HIGH DIMENSIONAL PROBABILITY

for Mathematicians and Data Scientists

Working draft – not for distribution

ROMAN VERSHYNIN¹

February 25, 2016

¹University of Michigan. Webpage: www.umich.edu/~romanv
## Contents

1 Preliminaries on random variables 1
   1.1 Basic quantities associated with random variables . . . . . . . 1
   1.2 Some classical inequalities . . . . . . . . . . . . . . . . . . . 2
   1.3 The Law of Large Numbers and the Central Limit Theorem . 4

2 Concentration of sums of independent random variables 7
   2.1 Why concentration inequalities? . . . . . . . . . . . . . . . . 7
   2.2 Hoeffding’s inequality . . . . . . . . . . . . . . . . . . . . . . 10
   2.3 Chernoff’s inequality and Poisson tails . . . . . . . . . . . . 13
   2.4 Application: degrees of random graphs . . . . . . . . . . . . 17
   2.5 Sub-gaussian distributions . . . . . . . . . . . . . . . . . . . 19
   2.6 General Hoeffding’s and Khinchine’s inequalities . . . . . . 25
   2.7 Sub-exponential distributions . . . . . . . . . . . . . . . . . 29
   2.8 Bernstein’s inequality . . . . . . . . . . . . . . . . . . . . . . 32

3 Random vectors in high dimensions 37
   3.1 Concentration of the norm . . . . . . . . . . . . . . . . . . . 37
   3.2 Covariance matrices and isotropic distributions . . . . . . . 40
   3.3 Examples of high dimensional distributions . . . . . . . . . . 45
   3.4 Sub-gaussian distributions in higher dimensions . . . . . . . 51

4 Sub-gaussian random matrices 57
   4.1 Preliminaries on matrices . . . . . . . . . . . . . . . . . . . . 57
   4.2 Nets, covering numbers and packing numbers . . . . . . . . 61
   4.3 Upper bounds on sub-gaussian matrices . . . . . . . . . . . . 66
   4.4 Application: community detection in networks . . . . . . . . 71
   4.5 Two-sided bounds on sub-gaussian matrices . . . . . . . . . . 75
   4.6 Application: covariance estimation and clustering . . . . . . 78
5 Concentration without independence 83
  5.1 Concentration of Lipschitz functions on the sphere . . . . . . 83
  5.2 Concentration on other metric measure spaces . . . . . . . . . 88
  5.3 Application: Johnson-Lindenstrauss Lemma . . . . . . . . . . 94
  5.4 Matrix Bernstein inequality . . . . . . . . . . . . . . . . . . . . 98
  5.5 Random matrices with general independent rows . . . . . . . . 98
  5.6 Application: covariance estimation for general distributions . 102
Chapter 1

Preliminaries on random variables

1.1 Basic quantities associated with random variables

In a basic course in probability theory, we learned about the two most important quantities associated with a random variable $X$, namely the expectation\(^1\) (also called mean), and variance. They will be denoted in this book by

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

Let us recall some other classical quantities and functions that describe probability distributions. The moment generating function of $X$ is defined as

$$M_X(t) = \mathbb{E} e^{tX}, \quad t \in \mathbb{R}.$$ 

For $p > 0$, the $p$-th moment of $X$ is defined to be $\mathbb{E}X^p$, and the absolute $p$-th moment is $\mathbb{E}|X|^p$.

It is useful to take $p$-th root of the moments, which leads to the quantity called the $L^p$ norm:

$$\|X\|_p = (\mathbb{E}|X|^p)^{1/p}, \quad p \in (0, \infty).$$

\(^1\)If you studied measure theory, you will recall that the expectation $\mathbb{E}X$ of a random variable $X$ on a probability space $(\Omega, \Sigma, \mathbb{P})$ is, by definition, the Lebesgue integral of the function $X : \Omega \to \mathbb{R}$. This makes all theorems on Lebesgue integration applicable in probability theory, for expectations of random variables.
CHAPTER 1. PRELIMINARIES ON RANDOM VARIABLES

This definition can be extended to $p = \infty$ by the essential supremum of $|X|:

$$\|X\|_{\infty} = \text{ess sup } |X|.$$  

For fixed $p$ and a given probability space $(\Omega, \Sigma, \mathbb{P})$, the classical space $L^p = L^p(\Omega, \Sigma, \mathbb{P})$ consists of all random variables $X$ on $\Omega$ with finite $L^p$ norm, that is

$$L^p = \{ X : \|X\|_p < \infty \}.$$  

For $p \in [1, \infty]$ the quantity $\|X\|_p$ is a norm and $L^p$ is a Banach space. This fact follows from Minkowski’s inequality, which we will recall in (1.1). For $p < 1$, the triangle inequality fails and $\|X\|_p$ is not a norm.

The exponent $p = 2$ is special: $L^2$ is a **Hilbert space**. The inner product on $L^2$ is given by

$$\mathbb{E}(X,Y) = \mathbb{E}XY.$$  

Then the **standard deviation** of $X$ can be expressed as

$$\|X - \mathbb{E}X\|_2 = \sqrt{\text{Var}(X)} = \sigma(X).$$  

Similarly, we can express the **covariance** of random variables of $X$ and $Y$ can be expressed as

$$\text{cov}(X,Y) = \mathbb{E}(X - \mathbb{E}X)(Y - \mathbb{E}Y) = \langle X - \mathbb{E}X, Y - \mathbb{E}Y \rangle.$$  

When we consider random variables as vectors in the Hilbert space $L^2$, we may interpret covariance in a geometric way. The more the vectors $X - \mathbb{E}X$ and $Y - \mathbb{E}Y$ are aligned with each other, the bigger their inner product and covariance are.

1.2 Some classical inequalities

Let us review some basic inequalities on random variables. Most of them are usually covered in advanced calculus or basic probability courses.  

**Jensen’s inequality** states that for any random variable $X$ and a convex\(^2\) function $\varphi : \mathbb{R} \to \mathbb{R}$, we have

$$\varphi(\mathbb{E}X) \leq \mathbb{E}\varphi(X).$$

\(^2\)By definition, a function $\varphi$ is convex if $\varphi(\lambda x + (1 - \lambda)y) \leq \lambda \varphi(x) + (1 - \lambda)\varphi(y)$ for all $t \in [0, 1]$ and all $x, y$ in the domain of $\varphi$.  

As a simple consequence of Jensen’s inequality, we may note that \( \|X\|_p \) is an increasing function in \( p \), that is

\[
\|X\|_p \leq \|X\|_q \quad \text{for any } 0 \leq p \leq q = \infty.
\]

This inequality follows since \( \phi(x) = x^{q/p} \) is a convex function if \( q/p \geq 1 \).

**Minkowski’s inequality** states that for any \( p \in [1, \infty] \) and any random variables \( X, Y \in L_p \), we have

\[
\|X + Y\|_p \leq \|X\|_p + \|Y\|_p.
\]  

(1.1)

This can be viewed as the **triangle inequality**, which implies that \( \| \cdot \|_p \) is a norm when \( p \in [1, \infty] \).

**Cauchy-Schwarz inequality** states that for any random variables \( X, Y \in L_2 \), we have

\[
\mathbb{E} XY \leq \|X\|_2 \|Y\|_2.
\]

The more general **Hölder’s inequality** states that if \( p, p' \in (1, \infty) \) are conjugate exponents, that is \( 1/p + 1/q = 1 \), then random variables \( X \in L_p \) and \( Y \in L_q \) satisfy

\[
\mathbb{E} XY \leq \|X\|_p \|Y\|_q.
\]

This inequality also holds for the pair \( p = 1, q = \infty \).

As we recall from a basic probability course, the **distribution** of a random variable \( X \) is, intuitively, the information about what values \( X \) takes with what probabilities. More rigorously, the distribution of \( X \) is determined by the **cumulative distribution function** (CDF) of \( X \), defined as

\[
F_X(t) = \mathbb{P} \{X \leq t\} \quad \text{for } t \in \mathbb{R}.
\]

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

\[
\mathbb{P} \{X > t\} = 1 - F_X(t).
\]

There is an important connection between the tails and the expectation (and more generally, the moments) of a random variable. The following identity is typically used to bound the expectation by tails.

**Lemma 1.2.1** (Integral Identity). **For any non-negative random variable** \( X \), **we have**

\[
\mathbb{E} X = \int_0^\infty \mathbb{P} \{X > t\} \, dt.
\]
Exercise 1.2.2. Prove the Expectation Integral Identity.

Another classical fact, Markov’s inequality, bounds the tails in terms of expectation.

Proposition 1.2.3 (Markov’s Inequality). For any non-negative random variable \(X\) and \(t > 0\), we have

\[
P \{ X \geq t \} \leq \frac{\mathbb{E} X}{t}.
\]

A well-known consequence of Markov’s inequality is the following Chebyshev’s inequality. It offers a better, quadratic dependence on \(t\), and instead of the plain tails, it quantifies the concentration of \(X\) about its mean.

Corollary 1.2.4 (Chebyshev’s inequality). Let \(X\) be a random variable with mean \(\mu\) and variance \(\sigma^2\). Then, for any \(t > 0\), we have

\[
P \{ |X - \mu| \geq t \} \leq \frac{\sigma^2}{t^2}.
\]

Exercise 1.2.5. Deduce Chebyshev’s inequality by squaring both sides of the bound \(|X - \mu| \geq t\) and applying Markov’s inequality.

In Proposition 2.5.2 we will relate together the three basic quantities recalled here – the moment generating functions, the \(L^p\) norms, and the tails.

1.3 The Law of Large Numbers and the Central Limit Theorem

Let us recall the two arguably most important results in classical probability theory, the Law of Large Numbers and the Central Limit Theorem.

Theorem 1.3.1 (Strong Law of Large Numbers). Let \(X_1, X_2, \ldots\) be a sequence of independent, identically distributed (i.i.d.) random variables with mean \(\mu\). Consider the sum

\[
S_N = X_1 + \cdots + X_N.
\]

Then, as \(N \to \infty\),

\[
\frac{S_N}{N} \to \mu \quad \text{almost surely.}
\]
1.3. LLN AND CLT

The next result, Central Limit Theorem, states that the limiting distribution of the properly scaled sum of \( X_i \) is the normal distribution, sometimes also called Gaussian distribution. Recall that the standard normal distribution, denoted \( N(0,1) \), has density

\[
f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad x \in \mathbb{R}. \tag{1.2}
\]

**Theorem 1.3.2** (Lindeberg-Lévy Central Limit Theorem). Let \( X_1, X_2, \ldots \) be a sequence of independent, identically distributed (i.i.d.) random variables with mean \( \mu \) and variance \( \sigma^2 \). Consider the sum

\[ S_N = X_1 + \cdots + X_N \]

and normalize it to obtain a random variable with zero mean and unit variance as follows:

\[
Z_N := \frac{S_N - \mathbb{E}S_N}{\sqrt{\text{Var}(S_N)}} = \frac{1}{\sigma \sqrt{N}} \sum_{i=1}^{N} (X_i - \mu).
\]

Then, as \( N \to \infty \),

\[ Z_N \to N(0,1) \quad \text{in distribution.} \]

The convergence in distribution means that the CDF of the normalized sum converges pointwise to the CDF of the standard normal distribution. We can expressed this in terms of tails as follows. Then for every \( t \in \mathbb{R} \), we have

\[
P \{Z_N \geq t\} \to P \{g \geq t\} = \frac{1}{\sqrt{2\pi}} \int_{t}^{\infty} e^{-x^2/2} \, dx
\]

as \( N \to \infty \), where \( g \sim N(0,1) \) is a standard normal random variable.
Chapter 2

Concentration of sums of independent random variables

2.1 Why concentration inequalities?

Concentration inequalities quantify how a random variable $X$ deviates around its mean $\mu$. They usually take the form of two-sided bounds for the tails of $X - \mu$, such as

$$\mathbb{P} \{|X - \mu| \geq t\} \leq \text{something small}.$$ 

The simplest concentration inequality is Chebyshev’s inequality, Corollary 1.2.4. Unfortunately, it is usually too weak to capture the true order of deviation. Let us illustrate it with the example of the binomial distribution.

**Question 2.1.1.** Toss a fair coin $N$ times. What is the probability that we get at least $\frac{3}{4}N$ heads?

The number of heads $S_N$ is a binomial random variable. We can represent as $S_N = \sum_{i=1}^N X_i$ where $X_i$ are the indicators of heads. Thus $X_i$ are independent Bernoulli random variables with parameter $1/2$, i.e. $\mathbb{P} \{X_i = 0\} = \mathbb{P} \{X_i = 1\} = 1/2$. Thus

$$\mathbb{E} S_N = \frac{N}{2}, \quad \text{Var}(S_N) = \frac{N}{4}.$$ 

Applying Chebyshev’s inequality, we can bound the probability of at
least $\frac{3}{4}N$ heads as follows:

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

So the probability converges to zero at least linearly in $N$.

Is this the right rate of decay, or should we expect something faster? Let us approach the same question using the Central Limit Theorem. It states that the distribution of the normalized number of heads

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

converges to $N(0,1)$. Thus we should anticipate that for large $N$,

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

where $g \sim N(0,1)$. To understand how this quantity decays in $N$, we need a good bound on the tails of the normal distribution.

**Proposition 2.1.2** (Tails of the normal distribution). Let $g \sim N(0,1)$. Then for all $t \geq 1$

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

In particular, for $t \geq 1$ the tail is bounded by the density:

$$\mathbb{P}\left\{ g \geq t \right\} \leq \frac{1}{\sqrt{2\pi}} e^{-t^2/2}. \quad (2.2)$$

**Proof.** To obtain an upper bound on the tail

$$\mathbb{P}\left\{ g \geq t \right\} = \frac{1}{\sqrt{2\pi}} \int_t^\infty e^{-x^2/2} \, dx,$$

let us change variables $x = t + y$. This gives

$$\mathbb{P}\left\{ g \geq t \right\} = \frac{1}{\sqrt{2\pi}} \int_0^\infty e^{-t^2/2} e^{-ty} e^{-y^2/2} \, dy \leq \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \int_0^\infty e^{-ty} \, dy,$$

where we used that $e^{-y^2/2} \leq 1$. Since the last integral equals $1/t$, the desired upper bound on the tail follows.

The lower bound follows from the identity

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

This completes the proof. \qed
2.1. WHY CONCENTRATION INEQUALITIES?

Returning to (2.1), we see that we should expect the probability of having at least \( \frac{3}{4}N \) heads to be smaller than

\[
\frac{1}{\sqrt{2\pi}} e^{-N/8}, \tag{2.3}
\]

a quantity that decays to zero exponentially fast in \( N \).

Unfortunately, this can not be deduced rigorously from the Central Limit Theorem. Although the approximation by the normal density in (2.1) is valid, the error of approximation can not be ignored. And, unfortunately, the error decays to zero too slow, even slower than linear in \( N \). We can see this from the following quantitative version of the Central Limit Theorem.

**Theorem 2.1.3** (Berry-Esseen Central Limit Theorem). *In the setting of Theorem 1.3.2, for every \( N \) and every \( t \in \mathbb{R} \) we have*

\[
|\mathbb{P}\{Z_N \geq t\} - \mathbb{P}\{g \geq t\}| \leq \frac{\rho}{\sqrt{N}}.
\]

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

So, once we take into account in (2.1) the approximation error, which is order \( 1/\sqrt{N} \), it ruins the desired exponential decay (2.3).

Can we improve the approximation error in Central Limit Theorem? In general, no. If \( N \) is even, then the number of getting exactly \( N/2 \) heads is

\[
\mathbb{P}\left\{ S_N = \frac{N}{2} \right\} = 2^{-N} \binom{N}{N/2} \sim \frac{1}{\sqrt{N}};
\]

the last estimate can be obtained using Stirling's approximation. (Do it!) On the other hand, since the normal distribution is continuous, we have \( \mathbb{P}\{g = \frac{N}{2}\} = 0 \). Thus the approximation error here has to be of order \( 1/\sqrt{N} \).

Let us summarize our situation. The Central Limit theorem offers an approximation of a sum of independent random variables \( S_N = X_1 + \ldots + X_N \) by normal distribution. The normal distribution is especially nice since it has very light, exponentially decaying tails. At the same time, the error of approximation in Central Limit Theorem decays too slow, even slower than linear. This big error ruins desired concentration properties for \( S_N \), which should guarantee light, exponentially decaying tails for \( S_N \).

In order to resolve this issue, we need different ways to obtain concentration inequalities for \( S_N \). We will now develop a direct approach to concentration, which completely bypasses the Central Limit Theorem.
2.2 Hoeffding’s inequality

It will be more convenient to work here with symmetric Bernoulli random variables, which are already properly normalized:

**Definition 2.2.1** (Symmetric Bernoulli distribution). A random variable $X$ has symmetric Bernoulli distribution (also called Rademacher distribution) if it takes values $-1$ and $1$ with probabilities $1/2$ each, i.e.

$$P\{X = -1\} = P\{X = 1\} = \frac{1}{2}.$$ 

Clearly, a random variable $X$ has (usual) Bernoulli distribution with parameter $1/2$ if and only if $Z = 2X - 1$ has symmetric Bernoulli distribution.

**Theorem 2.2.2** (Hoeffding’s inequality). Let $X_1,\ldots,X_N$ be independent symmetric Bernoulli random variables, and $a = (a_1,\ldots,a_N) \in \mathbb{R}^N$. Then, for any $t > 0$, we have

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

*Proof.* By homogeneity, we can assume without loss of generality that $\|a\|^2 = 1$.

Let us recall how we deduced Chebyshev’s inequality (Corollary 1.2.4): we squared both sides and applied Markov’s inequality. Let us do something similar here. But instead of squaring both sides, let us multiply by a fixed parameter $\lambda > 0$ (to be chosen later) and exponentiate. This gives

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

(2.4)

In the last step we applied Markov’s inequality (Proposition 1.2.3).

We thus reduced the problem to bounding the moment generating function (MGF) of the sum $\sum_{i=1}^N a_i X_i$. As we recall from the basic probability course, the MGF of the sum is the product of the MGF’s of the terms; this follows immediately from independence. In other words,

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

(2.5)
Let us fix $i$. Since $X_i$ takes values $-1$ and $1$ with probabilities $1/2$ each, we have

$$
\mathbb{E}\exp(\lambda a_i X_i) = \frac{\exp(\lambda a_i) + \exp(-\lambda a_i)}{2} = \cosh(\lambda a_i).
$$

**Exercise 2.2.3** (Bounding the hyperbolic cosine). *Show that*

$$
cosh(x) \leq \exp(x^2/2) \quad \text{for} \quad x \in \mathbb{R}.
$$

*(Compare the Taylor’s expansions of both sides.)*

This bound shows that

$$
\mathbb{E}\exp(\lambda a_i X_i) \leq \exp(\lambda^2 a_i^2/2).
$$

Substituting this into (2.5) and then into (2.4), we obtain

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

In the last identity, we used the assumption that $\|a\|_2 = 1$.

This bound holds for arbitrary $\lambda > 0$. It remains to optimize in $\lambda$; the minimum is clearly attained for $\lambda = t$. With this choice, we obtain

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

This completes the proof of Hoeffding’s inequality. $\square$

We can view Hoeffding’s inequality as a concentration version of Central Limit Theorem. Indeed, the most we may expect from a concentration inequality is that the tail of $\sum a_i X_i$ match the tail for the normal distribution. And for all practical purposes, Hoeffding’s tail bound does that. With the normalization $\|a\|_2 = 1$, Hoeffding’s inequality provides the tail $e^{-t^2/2}$, which is exactly the same as the bound for the standard normal tail in (2.2). This is good news. We have been able to obtain the same exponentially light tails for sums as for normal distribution, even though the difference of these two distributions is not exponentially small.

Armed with Hoeffding’s inequality, we can now return to Question 2.1.1 of bounding the probability of at least $\frac{3}{4} N$ heads in $N$ tosses of a fair coin.
After rescaling from Bernoulli to symmetric Bernoulli, we obtain that this probability is \textit{exponentially small} in \(N\), namely
\[
P\left\{ \text{at least } \frac{3}{4}N \text{ heads} \right\} \leq \exp(-N/4).
\]
(Check this.)

It should be stressed that unlike the classical limit theorems of Probability Theory, Hoeffding’s inequality is \textit{non-asymptotic} in that it holds for all fixed \(N\) as opposed to \(N \to \infty\). The larger \(N\), the stronger inequality becomes. As we will see later, the non-asymptotic nature of concentration inequalities like Hoeffding makes them attractive in applications in data sciences, where \(N\) corresponds to \textit{sample size}.

We can easily derive a version of Hoeffding’s inequality for \textit{two-sided tails} \(P\{ |S| \geq t \}\) where \(S = \sum_{i=1}^{N} a_i X_i\). Indeed, applying Hoeffding’s inequality for \(-X_i\) instead of \(X_i\), we obtain a bound on \(P\{ -S \geq t \}\). Combining the two bounds, we obtain a bound on
\[
P\{ |S| \geq t \} = P\{ S \geq t \} + P\{ -S \geq t \}.
\]
Thus the bound doubles, and we obtain:

\textbf{Theorem 2.2.4} (Hoeffding’s inequality, two-sided). \textit{Let} \(X_1, \ldots, X_N\) \textit{be independent symmetric Bernoulli random variables, and} \(a = (a_1, \ldots, a_N) \in \mathbb{R}\). \textit{Then, for any} \(t > 0\), \textit{we have}
\[
P\left\{ \left| \sum_{i=1}^{N} a_i X_i \right| \geq t \right\} \leq 2 \exp\left( -\frac{t^2}{2||a||_2^2} \right).
\]

Our proof of Hoeffding’s inequality, which is based on bounding the moment generating function, is quite flexible. It applies far beyond the canonical example of symmetric Bernoulli distribution. For example, the following extension of Hoeffding’s inequality is valid for general bounded random variables.

\textbf{Theorem 2.2.5} (Hoeffding’s inequality for general bounded random variables). \textit{Let} \(X_1, \ldots, X_N\) \textit{be independent random variables. Assume that} \(X_i \in [m_i, M_i]\) \textit{almost surely for every} \(i\). \textit{Then, for any} \(t > 0\), \textit{we have}
\[
P\left\{ \sum_{i=1}^{N} a_i (X_i - \mathbb{E} X_i) \geq t \right\} \leq \exp\left( -\frac{2t^2}{\sum_{i=1}^{N} (M_i - m_i)^2} \right).
\]
2.3. CHERNOFF’S INEQUALITY AND POISSON TAILS

Theorem 2.2.2 is a partial case of this result with \( m_i = -|a_i| \) and \( M_i = |a_i| \).

**Exercise 2.2.6.** [Difficulty=5] Prove Theorem 2.2.5, possibly with some absolute constant instead of 2 in the tail.

2.3 Chernoff’s inequality and Poisson tails

We saw in the previous section that Hoeffding’s inequality is sharp for symmetric Bernoulli random variables. What about the general form of Hoeffding’s inequality, Theorem 2.2.5 – is it sharp for all distributions we may care about? Unfortunately, no. This bound is in terms of the worst possible, extreme values \( m_i \) and \( M_i \) the random variables \( X_i \) can take. These could be too large compared to more realistic average magnitudes, which could be quantified by the variances of \( X_i \).

An important example where these two bound could be very different is where \( X_i \) are Bernoulli random variables with very small parameters \( p_i \). If, for example, \( p_i = \mu/n \) for constant \( \mu \), then the sum \( S_N = \sum_{i=1}^N X_i \) has constant mean \( \mu \), and its distribution converges to Poisson distribution as \( N \to \infty \). (We will make this precise shortly). At the same time, Hoeffding’s inequality is completely insensitive to the magnitude of \( p_i \), and does not provide a useful tail bound for a Poisson-looking sum \( S_N \).

The following classical inequality provides a quite sharp result, which is sensitive to the true magnitudes of \( X_i \).

**Theorem 2.3.1** (Chernoff’s inequality). Let \( X_i \) be independent Bernoulli random variables with parameters \( p_i \). Consider their sum \( S_N = \sum_{i=1}^N X_i \) and denote its mean by \( \mu = \mathbb{E}S_N \). Then, for any \( t > \mu \), we have

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

In particular, for any \( t \geq e^2 \mu \) we have

\[
\mathbb{P}\{S_N \geq t\} \leq e^{-t}.
\] (2.6)

**Proof.** We will use the same method – based on moment generating function – as we did in the proof of Hoeffding’s inequality, Theorem 2.2.2. We repeat the first steps of that argument, leading to (2.4) and (2.5) – multiply both sides of the inequality \( S_N \geq t \) by a parameter \( \lambda \), exponentiate, and then use Markov’s inequality and independence. This yields

\[
\mathbb{P}\{S_N \geq t\} \leq e^{-\lambda t} \prod_{i=1}^N \mathbb{E}\exp(\lambda X_i).
\] (2.7)
It remains to bound the MGF of each Bernoulli random variable $X_i$ separately. Since $X_i$ takes value 1 with probability $p_i$ and 1 with probability 0, we have

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

In the last step, we used the numeric inequality $1 + x \leq e^x$. Consequently,

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

Substituting this into (2.7), we obtain

$$P\{S_N \geq t\} \leq e^{-\lambda t} \exp \left[ (e^{\lambda} - 1)\mu \right].$$

This bound holds for any $\lambda > 0$. Substituting the value $\lambda = \ln(t/\mu)$ which is positive by the assumption $t > \mu$ and simplifying the expression, we complete the proof.

**Exercise 2.3.2 (Chernoff’s inequality: lower tails).** [Difficulty=5] Modify the proof of Theorem 2.3.1 to obtain the following bound on the lower tail. For any $t < \mu$, we have

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

### 2.3.1 Poisson tails

We will see now that Chernoff’s inequality is related to the classical Poisson distribution – in the same spirit as Hoeffding’s inequality is related to the normal distribution. The connection is provided by the classical Poisson Limit Theorem. It is an analog of the Central Limit Theorem for sums of Bernoulli random variables $X_i \sim \text{Ber}(p_i)$ with extremely small parameters $p_i$, as opposed to constant parameters $p_i$ in Central Limit Theorem.

Recall that a random variable $X$ has Poisson distribution with parameter $\lambda$, denoted $X \sim \text{Pois}(\lambda)$, if it takes values in $\{0, 1, 2, \ldots\}$ with probabilities

$$P\{X = k\} = e^{-\lambda} \frac{\lambda^k}{k!}, \quad k = 0, 1, 2, \ldots$$ (2.8)
Theorem 2.3.3 (Poisson Limit Theorem). Consider a sequence of independent random variables \( X_i \sim \text{Ber}(p_i) \) and let \( S_N = \sum_{i=1}^{N} X_i \). Assume that, as \( N \to \infty \),
\[
\max_{i \leq N} p_i \to 0 \quad \text{and} \quad \mathbb{E} S_N \to \lambda.
\]
Then, as \( N \to \infty \),
\[
S_N \to \text{Pois}(\lambda) \quad \text{in distribution.}
\]

Poisson Limit Theorem allows us to immediately transfer Chernoff’s inequality from Bernoulli to Poisson distribution as follows. (Why?)

Corollary 2.3.4 (Poisson tails). Let \( X \sim \text{Pois}(\lambda) \). Then, for any \( t > \lambda \) we have
\[
\mathbb{P} \{ X \geq t \} \leq e^{-\lambda} \left( \frac{e \lambda}{t} \right)^t.
\]

In particular, for any \( t \geq e^2 \lambda \) we have
\[
\mathbb{P} \{ X \geq t \} \leq e^{-t}.
\]

How good are these bounds on Poisson tails? Let us compare them with (2.8). Using Stirling’s approximation \( k! \sim \sqrt{2\pi k} (k/e)^k \), we can approximate the Poisson probability mass function as
\[
\mathbb{P} \{ X = k \} \sim \frac{1}{\sqrt{2\pi k}} \cdot e^{-\lambda} \left( \frac{e \lambda}{k} \right)^k.
\] (2.9)

So our bound (2.9) on the entire tail of \( \text{Pois}(\lambda) \) has essentially the same form as the probability of hitting one value \( k \) (the smallest one) in that tail. The difference between these two quantities is the multiple \( \sqrt{2\pi k} \), which is negligible since both these quantities are exponentially small in \( k \).

Exercise 2.3.5. [Difficulty=5] How different are the magnitudes of the tail \( \mathbb{P} \{ X \geq k \} \) and the probability mass function \( \mathbb{P} \{ X = k \} \) for the Poisson distribution? Is our bound \( O(\sqrt{k}) \) on their ratio optimal?

2.3.2 Small deviations

It is sometimes useful to distinguish two regimes for concentration inequalities like \( \mathbb{P} \{ |X - \mu| \geq t \} \). In the large deviation regime, \( t \) is large, usually larger than the mean \( \mu \). In the small deviation regime, \( t \) is small. The exponential bound (2.6) is an example of a large deviation inequality. In the small deviation regime, Chernoff’s inequality is also very useful.
Exercise 2.3.6. [Difficulty=7] Deduce from Chernoff’s inequalities (Theorem 2.3.1 and Exercise 2.3.2) that, for $\delta \in (0, 1]$ we have
\[ P \{|X - \mu| \geq \delta \mu \} \leq 2e^{-c\delta^2}. \]
(Apply those results with $t = (1 + \delta)\mu$ and $t = (1 - \delta)\mu$, respectively, and analyze the bounds for small $\delta$.)

Changing variables to $t = \delta \mu$, we can restate this result so we can see the emergence of normal distribution there.

Corollary 2.3.7 (Chernoff’s inequality: small deviations). In the setting of Theorem 2.3.1, for any $t \in (0, \mu]$ we have
\[ P \{|S_N - \mu| \geq t\} \leq \exp \left(-\frac{ct^2}{\mu}\right). \]
In particular, by the Poisson Limit Theorem 2.3.3, a similar bound holds for $X \sim \text{Pois}(\lambda)$. For any $t \in (0, \lambda]$, we have
\[ P \{|X - \lambda| \geq t\} \leq \exp \left(-\frac{ct^2}{\lambda}\right). \]

The reader probably recognized here the same tails bound as we could derive from (2.2) for the normal distribution $N(\lambda, \sqrt{\lambda})$. This is should come as no surprise once we recall that Poisson distribution admits a normal approximation:

Theorem 2.3.8 (Normal approximation to Poisson). Let $X \sim \text{Pois}(\lambda)$. Then, as $\lambda \to \infty$, we have
\[ \frac{X - \lambda}{\sqrt{\lambda}} \to N(0, 1) \text{ in probability.} \]

This well known fact can be easily derived from Central Limit Theorem, using the fact that the sum of independent Poisson distributions is a Poisson distribution.

Exercise 2.3.9. Give a formal proof of Theorem 2.3.8.

Thus we can view Chernoff inequality for small deviations, Theorem 2.3.7, as a concentration version of the limit Theorem 2.3.8 – just like Hoeffding’s inequality provides a concentration version of the Central Limit Theorem.
2.3.3 Summary

Our findings about the tails of sums of Bernoulli random variables $S_N = \sum X_i$ with $\mathbb{E} S_N = \lambda$, and also for Poisson distribution $\text{Pois}(\lambda)$, can be summarized as follows. In the small deviation regime, in the interval of length $O(\mu)$ around the mean $\mu$, this distribution is similar to Gaussian $N(\lambda, \sqrt{\lambda})$. In the large deviation regime, the tail decay is $(\lambda/t)^t$, which is a bit lighter than exponential but heavier than Gaussian tail. Figure 2.1 illustrates these two tails.

![Figure 2.1: This is a sketch of the probability mass function of Poisson distribution Pois($\lambda$). In the small deviation regime, within $O(\lambda)$ from the mean $\lambda$, the tails of Pois($\lambda$) are like for the normal distribution $N(\lambda, \sqrt{\lambda})$. In the large deviation regime, the tails are heavier and decay like $(\lambda/t)^t$.](image)

2.4 Application: degrees of random graphs

We will give one application of Chernoff's inequality to the classical object in probability: random graphs.

The most thoroughly studied model of random graphs is the classical Erdős-Rényi model $G(n, p)$. A random graph $G \sim G(n, p)$ is constructed on $n$ vertices by connecting every pair of vertices by an edge independently and with probability $p$. Figure 2.2 shows a random graph generated according to Erdős-Rényi model. In applications, $G(n, p)$ with large $n$ often appears as the simplest stochastic model for large real world networks.

The degree of a vertex in the graph is the number of edges incident to it. The expected degree of every vertex in $G(n, p)$ is clearly

$$d := (n - 1)p.$$
Let us show that for relatively dense graphs where \( d \gtrsim \log n \), all degrees tend to concentrate near \( d \).

**Proposition 2.4.1 (Degrees of dense graphs concentrate).** Consider a random graph \( G \sim G(n, p) \) with expected degree satisfying \( d \geq C \log n \). Then, with high probability (for example, 0.9), the following occurs: all vertices of \( G \) have degrees between 0.9\( d \) and 1.1\( d \).

**Proof.** The argument is a combination of Chernoff’s inequality with a union bound. Let us fix a vertex \( i \) of the graph. The degree of \( i \), which we denote \( d_i \), is a sum of \( n \) independent random \( \text{Ber}(p) \) random variables. So we can control this degree using Chernoff’s inequality for small deviations, which we obtained in Exercise 2.3.6. It follows that

\[
\mathbb{P}\{|d_i - d| \geq 0.1d\} \leq 2e^{-cd}.
\]

This bound holds for each fixed vertex \( i \). Next, we can “unfix” it by taking the union bound over all \( n \) vertices. We obtain

\[
\mathbb{P}\{\exists i \leq n : |d_i - d| \geq 0.1d\} \leq \sum_{i=1}^{n} \mathbb{P}\{|d_i - d| \geq 0.1d\} \leq n \cdot 2e^{-cd}.
\]

If \( d \geq C \log n \) for a sufficiently large absolute constant \( C \), the probability is bounded by 0.1. This means that with probability 0.9, the complementary event occurs:

\[
\mathbb{P}\{\forall i \leq n : |d_i - d| < 0.1d\} \geq 0.9.
\]

This completes the proof.
For sparse graphs, those with bounded expected degrees \( d = O(1) \), the degrees do not concentrate about the mean. As we will see by solving the next exercise, such graphs are likely to have vertices with very small and very large degrees. This makes sparse random graphs more difficult to study, but also more interesting than dense graphs. Studying sparse random graphs is important in view of applications, since real world networks also tend to show big variation of degrees.

**Exercise 2.4.2** (Degrees of sparse random graphs do not concentrate). [Difficulty=7] Consider a random graph \( G \sim G(n, p) \) with expected degree satisfying \( d = O(1) \).

1. Show that with high probability (say, 0.9), \( G \) has at least one isolated vertex – a vertex with zero degree.
2. Show that with high probability, (say, 0.9), \( G \) has a vertex with degree larger than \( 10d \). (Here 10 can be replaced by any other constant).

Concerning the second part of Exercise 2.4.2, a more precise analysis reveals that the maximal degree of a sparse random graph \( G \) behaves asymptotically as \( (1 + o(1)) \log(n)/\log \log(n) \). The next exercise establishes an upper bound of this type.

**Exercise 2.4.3** (Degrees of sparse random graphs do not concentrate). [Difficulty=7] Consider a random graph \( G \sim G(n, p) \) with expected degree satisfying \( d = O(1) \). Show that with high probability (say, 0.9), the maximal degree of \( G \) is

\[
O\left(\frac{\log n}{\log \log n}\right).
\]

*(Hint: proceed similarly to the proof of Theorem 2.4.1. Use Chernoff’s inequality for large deviations, Theorem 2.3.1.)*

### 2.5 Sub-gaussian distributions

So far, we studied concentration inequalities that apply for a single class of distributions of \( X_i \), namely the Bernoulli distribution. This significantly limits the range of applications. It is natural to expect that Hoeffding’s inequality, which provides a quantitative view of the Central Limit Theorem, would apply at least to the normal distribution. So, what kind of random variables \( X_i \) can we expect to obey Hoeffding’s inequality in Theorem 2.2.4, namely

\[
\Pr \left\{ \left| \sum_{i=1}^{N} a_i X_i \right| \geq t \right\} \leq 2 \exp \left( -\frac{t^2}{2\|a\|_2^2} \right) ?
\]
Setting a single coefficient \( a_i \) to 1 and the other coefficients to zero, we find that \( X_i \) must have sub-gaussian tails:

\[
P \{ |X_i| > t \} \leq 2e^{-ct^2}.
\]

The class of such distributions, which we call sub-gaussian, deserves special attention. On the one hand, this class is sufficiently wide as it contains Gaussian, Bernoulli and all bounded distributions. On the other hand, as we will see, concentration results like Hoeffding’s inequality can be proved for all sub-gaussian distributions. This makes the class of sub-gaussian distributions a natural, and in many cases the canonical, class where one can develop various results in high dimensional probability theory and its applications.

We will now explore several equivalent approaches to sub-gaussian distributions, exploring the behavior of their tails, moments, and moment generating functions. To pave our way, let us recall how these quantities behave for the standard normal distribution.

Recalling (2.2) and using symmetry, we obtain the following bound on the tails of \( X \sim N(0, 1) \):

\[
P \{ |X| \geq t \} \leq 2e^{-t^2/2} \quad \text{for all } t \geq 0.
\]

(2.10)

In the next exercise, we obtain a bound on the absolute moments and \( L^p \) norms of the normal distribution.

**Exercise 2.5.1** (Moments of the normal distribution). Show that for each \( p \geq 1 \), the random variable \( X \sim N(0, 1) \) satisfies

\[
\|X\|_p = (\mathbb{E}|X|^p)^{1/p} = \sqrt{2} \left\{ \frac{\Gamma((1+p)/2)}{\Gamma(1/2)} \right\}^{1/p}.
\]

Deduce that

\[
\|X\|_p = O(\sqrt{p}) \quad \text{as } p \to \infty.
\]

(2.11)

Finally, the classical formula for the moment generating function of \( X \sim N(0, 1) \) is

\[
\mathbb{E}\exp(\lambda X) = e^{\lambda^2/2} \quad \text{for all } \lambda \in \mathbb{R}.
\]

(2.12)

### 2.5.1 Sub-gaussian properties

Now let \( X \) be a general random variable. The following proposition states that the properties we just considered are equivalent – a sub-gaussian
2.5. SUB-GAUSSIAN DISTRIBUTIONS

Tail decay as in (2.10), the growth of moments as in (2.5.1), and the growth of the moment generating function as in (2.12). The proof of this result is quite useful; it shows how to transform one type of information about random variables into another.

**Proposition 2.5.2 (Sub-gaussian properties).** Let $X$ be a random variable. Then the following properties are equivalent; the parameters $K_i > 0$ appearing in these properties differ from each other by at most an absolute constant factor.  

1. The tails of $X$ satisfy
   \[ P\{|X| \geq t\} \leq 2 \exp(-t^2/K_1^2) \quad \text{for all } t \geq 0. \]

2. The moments of $X$ satisfy
   \[ \|X\|_p = (E |X|^p)^{1/p} \leq K_2 \sqrt{p} \quad \text{for all } p \geq 1. \]

3. The MGF of $X^2$ is finite at some point, namely
   \[ E \exp(X^2/K_3^2) \leq 2. \]

Moreover, if $E X = 0$ then properties 1–3 are also equivalent to the following one.

4. The MGF of $X$ satisfies
   \[ E \exp(\lambda X) \leq \exp(\lambda^2 K_4^2) \quad \text{for all } \lambda \in \mathbb{R}. \]

**Proof.** $1 \Rightarrow 2$. Assume property 1 holds. By homogeneity, rescaling $X$ to $X/K_1$ we can assume that $K_1 = 1$. Applying the Integral Identity (Lemma 1.2.1) for $|X|^p$, we obtain

\[
E |X|^p = \int_0^\infty P\{|X|^p \geq u\} \, du \\
= \int_0^\infty P\{|X| \geq t\} \, pt^{p-1} \, dt \quad \text{(by change of variables } u = t^p) \\
\leq \int_0^\infty 2e^{-t^2} pt^{p-1} \, dt \quad \text{(by property 1)} \\
= p \Gamma(p/2) \quad \text{(set } t^2 = s \text{ and use definition of Gamma function)} \\
\leq p(p/2)^{p/2} \quad \text{(since } \Gamma(x) \leq x^x \text{ holds by Stirling’s approximation).}
\]

The precise meaning of this equivalence is the following. There exists an absolute constant $C$ such that property $i$ implies property $j$ with parameter $K_j \leq CK_i$ for any two properties $i, j = 1, 2, 3, 4$. 

---

\footnote{The precise meaning of this equivalence is the following. There exists an absolute constant $C$ such that property $i$ implies property $j$ with parameter $K_j \leq CK_i$ for any two properties $i, j = 1, 2, 3, 4$.}
Taking the $p$-th root yields property 2 with $K_2 \leq 2$.

2 $\Rightarrow$ 3. Assume property 2 holds. As before, by homogeneity we may assume that $K_2 = 1$. We will prove a bit more general bound than in property 3, namely

$$
\mathbb{E} \exp(\lambda^2 X^2) \leq \exp(K_2^2 \lambda^2) \quad \text{for all } \lambda \text{ satisfying } |\lambda| \leq \frac{1}{K_3}. \quad (2.13)
$$

Expanding the exponential function in Taylor series, we obtain

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

Property 2 guarantees that $\mathbb{E}[X^{2p}] \leq (2p)^p$, while Stirling’s approximation yields $p! \geq (p/e)^p$. Substituting these two bounds, we obtain

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

provided $2e\lambda^2 < 1$, in which case the geometric series above converges. To bound this quantity further, we can use the numeric inequality $\frac{1}{1-x} \leq e^{2x}$, which is valid for $x \in [0, 1/2]$. It follows that

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

This implies (2.13), a stronger version of property 3, with $K_3 = 1/2\sqrt{e}$.

3 $\Rightarrow$ 1. Assume property 3 holds. As before, we may assume that $K_3 = 1$. Then

$$
\mathbb{P}\{|X| > t\} = \mathbb{P}\{e^{X^2} \geq e^{t^2}\}
\leq e^{-t^2} \mathbb{E} e^{X^2} \quad (\text{by Markov’s inequality, Proposition 1.2.3})
\leq 2e^{-t^2} \quad (\text{by property 3}).
$$

This proves property 1 with $K_1 = 1$.

To prove the second part of the proposition, we will show that 4 $\Rightarrow$ 1 and 3 $\Rightarrow$ 4.

3 $\Rightarrow$ 4. Assume that property 3 holds; as before we can assume that $K_3 = 1$. Let us use the numeric inequality $e^x \leq x + e^{x^2}$, which is valid for all $x$. Then

$$
\mathbb{E} e^{\lambda X} \leq \mathbb{E} \left[ \lambda X + e^{\lambda^2 X^2} \right] = \mathbb{E} e^{\lambda^2 X^2},
$$
2.5. SUB-GAUSSIAN DISTRIBUTIONS

where we used the assumption that $\mathbb{E} X = 0$. Next, we apply the general form of property 3 obtained in (2.13), which states that

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

if $|\lambda| \leq 1$.

Thus we have proved property 4 in the range $|\lambda| \leq 1$.

Now let $|\lambda| \geq 1$. Here we can use the numeric inequality $\lambda X \leq \lambda^2 + X^2$, which is valid for all $\lambda$ and $X$. It follows that

$$\mathbb{E} e^{\lambda X} \leq e^{\lambda^2} \mathbb{E} e^{X^2} \leq 2e^{\lambda^2} \quad \text{(by property 3)}$$

$$\leq e^{2\lambda^2} \quad \text{(since $|\lambda| \geq 1$)}.$$ 

This proves property 4 with $K_4 = \sqrt{2}$.

4. $\Rightarrow$ 1. Assume property 4 holds; we can assume that $K_4 = 1$. We will use some ideas from the proof of Hoeffding’s inequality (Theorem 2.2.2). Let $\lambda > 0$ be a parameter to be chosen later. Then

$$\mathbb{P}\{X \geq t\} = \mathbb{P}\{e^{\lambda X} \geq e^{\lambda t}\}$$

$$\leq e^{-\lambda t} \mathbb{E} e^{\lambda X} \quad \text{(by Markov’s inequality)}$$

$$\leq e^{-\lambda t} e^{\lambda^2} \quad \text{(by property 4)}$$

$$= e^{-\lambda t + \lambda^2}.$$ 

Optimizing in $\lambda$ and thus choosing $\lambda = t/2$, we conclude that

$$\mathbb{P}\{X \geq t\} \leq e^{-t^2/4}.$$ 

Repeating this argument for $-X$, we also obtain $\mathbb{P}\{X \leq -t\} \leq e^{-t^2/4}$. Combining these two bounds we conclude that

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

Thus property 1 holds with $K_1 = 2$. The proposition is proved. $\square$

The constant 2 in properties 1 and 3 does not have any special meaning; they can be replaced by other absolute constants. (Why?)

Exercise 2.5.3. [Difficulty=3] Show that the condition $\mathbb{E} X = 0$ is necessary for property 4 to hold.

Exercise 2.5.4. [Difficulty=3] Property 3 and especially its stronger form (2.13) state that the MGF of $X^2$ is bounded in some constant neighborhood of zero. Show that for $X \sim N(0,1)$, the MGF of $X^2$ is infinite outside a constant neighborhood of zero.
2.5.2 Definition and examples of sub-gaussian distributions

**Definition 2.5.5** (Sub-gaussian random variables). A random variable $X$ that satisfies one of the equivalent properties 1 – 3 in Proposition 2.5.2 is called a sub-gaussian random variable. The sub-gaussian norm of $X$, denoted $\|X\|_{\psi_2}$, is defined to be the smallest $K_3$ in property 3. In other words,

$$\|X\|_{\psi_2} = \inf \{ t > 0 : E \exp(X^2/t^2) \leq 2 \}.$$  \hfill (2.14)

Proposition 2.5.2 states that every sub-gaussian random variable $X$ satisfies:

$$P\{|X| \geq t\} \leq 2 \exp(-ct^2/\|X\|_{\psi_2}^2) \text{ for all } t \geq 0;$$ \hfill (2.15)

$$\|X\|_p \leq C\|X\|_{\psi_2} \sqrt{p} \text{ for all } p \geq 1;$$ \hfill (2.16)

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

if $E X = 0$ then $E \exp(\lambda X) \leq \exp(C\lambda^2\|X\|_{\psi_2}^2)$ for all $\lambda \in \mathbb{R}$. \hfill (2.17)

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

**Example 2.5.6.** Classical examples of sub-gaussian random variables are Gaussian, Bernoulli and all bounded random variables.

1. (Gaussian): As we already noted, $X \sim N(0, 1)$ is a sub-gaussian random variable with $\|X\|_{\psi_2} \leq C$, where $C$ is an absolute constant. More generally, if $X$ is a centered normal random variable with variance $\sigma^2$, then $X$ is sub-gaussian with $\|X\|_{\psi_2} \leq C\sigma$. (Why?)

2. (Bernoulli): Consider a random variable $X$ with symmetric Bernoulli distribution (see Definition 2.2.1). Since $|X| = 1$, it follows that $X$ is a sub-gaussian random variable with $\|X\|_{\psi_2} = 1/\ln 2$.

3. (Bounded): More generally, any bounded random variable $X$ is sub-gaussian with

$$\|X\|_{\psi_2} \leq C\|X\|_\infty,$$ \hfill (2.18)

where $C = 1/\ln 2$.

**Exercise 2.5.7.** Show that Poisson, exponential, Pareto and Cauchy distributions are not sub-gaussian.
2.6. GENERAL HOEFFDING’S AND KHINCHINE’S INEQUALITIES

2.5.3 A more general view via Orlicz spaces

Sub-gaussian random variables can be included in the more general framework of Orlicz spaces. A function $\psi : [0, \infty) \to [0, \infty)$ is called an Orlicz function if $\psi$ is a convex, increasing, and satisfies

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

For a given Orlicz function $\psi$, the Orlicz norm of a random variable $X$ is defined as

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

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

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

Exercise 2.5.8. Show that $\|X\|_\psi$ is indeed a norm on the space $L_\psi$.

Thus $L_\psi$ is a normed space; it can also be shown that it is a Banach space.

To give a few concrete examples, consider the function $\psi(x) = x^p$, which is convex if $p \geq 1$. The resulting Orlicz space $L_\psi$ is the classical space $L_p$.

Another example is for the Orlicz function

$$\psi_2(x) := e^{x^2} - 1.$$ 

The resulting Orlicz norm is exactly the sub-gaussian norm $\| \cdot \|_{\psi_2}$ that we defined in (2.14), and Orlicz space $L_{\psi_2}$ consists of all sub-gaussian random variables.

Property 2 of Proposition 2.5.2 and (2.18) determine the place of the space $L_{\psi_2}$ of sub-gaussian random variables in the hierarchy of $L_p$ spaces:

$$L_\infty \subset L_{\psi_2} \subset L_p \quad \text{for every } p \geq 1.$$ 

Thus the space of sub-gaussian random variables $L_{\psi_2}$ is smaller than all $L_p$ spaces, but it is still larger than their limit, the space of bounded random variables $L_\infty$.

2.6 General Hoeffding’s and Khinchine’s inequalities

After all the work we did characterizing sub-gaussian distributions in the previous section, we can now easily extend Hoeffding’s inequality (Theorem 2.2.2) to general sub-gaussian distributions. But before we do this, let
us deduce an important and enlightening property of the sums of independent sub-gaussians, from which a general form of Hoeffding’s inequality will immediately follow.

In the first probability course, we learned that a sum of independent normal random variables $X_i$ is normal. Indeed, if $X_i \sim N(0, \sigma_i^2)$ are independent then

$$\sum_{i=1}^{N} X_i \sim N\left(0, \sum_{i=1}^{N} \sigma_i^2\right).$$  \hfill (2.19)

This fact follows from the rotation invariance of the normal distribution, see Section 3.3.2 below. Let us show that this property extends to general sub-gaussian distributions, albeit up to an absolute constant.

**Proposition 2.6.1** (Sums of independent sub-gaussians). Let $X_1, \ldots, X_N$ be independent, mean zero, sub-gaussian random variables. Then $\sum_{i=1}^{N} X_i$ is also a sub-gaussian random variable, and

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

where $C$ is an absolute constant.

**Proof.** Let us analyze the moment generating function of the sum. For any $\lambda \in \mathbb{R}$, we have

$$\mathbb{E} \exp \left( \lambda \sum_{i=1}^{N} X_i \right) = \prod_{i=1}^{N} \mathbb{E} \exp(\lambda X_i) \quad \text{by independence}$$

$$\leq \prod_{i=1}^{N} \exp\left( C\lambda^2 \|X_i\|_\psi^2 \right) \quad \text{by sub-gaussian property (2.17)}$$

$$= \exp(\lambda^2 K^2) \quad \text{where } K^2 := C \sum_{i=1}^{N} \|X_i\|_\psi^2.$$  

It remains to recall that this property of MGF characterizes sub-gaussian distributions. Indeed, the equivalence of properties 2 and 4 in Proposition 2.5.2 and Definition 2.5.5 imply that the sum $\sum_{i=1}^{N} X_i$ is sub-gaussian, and

$$\left\| \sum_{i=1}^{N} X_i \right\|_\psi \leq C_1 K$$

where $C_1$ is an absolute constant. The proof is complete. \qed
We can restate this result as a concentration inequality.

**Theorem 2.6.2** (General Hoeffding’s inequality). Let \( X_1, \ldots, X_N \) be independent, mean zero, sub-gaussian random variables. Then, for every \( t \geq 0 \), we have

\[
P\left\{ \left| \sum_{i=1}^{N} X_i \right| \geq t \right\} \leq 2 \exp \left( - \frac{ct^2}{\sum_{i=1}^{N} \|X_i\|_{\psi_2}^2} \right).
\]

To compare this general result with a specific case for Bernoulli distributions (Theorem 2.2.2), let us apply Theorem 2.6.3 for \( a_i X_i \) instead of \( X_i \).

**Theorem 2.6.3** (General Hoeffding’s inequality). Let \( X_1, \ldots, X_N \) be independent, mean zero, sub-gaussian random variables, and \( a = (a_1, \ldots, a_N) \in \mathbb{R}^N \). Then, for every \( t \geq 0 \), we have

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

where \( K = \max_i \|X_i\|_{\psi_2} \).

### 2.6.1 Khinchine’s inequality

As an application of sub-gaussian Hoeffding’s inequality, we can derive the so-called Khinchine’s inequality for the \( L_p \)-norms of sums of independent random variables. It is usually stated for symmetric Bernoulli random variables, but we can prove it for general sub-gaussian distributions with no extra work.

**Corollary 2.6.4** (Khinchine’s inequality). Let \( X_1, \ldots, X_N \) be independent sub-gaussian random variables with zero means and unit variances, and let \( a = (a_1, \ldots, a_N) \in \mathbb{R}^N \). Then, for every \( p \geq 2 \) we have

\[
\left( \sum_{i=1}^{N} a_i^2 \right)^{1/2} \leq \left\| \sum_{i=1}^{N} a_i X_i \right\|_p \leq CK \sqrt{p} \left( \sum_{i=1}^{N} a_i^2 \right)^{1/2}
\]

where \( K = \max_i \|X_i\|_{\psi_2} \) and \( C \) is an absolute constant.

**Proof.** Hoeffding’s inequality in (?) yields that the sum is sub-gaussian. Then (2.16) bounds the growth of the moments of the sum. This proves the upper bound on the \( L_p \) norm.
To obtain the lower bound, we bound
\[ \left\| \sum_{i=1}^{N} a_i X_i \right\|_p \geq \left\| \sum_{i=1}^{N} a_i X_i \right\|_2 \] (since \( p \geq 2 \))
\[ = \left[ \text{Var} \left( \sum_{i=1}^{N} a_i X_i \right) \right]^{1/2} \] (since \( \mathbb{E} X_i = 0 \))
\[ = \left( \sum_{i=1}^{N} \text{Var}(a_i X_i) \right)^{1/2} \] (by independence)
\[ = \left( \sum_{i=1}^{N} a_i^2 \right)^{1/2} \] (since \( \text{Var}(X_i) = 1 \)).

This completes the proof. \( \square \)

**Exercise 2.6.5.** Show that a version of Khinchine’s inequality holds also for all \( p \in (0, 2) \). In this case, an absolute constant factor will appear in the left hand side and not in the right hand side.

Provide a hint.

### 2.6.2 Centering

In results like Hoeffding’s inequality, and in many results in the future, we typically assume that the random variables \( X_i \) have zero means. If this is not the case, we can always center \( X_i \) by subtracting the mean. Centering does not harm the sub-gaussian property. Let us check this carefully.

One can quickly check the centering inequality for the \( L_2 \) norm, namely
\[ \| X - \mathbb{E} X \|_2 \leq \| X \|_2. \]
(Do this!) The next lemma provides a similar centering inequality for the sub-gaussian norm.

**Lemma 2.6.6 (Centering).** If \( X \) is a sub-gaussian random variable then \( X - \mathbb{E} X \) is sub-gaussian, too, and
\[ \| X - \mathbb{E} X \|_{\psi_2} \leq C \| X \|_{\psi_2}. \]

**Proof.** Recall that \( \| \cdot \|_{\psi_2} \) is a norm, so triangle inequality yields
\[ \| X - \mathbb{E} X \|_{\psi_2} \leq \| X \|_{\psi_2} + \| \mathbb{E} X \|_{\psi_2}. \] (2.20)
2.7. SUB-EXPONENTIAL DISTRIBUTIONS

Let us bound the second term. By (2.14), for any constant random variable $a$ we trivially have $\|a\|_{\psi_2} = C_1 |a|$ where $C_1 = 1/\ln 2$. Thus

$$\|\mathbb{E} X\|_{\psi_2} = C_1 |\mathbb{E} X|$$

$$\leq C_1 \mathbb{E} |X| \quad \text{(by Jensen’s inequality)}$$

$$= C_1 \|X\|_1$$

$$\leq C_2 \|X\|_{\psi_2} \quad \text{(by sub-gaussian moment property (2.16))}.$$  

Substituting this into (2.20), we complete the proof. $\square$

**Exercise 2.6.7** (Difficulty=8). What is the optimal constant $C$ in Lemma 2.6.6? Does it hold with $C = 1$?

**Exercise 2.6.8.** Using Theorem 2.6.3 and centering, deduce general Hoeffding’s inequality for general bounded random variables, Theorem 2.2.5, possibly with some absolute constant instead of 2 in the exponent.

2.7 Sub-exponential distributions

The class of sub-gaussian distributions is natural and quite wide. Nevertheless, it leaves out some important distributions whose tails are heavier than gaussian. Here is an example. Consider a standard normal random vector $g = (g_1, \ldots, g_N)$ in $\mathbb{R}^N$, whose coordinates $g_i$ are independent $N(0, 1)$ random variables. It is useful in many applications to have a concentration inequality for the Euclidean norm of $g$,

$$\|g\|_2 = \left(\sum_{i=1}^N g_i^2\right)^{1/2}.$$  

Here we find ourselves in a strange situation. On the one hand, $\|g\|_2^2$ is a sum of independent random variables $g_i^2$, so we should expect some concentration to hold. On the other hand, although $g_i$ are sub-gaussian random variables, $g_i^2$ are not. Indeed, recalling the behavior of Gaussian tails (Proposition 2.1.2) we have$^2$

$$\mathbb{P}\{g_i^2 > t\} = \mathbb{P}\{|g| > \sqrt{t}\} \sim \exp\left(-\frac{(\sqrt{t})^2}{2}\right) = \exp(-t/2).$$

The tails of $g_i^2$ are like for the exponential distribution, and are strictly heavier than sub-gaussian.

$^2$Here we ignored the pre-factor $1/t$, which does not make much effect on the exponent.
In this section we will focus on the class of distributions that have at least an exponential tail decay; we will call sub-exponential distributions. Our analysis here will be quite similar to what we did for sub-gaussian distributions in Section 2.5, and we will leave some details to the reader. In particular, it is not difficult to prove a version of Proposition 2.5.2 for sub-exponential (and more general) distributions.

Proposition 2.7.1 (Sub-exponential properties). Let $X$ be a random variable. Then the following properties are equivalent; the parameters $K_i > 0$ appearing in these properties differ from each other by at most an absolute constant factor.\footnote{The precise meaning of this equivalence is the following. There exists an absolute constant $C$ such that property $i$ implies property $j$ with parameter $K_j \leq CK_i$ for any two properties $i, j = 1, 2, 3, 4$.}

1. The tails of $X$ satisfy
   \[ \mathbb{P}\{|X| \geq t\} \leq 2 \exp(-t/K_1) \quad \text{for all } t \geq 0. \]

2. The moments of $X$ satisfy
   \[ \|X\|_p = (\mathbb{E}|X|^p)^{1/p} \leq K_2 p \quad \text{for all } p \geq 1. \]

3. The MGF of $|X|$ is finite at some point, namely
   \[ \mathbb{E}\exp(|X|/K_3) \leq 2. \]

Exercise 2.7.2 (Difficulty=5). 1. Prove Proposition 2.7.1 by modifying the proof of Proposition 2.5.2.

2. More generally, consider the class of distributions whose tail decay is of the type $\exp(-ct^\alpha)$ or faster. Here $\alpha = 2$ corresponds to sub-gaussian distributions, and $\alpha = 1$, to sub-exponential. State and prove a version of Proposition 2.7.1 for such distributions.

Definition 2.7.3 (Sub-exponential random variables). A random variable $X$ that satisfies one of the equivalent properties 1 – 3 in Proposition 2.7.1 is called a sub-exponential random variable. The sub-exponential norm of $X$, denoted $\|X\|_\psi_1$, is defined to be the smallest $K_3$ in property 3. In other words,

\[ \|X\|_\psi_1 = \inf \{ t > 0 : \mathbb{E}\exp(|X|/t) \leq 2 \}. \]
2.7. SUB-EXPONENTIAL DISTRIBUTIONS

Sub-gaussian and sub-exponential distributions are closely related. First, any sub-gaussian distribution is clearly sub-exponential. (Why?) Second, the square of a sub-gaussian random variable is sub-exponential:

**Lemma 2.7.4** (Sub-exponential is sub-gaussian squared). A random variable $X$ is sub-gaussian if and only if $X^2$ is sub-exponential. Moreover,

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

_Proof._ This follows easily from the definition. Indeed, $\|X^2\|_{\psi_1}$ is the infimum of the numbers $K > 0$ satisfying $\mathbb{E}\exp(X^2/K) \leq 2$, while $\|X\|_{\psi_2}$ is the infimum of the numbers $L > 0$ satisfying $\mathbb{E}\exp(X^2/L^2) \leq 2$. So these two become the same definition with $K = L^2$. \qed

More generally, the product of two sub-gaussian random variables is sub-exponential:

**Lemma 2.7.5** (Product of sub-gaussians is sub-exponential). Let $X$ and $Y$ be sub-gaussian random variables. Then $XY$ is sub-exponential. Moreover,

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

_Proof._ Denote $\|X\|_{\psi_2} = K$ and $\|Y\|_{\psi_2} = L$. The lemma claims that $\mathbb{E}\exp(|XY|/KL) \leq 2$. To prove this, let us use Young’s inequality

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

It yields

$$\mathbb{E}\exp\left(\frac{|XY|}{KL}\right) \leq \mathbb{E}\exp\left(\frac{X^2}{2K^2} + \frac{Y^2}{2L^2}\right)$$

$$= \mathbb{E}\left[\exp\left(\frac{X^2}{2K^2}\right) \exp\left(\frac{Y^2}{2L^2}\right)\right]$$

$$\leq \frac{1}{2} \mathbb{E}\left[\exp\left(\frac{X^2}{K^2}\right) + \exp\left(\frac{Y^2}{L^2}\right)\right] \text{ (by Young’s inequality)}$$

$$= \frac{1}{2} (2 + 2) = 2 \text{ (by definition of } K \text{ and } L).$$

The proof is complete. \qed

**Example 2.7.6.** Let us mention some natural examples of sub-exponential random variables. As we just learned, all sub-gaussian random variables and
their squares are sub-exponential, for example $g^2$ for $g \sim N(\mu, \sigma)$. Apart from that, sub-exponential distributions include the exponential and Poisson distributions. Recall that $X$ has exponential distribution with rate $\lambda > 0$, denoted $X \sim \text{Exp}(\lambda)$, if $X$ is a non-negative random variable with tails

$$
P\{X \geq t\} = e^{-\lambda t} \quad \text{for} \quad t \geq 0.
$$

The mean, standard deviation, and the sub-exponential norm of $X$ are all of order $1/\lambda$:

$$
\mathbb{E} X = \frac{1}{\lambda}, \quad \text{Var}(X) = \frac{1}{\lambda^2}, \quad \|X\|_{\psi_1} = \frac{C}{\lambda}.
$$

(Check this!)

### 2.8 Bernstein’s inequality

Our next goal is to prove a concentration inequality for sums of sub-exponential random variables. Just like in the proof of the previous concentration inequalities – Hoeffding’s and Chernoff’s – our argument will be based on the moment generating function. We will have to be a little more careful now, since the tails of sub-exponential distributions may not be light enough to make the MGF finite everywhere.

Indeed, consider the exponential random variable $X \sim \text{Exp}(1)$. A simple calculation shows that the MGF of the centered random variable $Z = X - \mathbb{E} X$ equals

$$
\mathbb{E} \exp(\lambda Z) = \frac{e^{-\lambda}}{1 - \lambda} \quad \text{for} \quad \lambda < 1
$$

and the MGF is infinite for $\lambda \geq 1$. (Check this!) More generally, the MGF of a sub-exponential distribution is always finite in some constant neighborhood of zero, and is similar there to the sub-gaussian MGF which we analyzed in (2.17). Let us first show this fact, and afterwards deduce from it a concentration inequality.

**Lemma 2.8.1** (MGF of sub-exponential distributions). Let $X$ be a mean zero, sub-exponential random variable. Then, for $\lambda$ such that $|\lambda| \leq c/\|X\|_{\psi_1}$, one has

$$
\mathbb{E} \exp(\lambda X) \leq \exp(C\lambda^2\|X\|_{\psi_1}^2).
$$
2.8. BERNSTEIN’S INEQUALITY

Proof. Without loss of generality we may assume that \( \|X\|_{\psi_1} = 1 \). (Why?) by replacing \( X \) with \( X/\|X\|_{\psi_1} \) and \( t \) with \( t\|X\|_{\psi_1} \). Expanding the exponential function in Taylor series, we obtain

\[
\mathbb{E} \exp(\lambda X) = \mathbb{E} \left[ 1 + \lambda X + \sum_{p=1}^{\infty} \frac{(\lambda X)^p}{p!} \right] = 1 + \sum_{p=1}^{\infty} \frac{\lambda^p \mathbb{E}[X^p]}{p!},
\]

where we used the assumption that \( \mathbb{E} X = 0 \). Property 2 in Proposition 2.7.1 guarantees that \( \mathbb{E}[X^p] \leq (Cp)^p \). (This is because \( K_2 \leq CK_3 = \|X\|_{\psi_1} = 1 \).) Moreover, Stirling’s approximation yields \( p! \geq (p/e)^p \). Substituting these two bounds, we obtain

\[
\mathbb{E} \exp(\lambda X) \leq 1 + \sum_{p=2}^{\infty} \frac{(C\lambda p)^p}{(p/e)^p} = 1 + \sum_{p=2}^{\infty} (C_1\lambda)^p.
\]

If \( |C_1\lambda| < 1/2 \), the geometric series converges and is dominated by the first term, so

\[
\mathbb{E} \exp(\lambda X) \leq 1 + 2(C_1\lambda)^2 \leq \exp(2(C_1\lambda)^2).
\]

This completes the proof. \( \square \)

Now we are ready to state and prove a concentration inequality for sums of independent sub-gaussian random variables.

**Theorem 2.8.2** (Bernstein’s inequality). Let \( X_1, \ldots, X_N \) be independent, mean zero, sub-exponential random variables. Then, for every \( t \geq 0 \), we have

\[
P \left\{ \left| \sum_{i=1}^{N} X_i \right| \geq t \right\} \leq 2 \exp \left[ -c \min \left( \frac{t^2}{\sum_{i=1}^{N} \|X_i\|_{\psi_1}^2}, \frac{t}{\max_{i} \|X_i\|_{\psi_1}} \right) \right].
\]

Proof. Without loss of generality, we assume that \( K = 1 \). (Why?) As in our proofs of some previous concentration inequalities for \( S_N = \sum_{i=1}^{N} X_i \), (e.g. Theorems 2.2.2 and 2.3.1), we multiply both sides of the inequality \( S_N \geq t \) by a parameter \( \lambda \), exponentiate, and then use Markov’s inequality and independence. This leads to the bound (2.7), which is

\[
P \{ S_N \geq t \} \leq e^{-\lambda t} \prod_{i=1}^{N} \mathbb{E} \exp(\lambda X_i).
\]  \( (2.22) \)
Lemma 2.8.1 can bound the MGF of each term $X_i$. If we choose $\lambda$ small enough so that

$$|\lambda| \leq \frac{c}{\max_i \|X_i\|_{\psi_1}}, \quad (2.23)$$

Then Lemma 2.8.1 yields $E \exp(\lambda X_i) \leq \exp \left( C \lambda^2 \|X_i\|_{\psi_1}^2 \right)$. Substituting this into (2.22), we obtain

$$P \left\{ S \geq t \right\} \leq \exp \left( -\lambda t + C \lambda^2 \sigma^2 \right), \quad \text{where} \quad \sigma^2 = \sum_{i=1}^{N} \|X_i\|_{\psi_1}^2.$$

Now we minimize this expression in $\lambda$ subject to the constraint (2.23). The optimal choice is $\lambda = \min(t/2C\sigma^2, c/\max_i \|X_i\|_{\psi_1})$, for which we obtain

$$P \left\{ S \geq t \right\} \leq \exp \left[ - \min \left( \frac{t^2}{4C\sigma^2}, \frac{ct}{2\max_i \|X_i\|_{\psi_1}} \right) \right].$$

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

To put Theorem 2.8.2 in a more convenient form, let us apply it for $a_i X_i$ instead of $X_i$.

**Theorem 2.8.3** (Bernstein’s inequality). Let $X_1, \ldots, X_N$ be independent, mean zero, sub-exponential random variables, and $a = (a_1, \ldots, a_N) \in \mathbb{R}^N$. Then, for every $t \geq 0$, we have

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

where $K = \max_i \|X_i\|_{\psi_1}$.

Let us state Bernstein’s inequality as a quantitative form of the Law of Large Numbers.

**Corollary 2.8.4** (Bernstein’s inequality: LLN). Let $X_1, \ldots, X_N$ be independent, mean zero, sub-exponential random variables. Then, for every $t \geq 0$, we have

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

where $K = \max_i \|X_i\|_{\psi_1}$. 
2.8. BERNSTEIN’S INEQUALITY

2.8.1 Summary

Let us compare Bernstein’s inequality (Theorem 2.8.2) with Hoeffding’s inequality (Theorem 2.6.2). The obvious difference is that Bernstein’s bound has two tails, as if the sum $S_N = \sum X_i$ were a mixture of sub-gaussian and sub-exponential distributions. The sub-gaussian tail is of course expected from the Central Limit Theorem. But the sub-exponential tails of the terms $X_i$ are too heavy to be able to produce a sub-gaussian tail everywhere, so the sub-exponential tail should be expected, too. In fact, the sub-exponential tail in Theorem 2.8.2 is produced by a single term $X_i$ in the sum, the one with the maximal sub-exponential norm. Indeed, this term alone has the tail of magnitude $\exp (-ct/\|X_i\|_{\psi_1})$.

We already saw a similar mixture of two tails, one for small deviations and the other for large deviations, in our analysis of Chernoff’s inequality (2.3.3). To put Bernstein’s inequality in the same perspective, let us normalize the sum as in the Central Limit Theorem and apply Theorem 2.8.3. We obtain

$$\P\left\{ \left| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} X_i \right| \geq t \right\} \leq \begin{cases} 2 \exp \left( -\frac{ct^2}{K^2} \right), & t \leq K\sqrt{N} \\ 2 \exp \left( -\frac{t}{t\sqrt{N}} \right), & t \geq K\sqrt{N} \end{cases}$$

(2.24)

where $K = \max_i \|X_i\|_{\psi_1}$ as before. Thus, in the small deviation regime where $t \leq K\sqrt{N}$, we have a sub-gaussian tail bound as if the sum had normal distribution $N(0,K)$. Note that this domain widens as $N$ increases and the Central Limit Theorem becomes more powerful. For large deviations where $t \geq K\sqrt{N}$, the sum has a heavier, sub-exponential tail bound, which can be explained by contribution of a single term $X_i$. We illustrate this in Figure 2.3.

Exercise 2.8.5 (Centering). Prove an analog of Centering Lemma 2.6.6 for sub-exponential random variables $X$:

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

Bernstein inequality can be made a bit stronger if we assume that the random variables $X_i$ are bounded, as opposed to sub-exponential.

Theorem 2.8.6 (Bernstein’s inequality for bounded distributions). Let $X_1, \ldots, X_N$ be independent, mean zero, bounded random variables. Then, for every $t \geq 0$, we have

$$\P\left\{ \left| \sum_{i=1}^{N} X_i \right| \geq t \right\} \leq 2 \exp \left( -\frac{-t^2/2}{\sigma^2 + CKt} \right).$$
Figure 2.3: Bernstein’s inequality for a sum of sub-exponential random variables $N^{-1/2} \sum_{i=1}^{N} X_i$ is a mixture of two tails, sub-gaussian for small deviations and sub-exponential for large deviations. The sub-gaussian tail emerges in the $O(\sqrt{N})$ neighborhood of zero, and it can be explained by Central Limit Theorem. The heavier, sub-exponential tail, is produced by a single term in the sum.

Here $K = \max_i \|X_i\|_\infty$.

**Exercise 2.8.7.** Prove Theorem 2.8.2. Deduce it by first proving the following version of Lemma 2.8.1. If $\|X\|_\infty \leq K$ then

$$
\mathbb{E} \exp(\lambda X) \leq \exp(g(\lambda)\sigma^2) \quad \text{where} \quad g(\lambda) = \frac{\lambda^2/2}{1 - CK\lambda}.
$$
Chapter 3

Random vectors in high dimensions

Why random vectors? Example: practitioners working with genetic data study the expressions of \( n \approx 60,000 \) genes in the human body. To study patterns in such genetic data in a given population, we can form a random vector \( X = (X_1, \ldots, X_n) \) by choosing a random person from the population and recording the expressions of his or her \( n \) genes.

Questions: are certain genes related to each other? How many people should be sampled from the population to see such relationships?

Talk about lots of room (volume) in higher dimensions. This leads to “the curse of high dimensions”. But probability can turn it into a blessing.

3.1 Concentration of the norm

Where in the space a random vector is likely to be located? Consider a random vector \( X = (X_1, \ldots, X_n) \) whose coordinates \( X_i \) are independent random variables with zero means and unit variances. What length do we expect \( X \) to have? Let us compute the expectation

\[
E \|X\|_2^2 = E \sum_{i=1}^{N} X_i^2 = \sum_{i=1}^{N} E X_i^2 = n.
\]

So we should expect that the length of \( X \) is

\[
\|X\|_2 \approx \sqrt{n}.
\]

We will see now that \( X \) is indeed very close to \( \sqrt{n} \) with high probability.
CHAPTER 3. RANDOM VECTORS IN HIGH DIMENSIONS

Theorem 3.1.1 (Concentration of the norm). Let \( X = (X_1, \ldots, X_n) \in \mathbb{R}^n \) 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 \( K = \max_i \|X_i\|_{\psi_2} \).

Proof. For simplicity, we will assume that \( K \geq 1 \). (Argue that you can make this assumption.) We shall apply Bernstein’s deviation inequality for the normalized sum of independent random variables

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

The terms \( X_i^2 - 1 \) are mean zero random variables with zero means. Since \( X_i \) are sub-gaussian, \( X_i^2 - 1 \) are sub-exponential; more precisely

\[
\|X_i^2 - 1\|_{\psi_1} \leq C\|X_i^2\|_{\psi_1} \quad \text{by Centering Lemma 2.6.6}
\]

\[
= C\|X_i\|_{\psi_2}^2 \quad \text{by Lemma 2.7.4}
\]

\[
\leq CK^2.
\]

Applying Bernstein’s inequality (Corollary 2.8.4), we obtain for any \( u \geq 0 \) that

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

(3.1)

(Here we used that \( K^4 \geq K^2 \) since \( K \) is bounded below by an absolute constant – why?)

To deduce from this a concentration inequality for \( \frac{1}{\sqrt{n}} \|X\|_2 - 1 \), we will use the following elementary observation

\[
|z - 1| \geq \delta \quad \text{implies} \quad \left| z^2 - 1 \right| \geq \max(\delta, \delta^2) \quad \text{for} \quad z \geq 0.
\]

Using this for \( z = \frac{1}{\sqrt{n}} \|X\|_2 \) together with (3.1) where \( u = \max(\delta, \delta^2) \), we obtain for any \( \delta \geq 0 \) that

\[
\mathbb{P} \left\{ \left| \frac{1}{\sqrt{n}} \|X\|_2 - 1 \right| \geq \delta \right\} \leq \mathbb{P} \left\{ \left| \frac{1}{n} \|X\|_2^2 - 1 \right| \geq \max(\delta, \delta^2) \right\}
\]

\[
\leq 2 \exp \left( -\frac{cn}{K^4} \cdot \delta^2 \right).
\]
3.1. CONCENTRATION OF THE NORM

Changing variables to $t = \delta \sqrt{n}$, we obtain the desired sub-gaussian tail

$$\mathbb{P}\{\|X\|_2 - \sqrt{n} > t\} \leq 2 \exp\left(-\frac{ct^2}{K^4}\right) \quad \text{for all } t \geq 0.$$  

The proof is complete. \hfill \Box

Is is convenient to restate Theorem 3.1.1 as the following concentration inequality, which is valid for all $t \geq 0$:

$$\mathbb{P}\{\|X\|_2 - \sqrt{n} \geq t\} \leq 2 \exp\left(-\frac{ct^2}{K^4}\right). \quad (3.2)$$

This inequality says that with high probability $X$ is located very close to the sphere of radius $\sqrt{n}$. Most of the time, $X$ even stays within constant distance from that big sphere.

Such small, constant, deviation could be surprising at the first sight. Let us explain this intuitively. The square of the norm, $S_N := \|X\|_2^2$, is a sum of $n$ independent, mean zero random variables. Not surprisingly, $S_N$ has mean $n$ and standard deviation of order $\sqrt{n}$, thus behaving exactly as we would expect of a sum. (Check this!) Now, the norm $\|X\|_2$ is the square root of $S_N$. So, as $S_n$ deviates by $O(\sqrt{n})$ from its mean $n$, the square root $\|X\|_2 = \sqrt{S_n}$ ought to have constant deviation around $\sqrt{n}$. This is because

$$\sqrt{n} \pm O(\sqrt{n}) = \sqrt{n} \pm O(1),$$

see Figure 3.1 for illustration.

![Figure 3.1](image.png)

Figure 3.1: Concentration of the norm of a random vector $X$ in $\mathbb{R}^n$. When $\|X\|_2^2$ deviates by $O(\sqrt{n})$ around $n$, the square root of this quantity, $\|X\|_2$, deviates by $O(1)$ around $\sqrt{n}$.

**Question 3.1.2.** Is the quadratic dependence on $K$ in Theorem 3.1.1 optimal? Can it be improved to the linear dependence?
Theorem 3.1.1 and Centering Lemma 2.6.6 imply the following concentration inequality of a random vector about its mean. Let $X = (X_1, \ldots, X_n) \in \mathbb{R}^n$ be a random vector with independent, sub-gaussian coordinates $X_i$ that satisfy $\text{Var}(X_i^2) = 1$. Then
\[
\left\| X - \mathbb{E}X \right\|_2 - \sqrt{n} \leq CK^2
\]
where $K = \max_i \|X_i\|_{\psi_2}$. To see this, apply Theorem 3.1.1 for $X - \mathbb{E}X$ instead of $X$, and use that $\|X_i - \mathbb{E}X_i\|_{\psi_2} \leq CK$ by centering.

**Exercise 3.1.3.** [Difficulty=5] Let $X$ be a random vector as in Theorem 3.1.1. Show that
\[
\text{Var}(\|X\|_2) = O(1).
\]

**Exercise 3.1.4** (Sub-gaussian concentration squared). [Difficulty=7] Let $X$ be a random variable with sub-gaussian concentration around its mean $\mu$, say
\[
\|X - \mu\|_{\psi_2} \leq 1.
\]
What kind of concentration does $X^2$ have around $\mu^2$? Give a tail bound.

### 3.2 Covariance matrices and isotropic distributions

In the last section we considered a special class of random variables, those with independent coordinates. Before we study more general situations, let us recall a few basic notions about high dimensional distributions the reader may have already seen in basic courses. We will thus be working with random vectors $X$ in $\mathbb{R}^n$, or equivalently with probability distributions in $\mathbb{R}^n$.

The concept of the *mean* of random variables generalizes in a straightforward way for random vectors. The notion of variance is replaced in high dimensions by the *covariance matrix* of a random vector $X$, defined as follows:
\[
\text{cov}(X) = \mathbb{E}(X - \mu)(X - \mu)^T = \mathbb{E}XX^T - \mu\mu^T
\]
where $\mu = \mathbb{E}X$. Thus $\text{cov}(X)$ is an $n \times n$ matrix. Note that the formula for covariance is a direct high-dimensional generalization of the definition of variance for random variables $Z$, which is
\[
\text{Var}(Z) = \mathbb{E}(Z - \mu)^2 = \mathbb{E}Z^2 - \mu^2, \quad \text{where } \mu = \mathbb{E}Z.
\]
The entries of $\Sigma$ are the covariances of the coordinates of $X = (X_1, \ldots, X_n)$:

$$
cov(X)_{ij} = \mathbb{E}(X_i - \mathbb{E}X_i)(Y_j - \mathbb{E}Y_j).
$$

It is sometimes useful to consider the second moment matrix of a random vector $X$, defined as

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

It is of course a high dimensional generalization of the second moment $\mathbb{E}Z^2$. By translation (replacing $X$ with $X - \mu$), in many problems we can assume that $X$ has zero mean, so

$$
\mu = 0 \quad \text{and thus} \quad \text{cov}(X) = \Sigma(X).
$$

So we will mostly focus on $\Sigma = \Sigma(X)$ rather than $\text{cov}(X)$ in the future.

The $n \times n$ matrix $\Sigma$ is symmetric and positive-semidefinite. (Check this!) The spectral theorem for such matrices says that all eigenvalues $s_i$ of $\Sigma$ are real and non-negative. Moreover, $\Sigma$ can be expressed via spectral decomposition as

$$
\Sigma = \sum_{i=1}^{n} s_i u_i u_i^T,
$$

where $u_i \in \mathbb{R}^n$ are the eigenvectors of $\Sigma$. We usually arrange this representation so that the eigenvalues $s_i$ are decreasing.

### 3.2.1 The Principal Component Analysis

The spectral decomposition of $\Sigma$ is of utmost importance in applications where the distribution of a random vector $X$ in $\mathbb{R}^n$ represents data, such as genetic data we mentioned on p. 37. The eigenvector $u_1$ corresponding to the largest eigenvalue $s_1$ indicates the direction in space in which the distribution is most extended. This principal direction best explains the variations in the data. The next eigenvector $u_2$ (corresponding to the next largest eigenvalue $s_2$) gives the next principal direction; it best explains the remaining variations in the data, and so on. This is illustrated in the Figure 3.2.

It often happens with real data that only a few eigenvalues $s_i$ are large and considered as informative; the remaining eigenvalues are small and considered as noise. In such situations, a few principal directions can explain the data. So, even though the data is presented in a high dimensional space $\mathbb{R}^n$, the data is essentially low dimensional. It clusters near the low-dimensional subspace $E$ spanned by the few principal components. This explains the
most basic data analysis algorithm, called the Principal Component Analysis (PCA). This algorithm projects the data in $\mathbb{R}^n$ onto the subspace $E$, which reduced the dimension of the data considerably. For example, if $E$ is two- or three-dimensional, the PCA allows to visualize the data.

### 3.2.2 Isotropy

We might remember from the basic probability course that is often convenient to assume that random variables in question have zero means and unit variances. The notion of isotropy is a high dimensional version of unit variance (more precisely, the unit second moment).

**Definition 3.2.1** (Isotropic random vectors). A random vector $X$ in $\mathbb{R}^n$ is called isotropic if

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

where $I_n$ denotes the identity matrix in $\mathbb{R}^n$.

Recall that any random variable $X$ with positive variance can be reduced to a random variable $Z$ with zero mean and unit variance by translation and dilation, namely

$$Z = \frac{X - \mu}{\sqrt{\text{Var}(X)}}.$$

The following exercise is a high dimensional version of this observation.

**Exercise 3.2.2** (Reduction to isotropy). [Difficulty=3] 1. Let $Z$ be a mean zero, isotropic random vector in $\mathbb{R}^n$. Let $\mu \in \mathbb{R}^n$ be a fixed vector and $\Sigma$ be an
3.2. COVARIANCE AND ISOTROPY

\( n \times n \) positive-semidefinite matrix. Then the random vector \( X := \mu + \Sigma^{1/2} Z \)
has mean \( \mu \) and covariance matrix \( \text{cov}(X) = \Sigma \).

2. Let \( X \) be a random vector with invertible covariance matrix \( \Sigma = \text{cov}(X) \). Then \( Z := \Sigma^{-1/2}(X - \mu) \) is an isotropic, mean zero random vector.

Thus in the future we can often assume that, without loss of generality, the distributions in question are isotropic and have zero means.

3.2.3 Properties of isotropic distributions

Lemma 3.2.3 (Characterization of isotropy). A random vector \( X \) 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. Recall that two \( n \times n \) matrices \( A \) and \( B \) are equal if and only if
\( x^T Ax = x^T Bx \) for all \( x \in \mathbb{R}^n \). (Check this!) Thus \( X \) is isotropic if and only if
\[
x^T (\mathbb{E} XX^T) x = x^T I_n x \quad \text{for all} \ x \in \mathbb{R}^n.
\]
The left side of this identity equals \( \mathbb{E} \langle X, x \rangle^2 \) and the right side, \( \|x\|_2^2 \). This completes the proof.

If \( x \) is a unit vector in Lemma 3.2.3, we can view \( \langle X, x \rangle \) as a one-dimensional marginal of the distribution of \( X \), obtained by projecting \( X \) onto the direction of \( x \). Then \( X \) is isotropic if and only if all one-dimensional marginals of \( X \) have unit variance. In plain words, an isotropic distribution is extended as evenly in all directions as possible.

Lemma 3.2.4. 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. To prove the first part, we have
\[
\mathbb{E} \|X\|_2^2 = \mathbb{E} XX^T = \mathbb{E} \text{tr}(X^T X) \quad \text{(viewing} \ X^T X \text{as a} \ 1 \times 1 \text{matrix)}
\]
\[
= \mathbb{E} \text{tr}(XX^T) \quad \text{(by invariance of trace under cyclic permutations)}
\]
\[
= \text{tr}(\mathbb{E} XX^T) \quad \text{(by linearity)}
\]
\[
= \text{tr}(I_n) \quad \text{(by isotropy)}
\]
\[
= n.
\]
CHAPTER 3. RANDOM VECTORS IN HIGH DIMENSIONS

To prove the second part, we use a conditioning argument. Fix a the realization of $Y$ and take the conditional expectation (with respect to $X$), which we denote $E_X$. The law of total expectation says that

$$E \langle X, Y \rangle^2 = E_Y E_X [\langle X, Y \rangle \mid Y],$$

where by $E_Y$ we of course denoted the expectation with respect to $Y$. To compute the inner expectation, we apply Lemma 3.2.3 with $x = Y$. It yields that the inner expectation equals $\|Y\|_2^2$. Thus

$$E \langle X, Y \rangle^2 = E_Y \|Y\|_2^2 = n \quad \text{(by the first part of lemma).}$$

The proof is complete. $\square$

3.2.4 Almost orthogonality of independent random vectors

Let us normalize the random vectors $X$ and $Y$ in Lemma 3.2.4, setting

$$\bar{X} := \frac{X}{\|X\|_2} \quad \text{and} \quad \bar{Y} := \frac{Y}{\|Y\|_2}.$$

Then we should expect from Lemma 3.2.4 that

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

with high probability. This shows that in high dimensional spaces, independent and isotropic random vectors are almost orthogonal, see Figure 3.3.

![Figure 3.3: Independent isotropic random vectors tend to be almost orthogonal in high dimensions $n$ but not in low dimensions.](image)

This could be surprising, since in low dimensions random vectors do not tend to be almost orthogonal. For example, the angle between two random independent directions on the plane has mean $\pi/4$. (Check!) But in higher dimensions, there is much more room, as we saw in the beginning of this chapter. This may be an intuitive reason why random directions in high dimensional spaces tend to be very far from each other – almost orthogonal.
3.3 Examples of high dimensional distributions

In this section we discuss several basic examples of isotropic high-dimensional distributions. It will be useful to keep them in mind when we will develop general theorems for such distributions.

3.3.1 Spherical and Bernoulli distributions

The coordinates of an isotropic random vector are uncorrelated (why?) but not necessarily independent. An example is the spherical distribution where a random vector is uniformly distributed on the unit Euclidean sphere in $\mathbb{R}^n$ with center at the origin and radius $\sqrt{n}$:

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

Exercise 3.3.1. Show that the spherically distributed random vector $X$ is isotropic. Argue that the coordinates of $X$ are not independent.

A good example of a discrete isotropic distribution in $\mathbb{R}^n$ is the symmetric Bernoulli distribution. We say that a random vector $X = (X_1, \ldots, X_n)$ is symmetric Bernoulli if the coordinates $X_i$ are independent, symmetric Bernoulli random variables. Equivalently, $X$ is uniformly distributed on the unit discrete cube in $\mathbb{R}^n$:

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

The symmetric Bernoulli distribution is isotropic. (Check!)

More generally, we may consider a random vector $X = (X_1, \ldots, X_n)$ whose coordinates $X_i$ are independent random variables with zero mean and unit variance. Then $X$ is an isotropic vector in $\mathbb{R}^n$. (Why?)

3.3.2 Multivariate normal

The most important high dimensional distribution is arguably the Gaussian, or multivariate normal. From the basic probability course we recall that a random vector $Z = (Z_1, \ldots, Z_n)$ has standard normal distribution in $\mathbb{R}^n$, denoted

$$Z \sim N(0, I_n),$$

if the coordinates $Z_i$ are independent standard normal random variables. The density of $Z$ is then the product of the $n$ standard normal densities
(1.2), which is
\[ f_Z(x) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-x_i^2/2} = \frac{1}{(2\pi)^{n/2}} e^{-\|x\|_2^2/2}, \quad x \in \mathbb{R}^n. \] (3.3)

The standard normal distribution is \textit{isotropic}. (Why?)

Note that the standard normal density is \textit{rotation invariant}. In particular, for any fixed unitary matrix \( U \),
\[ Z \sim \mathcal{N}(0, I_n) \implies UZ \sim \mathcal{N}(0, I_n). \]

**Exercise 3.3.2** (Sum of independent gaussians). [Difficulty=5] Let \( X_i \sim \mathcal{N}(0, \sigma_i^2) \) be independent random variables. Deduce from rotation invariance that the following property we mentioned in (2.19):
\[ n \sum_{i=1}^{n} X_i \sim \mathcal{N}(0, \sigma^2) \quad \text{where} \quad \sigma^2 = \sum_{i=1}^{n} \sigma_i^2. \]

*Hint:* consider the random vector \( Z \sim \mathcal{N}(0, I_n) \) with coefficients \( Z_i = X_i/\sigma_i \). Assume that \( u = (\sigma_1, \ldots, \sigma_n) \) is a unit vector without loss of generality. Then \( \sum X_i = \sum \sigma_i Z_i = \langle u, Z \rangle \). If \( U \) is a unitary matrix whose first row is \( u^\top \) then \( \langle u, Z \rangle \) is the first entry of the vector \( UZ \sim \mathcal{N}(0, I_n) \).

Let us also recall the notion of general normal distribution \( \mathcal{N}(\mu, \Sigma) \). Consider a vector \( \mu \in \mathbb{R}^n \) and an invertible \( n \times n \) positive-semidefinite matrix \( \Sigma \). According to Exercise 3.2.2, the random vector \( X := \mu + \Sigma^{-1/2}Z \) has mean \( \mu \) and covariance matrix \( \Sigma(X) = \Sigma \). Such \( X \) is said to have general normal distribution in \( \mathbb{R}^n \), denoted
\[ X \sim \mathcal{N}(\mu, \Sigma). \]

Summarizing, we have
\[ X \sim \mathcal{N}(\mu, \Sigma) \quad \text{iff} \quad Z := \Sigma^{-1/2}(X - \mu) \sim \mathcal{N}(0, I_n). \]

The density of \( X \sim \mathcal{N}(\mu, \Sigma) \) can be computed by change of variables formula, and it is
\[ f_X(x) = \frac{1}{(2\pi)^{n/2} \det(\Sigma)^{1/2}} e^{-(x-\mu)^\top \Sigma^{-1}(x-\mu)/2}, \quad x \in \mathbb{R}^n. \]

Figure 3.4 shows examples of the densities of multivariate normal distributions.
3.3. EXAMPLES OF HIGH DIMENSIONAL DISTRIBUTIONS

(a) The isotropic distribution \( N(0, I_2) \)

(b) A non-isotropic distribution \( N(0, \Sigma) \)

Figure 3.4: The densities of the multivariate distributions with different covariance matrices \( \Sigma \). The density of the standard normal distribution \( N(0, I_2) \) is rotation invariant, and its contour lines are circles. The contour lines of the non-isotropic distribution \( N(0, \Sigma) \) are ellipses.

3.3.3 Similarity of normal and spherical distributions

Contradicting our low dimensional intuition, the standard normal distribution \( N(0, I_n) \) in high dimensions is not concentrated close the origin where the density is maximal. Instead, it is concentrated in a thin spherical shell around the sphere of radius \( \sqrt{n} \), a shell of width \( O(1) \). Indeed, the concentration inequality (3.2) for the norm of \( X \sim N(0, I_n) \) states that

\[
P \{ \|X\|_2 - \sqrt{n} \geq t \} \leq 2 \exp(-ct^2) \quad \text{for all } t \geq 0.
\]

(3.4)

This suggests that the normal distribution should be quite similar to the uniform distribution on the sphere. Let us clarify the relation.

**Exercise 3.3.3** (Normal and spherical distributions). Let us represent \( g \sim N(0, I_n) \) in polar form as

\[ g = r\theta \]

where \( r = \|g\|_2 \) is the length and \( \theta = g/\|g\|_2 \) is the direction of \( X \). Prove the following.

1. The length \( r \) and direction \( \theta \) are independent random variables.
2. The direction \( \theta \) is uniformly distributed on the unit sphere \( S^{n-1} \).

Concentration inequality (3.4) for the length says that \( r \approx \sqrt{n} \) with high probability, so

\[ g \approx \sqrt{n}\theta \sim \text{Unif}(\sqrt{n}S^{n-1}). \]

This means that the spherical distribution considered in Section 3.3.1 and the standard normal distribution are approximately the same. We can write this heuristically as

\[ N(0, I_n) \approx \text{Unif}(\sqrt{n}S^{n-1}). \]

(3.5)
Figure 3.5 illustrates a shift of intuition about gaussian point clouds in high dimensions.

Figure 3.5: A Gaussian point cloud in two dimensions (left) and in high dimensions (right). In high dimensions, gaussian distribution is very close to the uniform distribution on the sphere of radius $\sqrt{n}$.

### 3.3.4 Frames

For an example of an extremely discrete distribution, consider a *coordinate random vector* $X$ uniformly distributed in the set $\{\sqrt{n}e_i\}_{i=1}^n$ where $\{e_i\}_{i=1}^n$ is the canonical basis of $\mathbb{R}^n$:

$$X \sim \text{Unif}\{\sqrt{n}e_i : i = 1, \ldots, n\}.$$ 

Then $X$ is an isotropic random vector in $\mathbb{R}^n$. (Check!)

Of all high dimensional distributions, Gaussian is often the most convenient to prove results for, so we may think of it as “the best” distribution. The coordinate distribution, the most discrete of all distributions, is “the worst”.

A general class of discrete, isotropic distributions arises in signal processing under the name of *frames*.

**Definition 3.3.4.** A frame is a set of vectors $\{u_i\}_{i=1}^N$ in $\mathbb{R}^n$ which obeys an approximate Parseval’s identity, i.e. there exist numbers $A, B > 0$ called frame bounds such that

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

If $A = B$ the set is called a tight frame.
Exercise 3.3.5. Show that \( \{u_i\}_{i=1}^N \) is a tight frame in \( \mathbb{R}^n \) with bound \( A \) if

\[
\sum_{i=1}^{N} u_i u_i^T = AI_n. \tag{3.6}
\]

Hint: Proceed similarly to the proof of Lemma 3.2.3.

Multiplying both sides of (3.6) by a vector \( x \), we see that

\[
\sum_{i=1}^{N} \langle u_i, x \rangle u_i = Ax \quad \text{for any } x \in \mathbb{R}^n. \tag{3.7}
\]

This is a frame expansion of a vector \( x \), and it should look familiar. Indeed, if \( \{u_i\} \) is an orthonormal basis, then (3.7) is just a classical basis expansion of \( x \), and it holds with \( A = 1 \).

We can think of tight frames as generalizations of orthogonal bases without the linear independence requirement. Any orthonormal basis in \( \mathbb{R}^n \) is clearly a tight frame. But so is the “Mercedez-Benz frame”, a set of three equidistant points on a circle in \( \mathbb{R}^2 \) shown on Figure 3.6.

In signal processing, tight frames are used as robust proxies of bases.

Figure 3.6: the Mercedes-Benz frame. A set of equidistant points on the circle form a tight frame in \( \mathbb{R}^2 \).

Now we are ready to connect the concept of frames to probability. We will show that tight frames correspond to isotropic distributions, and vice versa.

Lemma 3.3.6 (Tight frames and isotropic distributions). 1. Consider a tight frame \( \{u_i\}_{i=1}^N \) in \( \mathbb{R}^n \) scaled so that the bounds satisfy \( A = B = N \). Let \( X \) be a random vector that is uniformly distributed in the set of frame elements, i.e.

\[
X \sim \text{Unif} \{u_i : i = 1, \ldots, N\},
\]

Then \( X \) is an isotropic random vector in \( \mathbb{R}^n \).
2. Consider an isotropic random vector $X$ in $\mathbb{R}^n$ that takes a finite set of values $x_i$ with probabilities $p_i$ each, $i = 1, \ldots, N$. Then the vectors
\[ u_i := \sqrt{p_i} x_i, \quad i = 1, \ldots, N, \]
form a tight frame in $\mathbb{R}^N$ with bounds $A = B = 1$.

Proof. 1. The assumptions and (3.6) imply that
\[ \sum_{i=1}^{N} u_i u_i^T = N I_n. \]
Dividing both sides by $N$ and interpreting $\frac{1}{n} \sum_{i=1}^{n}$ as expectation, we conclude that $X$ is isotropic.

2. Isotropy of $X$ means that
\[ \mathbb{E} XX^T = \sum_{i=1}^{N} p_i x_i x_i^T = I_n. \]
Denoting $u_i := \sqrt{p_i} x_i$, we obtain (3.6) with $A = I_n$. \qed

### 3.3.5 Isotropic convex sets

Our last example of a high dimensional distribution comes from convex geometry. Consider a bounded convex set $K$ in $\mathbb{R}^n$ with non-empty interior; such sets are called convex bodies. Let $X$ be a random vector uniformly distributed in $K$, according to the probability measure given by normalized volume in $K$:
\[ X \sim \text{Unif}(K). \]
Denote the covariance matrix of $X$ by $\Sigma$. Then by Exercise 3.2.2, the random vector $Z := \Sigma^{-1/2} X$ is isotropic. Note that $Z$ is uniformly distributed in the linearly transformed copy of $K$:
\[ Z \sim \text{Unif}(\Sigma^{-1/2} K). \]
(Why?) Summarizing, we found a linear transformation $T := \Sigma^{-1/2}$ which makes the uniform distribution on $TK$ isotropic. The body $TK$ is sometimes called isotropic itself.

In algorithmic convex geometry, one thinks of the isotropic convex body $TK$ as a well conditioned version of $K$, with $T$ playing the role of a pre-conditioner, see Figure 3.7. Algorithms related to convex bodies $K$ (such
as computing the volume of $K$) work better for well-conditioned $K$. So in practice, it is useful to be able to compute or estimate the covariance matrix $\Sigma$ of $K$, since this allows one to transform $K$ into as well-conditioned convex body as possible.

Figure 3.7: A convex body $K$ on the left is transformed into an isotropic convex body $TK$ on the right. The pre-conditioner $T$ is computed from the covariance matrix $\Sigma$ of $K$ as $T = \Sigma^{-1/2}$.

3.4 Sub-gaussian distributions in higher dimensions

The concept of sub-gaussian distributions we introduced in Section 2.5 can be extended to higher dimensions. To see how, recall that the multivariate normal distribution $N(\mu, \Sigma)$ can be characterized through its one-dimensional marginals, or projections onto lines. A random vector $X$ has normal distribution in $\mathbb{R}^n$ if and only if the one-dimensional marginals $\langle X, x \rangle$ are normal for all $x \in \mathbb{R}^n$. Guided by this characterization, we can define multivariate sub-gaussian distributions as follows.

Definition 3.4.1 (Sub-gaussian random vectors). A random vector $X$ in $\mathbb{R}^n$ is called sub-gaussian if the one-dimensional marginals $\langle X, x \rangle$ are sub-gaussian random variables for all $x \in \mathbb{R}^n$. The sub-gaussian norm of $X$ is defined as

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

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

Lemma 3.4.2 (Sub-gaussian distributions with independent coordinates). Let $X = (X_1, \ldots, X_n) \in \mathbb{R}^n$ be a random vector with independent, mean zero, sub-gaussian coordinates $X_i$. Then $X$ is a sub-gaussian random vector, and

$$\|X\|_{\psi_2} \leq C \max_{i \leq n} \|X_i\|_{\psi_2}.$$
Proof. This is an easy consequence of the fact that the sum of independent sub-gaussian random variables is sub-gaussian, which we proved in Proposition 2.6.1. Indeed, for a fixed unit vector $x = (x_1, \ldots, x_n) \in S^{n-1}$ we have

$$\| (X, x) \|_{\psi_2}^2 = \left\| \sum_{i=1}^{n} x_i X_i \right\|_{\psi_2}^2 \leq C \sum_{i=1}^{n} x_i^2 \| X_i \|_{\psi_2}^2 \quad \text{(by Proposition 2.6.1)}$$

$$\leq C \max_{i \leq n} \| X_i \|_{\psi_2}^2 \quad \text{(using that } \sum_{i=1}^{n} x_i^2 = 1).$$

This completes the proof. \hfill \square

**Exercise 3.4.3.** [Difficulty=5] This exercise clarifies the role of independence of coordinates in Lemma 3.4.2.

1. Let $X = (X_1, \ldots, X_n) \in \mathbb{R}^n$ be a random vector with sub-gaussian coordinates $X_i$. Show that $X$ is a sub-gaussian random vector.

2. Nevertheless, find an example of a random vector $X$ with

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

Many important high-dimensional distributions are sub-gaussian, but some are not. We will now explore some basic distributions.

### 3.4.1 Gaussian and Bernoulli distributions

As we already noted, *multivariate normal distribution* $N(\mu, \Sigma)$ is sub-gaussian. Moreover, the standard normal random vector $X \sim N(0, I_n)$ has sub-gaussian norm of order $O(1)$:

$$\| X \|_{\psi_2} \leq C.$$ 

(Indeed, all one-dimensional marginals of $X$ are $N(0, 1).$)

Next, consider the multivariate *symmetric Bernoulli* distribution we introduced in Section 3.3.1. A random vector $X$ with this distribution has independent, symmetric Bernoulli coordinates. Then Lemma 3.4.2 yields that

$$\| X \|_{\psi_2} \leq C.$$ 

### 3.4.2 Discrete distributions

Let us now pass to discrete distributions. The extreme example we considered in Section 3.3.4 is the *coordinate distribution*. Recall that random
vector $X$ with coordinate distribution is uniformly distributed in the set \( \{ \sqrt{n}e_i : i = 1, \ldots, n \} \), where $e_i$ denotes the $n$-element set of the canonical basis vectors in $\mathbb{R}^n$.

Is $X$ sub-gaussian? Formally, yes. In fact, every distribution supported in a finite set is sub-gaussian. (Why?) But, unlike Gaussian and Bernoulli distributions, the coordinate distribution has a very large sub-gaussian norm:

\[
\|X\|_{\psi_2} \asymp \sqrt{n}.
\]

(To see this, note that $|\langle X, e_1 \rangle| = \sqrt{n}$ with probability one.) Such large norm makes it useless to think of $X$ as a sub-gaussian random vector.

More generally, discrete distributions do not make nice sub-gaussian distributions, unless they are supported on exponentially large sets:

**Exercise 3.4.4.** [Difficulty=10?] Let $X$ be an isotropic random vector supported in a finite set $T \subset \mathbb{R}^n$. Show that in order for $X$ to be sub-gaussian with $\|X\|_{\psi_2} = O(1)$, the cardinality of the set must be exponentially large in $n$:

\[
|T| \geq e^{cn}.
\]

In particular, this observation rules out frames (see Section 3.3.4) as good sub-gaussian distributions unless they have exponentially many terms (in which case they are mostly useless in practice).

### 3.4.3 Uniform distribution on the sphere

In all previous examples, good sub-gaussian random vectors had independent coordinates. This is not necessary. A good example is the uniform distribution on the sphere of radius $\sqrt{n}$, which we discussed in Section 3.4.3. We will show that it is sub-gaussian by reducing it to the Gaussian distribution $N(0, I_n)$.

**Theorem 3.4.5** (Uniform distribution on sphere is sub-gaussian). Let $X$ be a random vector uniformly distributed on the Euclidean sphere in $\mathbb{R}^n$ with center at the origin and radius $\sqrt{n}$:

\[
X \sim \text{Unif} \left( \sqrt{n} S^{n-1} \right).
\]

Then $X$ is sub-gaussian, and

\[
\|X\|_{\psi_2} \leq C.
\]
Proof. Consider a standard normal random vector \( g \sim N(0, I_n) \). As we noted in Section 3.3.3, the direction \( g/\|g\|_2 \) is uniformly distributed on the unit sphere \( S^{n-1} \). By rescaling, we can represent a random vector \( X \sim \text{Unif}(\sqrt{n} S^{n-1}) \) as \[ X = \sqrt{n} \frac{g}{\|g\|_2}. \]

We need to show that all one-dimensional marginals \( \langle X, x \rangle \) are subgaussian. By rotation invariance, we may assume that \( x = e_1 \). So it is enough to analyze \( \langle X, x \rangle = X_1 \) the first coordinate of \( X \). Thus we want to bound the tail probability \[ p(t) := \mathbb{P}\{|X_1| \geq t\} = \mathbb{P}\left\{ \left| \frac{g_1}{\|g\|_2} \right| \geq \frac{t}{\sqrt{n}} \right\}. \]

Heuristically, the concentration of norm (Theorem 3.1.1) implies that \( \|g\|_2 \approx \sqrt{n} \) with high probability.

This reduces the problem to bounding the tail \( \mathbb{P}\{|g_1| \geq t\} \), but as we know from (2.2), this tail is sub-gaussian.

Let us do this argument more carefully. The concentration of norm, Theorem 3.1.1, implies that \[ \|\|g\|_2 - \sqrt{n}\|_{\psi_2} \leq C. \]

Thus the event \[ \mathcal{E} := \left\{ \|g\|_2 \geq \frac{\sqrt{n}}{2} \right\} \]

is likely: by (2.15) its complement \( \mathcal{E}^c \) has probability \[ \mathbb{P}(\mathcal{E}^c) \leq 2 \exp(-cn). \]

Then the tail probability can be bounded as follows:

\[ p(t) \leq \mathbb{P}\left\{ \left| \frac{g_1}{\|g\|_2} \right| \geq \frac{t}{\sqrt{n}} \text{ and } \mathcal{E} \right\} + \mathbb{P}(\mathcal{E}^c) \]

\[ \leq \mathbb{P}\left\{ |g_1| \geq \frac{t}{2} \right\} + 2 \exp(-cn) \]

\[ \leq 2 \exp(-t^2/8) + 2 \exp(-cn) \quad \text{(using (2.2))}. \]

Consider two cases. If \( t \leq \sqrt{n} \) then \( 2 \exp(-cn) \leq 2 \exp(-ct^2/8) \), and we conclude that \[ p(t) \leq 4 \exp(-c't^2) \]
as desired. In the opposite case where \( t > \sqrt{n} \), the tail probability \( p(t) = \mathbb{P}\{|X_1| \geq t\} \) trivially equals zero, since we always have \( |X_1| \leq \|X\|_2 = \sqrt{n} \). This completes the proof.

\[ \square \]

**Exercise 3.4.6** (Uniform distribution on the Euclidean ball). [Difficulty=5]

Extend Theorem 3.4.5 for the uniform distribution on the Euclidean ball \( B(0, \sqrt{n}) \) in \( \mathbb{R}^n \) centered at the origin and with radius \( \sqrt{n} \). Namely, show that a random vector

\[ X \sim \text{Unif} \left( B(0, \sqrt{n}) \right) \]

is sub-gaussian, and

\[ \|X\|_{\psi^2} \leq C. \]

**Exercise 3.4.7.** [Difficulty=8] Prove Theorem 3.4.5 by reducing the spherical distribution to Gaussian. Use the similarity of these two distributions we explored in Section 3.3.3.

**Projective Limit Theorem**

Theorem 3.4.5 should be compared to the well known “Projective Central Limit Theorem”. It states that the marginals of the uniform distribution on the sphere become asymptotically normal as \( n \) increases, see Figure 3.8. Precisely, if \( X \sim \text{Unif} \left( \sqrt{n} S^{n-1} \right) \) then for any fixed unit vector \( x \) we have

\[ \langle X, x \rangle \rightarrow N(0, 1) \quad \text{in distribution as } n \rightarrow \infty. \]

So we can view Theorem 3.4.5 as a concentration version of the Projective Limit Theorem, in the same sense as we found Hoeffding’s inequality in Section 2.2 to be a concentration version of the classical Central Limit Theorem.

Figure 3.8: The Projective Central Limit Theorem. The projection of the uniform distribution on the sphere of radius \( \sqrt{n} \) onto a line converges to the normal distribution \( N(0, 1) \) as \( n \rightarrow \infty. \)
3.4.4 Uniform distribution on convex sets

To conclude this section, let us return to the class of uniform distributions on convex sets which we discussed in Section 3.3.5. Let $K$ be a convex body and

$$X \sim \text{Unif}(K)$$

be an isotropic random vector. Is $X$ always sub-gaussian?

For some bodies $K$ this is the case. Examples include the Euclidean ball of radius $\sqrt{n}$ (by Exercise 3.4.6) and the unit cube $[-1,1]^n$ (according to Lemma 3.4.2). For some other bodies, this is not true:

**Exercise 3.4.8.** [Difficulty=7] Consider a ball of the $\ell_1$ norm in $\mathbb{R}^n$:

$$K := \{x \in \mathbb{R}^n : \|x\|_1 \leq r \}.$$  

1. Show that the uniform distribution on $K$ is isotropic for $r \sim n$.
2. Show that this distribution is not sub-gaussian.

Nevertheless, a weaker result is possible to prove for a general isotropic convex body $K$. The random vector $X \sim \text{Unif}(K)$ has all sub-exponential marginals, and

$$\|\langle X, x \rangle\|_{\psi_1} \leq C$$

for all unit vectors $x$. This result follows from C. Borell’s lemma, which itself is a consequence of Brunn-Minkowski inequality; see [?, Section 2.2.b].

**Exercise 3.4.9.** [Difficulty=6] Show the concentration inequality in Theorem 3.1.1 may not hold for a general isotropic sub-gaussian random vector $X$. Thus, independence of the coordinates of $X$ is an essential requirement in that result.
Chapter 4

Sub-gaussian random matrices

4.1 Preliminaries on matrices

4.1.1 Singular value decomposition

The main object of our study will be an \( m \times n \) matrices \( A \) with real entries. Recall from a course in linear algebra that \( A \) admits them singular value decomposition (SVD), which we can write in the following form:

\[
A = \sum_{i=1}^{r} s_i u_i v_i^T, \quad \text{where} \quad r = \text{rank}(A).
\]

Here the non-negative numbers \( s_i = s_i(A) \) are called singular values of \( A \), the vectors \( u_i \in \mathbb{R}^m \) are called the left singular vectors of \( A \), and the vectors \( v_i \in \mathbb{R}^n \) are called the right singular vectors of \( A \).

Since for random matrices \( r = \text{rank}(A) \) is random, it is convenient to extend the sequence of singular values by setting \( s_i = 0 \) for \( r < i \leq n \). Also, for convenience we arrange them so that

\[
s_1 \geq s_2 \geq \cdots \geq s_n \geq 0.
\]

The vectors \( u_i \) are a set of orthonormal eigenvectors of \( AA^* \) and the vectors \( v_i \) are a set of orthonormal eigenvectors of \( A^*A \). The singular vectors \( s_i \) are the square roots of the eigenvalues \( \lambda_i \) of both \( AA^* \) and \( A^*A \):

\[
s_i(A) = \sqrt{\lambda_i(AA^*)} = \sqrt{\lambda_i(A^*A)}.
\]
In particular, if \( A \) is a symmetric matrix, the singular values of \( A \) are the absolute values of the eigenvalues \( \lambda_i \) of \( A \):

\[
s_i(A) = |\lambda_i(A)|,
\]

and both left and right singular vectors of \( A \) are the eigenvectors of \( A \).

### 4.1.2 Operator norm and the extreme singular values

The space of \( m \times n \) matrices can be equipped with several classical norms. We will mention two of them – operator and Frobenius norms – and emphasize their connection with the spectrum of \( A \).

When we think of the space \( \mathbb{R}^m \) along with the Euclidean norm \( \| \cdot \|_2 \) on it, we denote this Hilbert space \( \ell_2^m \). The matrix \( A \) acts as a linear operator from \( \ell_2^n \rightarrow \ell_2^m \). Its **operator norm** of \( A \), also called the **spectral norm**, is then defined as

\[
\|A\| := \|A : \ell_2^n \rightarrow \ell_2^m\| = \max_{x \in \mathbb{R}^n \setminus \{0\}} \frac{\|Ax\|_2}{\|x\|_2} = \max_{x \in S^{n-1}} \|Ax\|_2.
\]

Equivalently, the operator norm of \( A \) can be computed by maximizing the quadratic form \( x^T Ay = \langle Ax, y \rangle \) over all unit vectors \( x, y \):

\[
\|A\| = \max_{x \in S^{n-1}, y \in S^{m-1}} \langle Ax, y \rangle.
\]

In terms of spectrum, the operator norm of \( A \) is the same as the largest singular value of \( A \):

\[
s_1(A) = \|A\|.
\]

(Check!)

The smallest singular value \( s_n(A) \) also has a special meaning. By definition, it can only be non-zero for tall matrices where \( m \geq n \). In this case, \( A \) has full rank \( n \) if and only if \( s_n(A) > 0 \). Moreover, \( s_n(A) \) is a quantitative measure of non-degeneracy of \( A \). Indeed,

\[
s_n(A) = \frac{1}{\|A^\dagger\|}
\]

where \( A^\dagger \) is the pseudo-inverse of \( A \). Its norm \( \|A^\dagger\| \) is the norm of the operator \( A^{-1} \) restricted to the image of \( A \).
4.1.3 Frobenius norm
The Frobenius norm, also called Hilbert-Schmidt norm of a matrix $A$ with entries $A_{ij}$ is defined as
\[ \| A \|_F = \left( \sum_{i=1}^{m} \sum_{j=1}^{n} |A_{ij}|^2 \right)^{1/2}. \]

Thus Frobenius norm is the Euclidean norm on the space of matrices $\mathbb{R}^{m \times n}$. In terms of spectrum, the Frobenius norm can be computed as
\[ \| A \|_F = \left( \sum_{i=1}^{r} s_i(A)^2 \right)^{1/2}. \]

Let us now compare the operator and Frobenius norm. If we look at the vector $s = (s_1, \ldots, s_r)$ of singular values of $A$, these norms become the $\ell_\infty$ and $\ell_2$ norms, respectively:
\[ \| A \| = \| s \|_\infty, \quad \| A \|_F = \| s \|_2. \]

Using the inequality $\| s \|_\infty \leq \| s \|_2 \leq \sqrt{r} \| s \|_\infty$ for $s \in \mathbb{R}^n$ (check it!) we obtain the best possible relation between the operator and Frobenius norms:
\[ \| A \| \leq \| A \|_F \leq \sqrt{r} \| A \|. \]

4.1.4 Approximate isometries
The extreme singular values $s_1(A)$ and $s_r(A)$ have an important geometric meaning. They are respectively the smallest number $M$ and the largest number $m$ that make the following inequality true:
\[ m\| x \|_2 \leq \| Ax \|_2 \leq M\| x \|_2 \text{ for all } x \in \mathbb{R}^n. \]

(Check!) Applying this inequality for $x - y$ instead of $x$ and with the best bounds, we can rewrite as
\[ s_r(A)\| x - y \|_2 \leq \| Ax - Ay \|_2 \leq s_1(A)\| x - y \|_2 \text{ for all } x \in \mathbb{R}^n. \]

This means that the matrix $A$, acting as an operator from $\mathbb{R}^m$ to $\mathbb{R}^n$, change the distances between points by factors that lie between $s_r(A)$ and $s_1(A)$. Thus the extreme singular values control the distortion of the geometry of $\mathbb{R}^n$ under the action of $A$.

The best possible matrices in this sense, which preserve distances exactly, are called isometries. Let us recall their characterization, which can be proved using elementary linear algebra. (Do it!)
**Lemma 4.1.1** (Isometries). Let $A$ be an $m \times n$ matrix with $m \geq n$. Then the following are equivalent.

1. $A$ is an isometry, or isometric embedding of $\mathbb{R}^n$ into $\mathbb{R}^m$. This means that
   \[ \|Ax\|_2 = \|x\|_2 \quad \text{for all } x \in \mathbb{R}^n. \]
2. $A^TA = I_n$.
3. All singular values of $A$ equal 1; equivalently
   \[ s_n(A) = s_1(A) = 1. \]

Quite often the conditions of Lemma 4.1.1 hold only approximately, in which case we think of $A$ as an approximate isometry.

**Lemma 4.1.2** (Approximate isometries). Let $A$ be an $m \times n$ matrix and $\delta > 0$. Suppose that
\[ \|A^TA - I_n\| \leq \max(\delta, \delta^2). \]
Then
\[ (1 - \delta)\|x\|_2 \leq \|Ax\|_2 \leq (1 + \delta)\|x\|_2 \quad \text{for all } x \in \mathbb{R}^n. \tag{4.1} \]
Equivalently, all singular values of $A$ are between $1 - \delta$ and $1 + \delta$:
\[ 1 - \delta \leq s_n(A) \leq s_1(A) \leq 1 + \delta. \tag{4.2} \]

**Proof.** By assumption, we have
\[ \left| \left( A^TA - I_n \right)x, x \right| = \|Ax\|_2^2 - 1 \leq \max(\delta, \delta^2). \]
Applying the elementary inequality
\[ \max(|z - 1|, |z - 1|^2) \leq |z^2 - 1|, \quad z \geq 0 \]
for $z = \|Ax\|_2$, we conclude that
\[ \|Ax\|_2^2 - 1 \leq \delta. \]
This proves (4.1), which in turn implies (4.2) as we saw before. \hfill \Box

**Exercise 4.1.3** (Approximate isometries). Prove the following converse to Lemma 4.1.2. If $A$ is an approximate isometry, i.e. (4.2) holds, then
\[ \|A^TA - I_n\| \leq 2 \max(\delta, \delta^2). \]
Suppose $A$ is a fat $m \times n$ matrix, that is $m \leq n$, and
$$s_1(A) \approx s_m(A) \approx 1.$$ Then $A$ can be viewed as an *approximate projection* from $\mathbb{R}^m$ into $\mathbb{R}^n$. Thus $A$ is an approximate isometry if and only if $A^T$ is an approximate projection.

Canonical example of isometries and projections can be constructed from a fixed unitary matrix $U$. Any sub-matrix of $U$ obtained by selecting a subset of columns is an (exact) isometry, and any sub-matrix obtained by selecting a subset of rows is an (exact) projection in this sense.

### 4.2 Nets, covering numbers and packing numbers

In a course in analysis, you may have studied the notion of an $\varepsilon$-net. Let us recall it here.

**Definition 4.2.1 (\varepsilon-net).** Consider a subset $K$ of $\mathbb{R}^n$ and let $\varepsilon > 0$. A subset $\mathcal{N} \subseteq K$ is called an $\varepsilon$-net of $K$ if every point in $K$ is within distance $\varepsilon$ of some point of $\mathcal{N}$, i.e.
$$\forall x \in K \exists x_0 \in \mathcal{N} : \|x - x_0\|_2 \leq \varepsilon.$$ Equivalently, $\mathcal{N} \subseteq K$ is an $\varepsilon$-net of $K$ if and only if $K$ can be covered by balls with centers in $\mathcal{N}$ and radii $\varepsilon$, see Figure 4.1a.

An important result in analysis about *compactness* of sets is that $K \subset \mathbb{R}^n$ is pre-compact (which in $\mathbb{R}^n$ simply means that $K$ is bounded) if and only if $K$ has a *finite* $\varepsilon$-net for every $\varepsilon > 0$. More quantitatively, the smallest cardinality of an $\varepsilon$-net can be taken as a measure of compactness of $K$.

**Definition 4.2.2 (Covering numbers).** The smallest cardinality of an $\varepsilon$-net of $K$ is called the covering number of $K$ and is denoted $\mathcal{N}(K, \varepsilon)$. Equivalently, the $\mathcal{N}(K, \varepsilon)$ is the smallest number of closed balls with centers in $K$ and radii $\varepsilon$ whose union covers $K$.

Closely related to covering is the notion of *packing*.

**Definition 4.2.3 (Packing numbers).** The packing number $\mathcal{P}(K, \varepsilon)$ is the smallest number of open disjoint balls with centers in $K$ and radii $\varepsilon > 0$.

Figure 4.1b illustrates a packing of $K$ by balls centered at some points $x_i \in K$. 

---

[1] For a detailed discussion on nets, covering numbers, and packing numbers, see [Wells et al. (2015)](#).

---

[2] This definition and its implications can be further explored in [Stein and Shakarchi (2011)](#).

---

[3] For a comprehensive treatment of these concepts in the context of functional analysis, refer to [Brezis (2011)](#).

---


---

[5] Additional insights can be gained from [Keece and Saccà (2014)](#).
Exercise 4.2.4. A collection of points $x_i \in K$ are centers of balls that form a packing in $K$ if and only if

$$\|x_i - x_j\|_2 > 2\varepsilon \quad \text{for all } i \neq j.$$ 

The covering and packing numbers are equivalent up to a slight scaling of the radius:

Lemma 4.2.5 (Equivalence of covering and packing numbers). For any bounded set $K \subset \mathbb{R}^n$ and any $\varepsilon > 0$, we have

$$\mathcal{P}(K, \varepsilon) \leq \mathcal{N}(K, \varepsilon) \leq \mathcal{P}(K, \varepsilon/2).$$

Proof. Lower bound. Let $\mathcal{P} = \{x_i\}$ and $\mathcal{N} = \{y_i\}$ be the centers of $\varepsilon$-balls that form a packing and a covering of $K$, respectively. By Exercise 4.2.4, the centers of packing are $2\varepsilon$-separated:

$$\|x_i - x_j\|_2 > 2\varepsilon \quad \text{for all } i \neq j.$$ 

Since any $\varepsilon$-ball can not have a pair of $2\varepsilon$-separated points, each covering ball $B(y_i, \varepsilon)$ may contain at most one point $x_i$. It follows that

$$|\mathcal{P}| \leq |\mathcal{N}|.$$ 

This proves the lower bound.

Upper bound. Let $\mathcal{P} = \{x_i\}$ be a maximal packing of $K$ by $\varepsilon/2$-balls. Here by “maximal” we mean that an addition of any $\varepsilon/2$-ball to $\mathcal{P}$ will always
4.2. NETS, COVERING NUMBERS AND PACKING NUMBERS

destroy the packing property; Figure 4.1b shows an example of a maximal packing. Equivalently, by Exercise 4.2.4, \( \mathcal{P} \) is a maximal \( \varepsilon \)-separated set of points \( x_i \in K \):

\[
\|x_i - x_j\|_2 > \varepsilon \quad \text{for all } i \neq j.
\]

By maximality, \( \{x_i\} \) is an \( \varepsilon \)-net of \( K \). (Indeed, an addition of any point \( x \in K \) to the family \( \{x_i\} \) destroys its \( \varepsilon \)-separation property, which means that \( \|x - x_i\|_2 \leq \varepsilon \) for some \( i \).) Thus we constructed an \( \varepsilon \)-net of \( K \) of cardinality at most \( |\mathcal{P}| \). The upper bound in the lemma is proved.

Exercise 4.2.6 (Allowing the centers to be outside \( K \)). In our definition of covering numbers of \( K \), we required that the centers \( x_i \) of the balls \( B(x_i, \varepsilon) \) that form covering lie in \( K \). Relaxing this condition, define external covering number \( \mathcal{N}^{ext}(K, \varepsilon) \) similarly but without requiring that \( x_i \in K \). Prove that

\[
\mathcal{N}^{ext}(K, \varepsilon) \leq \mathcal{N}(K, \varepsilon) \leq \mathcal{N}^{ext}(K, \varepsilon/2).
\]

Exercise 4.2.7 (Monotonicity of covering numbers). Give a counterexample to the monotonicity property

\[
L \subset K \quad \text{implies} \quad \mathcal{N}(L, \varepsilon) \leq \mathcal{N}(K, \varepsilon).
\]

Prove an approximate version of monotonicity:

\[
L \subset K \quad \text{implies} \quad \mathcal{N}(L, \varepsilon) \leq \mathcal{N}(K, \varepsilon/2).
\]

4.2.1 Metric entropy and coding

Covering and packing numbers measure the size, or rather complexity, of a set \( K \), which makes them a useful tool in coding theory. The logarithm of the covering numbers \( \log \mathcal{N}(K, \varepsilon) \) is often called the metric entropy of \( K \). As we will see now, the metric entropy is equivalent to the number of digits needed to encode points in \( K \) as bit strings.

Proposition 4.2.8 (Metric entropy and coding). Let \( C(K, \varepsilon) \) denote the smallest number of bits sufficient to specify every point \( x \in K \) with accuracy \( \varepsilon \) in the Euclidean norm. Then

\[
\log_2 \mathcal{N}(K, \varepsilon) \leq C(K, \varepsilon) \leq \log_2 \mathcal{N}(K, \varepsilon/2).
\]

Proof. (Lower bound) Assume \( C(K, \varepsilon) \leq N \), so there is a way to represent every point \( x \in K \) with accuracy \( \varepsilon \) using a bit string of length \( N \). This induces a partition of \( K \) into at most \( 2^N \) subsets, which are obtained by
grouping points represented the same bit string; see Figure 4.2 for illustration. Each subset must have diameter at most $\varepsilon$, and thus it can be covered by a Euclidean ball centered in $K$ and with radius $\varepsilon$. (Why?) So $K$ is covered by at most $2^N$ balls with radii $\varepsilon$. This implies that $\mathcal{N}(K, \varepsilon) \leq 2^N$. Taking logarithm of both sides, we obtain the lower bound in the proposition.

(Upper bound) Assume $\log_2 \mathcal{N}(K, \varepsilon/2) \leq N$, so there exists an $(\varepsilon/2)$-net $\mathcal{N}$ of $K$ with cardinality $|\mathcal{N}| \leq 2^N$. To every point $x \in K$, let us assign a point $x_0 \in \mathcal{N}$ closest to $x$. Since there are at most $2^N$ such points, $N$ bits are sufficient to specify the point $x_0$. The encoding $x \mapsto x_0$ represents points in $K$ with accuracy $\varepsilon$. (Indeed, if both $x$ and $y$ are encoded by the same $x_0$ then $\|x - y\|_2 \leq \|x - x_0\|_2 + \|y - x_0\|_2 \leq \varepsilon$.) This shows that $\mathcal{C}(K, \varepsilon) \leq N$. 

4.2.2 Covering numbers and volume

If the covering numbers measure the size of $K$, how are they related to the most classical measure of size, the volume of $K$ in $\mathbb{R}^n$? There could not be a full equivalence between these two quantities, since “flat” sets have zero volume but non-zero covering numbers.

Still, there is a useful partial equivalence holds, which is often quite sharp. It is based on the notion of Minkowski sum of sets in $\mathbb{R}^n$.

Definition 4.2.9 (Minkowski sum). Let $A$ and $B$ be subsets of $\mathbb{R}^n$. The Minkowski sum $A + B$ is defined as

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

Figure 4.3 shows an example of Minkowski sum of two sets on the plane.
4.2. NETS, COVERING NUMBERS AND PACKING NUMBERS

\[ 4.2. \quad \text{NETS, COVERING NUMBERS AND PACKING NUMBERS} \]

Figure 4.3: Minkowski sum of a square and a circle is a square with rounded corners.

**Proposition 4.2.10** (Covering numbers and volume). Let \( K \) be a subset of \( \mathbb{R}^n \) and \( \varepsilon > 0 \). Then

\[
\frac{\text{Vol}(K)}{\text{Vol}(\varepsilon B_2^n)} \leq \mathcal{N}(K, \varepsilon) \leq \frac{\text{Vol}(K + (\varepsilon/2)B_2^n)}{\text{Vol}((\varepsilon/2)B_2^n)}.
\]

Here as usual \( B_2^n \) denotes the unit Euclidean ball in \( \mathbb{R}^n \), so \( \varepsilon B_2^n \) is the Euclidean ball with radius \( \varepsilon \).

**Proof.** (Lower bound) Let \( N := \mathcal{N}(K, \varepsilon) \). Then \( K \) can be covered by \( N \) balls with radii \( \varepsilon \). Comparing volume, we obtain

\[
\text{Vol}(K) \leq N \cdot \text{Vol}(\varepsilon B_2^n),
\]

which proves the lower bound in the proposition.

(Upper bound) By Lemma 4.2.5, it is enough to prove the same upper bound but for the packing number \( P(K, \varepsilon/2) =: P \). Consider \( P \) open disjoint balls \( B(x_i, \varepsilon/2) \) with centers \( x_i \in K \) and radii \( \varepsilon/2 \). While these balls do not need to fit entirely in \( K \) (see Figure 4.1b), they fit in a slightly inflated set, namely \( K + (\varepsilon/2)B_2^n \). (Why?) Comparing the volumes, we obtain

\[
P \cdot \text{Vol}((\varepsilon/2)B_2^n) \leq \text{Vol}(K + (\varepsilon/2)B_2^n),
\]

which leads to the upper bound in the proposition. \( \Box \)

Let us give some examples for the volumetric bound.

**Corollary 4.2.11** (Covering numbers of the Euclidean ball). The covering numbers of the unit Euclidean ball \( B_2^n \) satisfy the following for any \( \varepsilon > 0 \):

\[
\left( \frac{1}{\varepsilon} \right)^n \leq \mathcal{N}(B_2^n, \varepsilon) \leq \left( \frac{2}{\varepsilon} + 1 \right)^n.
\]

The same upper bound is true for the unit Euclidean sphere \( S^{n-1} \).
Proof. The lower bound follows immediately from Proposition 4.2.10, since the volume scales as \( \text{Vol}(\varepsilon B_2^n) = \varepsilon^n \cdot \text{Vol}(B_2^n) \). The upper bound follows from Proposition 4.2.10, too:

\[
N(K, \varepsilon) \leq \frac{\text{Vol}((1 + \varepsilon/2)B_2^n)}{\text{Vol}((\varepsilon/2)B_2^n)} = \left( \frac{2}{\varepsilon} + 1 \right)^n.
\]

The upper bound for the sphere can be proved in the same way. \( \square \)

To simplify the bound a bit, note the interesting range is \( \varepsilon \in (0, 1] \), where we have

\[
\left( \frac{1}{\varepsilon} \right)^n \leq N(B_2^n, \varepsilon) \leq \left( \frac{3}{\varepsilon} \right)^n.
\] (4.3)

In the trivial range where \( \varepsilon > 1 \), the unit ball can be covered by just one \( \varepsilon \)-ball, so \( N(B_2^n, \varepsilon) = 1 \).

The important message of (4.3) is that the covering numbers are exponential in the dimension \( n \). This should not be surprising if we recall the coding theory perspective we discussed in Section 4.2.1. Indeed, to encode a vector in dimension \( n \), one should be prepared to spend at least one bit per coefficient, so \( n \) bits total. This makes the metric entropy linear in \( n \), and the covering numbers exponential in \( n \).

Let us finish the discussion of covering and packing numbers with a more general outlook. First, these numbers can be defined in an arbitrary metric space, with a general metric replacing the Euclidean distance. Equivalently, one can replace covering by Euclidean balls in \( \mathbb{R}^n \) with covering by a translate of a general set \( D \).

**Exercise 4.2.12** (Covering and packing by general sets). Generalize the results of Sections 4.2 and 4.2.2 for covering and packing by translates of a general set \( D \) in place of the Euclidean balls.

### 4.3 Upper bounds on sub-gaussian matrices

#### 4.3.1 Computing the norm on a net

The notion of \( \varepsilon \)-nets can help us to simplify various problems involving dimensional sets. One such problem is the computation of the operator norm of an \( m \times n \) matrix \( A \). The operator norm was defined in Section 4.1.2 as

\[
\|A\| = \sup_{x \in S^{n-1}} \|Ax\|_2.
\]
Thus, to evaluate \(\|A\|\) one needs to control \(\|Ax\|\) uniformly over the sphere \(S^{n-1}\). We will show that instead of the entire sphere, it is enough to have a control just over an \(\epsilon\)-net of the sphere.

**Lemma 4.3.1** (Computing the operator norm on a net). Let \(A\) be an \(m \times n\) matrix and \(\epsilon \in [0, 1)\). Then, for any \(\epsilon\)-net \(\mathcal{N}\) of the sphere \(S^{n-1}\), we have

\[
\sup_{x \in \mathcal{N}} \|Ax\|_2 \leq \|A\| \leq \frac{1}{1-\epsilon} \cdot \sup_{x \in \mathcal{N}} \|Ax\|_2
\]

**Proof.** The lower bound in the conclusion is trivial since \(\mathcal{N} \subset S^{n-1}\). To prove the upper bound, fix a vector \(x \in S^{n-1}\) for which

\[
\|A\| = \|Ax\|_2
\]

and choose \(x_0 \in \mathcal{N}\) that approximates \(x\) so that

\[
\|x - x_0\|_2 \leq \epsilon.
\]

By the triangle inequality, this implies

\[
\|Ax - Ax_0\|_2 = \|A(x - x_0)\|_2 \leq \|A\|\|x - x_0\|_2 \leq \epsilon\|A\|.
\]

Using the triangle inequality again, we find that

\[
\|Ax_0\|_2 \geq \|Ax\|_2 - \|Ax - Ax_0\|_2 \geq \|A\| - \epsilon\|A\| = (1 - \epsilon)\|A\|.
\]

Dividing both sides of this inequality by \(1 - \epsilon\), we complete the proof.

Further, we may recall from Section 4.1.2 that the operator norm can be computed by maximizing a quadratic form:

\[
\|A\| = \max_{x \in S^{n-1}, y \in S^{m-1}} \langle Ax, y \rangle.
\]

Moreover, for symmetric matrices one can take \(x = y\). The following exercise shows that again, instead of controlling the quadratic form on the spheres, it suffices to have control just over the \(\epsilon\)-nets.

**Exercise 4.3.2.** 1. Let \(A\) be an \(m \times n\) matrix and \(\epsilon \in [0, 1/2)\). Show that for any \(\epsilon\)-net \(\mathcal{N}\) of the sphere \(S^{n-1}\) and any \(\epsilon\)-net \(\mathcal{M}\) of the sphere \(S^{m-1}\), we have

\[
\sup_{x \in \mathcal{N}, y \in \mathcal{M}} \langle Ax, y \rangle \leq \|A\| \leq \frac{1}{1 - 2\epsilon} \cdot \sup_{x \in \mathcal{N}, y \in \mathcal{M}} \langle Ax, y \rangle.
\]
2. Moreover, if $n = m$ and $A$ is symmetric, show that

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

Hint: Proceed similarly to the proof of Lemma 4.3.1 and use the identity $\langle Ax, y \rangle - \langle Ax_0, y_0 \rangle = \langle Ax - y, y_0 \rangle + \langle A(x - x_0), y_0 \rangle$.

**Exercise 4.3.3** (Bounding the norm deviation on a net). 1. Let $A$ be an $m \times n$ matrix, $\mu \in \mathbb{R}$ and $\varepsilon \in [0, 1/2)$. Show that for any $\varepsilon$-net $\mathcal{N}$ of the sphere $S^{n-1}$, we have

$$\sup_{x \in S^{n-1}} \|Ax\|_2 - \mu \leq C \frac{1 - 2\varepsilon}{1 - 2\varepsilon} \cdot \sup_{x \in \mathcal{N}} \|Ax\|_2 - \mu.$$ 

Hint: Assume that $\mu = 1$ without loss of generality. Represent $\|Ax\|_2^2 - 1$ as a quadratic form $\langle Rx, x \rangle$ where $R = A^T A - I_n$. Use Exercise 4.3.2 to compute the maximum of this quadratic form on a net.

### 4.3.2 The norms of sub-gaussian random matrices

We are ready for the first result on random matrices. It states that an $m \times n$ random matrix $A$ with independent sub-gaussian entries satisfies

$$\|A\| \lesssim \sqrt{m} + \sqrt{n}$$

with high probability.

**Theorem 4.3.4** (Norm of matrices with sub-gaussian entries). Let $A$ be an $m \times n$ random matrix whose entries $A_{ij}$ are independent, mean zero, sub-gaussian random variables. Then, for any $t > 0$ we have

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

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

**Proof.** This proof is a good example of an $\varepsilon$-net argument. We need to control $\langle Ax, y \rangle$ for all vectors $x$ and $y$ on the unit sphere. To this end, we will discretize the sphere using a net (approximation step), establish a tight control of $\langle Ax, y \rangle$ for fixed vectors $x$ and $y$ from the net (concentration step), and finish by taking a union bound over all $x$ and $y$ in the net.

**Step 1: Approximation.** Choose $\varepsilon = 1/4$. Using Corollary 4.2.11, we can find an $\varepsilon$-net $\mathcal{N}$ of the sphere $S^{n-1}$ and $\varepsilon$-net $\mathcal{M}$ of the sphere $S^{m-1}$ with cardinalities

$$|\mathcal{N}| \leq 9^n \quad \text{and} \quad |\mathcal{M}| \leq 9^m.$$  (4.4)
4.3. UPPER BOUNDS ON SUB-GAUSSIAN MATRICES

By Exercise 4.3.2, the operator norm of $A$ can be bounded using these nets as follows:

$$
\|A\| \leq 2 \max_{x \in \mathcal{N}, y \in \mathcal{M}} \langle Ax, y \rangle. \tag{4.5}
$$

**Step 2: Concentration.** Fix $x \in \mathcal{N}$ and $y \in \mathcal{M}$. Then the quadratic form

$$
\langle Ax, y \rangle = \sum_{i=1}^{n} \sum_{j=1}^{m} A_{ij} x_i y_j
$$

is a sum of independent, sub-gaussian random variables. Proposition 2.6.1 states that the sum is sub-gaussian, and

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

Recalling (2.15), we can restate this as the tail bound

$$
P \{ \langle Ax, y \rangle \geq u \} \leq 2 \exp(-cu^2/K^2), \quad u \geq 0. \tag{4.6}
$$

**Step 3: Union bound.** Next, we will unfix $x$ and $y$ using a union bound. Suppose the event $\max_{x \in \mathcal{N}, y \in \mathcal{M}} \langle Ax, y \rangle \geq u$ occurs. Then there exist $x \in \mathcal{N}$ and $y \in \mathcal{M}$ such that $\langle Ax, y \rangle \geq u$. Thus the union bound yields

$$
P \left\{ \max_{x \in \mathcal{N}, y \in \mathcal{M}} \langle Ax, y \rangle \geq u \right\} \leq \sum_{x \in \mathcal{N}, y \in \mathcal{M}} P \{ \langle Ax, y \rangle \geq u \}.
$$

Using the tail bound (4.6) and the estimate (4.4) on the sizes of $\mathcal{N}$ and $\mathcal{M}$, we bound the probability above by

$$
9^{n+m} \cdot 2 \exp(-cu^2/K^2). \tag{4.7}
$$

Choose

$$
u = CK(\sqrt{n} + \sqrt{m} + t). \tag{4.8}
$$

Then $u^2 \geq C^2 K^2 (n + m + t)$, and if constant $C$ is chosen sufficiently large, the exponent in (4.7) is large enough, say $cu^2/K^2 \geq 3(n + m) + t^2$. Thus

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

Finally, combining this with (4.5), we conclude that

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

Recalling our choice of $u$ in (4.8), we complete the proof. \qed
Optimality

Theorem 4.3.4 states that

$$\|A\| \lesssim \sqrt{m} + \sqrt{n} \tag{4.9}$$

with high probability. Is this bound optimal?

The operator norm of a matrix is always bounded below by the norms of any column and row, and in particular the first column and row. (Check!) Suppose the entries of $A$ have unit variances. Then, by Theorem 3.1.1, the Euclidean norm of the first column of $A$ is concentrated around $\sqrt{m}$, and the Euclidean norm of the first row of $A$ is concentrated around $\sqrt{n}$. Therefore

$$\|A\| \geq \max(\sqrt{m}, \sqrt{n}) \geq \frac{1}{2}(\sqrt{m} + \sqrt{n}),$$

so the upper bound (4.9) has the optimal form.

Symmetric matrices

Theorem 4.3.4 can be easily extended for symmetric matrices, and the bound for them is

$$\|A\| \lesssim \sqrt{n}$$

with high probability.

**Corollary 4.3.5** (Norm of symmetric matrices with sub-gaussian entries). Let $A$ be an $n \times n$ symmetric random matrix whose entries $A_{ij}$ on and above diagonal are independent, mean zero, sub-gaussian random variables. Then, for any $t > 0$ we have

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

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

**Proof.** Decompose $A$ into the upper-triangular part $A^+$ and lower-triangular part $A^-$. It does not matter where the diagonal goes; let us include it into $A^+$ to be specific. Then

$$A = A^+ + A^-.$$

Theorem 4.3.4 applies for each part $A^+$ and $A^-$ separately. By union bound, we have simultaneously

$$\|A^+\| \leq CK (\sqrt{n} + t) \quad \text{and} \quad \|A^+\| \leq CK (\sqrt{n} + t)$$

with probability at least $1 - 4 \exp(-t^2)$. Since by triangle inequality $\|A\| \leq \|A^+\| + \|A^-\|$, the proof is complete. \qed
4.4 Application: community detection in networks

We are going to illustrate Corollary 4.3.5 with an application to the analysis of networks. Real-world networks tend to have communities, or clusters, of tightly connected vertices. Finding the communities accurately and efficiently is one of the main problems in network analysis.

4.4.1 Stochastic Block Model

We will address this problem for a basic probabilistic model of a network with two communities. It is a simple extension of the Erdős-Rényi model of random graphs, which we described in Section 2.4.

**Definition 4.4.1 (Stochastic block model).** Divide \( n \) vertices into two sets (“communities”) of sizes \( n/2 \) each. Construct a random graph \( G \) by connecting every pair of vertices independently with probability \( p \) if they belong to the same community and \( q \) if they belong to different communities. This distribution on graphs is called the stochastic block model\(^1\) and is denoted \( G(n, p, q) \).

In the partial case where \( p = q \) we obtain the Erdős-Rényi model \( G(n, p) \). But we will assume that \( p > q \) here. In this case, edges are more likely to occur within than across communities. This gives the network a community structure; see Figure 4.4.

4.4.2 Expected adjacency matrix

Consider a random graph \( G \sim G(n, p, q) \). The adjacency matrix \( A \) of \( G \) is defined as the \( n \times n \) matrix with zero-one entries, thus \( A_{ij} = 1 \) if the vertices \( i \) and \( j \) are connected by an edge and \( A_{ij} = 0 \) otherwise. The adjacency matrix \( A \) of a random graph \( G \) is thus a random matrix, and we will examine \( A \) using the tools we developed in this chapter.

It is enlightening to split \( A \) into deterministic and random parts,

\[
A = D + R,
\]

where \( D \) is the expectation of \( A \). It is useful to think about \( D \) as an informative part (the “signal”) and \( R \) as a “noise”.

---

\(^1\)The term stochastic block model can also refer to a more general model of random graphs with multiple communities of variable sizes.
Figure 4.4: A random graph generated according to the stochastic block model $G(n,p,q)$.

To see why $D$ is informative, let us compute its eigenstructure. The entries $A_{ij}$ are $\text{Ber}(p)$ or $\text{Ber}(q)$ depending on community membership of vertices $i$ and $j$. Thus the entries of $D$ are either $p$ or $q$, depending on the membership. For illustration, if we group the vertices that belong to the same community together, then for $n = 4$ the matrix $D$ will look like this:

$$D = \mathbb{E} \, \text{A} = \begin{bmatrix} p & p & q & q \\ p & p & q & q \\ q & q & p & p \\ q & q & p & p \end{bmatrix}$$

**Exercise 4.4.2.** The matrix $D$ has rank 2. Check that the non-zero eigenvalues $\lambda_i$ and the corresponding eigenvectors $u_i$ of $D$ are

$$\lambda_1 = \left( \frac{p + q}{2} \right) n, \quad u_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}; \quad \lambda_2 = \left( \frac{p - q}{2} \right) n, \quad u_2 = \begin{bmatrix} 1 \\ -1 \\ 1 \\ -1 \end{bmatrix}. \quad (4.10)$$

The important object here is the second eigenvector $u_2$. It contains all information about community structure. If we knew $u_2$, we would identify the communities precisely based on the sizes of coefficients of $u_2$.

But we do not know $D = \mathbb{E} \, \text{A}$ and so we do not have access to $u_2$. Instead, we know $A = D + R$, a noisy version of $D$. The level of the signal
4.4. APPLICATION: COMMUNITY DETECTION IN NETWORKS

\[ \|D\| = \lambda_1 \sim n \]

while the level of the noise \( R \) can be estimated using Corollary 4.3.5:

\[ \|R\| \leq C\sqrt{n} \quad \text{with probability at least } 1 - 4e^{-n}. \quad (4.11) \]

So for large \( n \), the signal to noise ratio is quite large, which should allow us to use \( A + D \) instead of \( D \) to extract the community information. This can be justified using the classical perturbation theory for matrices.

4.4.3 Perturbation theory

Perturbation theory describes how the eigenvalues and eigenvectors change under matrix perturbations. For eigenvalues, a simple argument shows that symmetric matrices \( S \) and \( T \) satisfy

\[ \max_i |\lambda_i(S) - \lambda_i(T)| \leq \|S - T\|. \]

Thus, operator norm of the perturbation controls the stability of spectrum.

A similar result holds for eigenvectors. We need to be careful to track the same eigenvector before and after perturbation. If the eigenvalues \( \lambda_i(S) \) and \( \lambda_{i+1}(S) \) are too close to each other, the perturbation can swap their order and force us to compare wrong eigenvectors. To prevent this from happening, we need to assume that eigenvalues of \( S \) are well separated.

**Theorem 4.4.3 (Davis-Kahan).** Let \( S \) and \( T \) be symmetric matrices with the same dimensions. Fix \( i \) and assume that the \( i \)-th largest eigenvalue of \( S \) is well separated from the rest of the spectrum:

\[ \min (|\lambda_i(S) - \lambda_j(S)| : j \neq i) = \delta > 0. \]

Then the angle between the eigenvectors of \( S \) and \( T \) corresponding to the \( i \)-th largest eigenvalues (as a number between 0 and \( \pi/2 \)) satisfies

\[ \sin \angle (v_i(S), v_i(T)) \leq \frac{C\|S - T\|}{\delta}. \]

In particular, the conclusion of Davis-Kahan Theorem implies that if the eigenvectors \( v_i(S) \) and \( v_i(T) \) have unit norm, then they are close to each other (up to a sign):

\[ \exists \theta \in \{-1, 1\} : \|v_i(S) - \theta v_i(T)\|_2 \leq \frac{C\|S - T\|}{\delta}. \quad (4.12) \]
4.4.4 Spectral Clustering

Let us apply Davis-Kahan Theorem for $S = D$ and $T = A = D + R$, and for the second largest eigenvalue. We need to check that $\lambda_2$ is well separated from the rest of the spectrum of $D$, that is from 0 and $\lambda_1$. The distance is

$$\delta = \min(\lambda_2, \lambda_1 - \lambda_2) = \min \left( \frac{p-q}{2}, q \right) n =: \mu n.$$ 

Recalling the bound (4.11) on $R = T - S$ and applying (4.12), we can bound the distance between the normalized eigenvectors of $D$ and $A$. There exists a sign $\theta \in \{-1, 1\}$ such that

$$\|v_2(D) - v_2(A)\|_2 \leq \frac{C \sqrt{n}}{\mu n} = \frac{C}{\mu \sqrt{n}},$$

with probability at least $1 - 4e^{-n}$. We computed the eigenvectors of $D$ in (4.10), but there they had norm $\sqrt{n}$. So, multiplying both sides by $\sqrt{n}$, we obtain in this normalization that

$$\|u_2(D) - \theta u_2(A)\|_2 \leq \frac{C}{\mu}.$$ 

It follows from this that that the signs of most coefficients of $\theta v_2(A)$ and $v_2(D)$ must agree. Indeed, we know that

$$\sum_{j=1}^{n} |u_2(D)_j - \theta u_2(A)_j|^2 \leq \frac{C}{\mu^2}. \quad (4.13)$$

and we also know from (4.10) that the coefficients $u_2(D)_j$ are all $\pm 1$. So, every coefficient $j$ on which the signs of $\theta v_2(A)_j$ and $v_2(D)_j$ disagree contributes at least 1 to the sum in (4.13). Thus the number of disagreeing signs must be bounded by

$$\frac{C}{\mu^2}.$$ 

Summarizing, we can use the vector $v_2(A)$ that can be computed from the network to accurately estimate the vector $v_2 = v_2(D)$ in (4.10), whose signs identify the two communities. This method for community detection is usually called em spectral clustering. Let us explicitly state this method and the guarantees that we just obtained.
4.5. TWO-SIDED BOUNDS ON SUB-GAUSSIAN MATRICES

\begin{center}
\textbf{Spectral Clustering Algorithm}
\end{center}

\textbf{Input:} graph $G$

\textbf{Output:} a partition of the vertices of $G$ into two communities

1: Compute the adjacency matrix $A$ of the graph.
2: Compute the eigenvector $v_2(A)$ corresponding to the second largest eigenvalue of $A$.
3: Partition the vertices into two communities based on the signs of the coefficients of $v_2(A)$. (To be specific, if $v_2(A)_j > 0$ put vertex $j$ into first community, otherwise in the second.)

\textbf{Theorem 4.4.4} (Spectral clustering of the stochastic block model). Let $G \sim G(n,p,q)$ with $p > q$, and $\min(q,p-q) = \mu > 0$. Then, with probability at least $1 - 4e^{-n}$, the Spectral Clustering Algorithm identifies the communities of $G$ correctly up to $C/\mu^2$ misclassified vertices.

Summarizing, the Spectral Clustering correctly classifies all but a constant number of vertices, provided the random graph is dense enough ($q \geq \text{const}$) and the probabilities of within- and across-community edges are well separated ($p - q \geq \text{const}$).

4.5 Two-sided bounds on sub-gaussian matrices

Let us return to Theorem 4.3.4, which gives an upper bound on the spectrum of an $n \times m$ matrix $A$ with independent sub-gaussian entries:

$$s_1(A) \leq C(\sqrt{m} + \sqrt{n})$$

with high probability. We will now improve this result in two important ways.

First, we are going to prove sharper, \textit{two-sided bounds} on the entire spectrum of $A$:

$$\sqrt{m} - C\sqrt{n} \leq s_i(A) \leq \sqrt{m} + C\sqrt{n}.$$ 

Thus a tall random matrix (with $m \ll n$) is an approximate isometry in the sense of Section 4.1.4.

Second, the independence of entries is going to be relaxed to just \textit{independence of rows}. We will still require sub-gaussian tails, so this time we assume that the rows are sub-gaussian random vectors. (We studied such vectors in Section 3.4). This relaxation of independence is important in some applications to data sciences, in particular where the rows of $A$ are samples.
from a high-dimensional distribution. The samples are usually independent, and so are the rows of \( A \). But there is no reason to assume independence of columns of \( A \), since the coordinates of the distribution (the “parameters”) are not usually independent.

**Theorem 4.5.1** (Two-sided bound on sub-gaussian matrices). Let \( A \) be an \( m \times n \) matrix whose rows \( A_i \) are independent, mean zero, sub-gaussian isotropic random vectors in \( \mathbb{R}^n \). Then for any \( t \geq 0 \) we have

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

(4.14) with probability at least \( 1 - 2\exp(-t^2) \). Here \( K = \max_i \|A_i\|_{\psi_2} \).

Before we prove this theorem, let us note that by Lemma 4.1.2, we can equivalently restate the conclusion (4.14) in the following form:

\[
\|1/m A^T A - I_m\| \leq \max(\delta, \delta^2)
\]

where

\[
\delta = CK^2\left(\frac{\sqrt{n}}{m} + \frac{t}{\sqrt{m}}\right).
\]

(4.15)

**Proof.** We will prove (4.15) using an \( \varepsilon \)-net argument. This will be similar to the proof of Theorem 4.3.4, but we will now use Bernstein’s concentration inequality instead of Hoeffding’s.

**Step 1: Approximation.** Using Corollary 4.2.11, we can find an \( \frac{1}{4} \)-net \( \mathcal{N} \) of the unit sphere \( S^{n-1} \) with cardinality

\[
|\mathcal{N}| \leq 9^n.
\]

Using Lemma 4.3.1, we can evaluate the operator norm in (4.15) on the \( \mathcal{N} \):

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

To complete the proof of (4.15) it suffices to show that, with the required probability,

\[
\max_{x \in \mathcal{N}} \left| \frac{1}{m} \|Ax\|_2^2 - 1 \right| \leq \frac{\varepsilon}{2} \quad \text{where} \quad \varepsilon := \max(\delta, \delta^2).
\]

**Step 2: Concentration.** Fix \( x \in S^{n-1} \) and express \( \|Ax\|_2^2 \) as a sum of independent random variables:

\[
\|Ax\|_2^2 = \sum_{i=1}^m \langle A_i, x \rangle^2 =: \sum_{i=1}^m X_i^2
\]

(4.16)
4.5. TWO-SIDED BOUNDS ON SUB-GAUSSIAN MATRICES

where $A_i$ denote the rows of $A$. By assumption, $A_i$ are independent, isotropic, and sub-gaussian random vectors with $\|A_i\|_{\psi_2} \leq K$. Thus $X_i = \langle A_i, x \rangle$ are independent sub-gaussian random variables with $\mathbb{E} X_i^2 = 1$ and $\|X_i\|_{\psi_2} \leq K$. Therefore $X_i^2 - 1$ are independent, mean zero, and sub-exponential random variables with

$$\|X_i^2 - 1\|_{\psi_1} \leq CK^2.$$  

(Check this; we did a similar computation in the proof of Theorem 3.1.1.)

Thus we can use Bernstein’s inequality (Corollary 2.8.4) and obtain

$$P \left\{ \left| \frac{1}{m} \|Ax\|_2^2 - 1 \right| \geq \frac{\varepsilon}{2} \right\} = \mathbb{P} \left\{ \left| \frac{1}{m} \sum_{i=1}^{m} X_i^2 - 1 \right| \geq \frac{\varepsilon}{2} \right\}$$

$$\leq 2 \exp \left[ - \frac{c_1}{K^4} \min(\varepsilon^2, \varepsilon)m \right]$$

$$= 2 \exp \left[ - \frac{c_1}{K^4} \delta^2 m \right] \quad \text{(since } \varepsilon = \max(\delta, \delta^2))$$

$$\leq 2 \exp \left[ - c_1 C^2 (n + t^2) \right].$$

The last bound follows from the definition of $\delta$ in (4.15) and using the inequality $(a + b)^2 \geq a^2 + b^2$ for $a, b \geq 0$.

**Step 3: Union bound.** Now we can unfix $x \in \mathcal{N}$ using a union bound. Recalling that $\mathcal{N}$ has cardinality bounded by $9^n$, we obtain

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

As we noted in Step 1, this completes the proof of the theorem. \qed

**Exercise 4.5.2.** Give a simpler proof of Theorem 4.5.1, using Theorem 3.1.1 to obtain a concentration bound for $\|Ax\|_2$ and Exercise 4.3.3 to reduce to a union bound over a net.

Let us emphasize the alternative form (4.15) of Theorem 4.5.1, which is important on its own. Replacing there $t$ with $t\sqrt{n}$ and assuming that $t \geq 1$, we can restate it as follows:

$$\left\| \frac{1}{m} A^T A - I_n \right\| \leq \max(\delta, \delta^2)$$

where $\delta = CK^2 t \sqrt{\frac{n}{m}}$ \hspace{1cm} (4.17)

with probability at least $1 - 2 \exp(-t^2n)$.

**Exercise 4.5.3** (Non-isotropic distributions). Prove the following version of (4.17) for non-isotropic distributions. Let $A$ be an $m \times n$ matrix whose
CHAPTER 4. SUB-GAUSSIAN RANDOM MATRICES

rows $A_i$ are independent, mean zero, sub-gaussian random vectors in $\mathbb{R}^n$ with the same covariance matrix

$$\Sigma = \mathbb{E} A_i A_i^T.$$  

Then for any $t \geq 0$ we have

$$\| \frac{1}{m} A^T A - \Sigma \| \leq \max(\delta, \delta^2) \quad \text{where} \quad \delta = CK^2 t \sqrt{\frac{n}{m}}$$

with probability at least $1 - 2 \exp(-t^2 n)$. Here $K = \max_i \| A_i \|_{\psi_2}$.

4.6 Application: covariance estimation and clustering

Suppose we are analyzing high dimensional data, which is represented as points $X_1, \ldots, X_m$ sampled from an unknown distribution in $\mathbb{R}^n$. The most basic method is Principal Component Analysis (PCA), which we discussed briefly in Section 3.2.1. The goal of PCA is to identify the principal components the eigenvectors of the covariance matrix of the distribution.

Since we do not have access to the full distribution but only to the finite sample $\{X_1, \ldots, X_m\}$, we can only expect to compute the covariance matrix and its eigenvectors approximately. How can we do this? Let $X$ denote the random vector drawn from the (unknown) distribution. Assume for simplicity that $X$ have zero mean, and let us denote the covariance matrix

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

To estimate $\Sigma$, we can use the sample covariance matrix $\Sigma_m$ that is computed from the sample $X_1, \ldots, X_m$ as follows:

$$\Sigma_m = \frac{1}{m} \sum_{i=1}^{m} X_i X_i^T.$$  

(Basically, we replace the expectation over the entire distribution by expectation over the sample.)

Since $X_i$ and $X$ are identically distributed, our estimate is unbiased, that is

$$\mathbb{E} \Sigma_m = \Sigma.$$  

Moreover, by the Law of Large Numbers (Theorem 1.3.1),

$$\Sigma_m \rightarrow \Sigma \quad \text{almost surely}$$
as the sample size \( m \) increases to infinity. (Justify that we can apply the Law of Large Numbers to matrices.)

This leads to the quantitative question: how many sample points \( m \) are needed to guarantee that

\[
\Sigma_m \approx \Sigma
\]

with high probability? For dimension reasons, we need at least \( m \gtrsim n \) sample points. (Why?) And we will now show that \( m \sim n \) sample points suffice.

**Theorem 4.6.1** (Covariance estimation). Consider a sub-gaussian random vector \( X \) in \( \mathbb{R}^n \) with zero mean and covariance matrix \( \Sigma \), and let \( \varepsilon \in (0, 1) \) and \( t \geq 1 \). Suppose the sample size satisfies

\[
m \geq CK^4(t/\varepsilon)^2n.
\]

Then the sample covariance matrix \( \Sigma_m \) satisfies

\[
\|\Sigma_m - \Sigma\| \leq \varepsilon
\]

with probability at least \( 1 - 2 \exp(-t^2n) \). Here \( K = \|X\|_{\psi_2} \).

**Proof.** Consider the \( m \times n \) matrix \( A \) whose rows are the sample points \( X_i^T \). Then the sample covariance matrix \( \Sigma \) can be represented as

\[
\Sigma_m = \sum_{i=1}^{m} X_i X_i^T = \frac{1}{m} A^T A.
\]

So we can apply the non-isotropic form of Theorem 4.5.1 stated in Exercise 4.5.3 to bound the error \( \|\Sigma_m - \Sigma\| \). Setting the error bound \( \max(\delta, \delta^2) \) to \( \varepsilon \) and solving for \( m \), we complete the proof.

Let us emphasize the meaning of this result. The covariance matrix can be estimated accurately by the sample covariance matrix, if the size of the sample \( m \) is proportional to the dimension \( n \).

**4.6.1 Application: clustering of point sets**

We are going to illustrate Theorem 5.6.1 with an application to clustering. The problem here will be similar to the community detection problem we studied in Section 4.4. Like before, we will try to partition data into clusters, but the nature of data will be different. Instead of networks, we will now be working with point sets in \( \mathbb{R}^n \). The general goal is to partition the points
into few subsets (“clusters”). What exactly constitutes cluster is not well defined in data sciences. But the common sense suggests that the points in the same cluster should tend to be closer to each other than points taken from different clusters.

Just like we did for networks, we will design a basic probabilistic model of point sets in $\mathbb{R}^n$ with two communities, and we will study the clustering problem for that model.

**Definition 4.6.2 (Gaussian mixture model).** Generate $m$ random points in $\mathbb{R}^n$ as follows. Flip a fair coin. If we get heads, draw a point from $N(\mu, I_n)$, and if we get tails, from $N(-\mu, I_n)$. We call this a Gaussian mixture model with means $\mu$ and $-\mu$.

Equivalently, we may consider a random vector

$$X = \theta \mu + g$$

where $\theta$ is a symmetric Bernoulli random variable, $g \in N(0, I_n)$, and $\theta$ and $g$ are independent. Draw a sample $X_1, \ldots, X_m$ of independent random vectors identically distributed with $X$. Then the sample is distributed according to the Gaussian mixture model. Figure 4.5 illustrates a simulation of Gaussian mixture model.

Figure 4.5: A simulation of points generated according to the Gaussian mixture model, which has two clusters with different means.

A basic clustering method, called *spectral clustering*, is based on the Principal Component Analysis (PCA) of the data, which we outlined in Section 3.2.1. The distribution of $X$ is not isotropic; it is stretched in the direction of $\mu$. The first principal component of the data should also be close to $\mu$, and thus one should be able classify the data points by projecting them onto that first principal component. This is formalized in the following basic clustering algorithm.
Spectral Clustering Algorithm

Input: points $X_1, \ldots, X_m$ in $\mathbb{R}^n$

Output: a partition of the points into two clusters

1. Compute the sample covariance matrix $\Sigma_m = \frac{1}{m} \sum_{i=1}^{m} X_i X_i^T$.
2. Compute the eigenvector $v = v_1(\Sigma_m)$ corresponding to the largest eigenvalue of $\Sigma_m$.
3. Partition the vertices into two communities based on the signs of the inner product of $v$ with the data points. (To be specific, if $\langle v, X_i \rangle > 0$ put point $X_i$ into first community, otherwise in the second.)

Theorem 4.6.3 (Spectral clustering of the Gaussian mixture model). Let $X_1, \ldots, X_m$ be points in $\mathbb{R}^n$ drawn from the Gaussian mixture model as above, i.e. there are two communities with means $\mu$ and $-\mu$, and let $\varepsilon, t > 0$. Suppose the sample size satisfies

$$m \geq \text{poly}(n, \frac{1}{\varepsilon}, \frac{1}{\|\mu\|_2}).$$

Then, with probability at least $1 - 4e^{-n}$, the Spectral Clustering Algorithm identifies the communities correctly up to $\varepsilon n$ misclassified vertices.

Exercise 4.6.4. Prove Theorem 4.6.3 along the following lines.

1. Compute the covariance matrix $\Sigma$ of $X$ and note that the eigenvector corresponding to the largest eigenvalue is parallel to $\mu$.
2. Use results about covariance estimation (such as Exercise 4.5.3) to show that the sample covariance matrix $\Sigma_m$ is close to $\Sigma$.
3. Use Davis-Kahan Theorem 4.4.3 to deduce that the eigenvector $v = v_1(\Sigma_m)$ is close to $\mu$.
4. If conclude that the signs of $\langle \mu, X_i \rangle$ predict well which community $X_i$ belongs to.
5. Since $v \approx \mu$, conclude the same for $v$. 
Chapter 5

Concentration without independence

Our approach to concentration was crucially based on independence of random variables. This was clearly the case for sums of independent random variables we studied in Chapter 2: our later results were based on concentration for these sums. We will now develop alternative approaches to concentration that are not based on independence.

5.1 Concentration of Lipschitz functions on the sphere

Consider a Gaussian random vector $X \sim N(0,I_n)$ and a function $f : \mathbb{R}^n \to \mathbb{R}$. When does the random vector $f(X)$ concentrate about its mean, i.e.

$$f(X) \approx \mathbb{E} f(X)$$

with high probability?

This question is easy for linear functions $f$. Indeed, in this case $f(X)$ has normal distribution, and it concentrates around its mean well.

In this section, we will study concentration of non-linear functions $f(X)$ of random vectors $X \sim \text{Unif}(S^{n-1})$ and $X \sim N(0,I_n)$. While we can not expect concentration for completely arbitrary $f$ (why?), the Lipschitz requirement for $f$ will be enough.

5.1.1 Lipschitz functions

**Definition 5.1.1** (Lipschitz functions). Let $(X,d_X)$ and $(Y,d_Y)$ be metric spaces. A function $f : X \to Y$ is called Lipschitz if there exists $L \in \mathbb{R}$ such
that
\[ d_X(f(u), f(v)) \leq d_Y(u, v) \quad \text{for every } u, v \in X. \]

The infimum of all \( L \) in this definition is called the Lipschitz norm of \( f \) and is denoted \( \| f \|_{\text{Lip}} \).

Lipschitz functions with \( \| f \|_{\text{Lip}} \leq 1 \) are usually called contractions.

**Exercise 5.1.2.**

1. \( f(x) = |x| \) is a Lipschitz a function on \( \mathbb{R} \), while \( f(x) = \sqrt{x} \) and \( f(x) = x^2 \) are not.
2. \( f(x) = \|x\|_2 \) is a Lipschitz function on \( \mathbb{R}^n \), and \( \| f \|_{\text{Lip}} = 1 \).
3. Every differentiable function \( f : \mathbb{R}^n \to \mathbb{R} \) is Lipschitz, and
   \[ \| f \|_{\text{Lip}} \leq \| \nabla f \|_{\infty}. \]
4. For a fixed \( \theta \in \mathbb{R}^n \), the linear functional \( f(x) = \langle x, \theta \rangle \) is a Lipschitz function on \( \mathbb{R}^n \), and \( \| f \|_{\text{Lip}} = \| \theta \|_2 \).
5. More generally, an \( m \times n \) matrix \( A \) acting as a linear operator between \( A : (\mathbb{R}^n, \| \cdot \|_2) \to (\mathbb{R}^m, \| \cdot \|_2) \) is Lipschitz, and
   \[ \| A \|_{\text{Lip}} = \| A \|. \]
6. Any norm \( f(x) = \|x\| \) on \( \mathbb{R}^n \) is a Lipschitz function. The Lipschitz norm of \( f \) is the smallest \( L \) such that
   \[ \|x\| \leq L \|x\|_2 \quad \text{for all } x \in \mathbb{R}^n. \]

### 5.1.2 Concentration via isoperimetric inequalities

The main result of this section is that any Lipschitz function on the sphere concentrates well.

**Theorem 5.1.3 (Concentration of Lipschitz functions on the sphere).** Consider a random vector \( X \sim \text{Unif}(\sqrt{n}S^{n-1}) \) and a Lipschitz function\(^1\) \( f : \sqrt{n}S^{n-1} \to \mathbb{R} \). Then
\[ \| f(X) - \mathbb{E} f(X) \|_{\psi_2} \leq C \| f \|_{\text{Lip}}. \]

---

\(^1\)This theorem is valid for both the geodesic metric on the sphere (where \( d(x, y) \) is the length of the shortest arc connecting \( x \) and \( y \)) and the Euclidean metric \( d(x, y) = \|x - y\|_2 \). We will prove the theorem for the Euclidean metric; Exercise ?? extends it to geodesic metric.
5.1. **CONCENTRATION OF LIPSCHITZ FUNCTIONS ON THE SPHERE**

Equivalently, Theorem 5.1.3 states that for every $t \geq 0$, we have

$$\mathbb{P} \{|f(X) - \mathbb{E} f(X)| \geq t\} \leq 2 \exp \left(- \frac{ct^2}{\|f\|_{Lip}^2}\right)$$

We already know that Theorem 5.1.3 holds for linear functions. Indeed, Theorem 3.4.5 states that $X \sim \text{Unif}(\sqrt{n}S^{n-1})$ is a sub-gaussian random vector. By definition, this means that any linear function of $X$ is a sub-gaussian random variable.

To prove Theorem 5.1.3 in full generality, we will argue that non-linear functions must concentrate at least as good as linear functions. To compare non-linear to linear functions, it is enough to compare their sub-level sets: arbitrary sets on the sphere and the spherical caps. Such comparison will be based on a geometric principle called *isoperimetric inequality*.

The classical isoperimetric inequality in $\mathbb{R}^3$ (and also in $\mathbb{R}^n$) states that among all sets with given volume, the area is minimal for the Euclidean balls. A similar isoperimetric inequality holds on the sphere, and the minimizers are the spherical caps. To state it carefully, we will denote the normalized area ($n-1$-dimensional Lebesgue measure) on the sphere $S^{n-1}$ by $\sigma_{n-1}$. An $\varepsilon$-neighborhood of a set $A \subset S^{n-1}$ is defined as

$$A_\varepsilon := \{x \in S^{n-1} : \exists y \in A, \|x - y\|_2 \leq \varepsilon\} = (A + \varepsilon B^n_2) \cap S^{n-1}. \quad (5.1)$$

Here we used the notation for Minkowski sum introduced in Definintion 4.2.9. Figure 5.1 illustrates the $\varepsilon$-neighborhood of $A$. The perimeter of $A$ is the $(n-2)$-dimensional area of boundary $\partial A$, and it can be defined as

$$\text{Area}(\partial A) := \lim_{\varepsilon \to 0} \frac{\sigma_{n-1}(A_\varepsilon) - \sigma_{n-1}(A)}{\varepsilon}.$$
Theorem 5.1.4 (Isoperimetric inequality on the sphere). 1. Let $\varepsilon > 0$. Then, among all sets $A \subset S^{n-1}$ with fixed area $\sigma_{n-1}(A)$, the spherical caps minimize the area of the neighborhood $\sigma_{n-1}(A_{\varepsilon})$.

2. Among all sets $A \subset S^{n-1}$ with fixed area $\sigma_{n-1}(A)$, the spherical caps minimize the perimeter $\text{Area}(\partial A)$.

We will not prove this theorem but just note that Part 2 follows from Part 1 by letting $\varepsilon \to 0$. (Check!)

5.1.3 Blow-up of sets on the sphere

Now we will deduce from isoperimetric a statement that may sound counter-intuitive. If $A$ makes up at least half of the sphere (in terms of volume) then $A_{\varepsilon}$ will make up most of the sphere. This fact nevertheless is simple to check for a hemisphere, and then extend to general sets using isoperimetric inequality. In view of Theorem 5.1.3, it will be convenient to work with the sphere that is scaled by the factor $\sqrt{n}$.

Lemma 5.1.5 (Blow-up of neighborhoods on the sphere). Let $A$ be a subset of the sphere $\sqrt{n}S^{n-1}$, and let $\sigma$ denote the normalized area on that sphere. If $\sigma(A) \geq 1/2$ then, for every $t \geq 0$,

$$\sigma(A_t) \geq 1 - 2 \exp(-ct^2).$$

Proof. Consider the hemisphere

$$H := \{x \in \sqrt{n}S^{n-1}: x_1 \leq 0\}.$$

Then $\sigma(A) \geq \sigma(H) = 1/2$, and the isoperimetric inequality (Theorem 5.1.4) implies that

$$\sigma(A_t) \geq \sigma(H_t). \quad (5.2)$$

The set $H_t$ is a spherical cap, and it should be easy to compute its area. It is even easier to use Theorem 3.4.5 instead, which states a random vector

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

is sub-gaussian, and $\|X\|_{\psi_2} \leq C$. Since $\sigma$ is the uniform probability measure on the sphere, it follows that

$$\sigma(H_t) = \mathbb{P}\{X \in H_t\}.$$

Now, the definition of the neighborhood (5.1) implies that

$$H_t \supset \left\{x \in \sqrt{n}S^{n-1}: x_1 \leq \sqrt{2t}\right\}.$$
5.1. Concentration of Lipschitz Functions on the Sphere

Thus
\[ \sigma(H_t) \geq \mathbb{P}\{X_1 \leq \sqrt{2t}\} \geq 1 - 2\exp(-ct^2). \]
The last inequality holds because \( \|X_1\|_{\psi_2} \leq \|X\|_{\psi_2} \leq C. \) In view of (5.2), the lemma is proved.

5.1.4 Proof of Theorem 5.1.3

Without loss of generality, we can assume that \( \|f\|_{\text{Lip}} = 1. \) (Why?) Let \( M \) denote a median of \( f(X) \), which by definition is a number satisfying\(^2\)
\[ \mathbb{P}\{f(X) \leq M\} \geq \frac{1}{2} \quad \text{and} \quad \mathbb{P}\{f(X) \geq M\} \geq \frac{1}{2}. \]
Consider the sub-level set
\[ A := \{ x \in \sqrt{n}S^{n-1} : f(x) \leq M \}. \]
Since \( \mathbb{P}\{X \in A\} \geq 1/2, \) Lemma 5.1.5 implies that
\[ \mathbb{P}\{X \in A_t\} \geq 1 - 2\exp(-ct^2). \quad (5.3) \]
On the other hand, we claim that
\[ \mathbb{P}\{X \in A_t\} \leq \mathbb{P}\{f(X) \leq M + t\}. \quad (5.4) \]
Indeed, if \( X \in A_t \) then \( \|X - y\|_2 \leq t \) for some point \( y \in A \). By definition, \( f(y) \leq M. \) Since \( f \) Lipschitz with \( \|f\|_{\text{Lip}} = 1, \) it follows that
\[ f(X) \leq f(y) + \|X - y\|_2 \leq M + t. \]
This proves our claim (5.4).

Combining (5.3) and (5.4), we conclude that
\[ \mathbb{P}\{f(X) \leq M + t\} \geq 1 - 2\exp(-ct^2). \]
Repeating the argument for \(-f\), we obtain a similar bound for the probability that \( f(X) \geq M - t. \) Combining the two, we obtain a similar bound for the probability that \( |f(X) - M| \leq t, \) showing that
\[ \|f(X) - M\|_{\psi_2} \leq C. \]
It remains to replace the median \( M \) by the expectation \( \mathbb{E}f. \) This can be done automatically by applying the Centering Lemma 2.6.6. (Do this!) The proof of Theorem 5.1.3 is now complete. \( \square \)

\(^2\)The median may not be unique. However, for continuous and one-to-one functions \( f, \) the median is unique. (Check!)
Exercise 5.1.6 (Geodesic metric). We proved Theorem 5.1.3 for functions $f$ that are Lipschitz with respect to the Euclidean metric $\|x - y\|_2$ on the sphere. Argue that the same result holds for the geodesic metric (the length of the shortest arc connecting $x$ and $y$.)

Exercise 5.1.7 (Concentration on the unit sphere). We stated Theorem 5.1.3 for the scaled sphere $\sqrt{n}S^{n-1}$. Deduce that a Lipschitz function $f$ on the unit sphere $S^{n-1}$ satisfies
\[\|f(X) - \mathbb{E} f(X)\|_{\psi_2} \leq \frac{C\|f\|_{\text{Lip}}}{\sqrt{n}}.\] (5.5)

where $X \sim \text{Unif}(S^{n-1})$. Equivalently, for every $t \geq 0$, we have
\[\mathbb{P}\{|f(X) - \mathbb{E} f(X)| \geq t\} \leq 2 \exp\left(-\frac{c n t^2}{\|f\|_{\text{Lip}}^2}\right)\] (5.6)

Exercise 5.1.8 (Exponential set of mutually almost orthogonal points). Fix $\varepsilon \in (0, 1)$. Show that there exists a set $\{x_1, \ldots, x_N\}$ of unit vectors in $\mathbb{R}^n$ which are mutually almost orthogonal:
\[|\langle x_i, x_j \rangle| \leq \varepsilon \quad \text{for all } i \neq j,\]
and the set is exponentially large in $n$:
\[N \geq \exp(c(\varepsilon)n).\]

Hint: Construct the points $x_i \in S^{n-1}$ one at a time. Note that the set of points on the sphere that are almost orthogonal with a given point $x_0$ form a spherical cap. Show that the normalized area of that cap is exponentially small.

5.2 Concentration on other metric measure spaces

In this section, we will extend the concentration for the sphere to other spaces. To do this, note that our proof of Theorem 5.1.3 was based on two main ingredients:
(a) an isoperimetric inequality;
(b) a blow-up of the minimizers for the isoperimetric inequality.

The sphere is not the only space where these two ingredients are in place. In this section, we will briefly survey several other metric measure spaces where (a) and (b) can be shown, and thus concentration of Lipschitz functions almost automatically follows.
5.2. CONCENTRATION ON OTHER METRIC MEASURE SPACES

5.2.1 Gauss space

Theorem 5.1.3 can be proved for the normal random vector \( X \sim N(0, I_n) \).

**Theorem 5.2.1** (Concentration on the Gauss space). Consider a random vector \( X \sim N(0, I_n) \) and a Lipschitz function \( f : \mathbb{R}^n \to \mathbb{R} \) (with respect to the Euclidean metric). Then

\[
\| f(X) - \mathbb{E} f(X) \|_{\psi_2} \leq C \| f \|_{\text{Lip}}. \tag{5.7}
\]

The proof is similar to Theorem 5.1.3, and it is based on the following isoperimetric inequality in the Gauss space\(^3\) \((\mathbb{R}^n, \| \cdot \|_2, \gamma_n)\).

**Theorem 5.2.2** (Isoperimetric inequality on the Gauss space). Let \( \varepsilon > 0 \). Then, among all sets \( A \subset \mathbb{R}^n \) with fixed Gaussian measure \( \gamma_n(A) \), the half spaces minimize the area of the neighborhood \( \gamma_n(A_\varepsilon) \).

**Exercise 5.2.3.** Deduce Theorem 5.2.1. Hint: The \( \varepsilon \)-neighborhood of a half-space is still a half-space, and its Gaussian measure should be easy to compute.

**Remark 5.2.4.** We came across two partial cases of Theorem 5.2.1 before.

1. For linear functions \( f \), concentration follows from the fact the normal distribution \( N(0, I_n) \) is sub-gaussian.
2. For the Euclidean norm \( f(x) = \|x\|_2 \), concentration follows from Theorem 3.1.1.

**Exercise 5.2.5** (Replacing expectation by \( L_p \) norm). Prove that in the concentration results for sphere and Gauss space (Theorems 5.1.3 and 5.2.1), the expectation \( \mathbb{E} f(X) \) can be replaced by the \( L_p \) norm \( (\mathbb{E} f^p)^{1/p} \) for any \( p > 0 \) and for any non-negative function \( f \). The constants may depend on \( p \).

5.2.2 Discrete cube

A similar method based on isoperimetry yields concentration on many other metric measure spaces. One of them is the discrete cube

\[
(\{0, 1\}^n, d, \mathbb{P}).
\]

Here \( d(x, y) \) is the Hamming distance, which is defined for binary strings \( x, y \in \{0, 1\}^n \) as the fraction of the digits where \( x \) and \( y \) disagree:

\[
d(x, y) = \frac{1}{n} \left| \{i : x_i \neq y_i\} \right|.
\]

\(^3\)Here the measure \( \gamma_n \) has the standard normal density (3.3).
The measure $\mathbb{P}$ is the uniform probability measure on the discrete cube. So, the coordinates of the random vector

$$X \sim \text{Unif} \left( \{0, 1\}^n \right)$$

are independent $\text{Ber}(1/2)$ random variables.

The concentration inequality for the discrete cube states that

$$\|f(X) - \mathbb{E} f(X)\|_{\psi_2} \leq \frac{C\|f\|_{\text{Lip}}}{\sqrt{n}}$$

(5.8)

for any function $f : \{0, 1\}^n \to \mathbb{R}$, see [4, Section 6.2]. This result can be deduced from the isoperimetric inequality on the cube, whose minimizers are known to be the Hamming cubes – the neighborhoods of single points with respect to the Hamming distance.

### 5.2.3 Symmetric group

The permutation group $S_n$ consists of permutations of $n$ symbols, which we choose to be $\{1, \ldots, n\}$ to be specific. We can view the symmetric group as a metric measure space

$$(S_n, d, \mathbb{P}).$$

Here $d(\pi, \rho)$ is the Hamming distance – the fraction of the symbols where the permutations $\pi$ and $\rho$ disagree:

$$d(\pi, \rho) = \frac{1}{n} |\{i : \pi(i) \neq \rho(i)\}|.$$

The measure $\mathbb{P}$ is the uniform probability measure on $S_n$.

Then concentration inequality (5.8) holds for any function $f : S_n \to \mathbb{R}$, see [4, Section 6.3].

### 5.2.4 Riemannian manifolds with strictly positive curvature

In addition of arguments based on isoperimetry, there are several other methods to prove concentration of Lipshitz functions. For a thorough treatment of this topic, we refer the reader to the books [2, 1, 4, 3]; here we will briefly survey some of these results.

A very general class of examples is covered by the notion of a Riemannian manifold. We refer the reader to ... for necessary background in differential geometry, and here we will just mention a relevant concentration result.

Let $(M, g)$ be a compact connected smooth Riemannian manifold. The canonical distance $d(x, y)$ on $M$ is defined as the the arclength (with respect
to the Riemannian tensor \( g \) of a minimizing geodesic connecting \( x \) and \( y \). The Riemannian manifold can be viewed as a metric measure space

\[(M, d, \mathbb{P})\]

where \( \mathbb{P} = \frac{dv}{V} \) is the probability measure on \( M \) obtained from the Riemann volume element \( dv \) by normalization (here \( V \) is the total volume of \( M \)).

Let \( c(M) \) denote the infimum of the Ricci curvature tensor over all tangent vectors. Assuming that \( c(M) > 0 \), it can be proved that

\[
\| f(X) - \mathbb{E} f(X) \|_{\psi^2} \leq \frac{C \| f \|_{\text{Lip}}}{\sqrt{c(M)}}
\]

(5.9)

for any Lipschitz function \( f : M \to \mathbb{R} \). This concentration inequality can be proved by semigroup tools, see [2, Section 2.3].

To give an example, it is known that \( c(S^{n-1}) = n - 1 \). Thus (5.9) gives an alternative approach to proving the concentration inequality (5.5) for the sphere \( S^{n-1} \). We will give several other examples next.

### 5.2.5 Special orthogonal group

The special orthogonal group \( SO(n) \) consists of all distance preserving linear transformations on \( \mathbb{R}^n \). Equivalently, the elements of \( SO(n) \) are \( n \times n \) orthogonal matrices whose determinant equals 1. We will view the special orthogonal group as a metric measure space

\[(SO(n), \| \cdot \|_F, \mathbb{P})\]

where the distance is the Frobenius norm \( \| A - B \|_F \) (see Section 4.1.3) and \( \mathbb{P} \) is the uniform probability measure on \( SO(n) \).

Technically, \( \mathbb{P} \) is the Haar measure on \( SO(n) \) – the unique probability measure that is invariant under the action on the group;\(^4\) see ... . This Cite measure allows us to talk about random orthogonal matrices

\[X \sim \text{Unif}(O(n))\]

and discuss concentration inequalities for \( f(X) \) where \( f \) is a Lipschitz function on \( SO(n) \).

Alternatively, a random orthogonal matrix \( U \sim \text{Unif}(SO(n)) \) (and thus the Haar measure on the special orthogonal group) can be constructed by

\(^4\)A measure \( \mu \) on \( SO(n) \) is rotation invariant if for any measurable set \( E \subset SO(n) \) and any \( T \in SO(n) \), one has \( \mu(E) = \mu(T(E)) \).
computing the singular value decomposition $G = U G V^T$ of a random Gaussian matrix $G$ with i.i.d. $N(0,1)$ entries. (Why? Check rotation invariance.)

The concentration inequality (5.8) holds for for any Lipschitz function $f : SO(n) \to \mathbb{R}$, see [4, Section 6.5.1]. This result follows from concentration on general Riemannian manifolds discussed in Section 5.2.4.

5.2.6 Grassmann manifold

The Grassmann manifold $G_{n,m}$ consists of all $m$-dimensional subspaces of $\mathbb{R}^n$. In the special case where $m = 1$, the Grassman manifold can be identified with the sphere $S^{n-1}$ (how?), so the result below will generalize concentration on the sphere.

We can view the Grassmann manifold as a metric measure space

$$(G_{n,m}, d, P).$$

The distance between subspaces $E$ and $F$ can be defined as the operator norm

$$d(E, F) = \|P_E - P_F\|$$

where $P_E$ and $P_F$ are the orthogonal projections onto $E$ and $F$, respectively.

The probability $P$ is, like before, the uniform (Haar) probability measure on $G_{n,m}$. This measure allows us to talk about random $m$-dimensional subspaces of $\mathbb{R}^n$

$$E \sim \text{Unif}(G_{n,m}),$$

and discuss concentration inequalities for $f(E)$ where $f$ is a Lipschitz function on $G_{n,m}$.

Alternatively, a random subspace $E \sim \text{Unif}(G_{n,m})$, and thus the Haar measure on the Grassmann manifold, can be constructed by computing the column span (i.e. the image) of a random $n \times m$ Gaussian random matrix $G$ with i.i.d. $N(0,1)$ entries. (Why? Check rotation invariance.)

The concentration inequality (5.8) holds for for any Lipschitz function $f : G_{n,m} \to \mathbb{R}$, see [4, Section 6.7.2]. This result can be deduced from concentration on the special orthogonal group discussed in Section 5.2.5. Indeed, one expresses Grassmann manifold as a quotient $G_{n,k} = SO(n)/(SO_m \times SO_{n-m})$ and note that concentration passes on to quotients, see [4, Section 6.6].

Exercise 5.2.6. State and prove a concentration inequality for Lipschitz functions on the set of all $n \times m$ matrices with orthonormal columns.
5.2. CONCENTRATION ON OTHER METRIC MEASURE SPACES

5.2.7 Euclidean ball
The same concentration inequality (5.5) that holds for the unit Euclidean sphere $S^{n-1}$ also holds for the unit Euclidean ball

$$(B_2^n, \| \cdot \|_2, \mathbb{P})$$

equipped with the uniform probability measure. This can be deduced from concentration in Gauss space, see [2, Proposition 2.9]. Exercise?

5.2.8 Continuous cube
Consider the continuous cube

$$([0,1]^n, \| \cdot \|_2, \mathbb{P})$$
as a metric measure space, equipped with the Euclidean distance and the uniform probability measure. Thus the coordinates of a random vector

$$X \sim \text{Unif}([0,1]^n)$$

are independent random variables uniformly distributed in the interval $[0,1]$. Then concentration inequality (5.7) holds for any Lipschitz function $f : [0,1]^n \to \mathbb{R}$. This result can be deduced from concentration in Gauss space, see [2, Proposition 2.8]. Exercise?

Other than for Euclidean balls and cubes, sub-gaussian concentration inequalities hold for many other convex bodies, but not all of them. (A unit ball in $\ell_1$ norm is a counterexample.) A weaker sub-exponential concentration can be proved for general convex bodies using C. Borell’s inequality, see [4, Section III.3].

5.2.9 Densities $e^{-U(x)}$
Concentration inequality for Gaussian distribution we proved in Theorem 5.2.1 can be extended for distributions more general densities. Let $X$ be a random vector in $\mathbb{R}^n$ whose density has the form

$$f(x) = e^{-U(x)}, \quad x \in \mathbb{R}^n,$$

for some function $U : \mathbb{R}^n \to \mathbb{R}$. As an example, if $X \sim N(0, I_n)$ then the normal density (3.3) gives

$$U(x) = \|x\|_2^2 + c$$
where $c$ is a constant (that depends on $n$ but not $x$).

Suppose $U$ has curvature like $\|x\|_2^2$ or better. More rigorously, we require that the Hessian of $U$ be lower bounded on all of the space. So, suppose there exists $\kappa > 0$ such that

$$\text{Hess } U(x) \succeq \kappa I_n \quad \text{for all } x \in \mathbb{R}^n.$$ 

Then the concentration inequality

$$\|f(X) - \mathbb{E} f(X)\|_{\psi_2} \leq \frac{C\|f\|_{\text{Lip}}}{\sqrt{\kappa}}$$

for any Lipschitz function $f : \mathbb{R}^n \to \mathbb{R}$.

Note a similarity of this result with the concentration inequality (5.9) for Riemannian manifolds. Both of them can be proved using semigroup tools \[2, \text{Proposition 2.18}].

5.2.10 Random vectors with independent bounded coordinates

In Section 5.2.8, we mentioned a concentration inequality for random vectors $X = (X_1, \ldots, X_n)$ whose coordinates are independent random variables uniformly distributed in $[0, 1]$. We may wonder if this can be extended from uniform to more general distributions.

This indeed can be done. Suppose the coordinates of $X = (X_1, \ldots, X_n)$ are independent bounded random variables; to be specific, assume that

$$|X_i| \leq 1 \quad \text{almost surely for every } i.$$

Then concentration inequality (5.7) holds for any convex Lipschitz function $f : [0, 1]^n \to \mathbb{R}$. In particular, this holds for any norm on $\mathbb{R}^n$. This result is due to M. Talagrand; see \[2, \text{Corollary 4.10}].

5.2.11 Bounded differences inequality?

5.3 Application: Johnson-Lindenstrauss Lemma

Suppose we have $N$ data points in $\mathbb{R}^n$ where $n$ is very large. We would like to reduce dimension of the data without sacrificing too much of its geometry. The simplest form of dimension reduction is to project the data points onto a low-dimensional subspace

$$E \subset \mathbb{R}^n, \quad \dim(E) := m \ll n.$$
see Figure ?? for illustration. How shall we choose the subspace $E$, and how small its dimension $m$ can be?

Johnson-Lindenstrauss Lemma states that the geometry of data is well preserved if we choose $E$ to be a random subspace of dimension

$$m \sim \log N.$$  

We already came across the notion of a random subspace in Section ???. Let us recall it here. The Grassmann manifold $G_{n,m}$ is the set of all $m$-dimensional subspaces in $\mathbb{R}^n$. It is equipped with Haar measure, which is the unique rotation-invariant probability measure on $G_{n,m}$. This measure allows us to talk about random $m$-dimensional subspaces uniformly distributed in the Grassman manifold

$$E \sim \text{Unif}(G_{n,m}).$$

**Theorem 5.3.1** (Johnson-Lindenstrauss Lemma). Let $X$ be a set of $N$ points in $\mathbb{R}^n$ and $\varepsilon > 0$. Assume that

$$m \geq (C/\varepsilon^2) \log N.$$  

Consider a random $m$-dimensional subspace $E$ in $\mathbb{R}^n$ uniformly distributed in $G_{n,m}$. Denote the orthogonal projection onto $E$ by $P$. Then, with probability

---

Rotation invariance means that for any measurable set $\mathcal{E} \in G_{n,m}$ and any orthogonal matrix $U \in O(n)$, the set $U(\mathcal{E})$ has the same measure as $\mathcal{E}$. 

---
at least \(1 - 2 \exp(-c\varepsilon^2 m)\), the scaled projection
\[ A := \sqrt{\frac{n}{m}} P \]
is an approximate isometry on \(X\):
\[
(1 - \varepsilon)\|x - y\|_2 \leq \|Ax - Ay\|_2 \leq (1 + \varepsilon)\|x - y\|_2 \quad \text{for all } x, y \in X.
\]

The proof of Johnson-Lindenstrauss Lemma is based on concentration of Lipschitz functions on the sphere, which we studied in Section 5.1. We will use it to first examine the action of the random projection \(P\) on a fixed vector \(x - y\), and then take union bound over all \(N^2\) vectors \(x - y\).

**Lemma 5.3.2** (Random projection). Let \(P\) be a projection in \(\mathbb{R}^n\) onto a random \(m\)-dimensional subspace uniformly distributed in \(G_{n,m}\). Let \(x \in \mathbb{R}^n\) be a (fixed) point and \(\varepsilon > 0\). Then:

1. \((\mathbb{E}\|Pz\|_2^2)^{1/2} = \sqrt{\frac{m}{n}}\|z\|_2\).

2. 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.** Without loss of generality, we may assume that \(\|z\|_2 = 1\). (Why?) Next, we will change the model: instead of a random projection \(P\) acting on a fixed vector \(x - y\), we will consider a fixed projection \(P\) acting on a random vector \(z\). Specifically, the distribution of \(\|Pz\|_2\) will not change if we let \(P\) be fixed and \(z \sim \text{Unif}(S^{n-1})\).

(Show this using rotation invariance.)

Using rotation invariance again, we may assume without loss of generality that \(P\) is the coordinate projection onto \(\mathbb{R}^m\) viewed as a subspace of \(\mathbb{R}^n\). Thus
\[
\mathbb{E}\|Pz\|_2^2 = \mathbb{E}\sum_{i=1}^m z_i^2 = \sum_{i=1}^m \mathbb{E} z_i^2 = m \mathbb{E} z_1^2
\]
since the coordinates \(z_i\) of the random vector \(z \sim \text{Unif}(S^{n-1})\) are identically distributed. To compute \(\mathbb{E} z_1^2\), note that \(1 = \|z\|_2^2 = \sum_{i=1}^m z_i^2\), Taking expectations of both sides, we obtain
\[
1 = \sum_{i=1}^m \mathbb{E} z_i^2 = m \mathbb{E} z_1^2,
\]
5.3. APPLICATION: JOHNSON-LINDENSTRAUSS LEMMA

which yields
\[ E z_1^2 = \frac{1}{n}. \]

Putting this into (5.10), we get
\[ E \| Pz \|^2_2 = \frac{m}{n}. \]

This proves the first part of the lemma.

The second part follows from concentration of Lipschitz functions on the sphere. Indeed,
\[ f(x) := \| Px \|_2 \]
is a Lipschitz function on \( S^{n-1} \), and \( \| f \|_{\text{Lip}} = 1 \). (Why?) Then concentration inequality (5.6) yields
\[ \mathbb{P} \left\{ \left| \| Px \|_2 - \sqrt{\frac{m}{n}} \right| \geq t \right\} \leq 2 \exp(-cnt^2). \]

(Here we also used Exercise 5.2.5 to replace \( \mathbb{E} \| x \|_2 \) by the \( (\mathbb{E} \| x \|_2^2)^{1/2} \) in the concentration inequality.) Choosing \( t := \varepsilon \sqrt{m/n} \), we complete the proof of the lemma.

\( \square \)

**Proof of Johnson-Lindenstrauss Lemma.** Consider the difference set
\[ \mathcal{X} - \mathcal{X} = \{ x - y : x, y \in \mathcal{X} \}. \]

We would like to show that, with required probability, the inequality
\[ (1 - \varepsilon) \| z \|_2 \leq \| Az \|_2 \leq (1 + \varepsilon) \| z \|_2 \]
holds for all \( z \in \mathcal{X} - \mathcal{X} \). Since \( A = \sqrt{n/m} P \), this is inequality is equivalent to
\[ (1 - \varepsilon) \sqrt{\frac{m}{n}} \| z \|_2 \leq \| Pz \|_2 \leq (1 + \varepsilon) \sqrt{\frac{m}{n}} \| z \|_2. \]  \hspace{1cm} (5.11)

For any fixed \( z \), Lemma 5.3.2 states that (5.11) holds with probability at least \( 1 - 2 \exp(-c\varepsilon^2 m) \). It remains to take a union bound over \( z \in \mathcal{X} - \mathcal{X} \). By doing this, we make the inequality (5.11) hold with probability at least
\[ 1 - |\mathcal{X} - \mathcal{X}| \cdot 2 \exp(-c\varepsilon^2 m) \geq 1 - N^2 \cdot 2 \exp(-c\varepsilon^2 m). \]

If we choose \( m \geq (C/\varepsilon^2) \log N \) then this probability is at least \( 1 - 3 \exp(-c\varepsilon^2 m/2) \), as claimed. Johnson-Lindenstrauss Lemma is proved.

\( \square \)
A remarkable feature of Johnson-Lindenstrauss lemma is dimension reduction map $A$ is non-adaptive, it does not depend on the data. It is also interesting that the ambient dimension $n$ of the data plays no role in this result.

Johnson-Lindenstrauss Lemma uses a random projection as a means of dimension reduction. Other linear and non-linear maps are possible to use, too:

**Exercise 5.3.3** (Johnson-Lindenstrauss with sub-gaussian matrices). Let $G$ be an $m \times n$ random matrix whose rows are independent, mean zero, sub-gaussian isotropic random vectors in $\mathbb{R}^n$. Show that the conclusion of Johnson-Lindenstrauss lemma holds for $A = \frac{1}{\sqrt{m}} G$.

### 5.4 Matrix Bernstein inequality

### 5.5 Random matrices with general independent rows

We will illustrate the matrix Bernstein inequality with an application to random matrices. In this section, we will prove a remarkably general version of Theorem 4.5.1 which holds for random matrices with arbitrary, not necessarily sub-gaussian distributions of rows.

Such generality is important because, as we noted in Section 3.4.2, discrete distributions are poorly sub-gaussian. So Theorem 4.5.1 does not usually give a satisfactory result for random matrices whose rows have discrete distributions. The next result will be more useful.

**Theorem 5.5.1** (Random matrices with general rows). Let $A$ be an $m \times n$ matrix whose rows $A_i$ are independent isotropic random vectors in $\mathbb{R}^n$. Assume that for some $L \geq 0$,

$$\|A_i\|_2 \leq L\sqrt{n} \quad \text{almost surely for every } i. \quad (5.12)$$

Then for every $t \geq 0$, one has

$$\sqrt{m} - tL\sqrt{n} \leq s_n(A) \leq s_1(A) \leq \sqrt{m} + tL\sqrt{n} \quad (5.13)$$

with probability at least $1 - 2n \cdot \exp(-ct^2)$.

Before we prove this theorem, let us pause to make two remarks. First, why does the boundedness assumption (5.12) have this form? The isotropy
5.5. RANDOM MATRICES WITH GENERAL INDEPENDENT ROWS

of random vectors $A_i$ implies by Lemma 3.2.4 that

$$\mathbb{E} \|A_i\|^2 = n.$$  

By Markov’s inequality, it follows that with high probability (let’s say, 0.99), we have

$$\|A_i\|_2 = O(\sqrt{n}).$$

So in applications, we expect that the boundedness assumption (5.12) hold with

$$L = O(1).$$

Next, let us note that by Lemma 4.1.2, we can equivalently restate the conclusion (5.13) in the following form:

$$\left\| \frac{1}{m} A^T A - I_n \right\| \leq \max(\delta, \delta^2) \quad \text{where} \quad \delta = tL \sqrt{\frac{n}{m}}. \quad (5.14)$$

Proof. Our proof of (5.14) will be based on matrix Bernstein’s inequality, Theorem 3.2. We use this result, we express the matrix in question as a sum of independent random matrices:

$$\frac{1}{m} A^T A - I_n = \frac{1}{m} \sum_{i=1}^{m} A_i A_i^T - I_n = \sum_{i=1}^{N} X_i \quad \text{where} \quad X_i := \frac{1}{m} (A_i A_i^T - I_n).$$

Note that $X_i$ are independent symmetric $n \times n$ random matrices, and they have zero means by the isotropy assumption. So, in order to apply Theorem 3.2, it remains to bound the range $\|X_i\|$ and the norm of the matrix variance $\sum_{i=1}^{m} \mathbb{E} X_i^2$.

To bound the range, we can use boundedness assumption (5.12), where we necessarily have $L \geq 1$ by isotropy. (Why?) By triangle inequality, we have

$$\|X_i\| \leq \frac{1}{m} (\|A_i A_i^T\| + 1) = \frac{1}{m} (\|A_i\|^2 + 1)$$

$$\leq \frac{1}{m} (L^2 n + 1) \quad \text{(by the boundedness assumption)}$$

$$\leq \frac{2L^2 n}{m} \quad \text{(since} \quad L \geq 1)$$

$$=: K. \quad (5.15)$$

To estimate the norm of the matrix variance $\| \sum_{i=1}^{m} \mathbb{E} X_i^2 \|$, we first compute

$$X_i^2 = \frac{1}{m^2} \left[ (A_i A_i^T)^2 - 2(A_i A_i^T) + I_n \right].$$
Taking expectations of both sides and using isotropy of $A_i$, we obtain
\[ E X_i^2 = \frac{1}{m^2} \left[ E(A_i \otimes A_i)^2 - I_n \right]. \]

Refer to an exercise? Now, using boundedness assumption (5.12), we have \[ E(A_i A_i^T)^2 = E \|A_i\|^2 A_i A_i^T \]
\[ \leq E \left[ L^2 n \cdot A_i A_i^T \right] \quad \text{(by the boundedness assumption– check!)} \]
\[ = L^2 n \cdot I_n \quad \text{(by isotropy).} \]

Thus
\[ E X_i^2 \leq L^2 n \cdot I_n. \]

Summing up, we obtain a bound on the matrix variance:
\[ \sum_{i=1}^m E X_i^2 \leq L^2 n \cdot I_n. \]

Since the matrix variance is a positive semidefinite matrix, it follows that
\[ \left\| \sum_{i=1}^m E X_i^2 \right\| \leq \frac{L^2 n}{m} : = \sigma^2. \tag{5.16} \]

Now we are ready to apply the matrix Bernstein inequality (??). \[ \mathbb{P}\left\{ \left\| \frac{1}{m} A^T A - I_n \right\| \geq \varepsilon \right\} = \mathbb{P}\left\{ \left\| \sum_{i=1}^m X_i \right\| \geq \varepsilon \right\} \]
\[ \leq 2n \cdot \exp \left[ -c \min \left( \frac{\varepsilon^2}{\sigma^2}, \frac{\varepsilon}{K} \right) \right]. \]

Substituting $\sigma$ and $K$ from (5.16) and (5.15), we bound this further by
\[ 2n \cdot \exp \left[ -c \min(\varepsilon^2, \varepsilon) \cdot \frac{m}{2L^2 n} \right]. \tag{5.17} \]

Returning to (5.14), we set $\varepsilon := \max(\delta, \delta^2)$ and $\delta = tL \sqrt{n/m}$. Substituting these values into (5.17), we bound the probability that (5.14) fails by
\[ 2n \cdot \exp \left( -c \frac{\delta^2 m}{2L^2 n} \right) = 2n \cdot \exp(-ct^2/2). \]

This completes the proof. \qed
5.5. RANDOM MATRICES WITH GENERAL INDEPENDENT ROWS

5.5.1 The logarithmic factor

Note the pre-factor $n$ in the probability bound $1 - 2n \cdot \exp(-ct^2)$ in Theorem 5.5.1. The result is non-trivial when this probability is greater than zero, and this happens when $t \geq C\sqrt{\log n}$. Let us restate the conclusion of Theorem 5.5.1 in this useful range. For every $s \geq 1$, one has

$$\sqrt{m} - sL\sqrt{n \log n} \leq s_n(A) \leq s_1(A) \leq \sqrt{m} + sL\sqrt{n \log n} \quad (5.18)$$

with probability at least $1 - 2n^{-\alpha^2}$. (To obtain this, set $t = s\sqrt{\log n}$.)

Summarizing, we extended Theorem 4.5.1 from sub-gaussian to general distributions, and we paid just a logarithmic factor for that.

The logarithmic factor cannot be removed from (5.18) in general. This usually happens for discrete distributions. For example, let the rows of the matrix $A$ have the coordinate distribution in $\mathbb{R}^n$, which we introduced in Section 3.3.4. This means that for each $i$,

$$A_i \sim \text{Unif}\{\sqrt{n}e_j : j = 1, \ldots, n\}.$$

where $e_j$ is a canonical basis in $\mathbb{R}^n$.

Let us rethink this matrix in terms of the classical coupon collector’s problem. In order for the inequality $s_n(A) > 0$ to hold, the matrix $A$ must have full rank, and this in turn means that $A$ must not have zero columns. Equivalently, each of the $n$ “coupons” $\sqrt{n}e_1, \ldots, \sqrt{n}e_n$ must be picked at least once in $m$ independent trials. (So each row $A_1, \ldots, A_m$ picks an independent coupon.)

By the known result on the coupon collector’s problem, one must make at least $m \geq Cn \log n$ trials to pick each of the $n$ coupons with high (and even constant) probability. Summarizing, in order to have $s_n(A) > 0$, we must have

$$m \gtrsim n \log n.$$

This shows that the logarithmic factor $\log n$ is necessary in the left hand side of (5.18).

We can track the source of this logarithmic factor to matrix Bernstein inequality (Theorem ??). As we saw, its only weakness compared with the scalar Bernstein’s inequality is the pre-factor $n$ in the probability bound. Our analysis then shows that this pre-factor is needed, and is optimal.

Exercise 5.5.2 (Non-isotropic distributions). **Prove the following version of (5.14) for non-isotropic distributions.** Let $A$ be an $m \times n$ matrix which
satisfies all assumptions of Theorem 5.5.1 except the isotropy. Assume that the rows have the same covariance matrix

$$\Sigma = \mathbb{E} A_i A_i^T.$$ 

Then, for every $t \geq 0$, the following inequality holds with probability at least $1 - 2n \cdot \exp(-ct^2)$:

$$\left\| \frac{1}{m} A^T A - \Sigma \right\| \leq \max(\|\Sigma\|^{1/2}\delta, \delta^2) \quad \text{where} \quad \delta = tL \sqrt{\frac{n}{m}}. \quad (5.19)$$

### 5.6 Application: covariance estimation for general distributions

**Theorem 5.6.1 (Covariance estimation).** Consider a random vector $X$ in $\mathbb{R}^n$ with zero mean and covariance matrix $\Sigma$. Assume that for some $L > 0$,

$$\|X\|_2 \leq K \sqrt{\|\Sigma\|n} \quad \text{almost surely.} \quad (5.20)$$

Let $\varepsilon \in (0,1)$ and $t \geq 1$. Suppose the sample size satisfies

$$m \geq C(Kt/\varepsilon)^2n \log n.$$ 

Then the sample covariance matrix $\Sigma_m$ satisfies

$$\|\Sigma_m - \Sigma\| \leq \varepsilon \|\Sigma\|$$

with probability at least $1 - 2n \cdot t^2$.

**Proof.** Consider the $m \times n$ matrix $A$ whose rows are the sample points $X_i^T$. Then the sample covariance matrix $\Sigma$ can be represented as

$$\Sigma_m = \sum_{i=1}^{m} X_i X_i^T = \frac{1}{m} A^T A.$$ 

So, to bound the error $\|\Sigma_m - \Sigma\|$ we can apply the non-isotropic form of Theorem 4.5.1 stated in Exercise 5.5.2. It states that with probability at least $1 - 2n \cdot \exp(-cs^2)$, we have

$$\|\Sigma_m - \Sigma\| \leq \max(\|\Sigma\|^{1/2}\delta, \delta^2) \quad \text{where} \quad \delta = sK \sqrt{\frac{\|\Sigma\|n}{m}}.$$
5.6. APPLICATION: COVARIANCE ESTIMATION FOR GENERAL DISTRIBUTIONS

After simplifying, this becomes

$$\|\Sigma_m - \Sigma\| \leq \max \left( sK \sqrt{\frac{n}{m}} \cdot \frac{s^2K^2n}{m} \right) \|\Sigma\|.$$ 

So, if \( m \geq (Ks/\varepsilon)^2n \) the right hand side is bounded by \( \varepsilon \|\Sigma\| \) as required. It remains to choose \( s = Ct\sqrt{\log n} \) to complete the proof. \( \square \)

Let us clarify the form of the boundedness assumption (5.20).

**Exercise 5.6.2.** Let \( X \) be a random vector in \( \mathbb{R}^n \) with \( \mathbb{E}XX^\top = \Sigma \). Show that

$$\mathbb{E} \|X\|_2^2 = \text{tr}(\Sigma).$$

Since we always have

$$\text{tr}(\Sigma) \leq \|\Sigma\|n$$

Markov’s inequality implies that with high probability (let’s say, 0.99), we have

$$\|X\|_2 = O(\sqrt{\|\Sigma\|n}).$$

So in applications, we expect that the boundedness assumption (5.20) hold with

$$K = O(1).$$
Bibliography


Index

Adjacency matrix, 71
Approximate isometry, 59, 60, 75
Approximate projection, 61
Bernoulli distribution, 7
Bernstein’s inequality, 32–34
Binomial distribution, 7
Cauchy-Schwarz inequality, 3
Centering, 28, 87
centering, 35
Central Limit Theorem
Berry-Esseen, 9
Lindeberg-Lévy, 5
projective, 55
Chebyshev’s inequality, 4
Chernoff’s inequality, 13, 14, 16
Clustering, 79
Convex body, 50
Coordinate distribution, 48, 52, 98
Coupon collector’s problem, 98
Covariance, 2, 40, 78, 79, 99
Covering number, 61, 64, 65
Davis-Kahan Theorem, 73, 81
Discrete cube, 89
Erdős-Rényi model, 17, 71
Expectation, 1
Exponential distribution, 32
Frame, 48, 53
Frobenius norm, 59
Gaussian distribution, 5
Gaussian mixture model, 80, 81
Grassmann manifold, 92, 95
Haar measure, 91, 92, 95
Hamming distance, 89
Hoeffding’s inequality, 10, 12, 27
Hölder’s inequality, 3
Indicator random variables, 7
Integral Identity, 3
Isoperimetric inequality, 85, 89
Isotropic random vectors, 42
Jensen’s inequality, 2
Johnson-Lindenstrauss Lemma, 94
Khinchine’s inequality, 27
Law of Large Numbers, 4, 34, 78
Lipschitz
   function, 83
   norm, 84
$L_p$ norm, 1
$L_{\psi_1}$ norm, 30
$L_{\psi_2}$ norm, 24
Markov’s inequality, 4
Metric entropy, 63
Metrix entropy, 63
Minkowski’s inequality, 3
Minkowski sum, 64
INDEX

Moment generating function, 10, 21, 24, 32
Net, 67, 68
net, 61
Network, 17
Normal distribution, 5, 45, 46, 51
Operator norm, 58, 66–68
Orlicz space, 25
Packing, 61
Packing number, 61
Perturbation of eigenvectors, 73
Poisson distribution, 14, 15, 32
Principal Component Analysis, 41, 78, 80
Random graph, 17
Riemannian manifold, 90
Sample covariance, 78
Second moment matrix, 41
Singular value decomposition, 57
Special orthogonal group, 91
Spectral Clustering, 74
Spectral decomposition, 41
Spherical distribution, 45
Standard deviation, 2
Stochastic block model, 71, 75
Sub-exponential distribution, 29, 30, 32
Sub-gaussian distribution, 21, 24, 25, 26, 51
Symmetric group, 90
Tails, 3
   normal, 8, 16
   Poisson, 15
Variance, 1