker-delta measure or point measure defined as

$$\delta_X(y) = \begin{cases} 1 & \text{if } X = y, \\ 0 & \text{if } y \neq X, \end{cases} \quad y \in \mathbb{R}^n.$$

We assume for simplicity that $\mathbb{E}[X] = 0$ and recall $\Sigma = \Sigma(X) = \mathbb{E}[XX^T]$.

Definition 4.19 Let $X \in \mathbb{R}^n$ with N random outcomes/random samples $X^{(1)}, \dots, X^{(N)}$.

(a) The *empirical measure* of $X^{(1)}, \dots, X^{(N)}$ is the probability measure on \mathbb{R}^n ,

$$L_N := \frac{1}{N} \sum_{i=1}^N \delta_{X^{(i)}}.$$

(b) The *empirical covariance* of $X^{(1)}, \dots, X^{(N)}$ is the random matrix

$$\Sigma_N := \frac{1}{N} \sum_{i=1}^N (X^{(i)})(X^{(i)})^T.$$

Note that $X_i \sim X$ implies that $\mathbb{E}[\Sigma_N] = \Sigma$. The law of large numbers yields

$$\Sigma_N \rightarrow \Sigma \quad \text{almost surely as } N \rightarrow \infty.$$

Theorem 4.20 (Covariance estimation) Let X be a sub-Gaussian random vector in \mathbb{R}^n , and assume that there exist $K \geq 1$ such that

$$\|\langle X, x \rangle\|_{\psi_2} \leq K \|\langle X, x \rangle\|_{L^2}, \quad \text{for all } x \in \mathbb{R}^n.$$

Then, for every $N \in \mathbb{N}$,

$$\mathbb{E}[\|\Sigma_N - \Sigma\|] \leq CK^2 \left(\sqrt{\frac{n}{N}} + \frac{n}{N} \right) \|\Sigma\|.$$

Proof. First note that

$$\|\langle X, x \rangle\|_{L^2}^2 = \mathbb{E}[|\langle X, x \rangle|^2] = \mathbb{E}[\langle X, x \rangle^2] = \langle \Sigma x, x \rangle.$$

We bring $X, X^{(1)}, \dots, X^{(N)}$ all into isotropic position. That is, there exist independent isotropic random vectors $Z, Z^{(1)}, \dots, Z^{(N)}$ such that

$$X = \Sigma^{1/2}Z \text{ and } X^{(i)} = \Sigma^{1/2}Z^{(i)}.$$

We have from our assumptions that

$$\|Z\|_{\psi_2} \leq K \text{ and } \|Z^{(i)}\|_{\psi_2} \leq K.$$

Then

$$\|\Sigma_N - \Sigma\| = \|\Sigma^{1/2}R_N\Sigma^{1/2}\| \leq \|R_N\|\|\Sigma\|,$$

where

$$R_N = \frac{1}{N} \sum_{i=1}^N (Z^{(i)})(Z^{(i)})^T - \mathbb{1}_n.$$

Suppose now that A is the $N \times n$ matrix whose rows are $(Z^{(i)})^T$, that is,

$$\frac{1}{N}A^T A - \mathbb{1}_n = R_N,$$

and Theorem 4.17 for A implies that

$$\mathbb{E}[\|R_N\|] \leq CK^2 \left(\sqrt{\frac{n}{N}} + \frac{n}{N} \right),$$

and we conclude with our statement. □

Remark 4.21 For all $\varepsilon \in (0, 1)$ we have

$$\mathbb{E}[\|\Sigma_N - \Sigma\|] \leq \varepsilon\|\Sigma\|,$$

if we take a sample of size $N \sim \varepsilon^{-2}n$. ◇

5 Concentration of measure - general case

We study now general concentration of measure phenomena and aim in particular to include cases where the random variables are not necessarily independent. The independence assumption made our concentration results relatively easy to develop. In the first section we summarise concentration results by entropy techniques before studying dependent random variables in the remaining sections.

5.1 Concentration by entropic techniques

In the following assume that $\varphi: \mathbb{R} \rightarrow \mathbb{R}$ is a convex function, X an \mathbb{R} -valued random variable such that $\mathbb{E}[X]$ and $\mathbb{E}[\varphi(X)]$ are finite unless otherwise stated. The random variable is a measurable map from some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ to \mathbb{R} and distributed with law $P_X = \mathbb{P} \circ X^{-1} \in \mathcal{M}_1(\mathbb{R})$.

Definition 5.1 (Entropy) The entropy of the random variable X for the convex function φ is

$$H_\varphi(X) = \mathbb{E}[\varphi(X)] - \varphi(\mathbb{E}[X]).$$

Corollary 5.2 By Jensen's inequality, $\varphi(\mathbb{E}[X]) \leq \mathbb{E}[\varphi(X)]$, we see that $H_\varphi(X) \geq 0$.

Example 5.3 (a) $\varphi(u) = u^2, u \in \mathbb{R}$, then the entropy of X ,

$$H_\varphi(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \text{Var}(X)$$

is the variance of X .

(b) $\varphi(u) = -\log u, u > 0$, and for X real-valued random variable we have that $Z := e^{\lambda X} > 0$ is a strictly positive real-valued random variable.

$$H_\varphi(Z) = -\lambda \mathbb{E}[X] + \log \mathbb{E}[e^{\lambda X}] = \log \mathbb{E}[e^{\lambda(X - \mathbb{E}[X])}].$$

(c) $\varphi(u) = u \log u, u > 0$, and $\varphi(0) := 0$. The function φ is convex function on \mathbb{R}_+ and continuous when we set $0 \log 0 = 0$. For any non-negative random variable $Z \geq 0$, the φ -entropy is

$$H_\varphi(Z) = \mathbb{E}[Z \log Z] - \mathbb{E}[Z] \log \mathbb{E}[Z].$$

♣

In the following we will drop the index φ for the entropy whenever we take $\varphi(u) = u \log u$ as in Example 5.3 (c). There are several reasons why this choice is particular useful. In the next remark we show some connection of that entropy to other entropy concepts in probability theory.

Remark 5.4 Suppose that Ω is a finite sample space, and let $p, q \in \mathcal{M}_1(\Omega)$ be two probability measures (vectors) such that $q(\omega) = 0$ implies $p(\omega) = 0$. Define

$$X: \Omega \rightarrow \mathbb{R}, \omega \mapsto X(\omega) = \begin{cases} \frac{p(\omega)}{q(\omega)} & \text{if } q(\omega) > 0, \\ 0 & \text{if } q(\omega) = p(\omega) = 0, \end{cases}$$

with distribution $q \in \mathcal{M}_1(\Omega)$. Then $X \geq 0$ with

$$\mathbb{E}[X] = \sum_{\omega \in \Omega} X(\omega)q(\omega) = \sum_{\omega \in \Omega} p(\omega) = 1.$$

$$\begin{aligned} H(X) &= \sum_{\omega \in \Omega} X(\omega)q(\omega) \log X(\omega) - \mathbb{E}[X] \log \mathbb{E}[X] = \sum_{\omega \in \Omega} p(\omega) \log \frac{p(\omega)}{q(\omega)} \\ &=: H(p|q) =: D(p||q), \end{aligned}$$

where $H(p|q)$ is the *relative entropy of p with respect to q* , a widely used function probability theory (e.g. large deviation theory) and information theory, and where $D(p||q)$ is the *Kullback-Leibler divergence* of p and q used in information theory.

◇

Definition 5.5 Let Ω be a finite (respectively discrete) sample space and denote $\mathcal{M}_1(\Omega)$ the set of probability measures (vectors).

(a) The *relative entropy* with respect to $q \in \mathcal{M}_1(\Omega)$ is defined as the mapping

$$H(\cdot|q): \mathcal{M}_1(\Omega) \rightarrow [0, \infty]; p \mapsto H(p|q) = \begin{cases} \sum_{\omega \in \Omega} p(\omega) \log \frac{p(\omega)}{q(\omega)} & \text{if } q(\omega) \Rightarrow p(\omega) = 0, \\ +\infty & \text{otherwise.} \end{cases}$$

(b) The *Shannon entropy* of a Ω -valued random variable X with probability density (distribution) $p \in \mathcal{M}_1(\Omega)$ is defined as

$$\mathcal{H}(X) \equiv \mathcal{H}(p) = - \sum_{\omega \in \Omega} p(\omega) \log p(\omega).$$

Lemma 5.6 Let Ω be a finite sample space and denote $\mathcal{M}_1(\Omega)$ the set of probability measures (vectors). Let $q \in \mathcal{M}_1(\Omega)$. Then the relative $H(\cdot|q)$ is strictly convex, continuous and

$$H(p|q) = 0 \Leftrightarrow p = q.$$

Proof. Exercise. □

Exercise 5.7 Let Ω be a finite sample space and denote $\mathcal{M}_1(\Omega)$ the set of probability measures (vectors). Let X be a Ω -valued random variable with distribution $p \in \mathcal{M}_1(\Omega)$. Its Shannon entropy is then $\mathcal{H}(X) = - \sum_{\omega \in \Omega} p(\omega) \log p(\omega)$. We shall explore the connection between the entropy functional H_φ with $\varphi(u) = u \log u$ and the Shannon entropy:

(a) Consider the random variable $Z := p(U)$, where $U \sim \text{Unif}(\Omega)$ is uniformly distributed over Ω . Show that

$$H_\varphi(Z) = \frac{1}{|\Omega|} (\log |\Omega| - \mathcal{H}(X)).$$

(b) Use part (a) and Lemma 5.6 to show that Shannon entropy for a discrete random variable is maximised by a uniform distribution. ☹

The following example motivates the definition of Shannon entropy and can be skipped for the following.

Example 5.8 (Shannon entropy) Let Ω be a finite sample space. Let $X^{(i)}, i = 1, \dots, N$, be independent identically distributed Ω -valued random variables and write $X = (X^{(1)}, \dots, X^{(N)})$. The empirical measure of the sample vector $X \in \Omega^N$ is then

$$L_N^X = \frac{1}{N} \sum_{i=1}^N \delta_{X^{(i)}}.$$

For a given $\mu \in \mathcal{M}_1(\Omega)$ with $N\mu(x) \in \mathbb{N}_0, x \in \Omega$, we define the number

$$\mathcal{N}_N(\mu) = \#\{\omega \in \Omega^N : L_N^\omega = \mu\} = \frac{N!}{\prod_{x \in \Omega} (N\mu(x))!},$$

where the second equality follows from multinomial distribution. For any $\mu \in \mathcal{M}_1(\Omega)$ pick $\mu_N \in \mathcal{M}_1(\Omega)$ with $N\mu_N(x) \in \mathbb{N}_0, x \in \Omega$, such that $\mu_N \rightarrow \mu$ as $N \rightarrow \infty$. Then, using Stirling's formula, one can show that

$$\mathcal{H}(\mu) = \lim_{N \rightarrow \infty} \frac{1}{N} \log \mathcal{N}_N(\mu_N). \quad (5.1)$$

♣

For any real-valued random variable X , $Z = \exp(\lambda X), \lambda \in \mathbb{R}$, is a positive random variable and the entropy can be written with the moment generating function of X , recall the definition of the moment generating function (MGF) in (1.4) (we assume again that the expectations are finite for all $\lambda \in \mathbb{R}$),

$$\mathcal{H}(e^{\lambda X}) = \lambda M'_X(\lambda) - M_X(\lambda) \log M_X(\lambda). \quad (5.2)$$

Example 5.9 (Entropy - Gaussian random variable) Suppose $X \sim \mathcal{N}(0, \sigma^2), \sigma > 0$. Then $M_X(\lambda) = \exp(\lambda^2 \sigma^2 / 2)$, and $M'_X(\lambda) = \lambda \sigma^2 M_X(\lambda)$, and therefore

$$\mathcal{H}(e^{\lambda X}) = \lambda^2 \sigma^2 M_X(\lambda) - \lambda^2 \sigma^2 / 2 M_X(\lambda) = \frac{1}{2} \lambda^2 \sigma^2 M_X(\lambda).$$

♣

A bound on the entropy leads to a bound of the centred moment generating function Φ , see (2.3), this is the content of the so-called *Herbst argument*.

Proposition 5.10 (Herbst argument) *Let X be a real-valued random variable and suppose that, for $\sigma > 0$,*

$$\mathcal{H}(e^{\lambda X}) \leq \frac{1}{2} \sigma^2 \lambda^2 M_X(\lambda)$$

for $\lambda \in I$ with interval I being either $[0, \infty)$ or \mathbb{R} . Then

$$\log \mathbb{E}[\exp(\lambda(X - \mathbb{E}[X]))] \leq \frac{1}{2} \lambda^2 \sigma^2 \quad \text{for all } \lambda \in I. \quad (5.3)$$

Remark 5.11 (a) If $I = \mathbb{R}$, then the bound (5.3) is equivalent to asserting that the centred random variable $X - \mathbb{E}[X]$ is sub-Gaussian with parameter $\sigma > 0$.

(b) For $I = [0, \infty)$, the bound (5.3) leads immediately to the one-sided tail estimate

$$\mathbb{P}(X \geq \mathbb{E}[X] + t) \leq \exp(-t^2 / 2\sigma^2), \quad t \geq 0,$$

and $I = \mathbb{R}$ provides the corresponding two-sided estimate.

◇

Proof of Proposition 5.10. Suppose that $I = [0, \infty)$. Using (5.2), our assumption turns into a differential inequality for the moment generating function,

$$\lambda M'_X(\lambda) - M_X(\lambda) \log M_X(\lambda) \leq \frac{1}{2} \sigma^2 \lambda^2 M_X(\lambda) \quad \text{for all } \lambda \geq 0. \quad (5.4)$$

Define now a function $G(\lambda) := \frac{1}{\lambda} \log M_X(\lambda)$ for $\lambda \neq 0$, and then extend the function to 0 by continuity (L'Hospital rule)

$$G(0) := \lim_{\lambda \rightarrow 0} G(\lambda) = \left. \frac{d}{d\lambda} \log M_X(\lambda) \right|_{\lambda=0} = \mathbb{E}[X].$$

Our assumptions on the existence of the MGF imply differentiability with respect to the parameter λ . Hence

$$G'(\lambda) = \frac{1}{\lambda} \frac{M'_X(\lambda)}{M_X(\lambda)} - \frac{1}{\lambda^2} \log M_X(\lambda),$$

and thus we can rewrite our differential inequality (5.4) as

$$G'(\lambda) \leq \frac{1}{2} \sigma^2 \quad \text{for all } \lambda \in I = [0, \infty).$$

For any $\lambda_0 > 0$ we can integrate both sides of the previous inequality to arrive at

$$G(\lambda) - G(\lambda_0) \leq \frac{1}{2} \sigma^2 (\lambda - \lambda_0).$$

Now, letting $\lambda_0 \downarrow 0$ and using the above definition of $G(0)$, we get

$$G(\lambda) - \mathbb{E}[X] = \frac{1}{\lambda} \left(\log M_X(\lambda) - \log e^{\lambda \mathbb{E}[X]} \right) \leq \frac{1}{2} \sigma^2 \lambda,$$

and we conclude with the statement in (5.3). □

Proposition 5.12 (Bernstein entropy bound) *Suppose there exist $B > 0$ and $\sigma > 0$ such that the real-valued random variable X satisfies the following entropy bound*

$$H(e^{\lambda X}) \leq \lambda^2 (B M'_X(\lambda) + M_X(\lambda) (\sigma^2 - B \mathbb{E}[X])) \quad \text{for all } \lambda \in [0, B^{-1}).$$

Then

$$\log \mathbb{E}[e^{\lambda(X - \mathbb{E}[X])}] \leq \sigma^2 \lambda^2 (1 - B\lambda)^{-1} \quad \text{for all } \lambda \in [0, B^{-1}). \quad (5.5)$$

Remark 5.13 As a consequence of the Chernoff argument, Proposition 5.12 implies that X satisfies the upper tail bound

$$\mathbb{P}(X \geq \mathbb{E}[X] + \delta) \leq \exp\left(-\frac{\delta}{4\sigma^2 + 2B\delta}\right), \quad \text{for all } \delta \geq 0. \quad (5.6)$$

◇

Exercise 5.14 Prove the tail bound (5.6). ☹

Proof of Proposition 5.12. We skip the proof of this statement as it employs similar techniques as in the proof of Proposition 5.10. \square

So far, the entropic method has not provided substantial new insight as all concentration results are done via the usual Chernoff bound. We shall now study concentration for functions of many random variables.

Definition 5.15 (a) A function $f: \mathbb{R}^n \rightarrow \mathbb{R}$ is *separately convex* if, for every index $k \in \{1, \dots, n\}$, the univariate function

$$f_k: \mathbb{R} \rightarrow \mathbb{R}, \\ y_k \mapsto f_k(y_k) := f(x_1, \dots, x_{k-1}, y_k, x_{k+1}, \dots, x_n),$$

is convex for each vector $(x_1, \dots, x_{k-1}, x_{k+1}, \dots, x_n) \in \mathbb{R}^{n-1}$.

(b) A function $f: X \rightarrow Y$ for metric spaces (X, d_X) and (Y, d_Y) is *Lipschitz continuous* (sometimes called *locally Lipschitz continuous*) if for every $x \in X$ there exists a neighbourhood $U \subset X$ such that $f|_U$ is globally Lipschitz continuous. Here $f|_U: U \rightarrow Y$ is the restriction of f to U .

(c) A function $f: X \rightarrow Y$ for metric spaces (X, d_X) and (Y, d_Y) is *L-Lipschitz continuous* (sometimes called *globally Lipschitz continuous*) if there exists $L \in \mathbb{R}$ such that

$$d_Y(f(x), f(y)) \leq L d_X(x, y) \quad \text{for all } x, y \in X. \quad (5.7)$$

The smallest constant $L > 0$ satisfying (5.7) is denoted $\|f\|_{\text{Lip}}$. In the following some statements hold for global Lipschitz continuity and some only for local Lipschitz continuity.

Theorem 5.16 (Tail-bound for Lipschitz functions) Let $X = (X_1, \dots, X_n) \in \mathbb{R}^n$ be a random vector with independent random coordinates X_i supported on the interval $[a, b]$, $a < b$, and let $f: \mathbb{R}^n \rightarrow \mathbb{R}$ be separately convex and L -Lipschitz continuous with respect to the Euclidean norm $\|\cdot\|_2$. Then, for every $t \geq 0$,

$$\mathbb{P}(f(X) \geq \mathbb{E}[f(X)] + t) \leq \exp\left(-\frac{t^2}{4L^2(b-a)^2}\right). \quad (5.8)$$

Example 5.17 (Operator norm of a random matrix) Let $M \in \mathbb{R}^{n \times d}$ be a $n \times d$ matrix with independent identically distributed mean-zero random entries $M_{ij}, i \in \{1, \dots, n\}, j \in \{1, \dots, d\}$ supported on the interval $[-1, 1]$.

$$\|M\| = \max_{\nu \in \mathbb{R}^d: \|\nu\|_2=1} \{\|M\nu\|_2\} = \max_{\substack{\nu \in \mathbb{R}^d: \\ \|\nu\|_2=1}} \max_{\substack{u \in \mathbb{R}^n: \\ \|u\|_2=1}} \{\langle u, M\nu \rangle\}.$$

$M \mapsto \|M\|$ is a function $f: \mathbb{R}^{n \times d} \rightarrow \mathbb{R}, f(M) = \|M\|$. To apply Theorem 5.16 above we shall show that f is Lipschitz and separately convex. The operator norm is the max-min/supremum of functions that are linear in the entries of the matrix M , and thus any such

function is convex and as such separately convex. Moreover, for $M, M' \in \mathbb{R}^{n \times d}$,

$$\left| \|M\| - \|M'\| \right| \leq \|M - M'\| \leq \|M - M'\|_F,$$

where $\|M\|_F := \sqrt{\sum_{i=1}^n \sum_{j=1}^d M_{ij}^2}$ is the Euclidean norm of all entries of the matrix, called the Frobenius norm of the matrix M . The first inequality shows that $f := \|\cdot\|$ is 1-Lipschitz continuous. Thus Theorem 5.16 implies that, for every $t \geq 0$,

$$\mathbb{P}(\|M\| \geq \mathbb{E}[\|M\|] + t) \leq e^{-t^2/16}.$$

♣

Key ingredients for the proof of Theorem 5.16 are the following two lemmas.

Lemma 5.18 (Entropy bound for univariate functions) *Let X and Y two independent, identically distributed \mathbb{R} -valued random variables. Denote by $\mathbb{E}_{X,Y}$ the expectation with respect to X and Y . For any function $g: \mathbb{R} \rightarrow \mathbb{R}$ the following statements hold:*

(a)

$$\mathbb{H}(e^{\lambda g(X)}) \leq \lambda^2 \mathbb{E}_{X,Y} \left[\left(g(X) - g(Y) \right)^2 e^{\lambda g(X)} \mathbb{1} \left\{ g(X) \geq g(Y) \right\} \right], \quad \text{for all } \lambda > 0.$$

(b) *If in addition the random variable X is supported on $[a, b]$, $a < b$, and the function g is convex and Lipschitz continuous, then*

$$\mathbb{H}(e^{\lambda g(X)}) \leq \lambda^2 (b - a)^2 \mathbb{E} \left[(g'(X))^2 e^{\lambda g(X)} \right], \quad \text{for all } \lambda > 0.$$

Lemma 5.19 (Tensorisation of the entropy) *Let X_1, \dots, X_n be independent real-valued random variables and $f: \mathbb{R}^n \rightarrow \mathbb{R}$ a given function. Then*

$$\mathbb{H}(e^{\lambda f(X_1, \dots, X_n)}) \leq \mathbb{E} \left[\sum_{k=1}^n \mathbb{H} \left(e^{\lambda f_k(X_k)} \middle| \bar{X}^k \right) \right], \quad \text{for all } \lambda > 0, \quad (5.9)$$

where f_k is the function introduced in Definition 5.15 and

$$\bar{X}^k = (X_1, \dots, X_{k-1}, X_{k+1}, \dots, X_n).$$

The entropy on the right hand side is computed with respect to X_k for $k = 1, \dots, n$, by holding the remaining \bar{X}^k fixed. That is,

$$\mathbb{H} \left(e^{\lambda f_k(X_k)} \middle| \bar{X}^k \right) = \mathbb{E}_{X_k} \left[e^{\lambda f_k(X_k)} \lambda f_k(X_k) \right] - \mathbb{E}_{X_k} \left[\exp(\lambda f_k(X_k)) \right] \log \mathbb{E}_{X_k} \left[\exp(\lambda f_k(X_k)) \right]$$

is still a function of \bar{X}^k and is integrated with respect to \mathbb{E} on the right hand side of (5.9).

We first finish the proof of theorem 5.16.

Proof of Theorem 5.16. For $k \in \{1, \dots, n\}$ and every vector (fixed) $\bar{X}^k \in \mathbb{R}^{n-1}$ the function f_k is convex, and hence Lemma 5.18 implies for all $\lambda > 0$ that for every fixed vector \bar{X}^k we have

$$\begin{aligned} \mathbb{H}\left(e^{\lambda f_k(X_k)} \middle| \bar{X}^k\right) &\leq \lambda^2(b-a)^2 \mathbb{E}_{X_k} \left[(f'_k(X_k))^2 e^{\lambda f_k(X_k)} \middle| \bar{X}^k \right] \\ \lambda^2(b-a)^2 \mathbb{E}_{X_k} \left[\left(\frac{\partial f(X_1, \dots, X_k, \dots, X_n)}{\partial x_k} \right)^2 \exp(\lambda f(X_1, \dots, X_k, \dots, X_n)) \middle| \bar{X}^k \right]. \end{aligned}$$

With Lemma 5.19 one obtains, writing $X = (X_1, \dots, X_n)$,

$$\begin{aligned} \mathbb{H}\left(e^{\lambda f(X)}\right) &\leq \lambda^2(b-a)^2 \mathbb{E} \left[\sum_{k=1}^n \mathbb{E}_{X_k} \left[\left(\frac{\partial f(X)}{\partial x_k} \right)^2 \exp(\lambda f(X)) \middle| \bar{X}^k \right] \right] \\ &= \mathbb{E} \left[\sum_{k=1}^n \left(\frac{\partial f(X)}{\partial x_k} \right)^2 \exp(\lambda f(X)) \right] \\ &\leq \lambda^2(b-a)^2 L^2 \mathbb{E}[e^{\lambda f(X)}], \end{aligned}$$

where the equality follows from the fact that the single coordinates $X_i, i = 1, \dots, n$, are independent and thus $\mathbb{E} = \mathbb{E}_{X_1} \otimes \dots \otimes \mathbb{E}_{X_n}$ and where we used the Lipschitz continuity of f leading to

$$\|\nabla f(X)\|_2^2 = \sum_{k=1}^n \left(\frac{\partial f(X)}{\partial x_k} \right)^2 \leq L^2 \quad \text{almost surely.}$$

The tail bound then follows from the Herbst argument in Proposition 5.10. \square

Proof of Lemma 5.18. Using the fact that X and Y are independent and identical distributed we have

$$\mathbb{H}(e^{\lambda g(X)}) = \mathbb{E}_X[\lambda g(X) e^{\lambda g(X)}] - \mathbb{E}_X[e^{\lambda g(X)}] \log \mathbb{E}_Y[e^{\lambda g(Y)}].$$

By Jensen's inequality,

$$\log \mathbb{E}_Y[e^{\lambda g(Y)}] \geq \mathbb{E}_Y[\lambda g(Y)],$$

and thus, using the symmetry between X and Y , we obtain (we write $\mathbb{E}_{X,Y}$ for the expectation with respect to both, X and Y , when we want to distinguish expectations with respect to the single random variables. Note that we can easily replace $\mathbb{E}_{X,Y}$ by \mathbb{E}),

$$\begin{aligned} \mathbb{H}(e^{\lambda g(X)}) &\leq \mathbb{E}_X \left[\lambda g(X) e^{\lambda g(X)} \right] - \mathbb{E}_{X,Y} \left[e^{\lambda g(X)} \lambda g(Y) \right] \\ &= \frac{1}{2} \mathbb{E}_{X,Y} \left[\lambda (g(X) - g(Y)) (e^{\lambda g(X)} - e^{\lambda g(Y)}) \right] \\ &\stackrel{\text{Symmetry}}{=} \lambda \mathbb{E} \left[(g(X) - g(Y)) (e^{\lambda g(X)} - e^{\lambda g(Y)}) \mathbf{1}_{\{g(X) \geq g(Y)\}} \right] \end{aligned} \tag{5.10}$$

For all $s, t \in \mathbb{R}$ we have $e^s - e^t \leq e^s(s - t)$. To see that, assume without loss of generality that $s \geq t$ and recall that $e^x \geq 1 + x$, to see that

$$e^s(1 - e^{t-s}) \leq e^s(1 - (1 + (t - s))) = e^s(s - t).$$

For $s \geq t$, we obtain therefore

$$(s - t)(e^s - e^t)\mathbb{1}\{s \geq t\} \leq (s - t)^2 e^s \mathbb{1}\{s \geq t\}.$$

Applying this bound with $s = \lambda g(X)$ and $t = \lambda g(Y)$ to the inequality (5.10) yields

$$H(e^{\lambda g(X)}) \leq \lambda^2 \mathbb{E}[(g(X) - g(Y))^2 e^{\lambda g(X)} \mathbb{1}\{g(X) \geq g(Y)\}].$$

If in addition the function g is convex, then we have the upper bound

$$g(x) - g(y) \leq g'(x)(x - y),$$

and hence, for $g(x) \geq g(y)$,

$$(g(x) - g(y))^2 \leq (g'(x))^2 (x - y)^2$$

which finishes the proof of the statement in Lemma 5.18. \square

Proof of Lemma 5.19. The key ingredient is the variational representation of the entropy (5.14), see Proposition 5.20 and Remark 5.21 below:

$$H(e^{\lambda f(X)}) = \sup_{g \in \mathcal{G}} \{\mathbb{E}[g(X)e^{\lambda f(X)}]\}, \quad (5.11)$$

where $\mathcal{G} = \{g: \Omega \rightarrow \mathbb{R}: e^g \leq 1\}$. The proof now amounts to some computation once the following notations are introduced.

For each $j \in \{1, \dots, n\}$ define $\bar{X}_j = (X_j, \dots, X_n)$, and for any $g \in \mathcal{G}$ define the functions $g^i, i = 1, \dots, n$, as follows using $X = (X_1, \dots, X_n)$ and $\mathbb{E}[\cdot | \bar{X}_j]$ denote expectation with respect to X conditions on fixing \bar{X}_j :

$$\begin{aligned} g^1(X_1, \dots, X_n) &:= g(X) - \log \mathbb{E}[e^{g(X)} | \bar{X}_2], \\ g^k(X_k, \dots, X_n) &:= \log \frac{\mathbb{E}[e^{g(X)} | \bar{X}_k]}{\mathbb{E}[e^{g(X)} | \bar{X}_{k+1}]}, \quad \text{for } k = 2, \dots, n. \end{aligned}$$

It is easy to see that by construction we have,

$$\sum_{k=1}^n g^k(X_k, \dots, X_n) = g(X) - \log \mathbb{E}[e^{g(X)}] \geq g(X), \quad (5.12)$$

and

$$\mathbb{E}[\exp(g^k(X_k, \dots, X_n) | X_{k+1})] = 1.$$

We use this decomposition within the variational representation (5.11) leading to the following chain of upper bounds:

$$\begin{aligned}
\mathbb{E}[g(X)e^{\lambda f(X)}] &\stackrel{(5.12)}{\leq} \sum_{k=1}^n \mathbb{E}[g^k(X_k, \dots, X_n)e^{\lambda f(X)}] \\
&= \sum_{k=1}^n \mathbb{E}_{\bar{X}^k} [\mathbb{E}_{X_k} [g^k(X_k, \dots, X_n)e^{\lambda f(X)} | \bar{X}^k]] \\
&\stackrel{(5.11)}{\leq} \sum_{k=1}^n \mathbb{E}_{\bar{X}^k} [\mathbb{H}(e^{\lambda f_k(X_k)} | \bar{X}^k)].
\end{aligned}$$

We conclude with the statement by optimising over the function $g \in \mathcal{G}$. □

Proposition 5.20 (Duality formula of the Entropy) *Let Y be a non-negative \mathbb{R} -valued random variable defined on a probability space (Ω, \mathcal{F}, P) such that $\mathbb{E}[\varphi(Y)] < \infty$, where $\varphi(u) = u \log u$ for $u \geq 0$. Then*

$$\mathbb{H}(Y) = \sup_{g \in \mathcal{U}} \{\mathbb{E}[gY]\}, \quad (5.13)$$

where $\mathcal{U} = \{g: \Omega \rightarrow \mathbb{R} \text{ measurable with } \mathbb{E}[e^g] = 1\}$.

Proof. Denote $Q = e^g P$ the probability measure with Radon-Nikodym density $\frac{dQ}{dP} = e^g$ with respect to P for some $g \in \mathcal{U}$. Denote \mathbb{E} the expectation with respect to P and \mathbb{E}_Q the expectation with respect to Q , and we write \mathbb{H}_Q when we compute the entropy with respect to the probability measure Q . Then

$$\begin{aligned}
\mathbb{H}_Q(Ye^{-g}) &= \mathbb{E}_Q[Ye^{-g}(\log Y - g)] - \mathbb{E}_Q[Ye^{-g}] \log \mathbb{E}_Q[Ye^{-g}] \\
&= \mathbb{E}[Y \log Y] - \mathbb{E}[Yg] - \mathbb{E}[Y] \log \mathbb{E}[Y] \\
&= \mathbb{H}(Y) - \mathbb{E}[Yg],
\end{aligned}$$

and as $\mathbb{H}_Q(Ye^{-g}) \geq 0$ (entropy is positive due to Corollary 5.2) we get that

$$\mathbb{H}(Y) \geq \mathbb{E}[Yg].$$

The equality in (5.13) follows by setting $e^g = Y/\mathbb{E}[Y]$, i.e., $g = \log Y - \log \mathbb{E}[Y]$. □

Remark 5.21 (Variational representation of the entropy) One can easily extend the variational formula (5.13) in Proposition to the set $\mathcal{G} = \{g: \Omega \rightarrow \mathbb{R}: e^g \leq 1\}$, namely

$$\mathbb{H}(Y) = \sup_{g \in \mathcal{G}} \{\mathbb{E}[gY]\}. \quad (5.14)$$

◇

5.2 Concentration via Isoperimetric Inequalities

We will see that Lipschitz functions concentrate well on $S^{(n-1)}$. In the following, when we consider the sphere $S^{(n-1)}$ or $\sqrt{n}S^{(n-1)}$ we use the Euclidean metric in \mathbb{R}^n instead of the geodesic metric of the spheres.

Theorem 5.22 (Concentration of Lipschitz functions on the sphere) *Let $f: \sqrt{n}S^{(n-1)} \rightarrow \mathbb{R}$ be a Lipschitz function and*

$$X \sim \text{Unif}(\sqrt{n}S^{(n-1)}).$$

Then

$$\|f(X) - \mathbb{E}[f(X)]\|_{\psi_2} \leq C\|f\|_{\text{Lip}},$$

for some absolute constant $C > 0$.

The statement of Theorem 5.22 amounts to the following concentration result, for every $t \geq 0$,

$$\mathbb{P}\left(|f(X) - \mathbb{E}[f(X)]| \geq t\right) \leq 2 \exp\left(-\frac{Ct^2}{\|f\|_{\text{Lip}}^2}\right), \quad (5.15)$$

for some absolute constant $C > 0$. We already know this statement for linear functions, see Theorem 3.21 saying that when $X \sim \sqrt{n}(S^{(n-1)})$ we have that X (or any linear map) is sub-Gaussian. To prove the extension to any nonlinear Lipschitz function we need two fundamental results, the so-called isoperimetric inequalities, which we can only state in order not to overload the lecture.

Definition 5.23 Suppose $f: \mathbb{R}^n \rightarrow \mathbb{R}$ is some function. Its *level-sets* (or *sub-level sets*) are

$$\mathcal{L}_f(c) := \{x \in \mathbb{R}^n : f(x) \leq c\}, \quad c \in \mathbb{R}.$$

Theorem 5.24 (Isoperimetric inequality on \mathbb{R}^n) *Among all subsets $A \subset \mathbb{R}^n$ with given volume, Euclidean balls have minimal surface area. Moreover, for any $\varepsilon > 0$, Euclidean balls minimise the volume of the ε -neighbourhood of A ,*

$$A_\varepsilon := \{x \in \mathbb{R}^n : \exists y \in A : \|x - y\|_2 \leq \varepsilon\} = A + \varepsilon B^{(n)},$$

where $B^{(n)}$ is the unit ball in \mathbb{R}^n .

Theorem 5.25 (Isoperimetric inequality on the sphere) *Let $\varepsilon > 0$. Then, among all $A \subset S^{(n-1)}$ with given area $\sigma_{n-1}(A)$, the spherical caps minimise the area of the ε -neighbourhood $\sigma_{n-1}(A_\varepsilon)$,*

$$A_\varepsilon := \{x \in S^{(n-1)} : \exists y \in A : \|x - y\|_2 \leq \varepsilon\}.$$

A spherical cap $C(a, \varepsilon)$ centred at a point $a \in S^{(n-1)}$ is the set

$$C(a, \varepsilon) := \{x \in S^{(n-1)} : \|x - a\|_2 \leq \varepsilon\}.$$

The following lemma is a crucial step in the proof of Theorem 5.22.

Lemma 5.26 (Blow-up) For any $A \subset \sqrt{n}S^{(n-1)}$, denote σ the normalised area on the sphere. if $\sigma(A) \geq \frac{1}{2}$, then, for every $t \geq 0$,

$$\sigma(A_t) \geq 1 - 2 \exp(-ct^2)$$

for some absolute constant $c > 0$.

Proof. For the hemisphere $H = \{x \in \sqrt{n}S^{(n-1)}: x_1 \leq 0\}$, we have $\sigma(A) \geq \frac{1}{2} = \sigma(H)$. The t -neighbourhood H_t of the hemisphere H is a spherical cap, and the isoperimetric inequality in Theorem 5.25 gives

$$\sigma(A_t) \geq \sigma(H_t).$$

We continue as in our proof of Theorem 3.21 noting that the normalised measure σ is the uniform probability measure on the sphere such that

$$\sigma(H_t) = \mathbb{P}(X \in H_t).$$

Recall that in that context, $X \sim \text{Unif}(\sqrt{n}S^{(n-1)})$, and thus X is Sub-Gaussian according to Theorem 3.21. Because of

$$H_t \supset \left\{ x \in \sqrt{n}S^{(n-1)}: x_1 \leq \frac{t}{\sqrt{2}} \right\}$$

we have

$$\sigma(H_t) \geq \mathbb{P}(X_1 \leq t/\sqrt{2}) = 1 - \mathbb{P}(X_1 > t/\sqrt{2}) \geq 1 - 2 \exp(-ct^2),$$

for some absolute constant $c > 0$. □

Proof of Theorem 5.22. Without loss of generality we assume that $\|f\|_{\text{Lip}} = 1$. Let M denote the median of $f(X)$, that is,

$$\mathbb{P}(f(X) \leq M) \geq \frac{1}{2} \quad \text{and} \quad \mathbb{P}(f(X) \geq M) \geq \frac{1}{2}.$$

The set $A = \{x \in \sqrt{n}S^{(n-1)}: f(x) \leq M\}$ is a level (sub-level) set of f with $\mathbb{P}(X \in A) \geq \frac{1}{2}$. Then Lemma 5.26 implies that $\mathbb{P}(X \in A_t) \geq 1 - 2 \exp(-Ct^2)$ for some absolute constant $C > 0$. We claim that, for every $t \geq 0$,

$$\mathbb{P}(X \in A_t) \leq \mathbb{P}(f(X) \leq M + t). \tag{5.16}$$

To see (5.16), note that $X \in A_t$ implies $\|X - y\|_2 \leq t$ for some point $y \in A$. By our definition of the set A , $f(y) \leq M$. As $\|f\|_{\text{Lip}} = 1$, we have

$$f(X) - f(y) \leq |f(X) - f(y)| \leq \|X - y\|_2$$

and thus

$$f(X) \leq f(y) + \|X - y\|_2 \leq M + t,$$

which implies (5.16). Hence

$$\mathbb{P}(f(X) \leq M + t) \geq 1 - 2 \exp(-Ct^2).$$

We now repeat our argument for $-f$: We define $\tilde{A} = \{x \in \sqrt{n}S^{(n-1)} : -f(x) \leq M\}$. Then $\mathbb{P}(X \in \tilde{A}) \geq 1/2$ and thus $\mathbb{P}(X \in \tilde{A}_t) \geq 1 - 2 \exp(-Ct^2)$. Now $X \in \tilde{A}$ implies $\|X - y\|_2 \leq t$ for some $y \in \tilde{A}$, and $f(y) \geq M$ by definition of \tilde{A} .

$$-f(X) - (-f(y)) \leq \|X - y\|_2 \leq t \Rightarrow f(X) \geq f(y) - t \geq M - t,$$

and thus $\mathbb{P}(f(X) \geq M - t) \geq 1 - 2 \exp(-\tilde{C}t^2)$ for some absolute constant $\tilde{C} > 0$. Combining our two estimates we obtain

$$\mathbb{P}(|f(X) - M| \leq t) \geq 1 - 2 \exp(-\hat{C}t^2)$$

for some absolute constant $\hat{C} > 0$, and thus the immediate tail estimate shows that

$$\|f(X) - M\|_{\psi_2} \leq C$$

for some absolute constant $C > 0$. To replace the median M by the expectation $\mathbb{E}[f(X)]$ note that although the median is not unique it is a fixed real number determined by the distribution of the random variable X and the function f . We use the centering Lemma 2.27 to get

$$\begin{aligned} \left| \|f(X)\|_{\psi_2} - \|M\|_{\psi_2} \right| &\leq \|f(X) - M\|_{\psi_2} \leq C \\ \|M\|_{\psi_2} &\leq \tilde{C} \Rightarrow -\tilde{C} + \|f(X)\|_{\psi_2} \leq C \\ &\Rightarrow \|f(X)\|_{\psi_2} \leq \tilde{C} + C \\ &\Rightarrow \|f(X) - \mathbb{E}[f(X)]\|_{\psi_2} \leq C. \end{aligned}$$

□

Definition 5.27 (Isoperimetric problem) Let (E, d) be a metric space and $\mathcal{B}(E)$ its Borel- σ -algebra, and $P \in \mathcal{M}_1(E, \mathcal{B}(E))$ some given probability measure and X an E -valued random variable.

Isoperimetric problem: Given $p \in (0, 1)$ and $t > 0$, find the sets A with $P(X \in A) \geq p$ for which $P(d(X, A) \geq t)$ is maximal, where

$$d(X, A) := \inf_{y \in A} \{d(X, y)\}.$$

Concentration function:

$$\alpha(t) := \sup_{\substack{A \subset E: \\ P(A) \geq 1/2}} \{P(d(X, A) \geq t)\} = \sup_{\substack{A \subset E: \\ P(A) \geq 1/2}} \{P(A_t^c)\},$$

where A_t is the t -blow-up of the set A , $A_t = \{x \in E : d(x, A) < t\}$. For a given function $f : E \rightarrow \mathbb{R}$ denote the median of $f(X)$ by $M_f(X)$.

There are many isoperimetric inequalities, we only mention the Gaussian isoperimetric inequality as it is widely used. Recall that the Gaussian (standard normal distribution) is given by the probability measure $\gamma_n \in \mathcal{M}_1(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$,

$$\gamma_n(A) = \int_A (2\pi)^{-n/2} e^{-\|x\|^2/2} dx_1 \cdots dx_n, \quad A \subset \mathcal{B}(\mathbb{R}^n).$$

The concentration function for the one-dimensional Gaussian ($n = 1$) is just $\alpha(t) = 1 - \Phi(t)$ with $\Phi(t) = \frac{1}{\sqrt{2\pi}} \int_t^\infty e^{-u^2/2} du$. In the following statement half-spaces are sets of the form

$$A = \{x \in \mathbb{R}^n : \langle x, u \rangle < \lambda\}, \quad u \in \mathbb{R}^n, \lambda \in \mathbb{R}, \quad \text{or} \quad A = \{x \in \mathbb{R}^n : x_1 \leq z\}, \quad z \in \mathbb{R}.$$

Theorem 5.28 (Gaussian isoperimetric inequality) *Let $\varepsilon > 0$ be given. Among all $A \subset \mathbb{R}^n$ with fixed Gaussian measure $\gamma_n(A)$, the half-spaces minimise the Gaussian measure $\gamma_n(A_\varepsilon)$ of the ε -neighbourhood A_ε .*

From this we can obtain the following concentration result. The proof is using similar steps as done above, and we leave the details as exercise for the reader.

Theorem 5.29 (Gaussian concentration) *Suppose $X \sim N(0, \mathbb{I}_n)$, and let $f: \mathbb{R}^n \rightarrow \mathbb{R}$ be a Lipschitz continuous function. Then*

$$\|f(X) - \mathbb{E}[f(X)]\|_{\psi_2} \leq C \|f\|_{\text{Lip}}$$

for some absolute constant $C > 0$.

5.3 Some matrix calculus and covariance estimation

In this section we are generalising our concentration to random matrices. The main focus is the following Bernstein type result for random matrices

Theorem 5.30 (Matrix Bernstein inequality) *Let X_1, \dots, X_N be independent mean-zero random $n \times n$ symmetric matrices such that $\|X_i\| \leq K$ almost surely for all $i = 1, \dots, N$. Then, for every $t \geq 0$, we have*

$$\mathbb{P}\left(\left\|\sum_{i=1}^N X_i\right\| \geq t\right) \leq 2n \exp\left(-\frac{t^2/2}{\sigma^2 + Kt/3}\right),$$

where $\sigma^2 = \left\|\sum_{i=1}^N \mathbb{E}[X_i^2]\right\|$ is the norm of the matrix variance of the sum.

For the proof we shall introduce a few well-known facts about matrix calculus.

Definition 5.31 (a) For any symmetric $n \times n$ matrix X with eigenvalues $\lambda_i = \lambda_i(X)$ and corresponding eigenvectors u_i the function of a matrix for any given $f: \mathbb{R} \rightarrow \mathbb{R}$ is defined as the $n \times n$ matrix

$$f(X) := \sum_{i=1}^n f(\lambda_i) u_i u_i^T.$$

(b) Suppose X is a $n \times n$ matrix. We write $X \succcurlyeq 0$ if X is positive-semidefinite. Equivalently, $X \succcurlyeq 0$ if all eigenvalues of X are positive, i.e., $\lambda_i \geq 0$. For $Y \in \mathbb{R}^{n \times n}$, we set $X \succcurlyeq Y$ and $Y \preccurlyeq X$ if $X - Y \succcurlyeq 0$.

We borrow the following trace inequalities from linear algebra. Recall the notion of the trace of a matrix.

Golden-Thompson inequality: For any $n \times n$ symmetric matrices A and B we have

$$\text{Trace}(e^{A+B}) \leq \text{Trace}(e^A e^B).$$

Lieb's inequality: Suppose H is a $n \times n$ symmetric matrix and define the function on matrices

$$f(X) := \text{Trace}(\exp(H + \log X)).$$

Then f is concave on the space of positive-definite $n \times n$ matrices.

In principle one can prove the Matrix Bernstein inequality in Theorem 5.30 with these two results from matrix analysis. If X is a random positive-definite matrix then Lieb's and Jensen's inequality imply that

$$\mathbb{E}[f(X)] \leq f(\mathbb{E}[X]).$$

We now apply this with $X = e^Z$ for some $n \times n$ symmetric matrix Z :

Lemma 5.32 (Lieb's inequality for random matrices) *Let H be a fixed $n \times n$ symmetric matrix and Z be a random $n \times n$ symmetric matrix. Then*

$$\mathbb{E}[\text{Trace}(\exp(H + Z))] \leq \text{Trace}(\exp(H + \log \mathbb{E}[e^Z])). \tag{5.17}$$

The proof of Theorem 5.30 follows below and is based on the following bound of the moment generating function (MGF).

Lemma 5.33 (Bound on MGF) *Let X be an $n \times n$ symmetric mean-zero random matrix such that $\|X\| \leq K$ almost surely. Then*

$$\mathbb{E}[\exp(\lambda X)] \leq \exp(g(\lambda) \mathbb{E}[X^2]), \quad \text{where } g(\lambda) = \frac{\lambda^2/2}{1 - |\lambda|K/3}, \quad \text{for } |\lambda| < 3/K.$$

Proof of Lemma 5.33. For $z \in \mathbb{C}$ we can easily obtain the following estimate

$$e^z \leq 1 + z + \frac{1}{1 - |z|/3} \frac{z^2}{2} \quad \text{for } |z| < 3.$$

This can be easily derived from the Taylor series of the exponential function in conjunction with the lower bound $p! \geq 2 \times 3^{p-2}$ and the geometric series. Details are left to the reader. Now with $z = \lambda x$, $|x| \leq K$, $|\lambda| < 3/K$, we then obtain

$$e^{\lambda x} \leq 1 + \lambda x + g(\lambda)x^2.$$

Now we transfer this inequality from scalar to matrices:

$$e^{\lambda X} \preceq \mathbb{1}_n + \lambda X + g(\lambda)X^2,$$

and then, after taking the expectation and using $\mathbb{E}[X] = 0$, to arrive at

$$\mathbb{E}[e^{\lambda X}] \leq 1 + g(\lambda)\mathbb{E}[X^2].$$

To finish, use the inequality $1 + z \leq e^z$ to conclude with

$$\mathbb{E}[e^{\lambda X}] \leq \exp(g(\lambda)\mathbb{E}[X^2]).$$

□

Proposition 5.34 (Expectation bound via the Bernstein Theorem) *Under all the assumptions in Theorem 5.30 we have the tail bound*

$$\mathbb{P}\left(\left\|\sum_{i=1}^N X_i\right\| \geq t\right) \leq 2n \exp\left(-\frac{t^2/2}{\sigma^2 + Kt/3}\right).$$

Then

$$\mathbb{E}\left[\left\|\sum_{i=1}^N X_i\right\|\right] \leq 2\sigma\left(\frac{\sqrt{2\pi}}{2} + \sqrt{\log(2n)}\right) + \frac{4}{3}K\left(1 + \log(2n)\right). \quad (5.18)$$

Proof. Define $b := \frac{K}{3}$. Then the right hand side in Theorem 5.30 reads as

$$2n \exp\left(-\frac{t^2}{2(\sigma^2 + bt)}\right).$$

We will use a crude union-type bound on the tail probability itself by observing that either $\sigma^2 \leq bt$ or $\sigma^2 \geq bt$. Define $Z := \sum_{i=1}^N X_i$. For every $t \geq 0$,

$$\begin{aligned} \mathbb{P}(\|Z\| \geq t) &\leq 2n \exp\left(-\frac{t^2}{2(\sigma^2 + bt)}\right) \leq 2n \max\left\{\exp\left(-\frac{t}{4b}\right), \exp\left(-\frac{t^2}{4\sigma^2}\right)\right\} \\ &\leq 2n \exp\left(-\frac{t}{4b}\right) + 2n \exp\left(-\frac{t^2}{4\sigma^2}\right). \end{aligned}$$

We shall combine this with the trivial inequality $\mathbb{P}(\|Z\| \geq t) \leq 1$. Thus

$$\mathbb{P}(\|Z\| \geq t) \leq 1 \wedge 2n \exp\left(-\frac{t}{4b}\right) + 1 \wedge 2n \exp\left(-\frac{t^2}{4\sigma^2}\right),$$

and

$$\mathbb{E}[\|Z\|] = \int_0^\infty \mathbb{P}(\|Z\| \geq t) dt =: I_1 + I_2,$$

with $I_1 = \int_0^\infty 1 \wedge 2n \exp(-t/(4b)) dt$ and $I_2 = \int_0^\infty 1 \wedge 2n \exp(-t^2/(4\sigma^2)) dt$. Solve $2n \exp(-t_1/(4b)) = 1$ to obtain $t_1 = 4b \log(2n)$, and then

$$I_1 = \int_0^{t_1} 1 dt + \int_{t_1}^\infty (2n) \exp(-t/(4b)) dt = t_1 + 4b = 4b(1 + \log(2n)).$$

Solve $2n \exp(-t^2/(4\sigma^2)) = 1$ to obtain $t_2 = 2\sigma\sqrt{\log(2n)}$, and then

$$\begin{aligned} I_2 &= \int_0^{t_2} 1 dt + \int_{t_2}^\infty (2n) \exp(-t^2/(4\sigma^2)) dt \stackrel{i=t/2\sigma}{=} t_2 + 2\sigma \int_{\sqrt{\log(2n)}}^\infty (2n) e^{-\tilde{t}^2/2} d\tilde{t} \\ &\leq 2\sigma(\sqrt{\log(2n)} + \sqrt{\pi}), \end{aligned}$$

where we used another transformation for the inequality, namely, $x = \tilde{t} - \sqrt{\log(2n)}$. Equivalently, we may use $2ne^{-\tilde{t}^2/2} \leq e^{-x^2}$ for all $t \geq \sqrt{\log(2n)}$. We arrive at

$$\mathbb{E}[\|Z\|] \leq 2\sigma \left(\frac{\sqrt{\pi}}{2} + \sqrt{\log(2n)} \right) + 4b(1 + \log(2n)).$$

□

Proof of Theorem 5.30.

Step 1: Define $S := \sum_{i=1}^N X_i$ and $\lambda_m(S) := \max_{1 \leq i \leq n} \lambda_i(S)$ the largest eigenvalue of S . Then $\|S\| = \max\{\lambda_m(S), \lambda_m(-S)\}$.

$$\mathbb{P}(\lambda_m(S) \geq t) = \mathbb{P}(e^{\lambda_m(S)} \geq e^{t}) \leq e^{-\lambda t} \mathbb{E}[e^{\lambda \lambda_m(S)}].$$

The eigenvalues of $e^{\lambda S}$ are $e^{\lambda \lambda_i(S)}$, and thus

$$E := \mathbb{E}[e^{\lambda \lambda_m(S)}] = \mathbb{E}[\lambda_m(e^{\lambda S})].$$

All eigenvalues of $e^{\lambda S}$ are positive, hence the maximal eigenvalue of $e^{\lambda S}$ is bounded by the sum of all eigenvalues, that is, by the trace of $e^{\lambda S}$. Henceforth

$$E \leq \mathbb{E}[\text{Trace}(e^{\lambda S})]. \quad (5.19)$$

Step 2: We now turn to bound the expectation of the trace using Lieb's inequality, in particular Lemma 5.32. We write the exponent of $e^{\lambda S}$ separately, namely splitting off the last term of the sum, writing

$$\text{Trace}(e^{\sum_{i=1}^{N-1} \lambda X_i + \lambda X_N}). \quad (5.20)$$

When we take the expectation of (5.20) we condition on X_1, \dots, X_{N-1} and apply Lieb's inequality (Lemma 5.32 for the fixed matrix $H := \sum_{i=1}^{N-1} \lambda X_i$), and obtain

$$\begin{aligned} E &\leq \mathbb{E} \left[\text{Trace} \left(e^{H + \lambda X_N} \right) \right] = \mathbb{E}_{X_1, \dots, X_{N-1}} \otimes \mathbb{E}_{X_N} \left[\text{Trace} \left(e^{H + \lambda X_N} \right) \middle| X_1, \dots, X_{N-1} \right] \\ &\leq \mathbb{E}_{X_1, \dots, X_{N-1}} \left[\text{Trace} \left(\exp \left(\sum_{i=1}^{N-1} \lambda X_i + \log \widehat{\mathbb{E}}_{X_N} [\exp(\lambda X_N)] \right) \right) \right], \end{aligned}$$

where $\widehat{\mathbb{E}}$ is the conditional expectation with respect to X_N conditioned on X_1, \dots, X_{N-1} . We repeat this application of Lemma 5.32 successively for $\lambda X_{N-1}, \dots, \lambda X_1$, to arrive at

$$E \leq \text{Trace} \left(\exp \left(\sum_{i=1}^N \log \mathbb{E}_{X_i} \left[e^{\lambda X_i} \right] \right) \right).$$

Step 3: We use Lemma 5.33 to bound $\mathbb{E}[e^{\lambda X_i}]$ for each $X_i, i = 1, \dots, N$,

$$\mathbb{E}_{X_i}[e^{\lambda X_i}] \leq \exp \left(g(\lambda) \mathbb{E}[X_i^2] \right), \quad i = 1, \dots, N.$$

Step 4: With Step 3 we get

$$E \leq \text{Trace} \left(\exp \left(g(\lambda) Z \right) \right),$$

where $Z := \sum_{i=1}^N \mathbb{E}[X_i^2]$. The trace of $\exp(g(\lambda)Z)$ is a sum of n positive eigenvalues,

$$E \leq n \lambda_m(\exp(g(\lambda)Z)) = n \exp(g(\lambda) \lambda_m(Z)) = n \exp(g(\lambda) \|Z\|) = n \exp(g(\lambda) \sigma^2).$$

Thus, with Step 1 and Lemma 5.33,

$$\mathbb{P}(\lambda_m(S) \geq t) \leq n \exp(-\lambda t + g(\lambda) \sigma^2) \quad \text{for } |\lambda| < 3/K.$$

The minimum in the exponent is attained for $\lambda = t/(\sigma^2 + 2Kt/3)$. We finally conclude with

$$\mathbb{P}(\lambda_m(S) \geq t) \leq n \exp \left(- \frac{t^2/2}{\sigma^2 + Kt/3} \right),$$

and repeating our steps for $\lambda_m(-S)$ we conclude with the statement of the theorem. \square

5.4 Application - Johnson-Lindenstrauss Lemma

Before we study an application of our results in the previous sections we recall our basic results. In Theorem 5.16 we have a concentration result for Lipschitz images of random vector with independent coordinates using entropy methods. Using isoperimetric inequalities in the Euclidean space we show in Theorem 5.22 a concentration result for $X \sim \sqrt{n}S^{(n-1)}$. Note that such a random vector does not have independent coordinates. In the following exercise it is easy to extend this result to the unit sphere.

Exercise 5.35 (Concentration for the unit sphere) Let f be a Lipschitz function on the unit sphere $S^{(n-1)}$. Show that for $X \sim \text{Unif}(S^{(n-1)})$ one has

$$\|f(X) - \mathbb{E}[f(X)]\|_{\psi_2} \leq \frac{C \|f\|_{\text{Lip}}}{\sqrt{n}}.$$

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

$$\mathbb{P}(|f(X) - \mathbb{E}[f(X)]| \geq t) \leq 2 \exp \left(- \frac{cnt^2}{\|f\|_{\text{Lip}}^2} \right).$$



In Theorem 5.29 we have a concentration result for Lipschitz images of normally distributed random vectors using the Gaussian isoperimetric inequality. One can find similar concentration results for other metric spaces. recall the Hamming cube and its metric $d_{\mathcal{H}}$ in Definition 4.8. We can define the uniform distribution on the Hamming cube \mathcal{H} as the probability measure $P(A) := |A|/2^n$ for any subset $A \subset \mathcal{H}$. If $X \sim \text{Unif}(\mathcal{H})$, then the coordinates X_i of X are Bernoulli distributed with parameter $1/2$. Then one can obtain the following concentration result.

Theorem 5.36 (Concentration for the Hamming cube) *Suppose $X \sim \text{Unif}(\mathcal{H})$ and $f: \mathcal{H} \rightarrow \mathbb{R}$. Then*

$$\|f(X) - \mathbb{E}[f(X)]\|_{\psi_2} \leq \frac{C\|f\|_{\text{Lip}}}{\sqrt{n}}.$$

We now want to discuss an important application of our concentration result for the Euclidean sphere. Suppose we have N data points in \mathbb{R}^n , i.e., a sample $X_i \in \mathbb{R}^n, i = 1, \dots, N$. We would like to reduce the dimension of the data without sacrificing too much of its geometry. Consider a subspace $E \subset \mathbb{R}^n$ with dimension $\dim(E) = m \ll n$. Denote $G_{n,m}$ the set (manifold) consisting of all m -dimensional subspaces of \mathbb{R}^n . If we choose $m = 1$ we can identify $G_{n,1}$ with the unit sphere $S^{(n-1)}$. To see this recall that any 1-dimensional subspace of \mathbb{R}^n can be generated by a direction vector $u \in S^{(n-1)}$, i.e.,

$$E = \{\alpha u: \alpha \in \mathbb{R}\}.$$

Then the set on the right hand side generate the subspace E with dimension $\dim(E) = 1$. So any concentration result for $G_{n,m}$ includes the concentration for the sphere as a special case. We need a metric and a probability measure for $G_{n,m}$. The distance (metric) between subspaces E and F can be defined as the operator norm

$$d(E, F) := \|P_E - P_F\|,$$

where $P_E(P_F)$ is the orthogonal projection onto E , i.e., $P_E: \mathbb{R}^n \rightarrow E, P_E(\mathbb{R}^n) = E$. If we define P to be the uniform (Haar) measure then we expect concentration results for $X \sim \text{Unif}(G_{n,m})$ in the metric space $(G_{n,m}, d, P)$.

Theorem 5.37 *Suppose that $X \sim \text{Unif}(G_{n,m})$ and let $f: G_{n,m} \rightarrow \mathbb{R}$ some function. Then*

$$\|f(X) - \mathbb{E}[f(X)]\|_{\psi_2} \leq \frac{C\|f\|_{\text{Lip}}}{\sqrt{n}}.$$

The proof of that theorem goes beyond what we can do in this lecture. It is based on concentration results for the special orthogonal group and the fact that $G_{n,m}$ can be written as quotient of orthogonal groups.

We now present the Johnson-Lindenstrauss Lemma for the N data points and prove the statement that the geometry of the given data is well preserved if we choose E to be a random subspace of dimension $\dim(E) = m \sim \log N$, where \sim means that in the limit $N \rightarrow \infty$ the quotient of both sides converges to 1. Thus we consider random subspaces $E \sim \text{Unif}(\mathbb{G}_{n,m})$ and the space $(\mathbb{G}_{n,m}, d, P)$ with P being the uniform (Haar) measure. Note that we have the following invariance.

$$P(E \in \mathcal{E}) = P(U(E) \in \mathcal{E}), \quad \text{for any subset } \mathcal{E} \subset \mathbb{G}_{n,m} \text{ and any orthogonal matrix } U.$$

Theorem 5.38 (Johnson-Lindenstrauss-Lemma) *Let $\mathcal{H} = \{X_1, \dots, X_N\}$, $X_i \in \mathbb{R}^n$, $i = 1, \dots, N$, be a set of N points in \mathbb{R}^n and pick $\varepsilon > 0$ and assume that $m \geq C/\varepsilon^2 \log N$ for some absolute constant $C > 0$. Suppose $E \sim \text{Unif}(\mathbb{G}_{n,m})$ and denote P the orthogonal projection onto E . Then, with probability at least $1 - 2 \exp(-c\varepsilon^2 m)$ the scaled projection $Q := \sqrt{\frac{n}{m}}P$ is an approximate isometry on the set \mathcal{H} , that is,*

$$(1 - \varepsilon)\|x - y\|_2 \leq \|Qx - Qy\|_2 \leq (1 + \varepsilon)\|x - y\|_2, \quad \text{for all } x, y \in \mathcal{H}.$$

We base the proof of the theorem on the concentration of Lipschitz functions for the sphere in Theorem 5.22 respectively its version on the unit sphere in Exercise 5.35. We consider first random projections P acting on a fixed vector $x - y$, and then we take the union bound over all N^2 differences $x - y$, $x, y \in \mathcal{H}$. For any fixed vector the next lemma gives the desired properties.

Lemma 5.39 *Let P be an orthogonal projection from \mathbb{R}^n onto a random m -dimensional subspace uniformly distributed in $\mathbb{G}_{n,m}$. Let $z \in \mathbb{R}^n$ be a fixed point and choose $\varepsilon > 0$. Then the following holds.*

(a)

$$(\mathbb{E}[\|Pz\|_2^2])^{1/2} = \sqrt{\frac{m}{n}}\|z\|_2.$$

(b) *With probability at least $1 - 2 \exp(-c\varepsilon^2 m)$, we have*

$$(1 - \varepsilon)\sqrt{\frac{m}{n}}\|z\|_2 \leq \|Pz\|_2 \leq (1 + \varepsilon)\sqrt{\frac{m}{n}}\|z\|_2.$$

Proof. (a) Without loss of generality we can assume that $\|z\|_2 = 1$. This is possible as for $z = 0$ the statement holds trivially and for $z \neq 0$ we can simply multiply by $1/\|z\|_2$. Instead of a random projection P we can fix a projection and consider $z \sim \text{Unif}(S^{n-1})$. According to the above mentioned invariance we can assume without loss of generality that P is the coordinate projection onto the first m coordinates in \mathbb{R}^n . Thus

$$\mathbb{E}[\|Pz\|_2^2] = \mathbb{E}\left[\sum_{i=1}^m z_i^2\right] = \sum_{i=1}^m \mathbb{E}[z_i^2] = m\mathbb{E}[z_1^2],$$

where the last equality follows from the fact that the coordinates z_i are identically distributed. As $z \sim \text{Unif}(S^{(n-1)})$ we have

$$1 = \sum_{i=1}^n \mathbb{E}[z_i^2] = n\mathbb{E}[z_1^2],$$

and thus $\mathbb{E}[z_1^2] = 1/n$ and therefore we get the statement that

$$\mathbb{E}[\|Pz\|_2^2] = \frac{m}{n},$$

and thus (a).

(b) Define the function $f: S^{(n-1)} \rightarrow \mathbb{R}$ by $f(x) := \|Px\|_2$. Then f is a Lipschitz function with $\|f\|_{\text{Lip}} = 1$:

$$|f(x) - f(y)| = |\|Px\|_2 - \|Py\|_2| \leq \|Px - Py\|_2 \leq \|P\|\|x - y\|_2 = \|x - y\|_2$$

as $\|P\| = 1$. Thus Exercise 5.35 (Theorem 5.22) give the concentration result

$$\mathbb{P}(|F(X) - \mathbb{E}[F(X)]| \geq t) \leq 2 \exp(-\tilde{C}nt^2). \quad (5.21)$$

Statement (5.21) is not quite (b) but we can replace the expectation $\mathbb{E}[f(X)]$ by $\sqrt{m/n}$ as follows. First note that due to Jensen

$$\mathbb{E}[F(X)] \leq \sqrt{\frac{m}{n}} = \mathbb{E}[f(X)^2]^{1/2}.$$

Then $-t \leq f(X) - \mathbb{E}[f(X)] \leq t$ is equivalent to

$$-t + \mathbb{E}[f(X)] - \sqrt{\frac{m}{n}} \leq f(X) - \sqrt{\frac{m}{n}} \leq t + \mathbb{E}[f(X)] - \sqrt{\frac{m}{n}}.$$

With a suitable constant $c > 0$ to accommodate the change in t we arrive at statement (b) by choosing $t = \varepsilon\sqrt{m/n}$. \square

Proof of Theorem 5.38. We consider the difference set

$$\mathcal{H} - \mathcal{H} = \{x - y : x, y \in \mathcal{H}\}.$$

We shall show, with the required minimal probability, that

$$(1 - \varepsilon)\|z\|_2 \leq \|Q\|_2 \leq (1 + \varepsilon)\|z\|_2$$

holds for all differences $z \in \mathcal{H} - \mathcal{H}$. Setting $Q := \sqrt{\frac{n}{m}}P$, this is equivalent to showing that

$$(1 - \varepsilon)\sqrt{\frac{m}{n}}\|z\|_2 \leq \|Pz\|_2 \leq (1 + \varepsilon)\sqrt{\frac{m}{n}}\|z\|_2. \quad (5.22)$$

For any fixed z , Lemma 5.39 states that (5.22) holds with probability at least $1 - 2 \exp(-c\varepsilon^2 m)$. All what remains is to take a union bound. First note that (5.22) holds simultaneously for all $z \in \mathcal{H} - \mathcal{H}$. with probability at least

$$1 - |\mathcal{H} - \mathcal{H}| 2 \exp(-c\varepsilon^2 m) \geq 1 - N^2 2 \exp(-c\varepsilon^2 m).$$

If we choose $m \geq (2/c\varepsilon^2) \log N + 1$, then we conclude with the statement of the theorem. \square

6 Basic tools in high-dimensional probability

6.1 Decoupling

Definition 6.1 Let X_1, \dots, X_n be independent real-valued random variables and $a_{ij} \in \mathbb{R}, i, j = 1, \dots, n$. The random quadratic form

$$\sum_{i,j=1}^n a_{ij} X_i X_j = X^T A X = \langle X, A X \rangle, \quad X = (X_1, \dots, X_n) \in \mathbb{R}^n, A = (a_{ij}),$$

is called *chaos* in probability theory.

For simplicity, we assume in the following that the random variables X_i have mean-zero and unit variances,

$$\mathbb{E}[X_i] = 0, \quad \text{Var}(X_i) = 1, i = 1, \dots, n.$$

Then

$$\mathbb{E}[\langle X, A X \rangle] = \sum_{i,j=1}^n a_{ij} \mathbb{E}[X_i X_j] = \sum_{i=1}^n a_{ii} = \text{Trace}(A).$$

We shall study concentration properties for chaos. This time we need to develop tools to overcome the fact that we have sums of not necessarily independent random variables. The idea is to use the *decoupling technique*. The idea is to study the following random quadratic form,

$$\sum_{i,j=1}^n a_{ij} X_i X'_j = X^T A X' = \langle X, A X' \rangle, \quad (6.1)$$

where $X' = (X'_1, \dots, X'_n)$ is a random vector independent of X but with the same distribution as X . Obviously, the bilinear form in (6.1) is easier to study, e.g., when we condition on X' we simply obtain a linear form for the random vector X . The vector X' is called an independent copy of X , and conditioning on X' ,

$$\langle X, A X' \rangle = \sum_{i=1}^n c_i X_i \quad \text{with} \quad c_i = \sum_{j=1}^n a_{ij} X'_j,$$

is a random linear form for X depending on the condition of the independent copy X' .

Lemma 6.2 Let Y and Z be independent random vectors in \mathbb{R}^n such that $\mathbb{E}_Y[Y] = \mathbb{E}_Z[Z] = 0$ (we write indices when we want to stress the expectation for a certain random variable). Then, for every convex function $F: \mathbb{R}^n \rightarrow \mathbb{R}$, one has

$$\mathbb{E}[F(Y)] \leq \mathbb{E}[F(Y + Z)].$$

Proof. Fix $y \in \mathbb{R}^n$ and use $\mathbb{E}_Z[Z] = 0$ and the convexity to get

$$F(y) = F(y + \mathbb{E}_Z[Z]) \leq \mathbb{E}_Z[F(y + Z)].$$

We choose $y = Y$ and take expectation with respect to Y ,

$$\mathbb{E}_Y[F(Y)] = \mathbb{E}_Y[F(Y + \mathbb{E}_Z[Z])] = \mathbb{E}_Y[F(\mathbb{E}_Z[Y + Z])] \leq \mathbb{E}_Y \otimes \mathbb{E}_Z[F(Y + Z)].$$

□

Theorem 6.3 (Decoupling) *Let A be an $n \times n$ diagonal-free (i.e., diagonal entries vanish) and $X = (X_1, \dots, X_n) \in \mathbb{R}^n$ be a random vector with independent mean-zero coordinates X_i , and X' an independent copy of X . Then, for every convex function $F: \mathbb{R} \rightarrow \mathbb{R}$, one has*

$$\mathbb{E}[F(\langle X, AX \rangle)] \leq \mathbb{E}[F(4\langle X, AX' \rangle)]. \quad (6.2)$$

Proof. The idea is to study partial chaos

$$\sum_{(i,j) \in I \times I^c} a_{ij} X_i X_j$$

with a randomly chosen subset $I \subset \{1, \dots, n\}$. Let $\delta_1, \dots, \delta_n$ be independent Bernoulli random variables with $\mathbb{P}(\delta_i = 0) = \mathbb{P}(\delta_i = 1) = \frac{1}{2}$. Then define the random set $I = \{i: \delta_i = 1\}$. We condition on the random variable X and obtain for $i \neq j$ using $a_{ii} = 0$ and $\mathbb{E}[\delta_i(1 - \delta_j)] = 1/4$,

$$\langle X, AX \rangle = \sum_{i \neq j} a_{ij} X_i X_j = 4\mathbb{E}_\delta \left[\sum_{i \neq j} \delta_i(1 - \delta_j) a_{ij} X_i X_j \right] = 4\mathbb{E}_I \left[\sum_{(i,j) \in I \times I^c} a_{ij} X_i X_j \right],$$

where \mathbb{E}_δ is expectation with respect to the Bernoulli random variables and \mathbb{E}_I is expectation with respect to the Bernoulli random variables for $I = \{i: \delta_i = 1\}$. Apply F on both sides and use Jensen (w.r.t. to \mathbb{E}_I) and and take the expectation w.r.t. to \mathbb{E}_X in conjunction with Fubini to get

$$\mathbb{E}_X[F(\langle X, AX \rangle)] \leq \mathbb{E}_I \mathbb{E}_X \left[F \left(4 \sum_{(i,j) \in I \times I^c} a_{ij} X_i X_j \right) \right].$$

There is a realisation of a random set I such that

$$\mathbb{E}_X[F(\langle X, AX \rangle)] \leq \mathbb{E}_X \left[F \left(4 \sum_{(i,j) \in I \times I^c} a_{ij} X_i X_j \right) \right]. \quad (6.3)$$

We fix such an realisation of I until the end. The random variables $(X_i)_{i \in I}$ and $(X_j)_{j \in I^c}$ are independent and thus the distribution of the sum on the right hand side will not change if we replace X_j by X'_j . Hence we replace $X_j, j \in I^c$, by X'_j on the right hand side of (6.3) to get

$$\mathbb{E}_X[F(\langle X, AX \rangle)] \leq \mathbb{E} \left[F \left(4 \sum_{(i,j) \in I \times I^c} a_{ij} X_i X'_j \right) \right]. \quad (6.4)$$

It remains to show that

$$\text{R.H.S. of (6.4)} \leq \mathbb{E} \left[F \left(4 \sum_{(i,j) \in [n] \times [n]} a_{ij} X_i X_j' \right) \right], \quad (6.5)$$

where $[n] = \{1, \dots, n\}$. We now split the argument of the function F on the right hand side of (6.5) into three terms,

$$\sum_{(i,j) \in [n] \times [n]} a_{ij} X_i X_j' =: Y + Z_1 + Z_2,$$

with

$$Y := \sum_{(i,j) \in I \times I^c} a_{ij} X_i X_j', \quad Z_1 := \sum_{(i,j) \in I \times I} a_{ij} X_i X_j', \quad Z_2 := \sum_{(i,j) \in I^c \times [n]} a_{ij} X_i X_j'.$$

We now condition on all random variables except $(X_j')_{j \in I}$ and $(X_i)_{i \in I^c}$. We denote the corresponding conditional expectation by $\tilde{\mathbb{E}}$. The conditioning already fixes Y . Furthermore, Z_1 and Z_2 have zero expectation. Applications of Lemma 6.2 leads to

$$F(4Y) \leq \tilde{\mathbb{E}}[F(4Y + 4Z_1 + 4Z_2)],$$

and thus

$$\mathbb{E}[F(4Y)] \leq \mathbb{E}[F(4Y + 4Z_1 + 4Z_2)].$$

This proves (6.5) and finishes the argument by taking the final expectation with respect to I . \square

Theorem 6.4 (Hanson-Wright inequality) *Let $X = (X_1, \dots, X_n) \in \mathbb{R}^n$ be a random vector with independent mean-zero sub-Gaussian coordinates and A be an $n \times n$ matrix. Then, for every $t \geq 0$,*

$$\mathbb{P} \left(|\langle X, AX \rangle - \mathbb{E}[\langle X, AX \rangle]| \geq t \right) \leq 2 \exp \left(-c \min \left\{ \left(\frac{t^2}{K^4 \|A\|_{\mathbb{F}}^2} \right), \left(\frac{t}{K^2 \|A\|} \right) \right\} \right),$$

where $K := \max_{1 \leq i \leq n} \{\|X_i\|_{\psi_2}\}$.

We prepare the proof with the following two lemmas.

Lemma 6.5 (MGF of Gaussian chaos) *Let $X, X' \sim \mathcal{N}(0, \mathbb{1}_n)$, X and X' be independent, and let A be an $n \times n$ matrix. Then*

$$\mathbb{E}[\exp(\lambda \langle X, AX' \rangle)] \leq \exp(C \lambda^2 \|A\|_{\mathbb{F}}), \quad \text{for all } |\lambda| \leq C/\|A\|,$$

for some absolute constant $C > 0$.

Proof. We use the singular value decomposition of the matrix A , see Definition 4.10:

$$\begin{aligned} A &= \sum_{i=1}^n s_i u_i v_i^T \\ \langle X, AX' \rangle &= \sum_{i=1}^n s_i \langle u_i, X \rangle \langle v_i, X' \rangle \\ \langle X, AX' \rangle &= \sum_{i=1}^n s_i Y_i Y_i', \end{aligned}$$

where $Y = (\langle u_1, X \rangle, \dots, \langle u_n, X \rangle) \sim \mathbf{N}(0, \mathbb{1}_n)$ and $Y' = (\langle v_1, X' \rangle, \dots, \langle v_n, X' \rangle) \sim \mathbf{N}(0, \mathbb{1}_n)$. Independence yields

$$\mathbb{E}[\exp(\lambda \langle X, AX' \rangle)] = \prod_{i=1}^n \mathbb{E}[\exp(\lambda s_i Y_i Y_i')].$$

For each $i \in \{1, \dots, n\}$ we compute taking the expectation with respect to Y' , i.e., the conditional expectation holding the Y_i 's fixed,

$$\mathbb{E}[\exp(\lambda s_i Y_i Y_i') | Y] = \mathbb{E}[\exp(\lambda^2 s_i^2 Y_i^2 / 2)] \leq \exp(C \lambda^2 s_i^2), \quad \text{provided } \lambda^2 s_i^2 \leq C,$$

where the first equality follows from the fact the MGF of $X \sim \mathbf{N}(0, 1)$ is $\mathbb{E}[\exp(\lambda X)] = \exp(\lambda^2/2)$, and the inequality follows from the fact that Y_i are Gaussian and thus sub-Gaussian and therefore the square Y_i^2 is sub-exponential and thus property (v) in Proposition 2.28 gives the bound.

$$\mathbb{E}[\exp(\lambda \langle X, AX' \rangle)] \leq \exp\left(C \lambda^2 \sum_{i=1}^n s_i^2\right), \quad \text{provided } \lambda \leq \frac{C}{\max_{1 \leq i \leq n} \{s_i^2\}}.$$

We conclude with $\sum_{i=1}^n s_i^2 = \|A\|_F^2$ and $\max_{1 \leq i \leq n} \{s_i\} = \|A\|$. \square

Lemma 6.6 (Comparison) *Let X and X' be independent mean-zero sub-Gaussian random vectors in \mathbb{R}^n with $\|X\|_{\psi_2} \leq K$ and $\|X'\|_{\psi_2} \leq K$. Furthermore, let Y, Y' be independent normally distributed random vectors $Y, Y' \sim \mathbf{N}(0, \mathbb{1}_n)$, and A be an $n \times n$ matrix. Then*

$$\mathbb{E}[\exp(\lambda \langle X, AX' \rangle)] \leq \mathbb{E}[\exp(C K^2 \lambda \langle Y, AY' \rangle)].$$

Proof. We condition on X' and take the expectation with respect to X (denoted by \mathbb{E}_X). Then $\langle X, AX' \rangle$ is conditionally sub-Gaussian and

$$\mathbb{E}_X[\exp(\lambda \langle X, AX' \rangle)] \leq \exp(C \lambda^2 K^2 \|AX'\|_2^2), \quad \lambda \in \mathbb{R}.$$

We now replace X by Y but still conditioning on X' (we replace one at a time),

$$E_Y[\exp(\mu \langle Y, AX' \rangle)] = \exp(\mu^2 \|AX'\|_2^2 / 2), \quad \mu \in \mathbb{R}.$$

Choosing $\mu = \sqrt{2C\lambda}K$, we can match our estimates to get

$$\mathbb{E}_X[\exp(\lambda\langle X, AX' \rangle)] \leq \mathbb{E}_Y[\exp(\mu\langle Y, AX' \rangle)] = \exp(C\lambda^2 K^2 \lambda \|AX'\|_2^2).$$

We can now take the expectation with respect to X' on both sides and see that we have successfully replaced X by Y . We can now repeat the same procedure for X' and Y' to obtain our statement. \square

Proof of Theorem 6.4. Without loss of generality, $K = 1$. It suffices to show the one-sided tail estimate. Denote

$$p = \mathbb{P}(\langle X, AX \rangle - \mathbb{E}[\langle X, AX \rangle] \geq t).$$

Write

$$\langle X, AX \rangle - \mathbb{E}[\langle X, AX \rangle] = \sum_{i=1}^n a_{ii}(X_i^2 - \mathbb{E}[X_i^2]) + \sum_{i,j: i \neq j} a_{ij}X_iX_j,$$

and thus the problem reduces to estimating diagonal and off-diagonal sums:

$$p \leq \mathbb{P}\left(\sum_{i=1}^n a_{ii}(X_i^2 - \mathbb{E}[X_i^2]) \geq t/2\right) + \mathbb{P}\left(\sum_{i,j: i \neq j} a_{ij}X_iX_j \geq t/2\right) =: p_1 + p_2.$$

Diagonal sum: $X_i^2 - \mathbb{E}[X_i^2]$ are independent mean-zero sub-exponential random variables. Thus, using centering (see Exercise 3.2) and Lemma 2.31 we have

$$\|X_i^2 - \mathbb{E}[X_i^2]\|_{\psi_1} \leq C\|X_i^2\|_{\psi_1} \leq C\|X_i\|_{\psi_2}^2 \leq C.$$

Then Bernstein's inequality (see Exercise 3.3 using it with $1/N$ replaced by A_{ii}) implies that

$$p_1 \leq \exp\left(-C \min\left\{\left(\frac{t^2}{\sum_{i=1}^n a_{ii}^2}\right), \left(\frac{t}{\max_{1 \leq i \leq n}\{a_{ii}\}}\right)\right\}\right).$$

Off-diagonal sum: $S := \sum_{i \neq j} a_{ij}X_iX_j$ and $\lambda > 0$. Then

$$\begin{aligned} \mathbb{E}[\exp(\lambda S)] &\leq \mathbb{E}[\exp(4\lambda\langle X, AX' \rangle)] ; \text{Decoupling - Theorem 6.3} \\ &\leq \mathbb{E}[\exp(C_1\lambda\langle Y, AY' \rangle)] ; \text{Comparison - Lemma 6.6} \\ &\leq \exp(C\lambda^2\|A\|_F^2) \quad \text{provided } |\lambda| \leq \bar{C}/\|A\| ; \text{Gaussian chaos - Lemma 6.5.} \end{aligned}$$

Thus

$$p_2 \leq \exp(-\lambda t/2)\mathbb{E}[\exp(\lambda S)] \leq \exp(-\lambda t/2 + C\lambda t^2\|A\|_F^2).$$

Optimising over $0 \leq \lambda \leq \bar{C}/\|A\|$, we get a solution $\lambda = \frac{t}{4C\|A\|_F^2}$ as long as $t \leq \frac{\bar{C}4C\|A\|_F^2}{\|A\|}$.

Inserting this solution gives the exponent $-t^2/(16C\|A\|_F^2)$. For $t > \frac{\bar{C}4C\|A\|_F^2}{\|A\|}$ we have $\lambda = \frac{\bar{C}}{\|A\|}$ and obtain the other exponent which is linear in t . Thus we have some absolute constant $C > 0$ such that

$$p_2 \leq \exp\left(-C \min\left\{\left(\frac{t^2}{\|A\|_F^2}\right), \left(\frac{t}{\|A\|}\right)\right\}\right).$$

\square

6.2 Concentration for Anisotropic random vectors

We now study anisotropic random vectors, which have the form BX where B is a fixed matrix and where X is an isotropic random vector.

Exercise 6.7 Let B be an $m \times n$ matrix and X be an isotropic random vector in \mathbb{R}^b . Show that

$$\mathbb{E}[\|BX\|_2^2] = \|B\|_F^2.$$



Theorem 6.8 (Concentration for random vectors) Let B be an $m \times n$ matrix and $X = (X_1, \dots, X_n) \in \mathbb{R}^n$ a random vector with independent mean-zero unit variance sub-Gaussian coordinates. Then

$$\left| \|BX\|_2 - \|B\|_F \right|_{\psi_2} \leq CK^2 \|B\|, \quad K := \max_{1 \leq i \leq n} \{\|X_i\|_{\psi_2}\},$$

for some absolute constant $C > 0$.

Remark 6.9 (a) We have, according to Exercise 6.7,

$$\mathbb{E}[\|BX\|_2^2] = \|B\|_F^2.$$

(b) Recall that we have already proved a version with $B = \mathbb{1}_n$ in Theorem 3.1, namely,

$$\left| \|X\|_2 - \sqrt{n} \right|_{\psi_2} \leq CK^2.$$



Proof of Theorem 6.8. Without loss of generality we assume that $K \geq 1$ and we write $A = B^T B$ and apply Hanson-Wright inequality (Theorem 6.4) with the following.

$$\begin{aligned} \langle X, AX \rangle &= \|BX\|_2^2, \\ \mathbb{E}[\langle X, AX \rangle] &= \|B\|_F^2; \|A\| = \|B\|^2, \\ \|B^T B\|_F &\leq \|B^T\| \|B\|_F = \|B\| \|B\|_F. \end{aligned}$$

Thus, for every $u \geq 0$,

$$\mathbb{P}\left(\left| \|BX\|_2^2 - \|B\|_F^2 \right| \geq u\right) \leq 2 \exp\left(-\frac{C}{K^4} \min\left\{\left(\frac{u^2}{\|B\|^2 \|B\|_F^2}\right), \left(\frac{u}{\|B\|^2}\right)\right\}\right).$$

Setting $u = \varepsilon \|B\|_F^2$ with $\varepsilon \geq 0$, we obtain

$$\mathbb{P}\left(\left| \|BX\|_2^2 - \|B\|_F^2 \right| \geq \varepsilon \|B\|_F^2\right) \leq 2 \exp\left(-C \min\{\varepsilon^2, \varepsilon\} \frac{\|B\|_F^2}{K^4 \|B\|^2}\right).$$

We now set $\delta^2 = \min\{\varepsilon^2, \varepsilon\}$, or, equivalently $\varepsilon = \max\{\delta, \delta^2\}$, and we recall our reasoning in the proof of Theorem 3.1, namely, that for $z \geq 0$, the bound $|z - 1| \geq \delta$ implies $|z^2 - 1| \geq \max\{\delta, \delta^2\}$.

$$\left| \|BX\|_2 - \|B\|_F \right| \geq \delta \|B\|_F \Rightarrow \left| \|BX\|_2^2 - \|B\|_F^2 \right| \geq \varepsilon \|B\|_F^2.$$

Thus

$$\mathbb{P}\left(\left|\|BX\|_2 - \|B\|_F\right| \geq \delta\|B\|_F\right) \leq 2 \exp\left(-C\delta^2 \frac{\|B\|_F^2}{K^4\|B\|^2}\right),$$

and setting $t = \delta\|B\|_F$, we obtain the tail-estimate

$$\mathbb{P}\left(\left|\|BX\|_2 - \|B\|_F\right| \geq t\right) \leq 2 \exp\left(-\frac{Ct^2}{K^4\|B\|^2}\right).$$

□

6.3 Symmetrisation

Definition 6.10 (symmetric random variables) A real-valued random variable X is called *symmetric* if X and $-X$ have the same distribution.

Exercise 6.11 Let X be a real-valued random variable independent of some symmetric Bernoulli random variable ε , i.e., $\mathbb{P}(\varepsilon = -1) = \mathbb{P}(\varepsilon = 1) = \frac{1}{2}$. Show the following statements.

- (a) εX and $\varepsilon|X|$ are symmetric random variables and εX and $\varepsilon|X|$ have the same distribution.
- (b) X symmetric \Rightarrow distribution of εX and $\varepsilon|X|$ equal the distribution of X .
- (c) Suppose X' is an independent copy of X , then $X - X'$ is a symmetric random variable.

Lemma 6.12 (Symmetrisation) Let X_1, \dots, X_N be independent mean-zero random vectors in a normed space $(E, \|\cdot\|)$ and let $\varepsilon_1, \dots, \varepsilon_N$ be independent symmetric Bernoulli random variables (that is, they are not only independent of each other but also of any $X_i, i = 1, \dots, N$). Show that

$$\frac{1}{2}\mathbb{E}\left[\left\|\sum_{i=1}^N \varepsilon_i X_i\right\|\right] \leq \mathbb{E}\left[\left\|\sum_{i=1}^N X_i\right\|\right] \leq 2\mathbb{E}\left[\left\|\sum_{i=1}^N \varepsilon_i X_i\right\|\right].$$

Proof.

Upper bound: Let (X'_i) be an independent copy of the random vector (X_i) . Since $\sum_{i=1}^N X'_i$ has zero mean,

$$p := \mathbb{E}\left[\left\|\sum_{i=1}^N X_i\right\|\right] \leq \mathbb{E}\left[\left\|\sum_{i=1}^N X_i - \sum_{i=1}^N X'_i\right\|\right] = \mathbb{E}\left[\left\|\sum_{i=1}^N (X_i - X'_i)\right\|\right],$$

where the inequality stems from an application of Lemma 6.2), namely, if $\mathbb{E}[Z] = 0$ then $\mathbb{E}[\|Y\|] \leq \mathbb{E}[\|Y + Z\|]$.

Since all $(X_i - X'_i)$ are symmetric random vectors, they have the same distribution as $\varepsilon_i(X_i - X'_i)$, see Exercise 6.11. Application of the triangle inequality and our assumptions conclude the upper bound

$$p \leq \mathbb{E} \left[\left\| \sum_{i=1}^N \varepsilon_i (X_i - X'_i) \right\| \right] \leq \mathbb{E} \left[\left\| \sum_{i=1}^N \varepsilon_i X_i \right\| \right] + \mathbb{E} \left[\left\| \sum_{i=1}^N \varepsilon_i X'_i \right\| \right] = 2\mathbb{E} \left[\left\| \sum_{i=1}^N \varepsilon_i X_i \right\| \right]$$

Lower bound: By conditioning on ε_i and the triangle inequality,

$$\begin{aligned} \mathbb{E} \left[\left\| \sum_{i=1}^N \varepsilon_i X_i \right\| \right] &\leq \mathbb{E} \left[\left\| \sum_{i=1}^N \varepsilon_i (X_i - X'_i) \right\| \right] = \mathbb{E} \left[\left\| \sum_{i=1}^N (X_i - X'_i) \right\| \right] \\ &\leq \mathbb{E} \left[\left\| \sum_{i=1}^N X_i \right\| \right] + \mathbb{E} \left[\left\| \sum_{i=1}^N X'_i \right\| \right] = 2\mathbb{E} \left[\left\| \sum_{i=1}^N X_i \right\| \right]. \end{aligned}$$

□

7 Random Processes

7.1 Basic concepts and examples

Definition 7.1 A *random process* is a collection of random variables $X := (X_t)_{t \in T}$ of random variables X_t defined on the same probability space, which are indexed by the elements t of some set T .

Example 7.2 (a) $T = \{1, \dots, n\}$, then $X = (X_1, \dots, X_n) \in \mathbb{R}^n$ random vector.

(b) $T = \mathbb{N}$, then $X = (X_n)_{n \in \mathbb{N}}$ with

$$X_n = \sum_{i=1}^n Z_i, \quad \text{with } Z_i \text{ are } \mathbb{R} - \text{valued and independent identically distributed,}$$

is the discrete time random walk.

(c) When the dimension of the index set is greater than one, e.g., $T \subset \mathbb{R}^n$ or $T \subset \mathbb{Z}^n$, we often call the process a *random field* $(X_t)_{t \in T}$.

(d) The most well-known continuous-time process is the *standard Brownian motion* $X = (X_t)_{t \geq 0}$, also called the *Wiener process*. We can characterise it as follows:

- (i) The process has continuous paths, i.e., $X : [0, \infty) \rightarrow \mathbb{R}, t \mapsto X_t$ is almost surely continuous.
- (ii) The increments are independent and satisfy $X_t - X_s \sim N(0, t - s)$ for all $t \geq s$.



Definition 7.3 Let $(X_t)_{t \in T}$ be a random process with $\mathbb{E}[X_t] = 0$ for all $t \in T$. The *covariance function* of the process is defined as

$$\Sigma(t, s) := \text{cov}(X_t, X_s), \quad t, s \in T.$$

The *increments* of the process are defined as

$$d(t, s) := \|X_t - X_s\|_{L^2} = (\mathbb{E}[(X_t - X_s)^2])^{1/2}, \quad t, s \in T.$$

Note that $\Sigma(t, s) = \mathbb{E}[X_t X_s]$, $t, s \in T$, when the process $X = (X_t)_{t \in T}$ has zero mean, i.e., $\mathbb{E}[X_t] = 0$ for all $t \in T$.

Example 7.4 (Increments of random walks) The increments of the discrete time random in Example [(b)]7.2 with $\mathbb{E}[Z_i^2] = 1$, $i \in \mathbb{N}$, are $d(n, m) = \sqrt{n - m}$ for all $n, m \in \mathbb{N}_0$, $n \geq m$. To see that, we compute

$$\begin{aligned} d(n, m)^2 &= \|X_n - X_m\|_{L^2}^2 = \mathbb{E}[(X_n - X_m)^2] = \mathbb{E}\left[\left(\sum_{k=m+1}^n Z_k\right)^2\right] = \sum_{k, j=m+1}^n \mathbb{E}[Z_k Z_j] \\ &= \sum_{k=m+1}^n \mathbb{E}[Z_k^2] = n - m, \end{aligned}$$

where we used the fact that the Z_i 's are independent with zero mean, i.e., $\mathbb{E}[Z_k Z_j] = \delta_{k, j}$, and our assumption $\mathbb{E}[Z_k^2] = 1$ for all $k \in \mathbb{N}$. ♣

Definition 7.5 (Gaussian process) (a) A random process $(X_t)_{t \in T}$ is called a *Gaussian process* if, for any finite subset $T_0 \subset T$, the random vector $(X_t)_{t \in T_0}$ has normal distribution. Equivalently, $(X_t)_{t \in T}$ is called a *Gaussian process* if every finite linear combination

$$\sum_{t \in T_0} a_t X_t, \quad a_t \in \mathbb{R},$$

is a normally distributed random variable.

(b) Suppose $T \subset \mathbb{R}^n$ and let $Y \sim N(0, \mathbb{1}_n)$, and define

$$X_t := \langle Y, t \rangle, \quad t \in T \subset \mathbb{R}^n.$$

We call the random process $(X_t)_{t \in T}$ the *canonical Gaussian process in \mathbb{R}^n* .

To compute the increments of a canonical Gaussian process, recall that $\langle Y, t \rangle \sim N(0, \|t\|_2^2)$ for any $t \in \mathbb{R}^n$. Then one show that (easy exercise) the increments are

$$d(t, s) = \|X_t - X_s\|_{L^2} = \|t - s\|_2 \quad t, s \in \mathbb{R}^n,$$

where $\|\cdot\|_2$ denotes the Euclidean norm in \mathbb{R}^n .

Lemma 7.6 *Let $X \in \mathbb{R}^n$ be a mean-zero Gaussian random vector. Then there exist points $t_1, \dots, t_n \in \mathbb{R}^n$ such that*

$$X := (\langle Y, t_i \rangle)_{i=1, \dots, n} \in \mathbb{R}^n, \quad \text{with } Y \sim N(0, \mathbb{1}_n).$$

Proof. Let Σ be the covariance matrix of X . Then

$$X = \Sigma^{1/2} Y \quad \text{where } Y \sim N(0, \mathbb{1}_n), \quad (7.1)$$

which follows from $X \sim N(0, \Sigma)$ if and only if $\Sigma^{-1/2} X \sim N(0, \mathbb{1}_n)$. This in turn can be seen by direct computation (change of variables) by recalling the probability density function of X , i.e.,

$$f_X(x) = \frac{1}{(2\pi)^{n/2} \det(\Sigma)^{1/2}} \exp\left(-\frac{1}{2} \langle x, \Sigma^{-1} x \rangle\right), \quad x \in \mathbb{R}^n.$$

The coordinates of $\Sigma^{1/2} Y$ are $\langle s_i, Y \rangle$, where the s_i denote the rows of the matrix $\Sigma^{1/2}$. □

7.2 Slepian's inequality and Gaussian interpolation

In many applications a uniform control of a stochastic process $X = (X_t)_{t \in T}$ is useful, i.e., to have a bound on

$$\mathbb{E}[\sup_{t \in T} X_t].$$

For general processes, even if they are Gaussian, obtaining such bounds is highly nontrivial. In this section we learn first how the expectation of the supremum of two processes compare to each other. In Section 7.3 below we obtain a lower and upper bound for expected supremum of a process.

Theorem 7.7 (Slepian's inequality) *Let $(X_t)_{t \in T}$ and $(Y_t)_{t \in T}$ be two mean-zero Gaussian processes. Assume that, for all $t, s \in T$, we have*

$$(i) \mathbb{E}[X_t^2] = \mathbb{E}[Y_t^2] \quad (ii) \mathbb{E}[(X_t - X_s)^2] \leq \mathbb{E}[(Y_t - Y_s)^2]. \quad (7.2)$$

Then, for every $\tau \in \mathbb{R}$, we have

$$\mathbb{P}\left(\sup_{t \in T} \{X_t\} \geq \tau\right) \leq \mathbb{P}\left(\sup_{t \in T} \{Y_t\} \geq \tau\right), \quad (7.3)$$

and, consequently,

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

Whenever (7.3) holds, we say that the process $(X_t)_{t \in T}$ is *stochastically dominated* by the process $(Y_t)_{t \in T}$. The proof of Theorem 7.7 follows by combining the two versions of Slepian's inequality which we will discuss below. To prepare these statements we introduce the method of *Gaussian interpolation* first.

Suppose T is finite, e.g., $|T| = n$. Let $X = (X_t)_{t \in T}$ and $Y = (Y_t)_{t \in T}$ be two Gaussian random vectors (without loss of generality we may assume that they X and Y are independent). We define the Gaussian interpolation as the following random vector in \mathbb{R}^n ,

$$Z(u) := \sqrt{u}X + \sqrt{1-u}Y, \quad u \in [0, 1]. \quad (7.4)$$

It is easy to see (following exercise) that the covariance matrix of the interpolation interpolates linearly between the covariance matrices of X and Y .

Exercise 7.8 Show that

$$\Sigma(Z(u)) = u\Sigma(X) + (1-u)\Sigma(Y) \quad t \in [0, 1].$$



For a function $f: \mathbb{R}^n \rightarrow \mathbb{R}$, we shall study how the expectation $\mathbb{E}[f(Z(u))]$ varies with $u \in [0, 1]$. Suppose, for example, that $f(x) := \mathbb{1}\{\max_{1 \leq i \leq n} \{x_i\} \leq u\}$, $x \in \mathbb{R}^n$. Then one can easily show that $\mathbb{E}[f(Z(u))]$ increases with u leading to

$$\mathbb{E}[f(Z(1))] \geq \mathbb{E}[f(Z(0))],$$

and henceforth

$$\mathbb{P}\left(\max_{1 \leq i \leq n} \{X_i\} < \tau\right) \geq \mathbb{P}\left(\max_{1 \leq i \leq n} \{Y_i\} < \tau\right),$$

leading to

$$\mathbb{P}\left(\sup_{t \in T} \{X_t\} \geq \tau\right) \leq \mathbb{P}\left(\sup_{t \in T} \{Y_t\} \geq \tau\right). \quad (7.5)$$

The following three lemmas collect basic facts about Gaussian random variables and their proofs will be left as an exercise (homework).

Lemma 7.9 (Gaussian integration by parts) Suppose $X \sim N(0, 1)$. Then, for any differentiable function $f: \mathbb{R} \rightarrow \mathbb{R}$,

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

Proof. Exercise for the reader. Solution see Lemma 7.2.3 in [Ver18]. □

Similarly, $X \sim N(0, \sigma^2)$, $\sigma > 0$ implies $\mathbb{E}[Xf(X)] = \sigma^2 \mathbb{E}[f'(X)]$.

Lemma 7.10 (Gaussian integration by parts) Suppose $X \sim N(0, \Sigma)$, $X \in \mathbb{R}^n$, Σ an $n \times n$ symmetric positive semi-definite matrix. Then, for any differentiable function $f: \mathbb{R}^n \rightarrow \mathbb{R}$,

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

Proof. Exercise for the reader (homework - Example Sheet 4). □

Lemma 7.11 (Gaussian interpolation) Consider two independent Gaussian random vectors $X \sim \mathcal{N}(0, \Sigma^X)$ and $Y \sim \mathcal{N}(0, \Sigma^Y)$ in \mathbb{R}^n . Define the interpolation Gaussian random vector as

$$Z(u) := \sqrt{u} X + \sqrt{1-u} Y, \quad u \in [0, 1]. \quad (7.6)$$

Then for any twice-differentiable function $f: \mathbb{R}^n \rightarrow \mathbb{R}$, we have

$$\frac{d}{du} \mathbb{E}[f(Z(u))] = \frac{1}{2} \sum_{i,j=1}^n (\Sigma_{ij}^X - \Sigma_{ij}^Y) \mathbb{E} \left[\frac{\partial^2 f}{\partial x_i \partial x_j} (Z(u)) \right]. \quad (7.7)$$

Proof. Exercise for the reader (homework - Example Sheet 4). Solution see Lemma 7.2.7 in [Ver18]. □

We now prove a first version of Slepian's inequality (Theorem 7.7).

Lemma 7.12 (Slepian's inequality, functional form) Let $X, Y \in \mathbb{R}^n$ be two mean-zero Gaussian random vectors. Assume that for all $i, j = 1, \dots, n$,

$$(i) \mathbb{E}[X_i^2] = \mathbb{E}[Y_i^2], \quad (ii) \mathbb{E}[(X_i - X_j)^2] \leq \mathbb{E}[(Y_i - Y_j)^2], \quad (7.8)$$

and let $f: \mathbb{R}^n \rightarrow \mathbb{R}$ be twice-differentiable such that

$$\frac{\partial^2 f}{\partial x_i \partial x_j} (x) \geq 0 \quad \text{for all } i \neq j.$$

Then

$$\mathbb{E}[f(X)] \geq \mathbb{E}[f(Y)].$$

Proof. We have, using (7.8) and our assumptions, that $\Sigma_{ii}^X = \Sigma_{ii}^Y$ and $\mathbb{E}[X_i X_j] \geq \mathbb{E}[Y_i Y_j]$. Thus $\Sigma_{ij}^X \geq \Sigma_{ij}^Y$. We assume without loss of generality that X and Y are independent. Then Lemma 7.11 shows that

$$\frac{d}{du} \mathbb{E}[f(Z(u))] \geq 0,$$

and hence that $\mathbb{E}[f(Z(u))]$ increases in u . Thus $\mathbb{E}[f(Z(1))] = \mathbb{E}[f(X)] \geq \mathbb{E}[f(Y)] = \mathbb{E}[f(Z(0))]$. □

We can now prove Theorem 7.7 with the following results for Gaussian vectors.

Theorem 7.13 Let X and Y be two mean-zero Gaussian random vectors in \mathbb{R}^n as in Lemma 7.12. Then, for every $\tau \in \mathbb{R}$, we have

$$\mathbb{P} \left(\max_{1 \leq i \leq n} \{X_i\} \geq \tau \right) \geq \mathbb{P} \left(\max_{1 \leq i \leq n} \{Y_i\} \geq \tau \right).$$

Consequently,

$$\mathbb{E} \left[\max_{1 \leq i \leq n} \{X_i\} \right] \leq \mathbb{E} \left[\max_{1 \leq i \leq n} \{Y_i\} \right].$$

Proof. The key idea is to use Lemma 7.12 for some appropriate approximation of the maximum. For that to work, let $h: \mathbb{R} \rightarrow [0, 1]$ be a twice-differentiable approximation of the indicator function of the interval $(-\infty, \tau)$, that is, $h(t) \approx \mathbb{1}_{(-\infty, \tau)}(t)$, $t \in \mathbb{R}$, and $h'(t) \leq 0$ (h is smooth non increasing function). Define $f(x) := h(x_1) \cdots h(x_n)$ for $x = (x_1, \dots, x_n) \in \mathbb{R}^n$. The second partial derivatives

$$\frac{\partial^2 f}{\partial x_i \partial x_j}(x) = h'(x_i)h'(x_j) \prod_{k \neq i, j} h(x_k)$$

are positive. It follows that $\mathbb{E}[f(X)] = \mathbb{E}[f(Z(1))] \geq \mathbb{E}[f(Y)] = \mathbb{E}[f(Z(0))]$. Thus, by the above approximations, according to Lemma 7.12,

$$\mathbb{P}\left(\max_{1 \leq i \leq n} \{X_i\} < \tau\right) \geq \mathbb{P}\left(\max_{1 \leq i \leq n} \{Y_i\} < \tau\right),$$

and thus, using the integral identity 1.9,

$$\mathbb{E}\left[\max_{1 \leq i \leq n} \{X_i\}\right] \leq \mathbb{E}\left[\max_{1 \leq i \leq n} \{Y_i\}\right],$$

and the statement. \square

The following theorem improves Slepian's inequality by using a different approximation in the proof.

Theorem 7.14 (Sudakov-Fernique inequality) *Let $(X_t)_{t \in T}$ and $(Y_t)_{t \in T}$ be two mean-zero Gaussian processes. Assume that, for all $t, s \in T$, we have*

$$(i) \quad \mathbb{E}[(X_t - X_s)^2] \leq \mathbb{E}[(Y_t - Y_s)^2].$$

Then

$$\mathbb{E}\left[\sup_{t \in T} \{X_t\}\right] \leq \mathbb{E}\left[\sup_{t \in T} \{Y_t\}\right].$$

Proof. We shall prove the statement for Gaussian random vectors $X, Y \in \mathbb{R}^n$ with the help of Theorem 7.13. The idea this time is to use an approximation of the maximum itself and not for the indicator function. For $\beta > 0$, define

$$f(x) := \frac{1}{\beta} \log \sum_{i=1}^n e^{\beta x_i} \quad x \in \mathbb{R}^n.$$

One can easily show that

$$f(x) \xrightarrow{\beta \rightarrow \infty} \max_{1 \leq i \leq n} \{x_i\}.$$

Inserting the function into the Gaussian interpolation formula (7.7) we can obtain after some tedious calculation that

$$\frac{d}{du} \mathbb{E}[f(Z(u))] \leq 0,$$

and conclude with our statement. \square

Exercise 7.15 For $\beta > 0$, define

$$f(x) := \frac{1}{\beta} \log \sum_{i=1}^n e^{\beta x_i} \quad x \in \mathbb{R}^n.$$

Show that

$$f(x) \xrightarrow{\beta \rightarrow \infty} \max_{1 \leq i \leq n} \{x_i\}.$$



Solution. Without loss of generality let $\exp(\beta x_k) = \max_{1 \leq i \leq n} \{\exp(\beta x_i)\}$, Then

$$\sum_{i=1}^n e^{\beta x_i} = e^{\beta x_k} \left(1 + \sum_{i \neq k} \exp(\beta(x_i - x_k)) \right),$$

and observing that

$$\sum_{i \neq k} \exp(\beta(x_i - x_k)) \leq n$$

we obtain that

$$0 \leq \lim_{\beta \rightarrow \infty} \frac{1}{\beta} \log \left(1 + \sum_{i \neq k} \exp(\beta(x_i - x_k)) \right) \leq \lim_{\beta \rightarrow \infty} \frac{1}{\beta} \log(1 + n) = 0,$$

and thus the statement, ☺

7.3 The supremum of a process

We now combine features of the index space with the random process and obtain bounds for $\mathbb{E}[\sup_{t \in T} X_t]$.

Definition 7.16 (Canonical metric) Suppose $X = (X_t)_{t \in T}$ is a random process with index set T . We define the *canonical metric of the process* by

$$d(t, s) := \|X_t - X_s\|_{L^2} = \left(\mathbb{E} \left[\left(X_t - X_s \right)^2 \right] \right)^{1/2} \quad t, s \in T.$$

We can now study the question if we can evaluate the expectation $\mathbb{E}[\sup_{t \in T} \{X_t\}]$ by using properties of the geometry, in particular using covering numbers of the index metric space (T, d) equipped with the canonical metric of the process. Via this approach we obtain a lower bound of the expectation of the supremum.

Theorem 7.17 (Sudakov's minorisation inequality) Let $X = (X_t)_{t \in T}$ be a mean-zero Gaussian process. Then, for any $\varepsilon \geq 0$, we have

$$\mathbb{E}[\sup_{t \in T} \{X_t\}] \geq C\varepsilon \sqrt{\log \mathcal{N}(T, d, \varepsilon)},$$

where d is the canonical metric of the process, $C > 0$ an absolute constant and $\mathcal{N}(T, d, \varepsilon)$ the covering number for T (recall Definition 4.1).

Proof. Assume that $\mathcal{N}(T, d, \varepsilon) = N < \infty$ is finite. When T is not compact, one can show that the expectation of the supremum is infinite (we skip this step). Let \mathcal{M} be a maximal ε -separated subset of T . Then \mathcal{M} is an ε -net according to Lemma 4.2, and thus $|\mathcal{M}| \geq N$. It suffices to show

$$\mathbb{E}[\sup_{t \in \mathcal{M}} \{X_t\}] \geq C\varepsilon\sqrt{\log N}. \quad (7.9)$$

To show (7.9), we compare the process $(X_t)_{t \in \mathcal{M}}$ with a simpler process $(Y_t)_{t \in \mathcal{M}}$.

$$Y_t := \frac{\varepsilon}{\sqrt{2}}G_t, \quad G_t \sim \mathcal{N}(0, 1), G_t \text{ independent of } G_s \text{ for all } t \neq s.$$

For $t, s \in \mathcal{M}$ fixed we have

$$\mathbb{E}[(X_t - X_s)^2] = d(t, s)^2 \geq \varepsilon^2,$$

while

$$\mathbb{E}[(Y_t - Y_s)^2] = \frac{\varepsilon^2}{2}\mathbb{E}[(G_t - G_s)^2] = \varepsilon^2.$$

Thus

$$\mathbb{E}[(X_t - X_s)^2] \geq \mathbb{E}[(Y_t - Y_s)^2] \quad \text{for all } t, s \in \mathcal{M}.$$

We now obtain with Theorem 7.14,

$$\mathbb{E}[\sup_{t \in \mathcal{M}} \{X_t\}] \geq \mathbb{E}[\sup_{t \in \mathcal{M}} \{Y_t\}] = \frac{\varepsilon}{\sqrt{2}}\mathbb{E}[\max_{i \in \mathcal{M}} \{G_i\}] \geq C\varepsilon\sqrt{\log N},$$

where we used Proposition 7.18 below. □

Proposition 7.18 (Maximum of normally distributed random variables) *Let $Y_i \sim \mathcal{N}(0, 1)$, $i = 1, \dots, N$, be independent normally distributed real-valued random variables. Then the following holds.*

(a)

$$\mathbb{E}\left[\max_{1 \leq i \leq N} \{Y_i\}\right] \leq \sqrt{2 \log N}.$$

(b)

$$\mathbb{E}\left[\max_{1 \leq i \leq N} \{|Y_i|\}\right] \leq \sqrt{2 \log(2N)}.$$

(c)

$$\mathbb{E}\left[\max_{1 \leq i \leq N} \{Y_i\}\right] \geq C\sqrt{2 \log N} \quad \text{for some absolute constant } C > 0.$$

Proof. (a) Let $\beta > 0$. Using Jensen's inequality, we obtain

$$\begin{aligned} \mathbb{E}[\max_{1 \leq i \leq n} Y_i] &= \frac{1}{\beta} \mathbb{E}[\log e^{\beta \max_{1 \leq i \leq n} \{Y_i\}}] \leq \frac{1}{\beta} \mathbb{E}[\log \sum_{i=1}^N e^{\beta Y_i}] \leq \frac{1}{\beta} \log \sum_{i=1}^N \mathbb{E}[e^{\beta Y_i}] \\ &= \frac{1}{\beta} \log \left(N e^{\beta^2/2} \right) = \frac{\beta}{2} + \frac{\log N}{\beta}. \end{aligned}$$

The claim is proved by taking $\beta = \sqrt{2 \log N}$.

(b) We repeat the steps in (a),

$$\mathbb{E}[\max_{1 \leq i \leq n} \{|Y_i|\}] = \dots \leq \frac{1}{\beta} \log \sum_{i=1}^N \mathbb{E}[e^{\beta Y_i} + e^{-\beta Y_i}] = \frac{\beta}{2} + \frac{\log(2N)}{\beta}.$$

Taking $\beta = \sqrt{2 \log(2N)}$, we conclude with the statement.

(c) The lower bound is slightly more involved and needs some preparation:

(i) The Y_i 's are symmetric (Gaussian), and hence, by symmetry,

$$\mathbb{E}[\max_{1 \leq i, j \leq N} \{|Y_i - Y_j|\}] = \mathbb{E}[\max_{1 \leq i, j \leq N} \{(Y_i - Y_j)\}] = 2\mathbb{E}[\max_{1 \leq i \leq N} \{Y_i\}].$$

(ii) For every $k \in \{1, \dots, N\}$, using (i),

$$\begin{aligned} \mathbb{E}[\max_{1 \leq i \leq N} \{Y_i\}] &\leq \mathbb{E}[\max_{1 \leq i \leq N} \{|Y_i|\}] \leq \mathbb{E}[|Y_k|] + \mathbb{E}[\max_{1 \leq i, j \leq N} \{|Y_i - Y_j|\}] \\ &= \mathbb{E}[|Y_k|] + 2\mathbb{E}[\max_{1 \leq i \leq N} \{Y_i\}]. \end{aligned}$$

To obtain a lower bound we exploit the fact that our Gaussian random variables Y_i are independent and identically distributed. Then, for every $\delta > 0$, noting that $1 - (1 - \mathbb{P}(|Y_1| > t))^N$ is the probability that one of the random variables $|Y| = i, i = 1, \dots, N$, is larger than t ,

$$\mathbb{E}[\max_{1 \leq i \leq N} \{|Y_i|\}] \geq \int_0^\delta [1 - (1 - \mathbb{P}(|Y_1| > t))^N] dt \geq \delta [1 - (1 - \mathbb{P}(|Y_1| > \delta))^N],$$

where the first inequality follows from the integral identity for the expectation and the second inequality is just the interval length times the minimal value of the integrand. Now, we obtain with the lower tail bound of the normal distribution (see Proposition 2.1),

$$\mathbb{P}(|Y_1| > \delta) = \sqrt{\frac{2}{\pi}} \int_\delta^\infty \exp(-t^2/2) dt \geq \frac{1}{\pi} \left(\frac{1}{\delta} - \frac{1}{\delta^3}\right) \exp(-\delta^2/2).$$

Now we choose $\delta = \sqrt{\log N}$ with N large enough so that

$$\mathbb{P}(|Y_1| > \delta) \geq \frac{1}{N},$$

and hence

$$\mathbb{E}[\max_{1 \leq i \leq N} \{Y_i\}] \geq \delta [1 - (1 - \frac{1}{N})^N] \geq \delta (1 - \frac{1}{e}).$$

We conclude with statement (c) using

$$\mathbb{E}[\max_{1 \leq i \leq N} \{|Y_i|\}] \leq \mathbb{E}[|Y_1|] + 2\mathbb{E}[\max_{1 \leq i \leq N} \{Y_i\}].$$

□

We have seen that the expected supremum of the canonical Gaussian process on some set $T \subset \mathbb{R}^n$,

$$\mathbb{E}[\sup_{t \in T} \langle Y, t \rangle],$$

where $Y \sim N(0, \mathbb{I}_n)$ plays an important role. In many application this geometric quantity is an important parameter. This leads to the following definition.

Definition 7.19 The *Gaussian width* of a subset $T \subset \mathbb{R}^n$ is defined

$$W(T) := \mathbb{E} \left[\sup_{x \in T} \langle Y, x \rangle \right] \quad \text{with } Y \sim N(0, \mathbb{I}_n).$$

Exercise 7.20 Suppose $X_i, i = 1, \dots, N$, are sub-Gaussian random variables, and $K = \max_{1 \leq i \leq N} \|X_i\|_{\psi_2}$. Show that, for any $N \geq 2$,

$$\mathbb{E} \left[\max_{1 \leq i \leq N} \{|X_i|\} \right] \leq CK \sqrt{\log N}.$$



We summarise a few simple properties of the Gaussian width. We skip the proof as it involves mostly elementary properties of the norm and the Minkowski sum. Recall that the diameter of a set $T \subset \mathbb{R}^n$ with respect to the Euclidean norm is

$$\text{diam}(T) = \sup_{x, y \in T} \|x - y\|_2.$$

Proposition 7.21 (Properties of Gaussian width) (a) $W(T)$ is finite $\Leftrightarrow T$ is bounded.

(b) $W(T) = W(UT)$ for any orthogonal $n \times n$ matrix U .

(c) $W(T + S) = W(T) + W(S)$ $S, T \subset \mathbb{R}^n$ and $W(aT) = |a|W(T)$ for every $a \in \mathbb{R}$.

(d)

$$W(T) = \frac{1}{2}W(T - T) = \frac{1}{2}\mathbb{E} \left[\sup_{x, y \in T} \langle Y, x - y \rangle \right].$$

(e)

$$\frac{1}{\sqrt{2\pi}} \text{diam}(T) \leq W(T) \leq \frac{\sqrt{n}}{2} \text{diam}(T).$$

Proof.

(a) Cauchy-Schwarz inequality gives

$$|\langle Y, x \rangle| \leq \|Y\|_2 \|x\|_2 \quad \text{for all } x \in T.$$

If $W(T) < \infty$, then this implies that $\|x\|_2 < \infty$ for all $x \in T$, and henceforth T is bounded. Conversely, if T is a bounded set we have that $\|x\|_2 \leq C$ for all $x \in T$ and some $C > 0$. Thus

$$\mathbb{E}[\langle Y, x \rangle] \leq \mathbb{E}[\|Y\|_2] C \leq \sqrt{n} C < \infty$$

implies that $W(T) < \infty$.

(b) We simply use the rotation invariance of the normal distribution, i.e., $Y \sim N(0, \mathbb{1}_n)$ implies that $UY \sim N(0, \mathbb{1}_n)$ for any orthogonal matrix U .

(c) Recalling the definition of the Minkowski sum of two sets we get

$$W(T + S) = \mathbb{E}[\sup_{x \in T, y \in S} \langle Y, x + y \rangle] = \mathbb{E}[\sup_{x \in T} \langle Y, x \rangle] + \mathbb{E}[\sup_{y \in S} \langle Y, y \rangle] = W(T) + W(S).$$

If $a \geq 0$ we have $a = |a|$ and $\langle Y, ax \rangle = |a| \langle Y, x \rangle$. If $a < 0$ we have $|a| = -a$ and using the fact that Y is symmetric, i.e., $-Y$ has same distribution as Y , we get

$$|a| \langle Y, x \rangle = -a \langle Y, x \rangle = \langle -Y, ax \rangle \sim \langle Y, ax \rangle,$$

and thus the statement.

(d) Using (c) we get

$$W(T) = \frac{1}{2}(W(T) + W(T)) = \frac{1}{2}(W(T) + W(-T)) = \frac{1}{2}W(T - T).$$

(e) Fix a pair $x, y \in T$. Then by definition $x - y, y - x \in T - T$. With part (d) we get a lower bound

$$W(T) \geq \frac{1}{2} \mathbb{E}[\max\{\langle Y, x - y \rangle, \langle Y, y - x \rangle\}] = \frac{1}{2} \mathbb{E}[|\langle Y, x - y \rangle|] = \frac{1}{2} \sqrt{\frac{2}{\pi}} \|x - y\|_2,$$

where the inequality follows from taking one pair $x, y \in T$ and the first equality follows from $\max\{a, -a\} = |a|$ for $a \in \mathbb{R}$. The second equality can be seen as follows. Recall that $\langle Y, x - y \rangle \sim N(0, \|x - y\|_2^2)$ and therefore

$$\langle Y, \frac{x - y}{\|x - y\|_2} \rangle \sim N(0, 1), \quad \text{and} \quad \mathbb{E}[|\langle Y, \frac{x - y}{\|x - y\|_2} \rangle|] = \sqrt{\frac{2}{\pi}}$$

follows from the calculation for any $X \sim N(0, 1)$,

$$\mathbb{E}[|X|] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} |x| e^{-x^2/2} dx = \frac{2}{\sqrt{2\pi}} \int_0^{\infty} x e^{-x^2/2} dx = \frac{2}{\sqrt{2\pi}} [-e^{-x^2/2}]_0^{\infty} = \sqrt{\frac{2}{\pi}}.$$

Taking now the supremum of all pairs $x, y \in T$ we obtain the lower bound. For the corresponding upper bound we use (d) and Cauchy-Schwarz inequality again.

$$\begin{aligned} W(T) &= \frac{1}{2} \mathbb{E} \left[\sup_{x, y \in T} \langle Y, x - y \rangle \right] \leq \frac{1}{2} \mathbb{E} \left[\sup_{x, y \in T} \|Y\|_2 \|x - y\|_2 \right] \leq \frac{1}{2} \mathbb{E} [\|Y\|_2] \text{diam}(T) \\ &\leq \frac{1}{2} \sqrt{n} \text{diam}(T), \end{aligned}$$

where we used $\mathbb{E}[\|Y\|_2] \leq \sqrt{n}$ which follows from $\mathbb{E}[\|Y\|_2] \leq (\mathbb{E}[\|Y\|_2^2])^{1/2} = \sqrt{n}$ and $\mathbb{E}[\|Y\|_2^2] = n$. \square

We discuss a few examples to obtain some understanding of the Gaussian width.

Example 7.22 (Gaussian width) (a) The Gaussian width of the unit sphere in n dimensions is

$$W(S^{(n-1)}) = \mathbb{E}[\|Y\|_2] = \sqrt{n} \pm C,$$

where the second equality follows from Exercise 6 - Example sheet 2 with some absolute constant $C > 0$ as an immediate consequence of our concentration of norm result in Theorem 3.1. To see the first equality, apply first Cauchy-Schwarz inequality to obtain an upper bound for any $x \in S^{(n-1)}$,

$$\mathbb{E}[\langle Y, x \rangle] \leq \mathbb{E}[\|Y\|_2].$$

Pick

$$x_i = \frac{Y_i}{\|Y\|_2}, \quad i = 1, \dots, N,$$

then $x = (x_1, \dots, x_n) \in S^{(n-1)}$.

(b) The Gaussian width for the cube $B_\infty^n = [-1, 1]^n$ with respect to the ℓ_∞ norm is

$$W(B_\infty^n) = \mathbb{E}[\|Y\|_1] = n\mathbb{E}[|Y_1|] = n\sqrt{\frac{2}{\pi}},$$

where the second equality follows from the isotropy of Y and the definition of the norm $\|Y\|_1 = \sum_{i=1}^n |Y_i|$. The third equality is just calculation, i.e.,

$$\mathbb{E}[|Y_1|] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} |y| e^{-y^2/2} dy = \frac{2}{\sqrt{2\pi}} \int_0^{\infty} y e^{-y^2/2} dy = \frac{2}{\sqrt{2\pi}} [-e^{-y^2/2}]_0^{\infty} = \sqrt{\frac{2}{\pi}}.$$

The first equality follows from Cauchy-Schwarz for an upper bound, i.e.,

$$\mathbb{E}[\langle Y, x \rangle] \leq \mathbb{E}[\|Y\|_1 \|x\|_\infty] = \mathbb{E}[\|Y\|_1],$$

and setting $x = (\text{sign}(Y_1), \dots, \text{sign}(Y_n)) \in B_\infty^n$ we obtain a lower bound and thus the equality.

(c) The unit ball B_1^n in \mathbb{R}^n with respect to the ℓ_1 norm is $B_1^n = \{x \in \mathbb{R}^n : \|x\|_1 \leq 1\}$, and its Gaussian width is

$$W(B_1^n) = \mathbb{E}[\|Y\|_\infty] = \mathbb{E}[\max_{1 \leq i \leq n} |Y_i|],$$

where the second equality is just the definition of the supremum norm and the first equality follows from Cauchy-Schwarz inequality. With Proposition 7.18 we get two absolute constants $c, C > 0$ such that

$$c\sqrt{\log n} \leq W(B_1^n) \leq C\sqrt{\log n}.$$



We finally present an upper bound for the expected supremum of a process in Theorem 7.24 below. The proof of that statement uses multi-scale approach in conjunction with the ε -net arguments, i.e., varying the threshold ε in a systematic and controlled way. This technique is called *chaining* and is a widely used tool in data science and high-dimensional probability. The whole proof goes beyond what we can do in this third year course, and we therefore only present the statement. The statement and its proof use the notion of Sub-Gaussian increments which we define first.

Definition 7.23 (Sub-Gaussian increments) Let $X = (X_t)_{t \in T}$ be a stochastic process on some metric space (T, d) . We say that the process X has *Sub-Gaussian increments* if there exists $K \geq 0$ such that

$$\|X_t - X_s\|_{\psi_2} \leq K d(t, s), \quad \text{for all } t, s \in T.$$

Theorem 7.24 (Dudley's inequality) Let $X = (X_t)_{t \in T}$ be a mean-zero stochastic process on a metric space (T, d) with Sub-Gaussian increments. Then

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

for some absolute constant $C > 0$.

Proof. The interested reader may check Chapter 8 in [Ver18]. □

7.4 Uniform law of large numbers

Definition 7.25 Let $(\Omega, \mathcal{B}(\Omega), \mu)$ be a probability space and denote $\mathcal{F} = \{f: \Omega \rightarrow \mathbb{R}\}$ a class of real-valued functions. Let X be a Ω -valued random variable with law $\mu \in \mathcal{M}_1(\Omega)$ and X_1, \dots, X_N be independent copies of X . The random process $X = (X_f)_{f \in \mathcal{F}}$ defined by

$$X_f := \frac{1}{N} \sum_{i=1}^N f(X_i) - \mathbb{E}[f(X)] \tag{7.10}$$

is called an *empirical process indexed by \mathcal{F}* .

Theorem 7.26 (Uniform law of large numbers) *Denote*

$$\mathcal{F} = \{f: [0, 1] \rightarrow \mathbb{R}: \|f\|_{\text{Lip}} \leq L\}$$

the class of Lipschitz function on $[0, 1]$, where $L > 0$ is a fixed number. Let X, X_1, \dots, X_N be independent identically distributed $[0, 1]$ -valued random variables. Then

$$\mathbb{E} \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{N} \sum_{i=1}^N f(X_i) - \mathbb{E}[f(X)] \right| \right] \leq \frac{CL}{\sqrt{N}} \quad (7.11)$$

for some absolute constant $C > 0$.

The proof of the theorem needs some bounds on the covering number of the class of Lipschitz function. We explore this in the next exercise before proving the theorem.

Exercise 7.27 Let $\mathcal{F} = \{f: [0, 1] \rightarrow [0, 1]: \|f\|_{\text{Lip}} \leq 1\}$. Show that

$$\mathcal{N}(\mathcal{F}, \|\cdot\|_{\infty}, \varepsilon) \leq (2/\varepsilon)^{2/\varepsilon} \quad \text{for any } \varepsilon > 0.$$



Solution. Recall the definition of the supremum norm $\|f\|_{\infty} = \sup_{x \in [0, 1]} |f(x)|$ and consider the square $\Lambda = [0, 1]^2$. We put a mesh of step (size) ε on Λ such that we get $(1/\varepsilon)^2$ squares of side length $1/\varepsilon$. Mesh-following functions f_0 are steps functions on the mesh taking one of $1/\varepsilon$ possible values on each interval of length $1/\varepsilon$. For every $f \in \mathcal{F}$ there is a mesh-following function f_0 such that $\|f - f_0\|_{\infty} \leq \varepsilon$. The number of all mesh-following functions is bounded by $(1/\varepsilon)^{1/\varepsilon}$. Recall that the covering number $\mathcal{N}(K, d, \varepsilon)$ is the smallest cardinality of closed ε -balls with centre in K whose union covers the set K . If we relax the assumption that the centres are in K we obtain the external covering number $\mathcal{N}^{\text{ext}}(K, d, \varepsilon)$. As we have done earlier, one can show that

$$\mathcal{N}(K, d, \varepsilon) \leq \mathcal{N}^{\text{ext}}(K, d, \varepsilon/2),$$

which we need for our case as the centres of our balls might not be element of \mathcal{F} . Thus

$$\mathcal{N}(\mathcal{F}, \|\cdot\|_{\infty}, \varepsilon) \leq (2/\varepsilon)^{2/\varepsilon}.$$



Proof of Theorem 7.26. Without loss of generality we put $L = 1$.

Step 1: We show that the empirical process has Sub-Gaussian increments. Fix a pair $f, g \in \mathcal{F}$ and consider

$$\|X_f - X_g\|_{\psi_2} = \frac{1}{N} \left\| \sum_{i=1}^N Z_i \right\|_{\psi_2},$$

where

$$Z_i := (f - g)(X_i) - \mathbb{E}[(f - g)(X)], \quad i = 1, \dots, N.$$

In the following we write \lesssim whenever we have \leq with some absolute constant $C > 0$ to avoid adapting the absolute constant in each single step. The Z_i 's are independent and mean-zero random numbers and thus Proposition 2.24 gives

$$\|X_f - X_g\|_{\psi_2} \lesssim \frac{1}{N} \left(\sum_{i=1}^N \|Z_i\|_{\psi_2}^2 \right)^{1/2}.$$

Using the centering Lemma 2.27 we have

$$\|Z_i\|_{\psi_2} \lesssim \|(f - g)(X_i)\|_{\psi_2} \lesssim \|f - g\|_{\infty},$$

and therefore

$$\|X_f - X_g\|_{\psi_2} \lesssim \frac{1}{N} N^{1/2} \|f - g\|_{\infty} = \frac{1}{\sqrt{N}} \|f - g\|_{\infty}.$$

Step 2: Application of Dudley's inequality. According to Step 1 the empirical process has Sub-Gaussian increments. The diameter of \mathcal{F} is one, that is,

$$\text{diam}(\mathcal{F}) = \sup_{f, g \in \mathcal{F}} \|f - g\|_{\infty} = \sup_{f, g \in \mathcal{F}} \sup_{x \in [0, 1]} |f(x) - g(x)| = 1.$$

Application of Theorem 7.24 gives the upper bound for the expected supremum where the integral runs between zero and the diameter. For the integral we use the bound on the covering number in Exercise 7.27. Thus we get

$$\mathbb{E}[\sup_{f \in \mathcal{F}} |X_f|] \lesssim \frac{1}{\sqrt{N}} \int_0^1 \sqrt{\log \mathcal{N}(\mathcal{F}, \|\cdot\|_{\infty}, \varepsilon)} \, d\varepsilon \lesssim \frac{1}{\sqrt{N}} \int_0^1 \sqrt{\frac{c}{\varepsilon} \log \frac{c}{\varepsilon}} \, d\varepsilon \lesssim \frac{1}{\sqrt{N}},$$

for some absolute constant $c > 0$. □

We finally introduce the VC dimension, which plays a major role in statistical learning theory as we see in Section 8 below. One can relate the VC dimension to covering numbers and via Dudley's inequality a uniform laws of large numbers involving the VC dimension follows. We can only give the definition and state the corresponding uniform law of large numbers without proof. The *Vapnik-Chervonenkis* (VC) dimension is a difficult concept and takes time to comprehend, roughly speaking, it is a measure of complexity for classes of Boolean functions $f: \Omega \rightarrow \{0, 1\}$ on some common domain Ω .

Definition 7.28 Suppose \mathcal{F} is a class of Boolean functions on some domain Ω . A subset $\Lambda \subset \Omega$ is *shattered* by \mathcal{F} if any Boolean function $g: \Lambda \rightarrow \{0, 1\}$ can be obtained by restricting some function $f \in \mathcal{F}$ on Λ . The *VC dimension of \mathcal{F}* , denoted $\text{VC}(\mathcal{F})$, is the largest cardinality of a subset $\Lambda \subset \Omega$ shattered by \mathcal{F} . If the largest cardinality does not exist, $\text{VC}(\mathcal{F}) \equiv \infty$.

We state just for information a uniform law of large numbers involving the VC dimension. It goes beyond the scope of this lecture to actually provide all details about the relationship between the covering numbers for Boolean functions and their VC dimension. We need the following result to appreciate the application example in Section 8.

Theorem 7.29 (Empirical processes via VC dimension) *Let $(\Omega, \mathcal{B}(\Omega), \mu)$ a probability space and $\mathcal{F} = \{f: \Omega \rightarrow \{0, 1\}\}$ a class of Boolean functions with $\text{VC}(\mathcal{F}) \in [1, \infty)$. Consider the i.i.d. samples X, X_1, \dots, X_N with law $\mu \in \mathcal{M}_1(\Omega)$. Then*

$$\mathbb{E} \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{N} \sum_{i=1}^N f(X_i) - \mathbb{E}[f(X)] \right| \right] \leq C \sqrt{\frac{\text{VC}(\mathcal{F})}{N}}$$

for some absolute constant $C > 0$.

8 Application: Statistical Learning theory

In statistical learning theory we are concerned with learning a *target function* $T: \Omega \rightarrow \mathbb{R}$ from empirical data X_1, \dots, X_N , where the X_i 's are independent and identical distributed according to some law $\mu \in \mathcal{M}_1(\Omega)$. We call the N pairs $(X_i, T(X_i))_{i=1, \dots, N}$ the *training data*. Our task is then to seek a good prediction of $T(X)$ for any data $X \notin \{X_1, \dots, X_N\}$. We focus in so-called classification problems where the target function $T: \Omega \rightarrow \{0, 1\}$ is a Boolean function. This target function classifies the data points in Ω into two classes depending on the label, i.e., the value of the target function.

Example 8.1 (Health study) We examine N patients and determine n health parameter, e.g., blood pressure heart rate, weight, etc. We then obtain the samples $X_i \in \mathbb{R}^n, i = 1, \dots, N$. Suppose we know whether each patient has a certain illness or not, e.g. diabetes. That is, we know $T(X_i) \in \{0, 1\}, i = 1, \dots, N$ with $T(X_i) = 1$ being sick and $T(X_i) = 0$ being healthy. We want to learn from the given training sample to diagnose diabetes, i.e., we want to learn the target function $T: \Omega \rightarrow \{0, 1\}$. This target function would output diagnosis for any person based on the n health parameter. ♣

A solution to the learning problem can be a function $f: \Omega \rightarrow \{0, 1\}$ which is as close as possible to the target function $T: \Omega \rightarrow \{0, 1\}$. We like to choose the function f which minimises the so-called *risk*.

Definition 8.2 Let $(\Omega, \mathcal{B}(\Omega), \mu)$ be a probability space. The risk of a function $f: \Omega \rightarrow \mathbb{R}$ in a learning problem with the target function $T: \Omega \rightarrow \mathbb{R}$ and class \mathcal{F} of functions is defined as

$$R(f) := \mathbb{E}[(f(X) - T(X))^2], \quad (8.1)$$

where the expectation is with respect to the probability measure $\mu \in \mathcal{M}_1(\Omega)$. The minimiser f^* of the risk is defined as

$$f^* := \mathbf{argmin}_{f \in \mathcal{F}} \{R(f)\}. \quad (8.2)$$

For our classification problem, i.e., learning of a Boolean target function $T: \Omega \rightarrow \{0, 1\}$ with \mathcal{F} a class of Boolean functions, we see that

$$R(f) = \mathbb{E}[f^2(X) - 2f(X)T(X) + T^2(X)] = \mathbb{P}(f(X) \neq T(X))$$

as $f^2(X) - 2f(X)T(X) + T^2(X)$ is zero when $f(X) = T(X)$ and 1 otherwise.

In any learning problem the choice of the class of functions is crucial. In this context we call the class of functions \mathcal{F} the hypothesis space. If we choose \mathcal{F} to be a class of simple functions like linear functions or Lipschitz function we might get easily calculations and estimates but we can be still off the real function, this is called *under fitting*. Conversely, if we consider too many different function types in \mathcal{F} we

end up with challenges in computing the VC dimension and obtaining estimates. In addition we might over fit the sample which happens if any normal fluctuation (noise) is reflected. If $T \in \mathcal{F}$, we get clearly $R(f) = 0$ as $f^* = T$ in that case. However, note that in general we cannot compute the risk $R(f)$ and thus f^* from the given empirical training data. Instead we can only approximate the risk $R(f)$ and its minimiser f^* given the empirical data.

Definition 8.3 (Empirical risk) Let $(\Omega, \mathcal{B}(\Omega), \mu)$ be a probability space and $f: \Omega \rightarrow \mathbb{R}$ an element of the hypothesis class \mathcal{F} of a learning problem with target function $T: \Omega \rightarrow \mathbb{R}$. Let X_1, \dots, X_N be Ω -valued i.i.d. samples with law $\mu \in \mathcal{M}_1(\Omega)$. The *empirical risk* of f given the sample is defined as

$$R_N(f) := \frac{1}{N} \sum_{i=1}^N (f(X_i) - T(X_i))^2, \quad (8.3)$$

and the minimiser $f_N^* \in \mathcal{F}$ of the empirical risk is

$$f_N^* := \mathbf{argmin}_{f \in \mathcal{F}} \{R_N(f)\}.$$

The *excess risk* is defined as the difference

$$R_N(f_N^*) - R(f^*).$$

Lemma 8.4 Let $(\Omega, \mathcal{B}(\Omega), \mu)$ be a probability space and $f: \Omega \rightarrow \mathbb{R}$ an element of the hypothesis class \mathcal{F} of a learning problem with target function $T: \Omega \rightarrow \mathbb{R}$. Let X_1, \dots, X_N be Ω -valued i.i.d. samples with law $\mu \in \mathcal{M}_1(\Omega)$. Then

$$R(f_N^*) - R(f^*) \leq 2 \sup_{f \in \mathcal{F}} |R_N(f) - R(f)|.$$

Proof. Define $\varepsilon := \sup_{f \in \mathcal{F}} |R_N(f) - R(f)|$, Then

$$\begin{aligned} R(f_N^*) &\leq R_N(f_N^*) + \varepsilon \quad (\text{since } f_N^* \in \mathcal{F} \text{ by definition}) \\ &\leq R_N(f^*) + \varepsilon \quad (\text{as } f_N^* \text{ minimises } R_N \text{ in } \mathcal{F}) \\ &\leq R(f^*) + 2\varepsilon \quad (\text{as } f^* \in \mathcal{F} \text{ by definition}). \end{aligned}$$

Subtracting $R(f^*)$ on both sides, we get the statement of the lemma. \square

Theorem 8.5 (Excess risk via VC dimension) Let $(\Omega, \mathcal{B}(\Omega), \mu)$ be a probability space and \mathcal{F} be a class of functions of a learning problem with target function $T: \Omega \rightarrow \mathbb{R}$ and $\text{VC}(\mathcal{F}) \geq 1$. Let X_1, \dots, X_N be Ω -valued i.i.d. samples with law $\mu \in \mathcal{M}_1(\Omega)$. Then

$$\mathbb{E}[R(f_N^*)] \leq R(f^*) + C \sqrt{\frac{\text{VC}(\mathcal{F})}{N}}$$

for some absolute constant $C > 0$.

Proof. According to Lemma 8.4 it suffices to show that

$$\mathbb{E}[\sup_{f \in \mathcal{F}} |R_N(f) - R(f)|] \lesssim \sqrt{\frac{\text{VC}(\mathcal{F})}{N}}. \quad (8.4)$$

We insert our definition for R_N and R and obtain

$$\text{L.H.S. of (8.4)} = \mathbb{E}[\sup_{\ell \in \mathcal{L}} |\frac{1}{N} \sum_{i=1}^N \ell(X_i) - \mathbb{E}[\ell(X)]|].$$

where $\mathcal{L} = \{(f - T)^2 : f \in \mathcal{F}\}$. Furthermore, an application of Theorem 7.29 and its proof gives

$$\text{L.H.S. of (8.4)} \lesssim \frac{1}{\sqrt{N}} \mathbb{E} \left[\int_0^1 \sqrt{\log \mathcal{N}(\mathcal{L}, L^2(\mu_N), \varepsilon)} d\varepsilon \right],$$

where $L^2(\mu_N)$ is the L^2 metric with respect to the empirical measure $\mu_N = \frac{1}{N} \sum_{i=1}^N \delta_{X_i}$. To proceed we need to relate the covering of \mathcal{L} and \mathcal{F} , that is, we show that

$$\mathcal{N}(\mathcal{L}, L^2(\mu_N), \varepsilon) \leq \mathcal{N}(\mathcal{F}, L^2(\mu_N), \varepsilon/4), \quad \text{for all } \varepsilon \in (0, 1). \quad (8.5)$$

To see (8.5) pick an ε -net $\{f_j\}_{j=1, \dots, J}$ of \mathcal{F} . For all $\ell \in \mathcal{L}$ there exist $\ell_j := (f_j - T)^2$ such that

$$\begin{aligned} \|\ell - \ell_j\|_{L^2(\mu_N)} &= \|f^2 - 2T(f - f_j) - f_j^2\|_{L^2(\mu_N)} = \|(f + f_j)(f - f_j) - 2T(f - f_j)\|_{L^2(\mu_N)} \\ &\leq 2\|f - f_j\|_{L^2(\mu_N)} + 2\|f - f_j\|_{L^2(\mu_N)} \leq \varepsilon \end{aligned}$$

whenever $\|f - f_j\|_{L^2(\mu_N)} \leq \varepsilon/4$. This shows our claim (8.5). Now we replace \mathcal{L} by \mathcal{F} and use Theorem 7.29 and its proof to see that

$$\log \mathcal{N}(\mathcal{F}, L^2(\mu_N), \varepsilon) \lesssim \text{VC}(\mathcal{F}) \log(2/\varepsilon)$$

to conclude with our statement. \square

If we want to bound the expected excess risk in our health study Example 8.1 by $\varepsilon > 0$, all we need to do is to take a sample of size

$$N \sim \varepsilon^{-2} \text{VC}(\mathcal{F}).$$

References

- [Dur19] R. DURRETT. *Probability - Theory and Examples*. Cambridge University Press, fifth ed., 2019.
- [Geo12] H.-O. GEORGII. *Stochastics*. De Gruyter Textbook. De Gruyter, 2nd rev. and ext. ed., 2012.
- [Ver18] R. VERSHYNIN. *High-Dimensional Probability*. Cambridge University Press, 2018.

Index

- ε -net, 45
- ε -separated, 45
- n -dimensional ball, 35
- n -dimensional sphere, 35
- Sub-exponential norm, 27
- Bernstein condition, 29
- canonical Gaussian process in \mathbb{R}^n , 82
- canonical metric of the process, 87
- centred moment generating function, 18
- chaining, 93
- Chaos, 74
- covariance function, 82
- covariance matrix, 9, 39
- covering number, 45
- cumulative distribution function (CDF),
3
- Decoupling, 74
- diameter, 90
- empirical covariance, 52
- empirical measure, 52
- empirical process, 93
- empirical risk, 98
- essential supremum, 5
- excess risk, 98
- Gamma function, 20
- Gaussian interpolation, 83
- Gaussian process, 82
- Gaussian width, 90
- global Lipschitz continuity, 58
- Hamming cube, 47
- Hamming distance, 47
- Herbst argument, 56
- increments, 82
- isotropic, 39
- Jensen's inequality, 4
- Johnson-Lindenstrauss Lemma, 72
- Kullback-Leibler divergence, 54
- Landau symbols, 36
- level sets, 63
- Lipschitz continuous, 58
- Minkowski sum, 46
- moment generation function, 3
- Multivariate Normal / Gaussian distribution, 42
- normal distribution, 9
- operator norm, 48
- packing number, 45
- Rademacher function, 29
- random field, 81
- random process, 81
- relative entropy, 54
- second moment matrix, 39
- separately convex, 58
- Shannon entropy, 55
- singular values, 48
- spherically distributed, 41
- standard Brownian motion, 81
- Stirling's formula, 20
- stochastically dominated, 83
- Sub-exponential random variable, 26
- Sub-exponential random variable, second definition, 28
- Sub-exponential random variables, first definition, 26
- Sub-Gaussian - first definition, 21
- Sub-Gaussian - second definition, 23
- Sub-Gaussian increments, 93
- Sub-Gaussian properties, 18
- sub-level sets, 63
- symmetric, 80
- symmetric Bernoulli distribution in \mathbb{R}^n ,
41
- tails of the normal distribution, 11

target function, 97

training data, 97

unit sphere, 41

VC dimension, 95

Young's inequality, 27