

---

## Concentration of sums of independent random variables

This chapter introduces the reader to the rich topic of concentration inequalities. After motivating the subject in Section 2.1, we prove some basic concentration inequalities: Hoeffding's in Sections 2.2 and 2.6, Chernoff's in Section 2.3 and Bernstein's in Section 2.8. Another goal of this chapter is to introduce two important classes of distributions: sub-gaussian in Section 2.5 and sub-exponential in Section 2.7. These classes form a natural "habitat" in which many results of high-dimensional probability and its applications will be developed. We give two quick applications of concentration inequalities for randomized algorithms in Section 2.2 and random graphs in Section 2.4. Many more applications are given later in the book.

### 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

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

The simplest concentration inequality is Chebyshev's inequality (Corollary 1.2.5). It is very general but often too weak. Let us illustrate this 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?*

Let  $S_N$  denote the number of heads. Then

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

Chebyshev's inequality bounds the probability of getting at least  $\frac{3}{4}N$  heads as follows:

$$\mathbb{P}\left\{S_N \geq \frac{3}{4}N\right\} \leq \mathbb{P}\left\{\left|S_N - \frac{N}{2}\right| \geq \frac{N}{4}\right\} \leq \frac{4}{N}. \quad (2.1)$$

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

Is this the right rate of decay, or we should expect something faster? Let us approach the same question using the central limit theorem. To do this, we represent  $S_N$  as a sum of independent random variables:

$$S_N = \sum_{i=1}^N X_i$$

where  $X_i$  are independent Bernoulli random variables with parameter  $1/2$ , i.e.  $\mathbb{P}\{X_i = 0\} = \mathbb{P}\{X_i = 1\} = 1/2$ . (These  $X_i$  are the indicators of heads.) De Moivre-Laplace central limit theorem (1.7) states that the distribution of the normalized number of heads

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

converges to the standard normal distribution  $N(0, 1)$ . Thus we should anticipate that for large  $N$ , we have

$$\mathbb{P}\left\{S_N \geq \frac{3}{4}N\right\} = \mathbb{P}\left\{Z_N \geq \sqrt{N/4}\right\} \approx \mathbb{P}\left\{g \geq \sqrt{N/4}\right\} \quad (2.2)$$

where  $g \sim N(0, 1)$ . To understand how this quantity decays in  $N$ , we now get 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  $t > 0$ , we have*

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

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

$$\mathbb{P}\{g \geq t\} \leq \frac{1}{\sqrt{2\pi}} e^{-t^2/2}. \quad (2.3)$$

*Proof* To obtain an upper bound on the tail

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

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

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

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. □

Returning to (2.2), we see that we should expect the probability of having at least  $\frac{3}{4}N$  heads to be smaller than

$$\frac{1}{\sqrt{2\pi}} e^{-N/8}. \quad (2.4)$$

This quantity decays to zero *exponentially* fast in  $N$ , which is much better than the linear decay in (2.1) that follows from Chebyshev's inequality.

Unfortunately, (2.4) does not follow rigorously from the central limit theorem. Although the approximation by the normal density in (2.2) is valid, the error of approximation can not be ignored. And, unfortunately, *the error decays too slow* – even slower than linearly in  $N$ . This can be seen from the following sharp 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*

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

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

Thus the approximation error in (2.2) is of order  $1/\sqrt{N}$ , which ruins the desired exponential decay (2.4).

Can we improve the approximation error in central limit theorem? In general, no. If  $N$  is even, then the probability of getting exactly  $N/2$  heads is

$$\mathbb{P}\{S_N = N/2\} = 2^{-N} \binom{N}{N/2} \asymp \frac{1}{\sqrt{N}};$$

the last estimate can be obtained using Stirling's approximation.<sup>1</sup> (Do it!) Hence,  $\mathbb{P}\{Z_N = 0\} \asymp 1/\sqrt{N}$ . On the other hand, since the normal distribution is continuous, we have  $\mathbb{P}\{g = 0\} = 0$ . 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  $S_N = X_1 + \dots + X_N$  by the normal distribution. The normal distribution is especially nice due to its 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 is a roadblock toward proving concentration properties for  $S_N$  with light, exponentially decaying tails.

In order to resolve this issue, we develop alternative, direct approaches to concentration, which bypass the central limit theorem.

<sup>1</sup> Our somewhat informal notation  $f \asymp g$  stands for the equivalence of functions (functions of  $N$  in this particular example) up to constant factors. Precisely,  $f \asymp g$  means that there exist positive constants  $c, C$  such that the inequality  $cf(x) \leq g(x) \leq Cf(x)$  holds for all  $x$ , or sometimes for all sufficiently large  $x$ . For similar one-sided inequalities that hold up to constant factors, we use notation  $f \lesssim g$  and  $f \gtrsim g$ .

**Exercise 2.1.4** (Truncated normal distribution). ☛ Let  $g \sim N(0, 1)$ . Show that for all  $t \geq 1$ , we have

$$\mathbb{E} g^2 \mathbf{1}_{\{g > t\}} = t \cdot \frac{1}{\sqrt{2\pi}} e^{-t^2/2} + \mathbb{P}\{g > t\} \leq \left(t + \frac{1}{t}\right) \frac{1}{\sqrt{2\pi}} e^{-t^2/2}.$$

**Hint:** Integrate by parts.

## 2.2 Hoeffding's inequality

We start with a particularly simple concentration inequality, which holds for sums of i.i.d. *symmetric Bernoulli* random variables.

**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.

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

Clearly, a random variable  $X$  has the (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  $X_1, \dots, X_N$  be independent *symmetric Bernoulli* random variables, and  $a = (a_1, \dots, a_N) \in \mathbb{R}^N$ . Then, for any  $t \geq 0$ , we have

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

*Proof* We can assume without loss of generality that  $\|a\|_2 = 1$ . (Why?)

Let us recall how we deduced Chebyshev's inequality (Corollary 1.2.5): 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

$$\begin{aligned} \mathbb{P}\left\{\sum_{i=1}^N a_i X_i \geq t\right\} &= \mathbb{P}\left\{\exp\left(\lambda \sum_{i=1}^N a_i X_i\right) \geq \exp(\lambda t)\right\} \\ &\leq e^{-\lambda t} \mathbb{E} \exp\left(\lambda \sum_{i=1}^N a_i X_i\right). \end{aligned} \quad (2.5)$$

In the last step we applied Markov's inequality (Proposition 1.2.4).

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 a basic probability course, the MGF of the sum is the product of the MGF's of the terms; this follows immediately from independence. Thus

$$\mathbb{E} \exp\left(\lambda \sum_{i=1}^N a_i X_i\right) = \prod_{i=1}^N \mathbb{E} \exp(\lambda a_i X_i). \quad (2.6)$$

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

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

**Hint:** 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 into (2.6) and then into (2.5), we obtain

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

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 the central limit theorem. Indeed, the most we may expect from a concentration inequality is that the tail of  $\sum a_i X_i$  behaves similarly to the tail of 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 (2.3). This is good news. We have been able to obtain the same *exponentially light* tails for sums as for the 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 *exponentially small* in  $N$ , namely

$$\mathbb{P} \left\{ \text{at least } \frac{3}{4}N \text{ heads} \right\} \leq \exp(-N/8).$$

(Check this.)

**Remark 2.2.4** (Non-asymptotic results). It should be stressed that unlike the classical limit theorems of Probability Theory, Hoeffding's inequality is *non-asymptotic* in that it holds for all fixed  $N$  as opposed to  $N \rightarrow \infty$ . 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  $N$  often corresponds to *sample size*.

We can easily derive a version of Hoeffding's inequality for *two-sided tails*  $\mathbb{P}\{|S| \geq t\}$  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  $\mathbb{P}\{-S \geq t\}$ . Combining the two bounds, we obtain a bound on

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

Thus the bound doubles, and we obtain:

**Theorem 2.2.5** (Hoeffding's inequality, two-sided). *Let  $X_1, \dots, X_N$  be independent symmetric Bernoulli random variables, and  $a = (a_1, \dots, a_N) \in \mathbb{R}^N$ . Then, for any  $t > 0$ , we have*

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

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.

**Theorem 2.2.6** (Hoeffding's inequality for general bounded random variables). *Let  $X_1, \dots, X_N$  be independent random variables. Assume that  $X_i \in [m_i, M_i]$  for every  $i$ . Then, for any  $t > 0$ , we have*

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

**Exercise 2.2.7.** 🍷🍷 Prove Theorem 2.2.6, possibly with some absolute constant instead of 2 in the tail.

**Exercise 2.2.8** (Boosting randomized algorithms). 🍷🍷 Imagine we have an algorithm for solving some decision problem (e.g. is a given number  $p$  a prime?). Suppose the algorithm makes a decision at random and returns the correct answer with probability  $\frac{1}{2} + \delta$  with some  $\delta > 0$ , which is just a bit better than a random guess. To improve the performance, we run the algorithm  $N$  times and take the majority vote. Show that, for any  $\varepsilon \in (0, 1)$ , the answer is correct with probability at least  $1 - \varepsilon$ , as long as

$$N \geq \frac{1}{2\delta^2} \ln\left(\frac{1}{\varepsilon}\right).$$

**Hint:** Apply Hoeffding's inequality for  $X_i$  being the indicators of the wrong answers.

**Exercise 2.2.9** (Robust estimation of the mean). 🍷🍷🍷 Suppose we want to estimate the mean  $\mu$  of a random variable  $X$  from a sample  $X_1, \dots, X_N$  drawn independently from the distribution of  $X$ . We want an  $\varepsilon$ -accurate estimate, i.e. one that falls in the interval  $(\mu - \varepsilon, \mu + \varepsilon)$ .

- (a) Show that a sample<sup>2</sup> of size  $N = O(\sigma^2/\varepsilon^2)$  is sufficient to compute an  $\varepsilon$ -accurate estimate with probability at least  $3/4$ , where  $\sigma^2 = \text{Var } X$ .

**Hint:** Use the sample mean  $\hat{\mu} := \frac{1}{N} \sum_{i=1}^N X_i$ .

- (b) Show that a sample of size  $N = O(\log(\delta^{-1})\sigma^2/\varepsilon^2)$  is sufficient to compute an  $\varepsilon$ -accurate estimate with probability at least  $1 - \delta$ .

**Hint:** Use the median of  $O(\log(\delta^{-1}))$  weak estimates from part 1.

**Exercise 2.2.10** (Small ball probabilities). 🍷🍷 Let  $X_1, \dots, X_N$  be *non-negative* independent random variables with continuous distributions. Assume that the densities of  $X_i$  are uniformly bounded by 1.

- (a) Show that the MGF of  $X_i$  satisfies

$$\mathbb{E} \exp(-tX_i) \leq \frac{1}{t} \quad \text{for all } t > 0.$$

- (b) Deduce that, for any  $\varepsilon > 0$ , we have

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

**Hint:** Rewrite the inequality  $\sum X_i \leq \varepsilon N$  as  $\sum (-X_i/\varepsilon) \geq -N$  and proceed like in the proof of Hoeffding's inequality. Use part 1 to bound the MGF.

## 2.3 Chernoff's inequality

As we noted, Hoeffding's inequality is quite sharp for symmetric Bernoulli random variables. But the general form of Hoeffding's inequality (Theorem 2.2.6) is sometimes too conservative and does not give sharp results. This happens, for example, when  $X_i$  are Bernoulli random variables with parameters  $p_i$  so small that we expect  $S_N$  to have approximately Poisson distribution according to Theorem 1.3.4. However, Hoeffding's inequality is not sensitive to the magnitudes of  $p_i$ , and the Gaussian tail bound it gives is very far from the true, Poisson, tail. In this section we study Chernoff's inequality, which is sensitive to the magnitudes of  $p_i$ .

**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*

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

<sup>2</sup> More accurately, this claim means that there exists an absolute constant  $C$  such that if  $N \geq C\sigma^2/\varepsilon^2$  then  $\mathbb{P} \{|\hat{\mu} - \mu| \leq \varepsilon\} \geq 3/4$ . Here  $\hat{\mu}$  is the sample mean; see the hint.

*Proof* We will use the same method – based on the 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 (2.5) and (2.6): multiply both sides of the inequality  $S_N \geq t$  by a parameter  $\lambda$ , exponentiate, and then use Markov's inequality and independence. This gives

$$\mathbb{P}\{S_N \geq t\} \leq e^{-\lambda t} \prod_{i=1}^N \mathbb{E} \exp(\lambda X_i). \quad (2.7)$$

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 value 0 with probability  $1 - p_i$ , we have

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


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

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


Substituting this into (2.7), we obtain

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

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.  $\square$

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

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

**Exercise 2.3.3** (Poisson tails).  Let  $X \sim \text{Pois}(\lambda)$ . Show that for any  $t > \lambda$ , we have

$$\mathbb{P}\{X \geq t\} \leq e^{-\lambda} \left( \frac{e\lambda}{t} \right)^t. \quad (2.8)$$

**Hint:** Combine Chernoff's inequality with Poisson limit theorem (Theorem 1.3.4).

**Remark 2.3.4** (Poisson tails). Note that the Poisson tail bound (2.8) is quite sharp. Indeed, the probability mass function (1.8) of  $X \sim \text{Pois}(\lambda)$  can be approximated via Stirling's formula  $k! \approx \sqrt{2\pi k}(k/e)^k$  as follows:

$$\mathbb{P}\{X = k\} \approx \frac{1}{\sqrt{2\pi k}} \cdot e^{-\lambda} \left( \frac{e\lambda}{k} \right)^k. \quad (2.9)$$

So our bound (2.8) on the *entire tail* of  $X$  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** (Chernoff's inequality: small deviations). ☕☕☕ Show that, in the setting of Theorem 2.3.1, for  $\delta \in (0, 1]$  we have

$$\mathbb{P}\{|S_N - \mu| \geq \delta\mu\} \leq 2e^{-c\mu\delta^2}$$

where  $c > 0$  is an absolute constant.

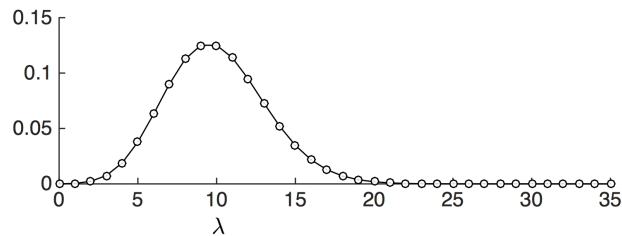
**Hint:** Apply Theorem 2.3.1 and Exercise 2.3.2  $t = (1 \pm \delta)\mu$  and analyze the bounds for small  $\delta$ .

**Exercise 2.3.6** (Poisson distribution near the mean). ☕ Let  $X \sim \text{Pois}(\lambda)$ . Show that for  $t \in (0, \lambda]$ , we have

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

**Hint:** Combine Exercise 2.3.5 with the Poisson limit theorem (Theorem 1.3.4).

**Remark 2.3.7** (Large and small deviations). Exercises 2.3.3 and 2.3.6 indicate two different behaviors of the tail of the Poisson distribution  $\text{Pois}(\lambda)$ . In the small deviation regime, near the mean  $\lambda$ , the tail of  $\text{Pois}(\lambda)$  is like that for the normal distribution  $N(\lambda, \lambda)$ . In the large deviation regime, far to the right from the mean, the tail is heavier and decays like  $(\lambda/t)^t$ ; see Figure 2.1.



**Figure 2.1** The probability mass function of the Poisson distribution  $\text{Pois}(\lambda)$  with  $\lambda = 10$ . The distribution is approximately normal near the mean  $\lambda$ , but to the right from the mean the tails are heavier.

**Exercise 2.3.8** (Normal approximation to Poisson). ☕☕ Let  $X \sim \text{Pois}(\lambda)$ . Show that, as  $\lambda \rightarrow \infty$ , we have

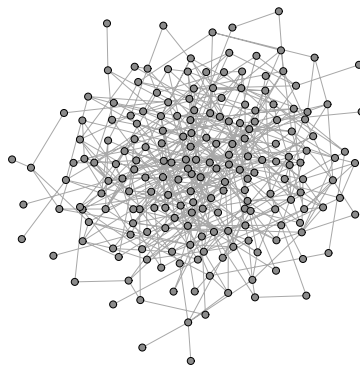
$$\frac{X - \lambda}{\sqrt{\lambda}} \rightarrow N(0, 1) \quad \text{in distribution.}$$

**Hint:** Derive this from the central limit theorem. Use the fact that the sum of independent Poisson distributions is a Poisson distribution.

## 2.4 Application: degrees of random graphs

We give an application of Chernoff's inequality to a classical object in probability: *random graphs*.

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



**Figure 2.2** A random graph from Erdős-Rényi model  $G(n, p)$  with  $n = 200$  and  $p = 1/40$ .

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

$$(n - 1)p =: d.$$

(Check!) We will show that relatively *dense graphs*, those where  $d \gtrsim \log n$ , are almost *regular* with high probability, which means that the degrees of all vertices approximately equal  $d$ .

**Proposition 2.4.1** (Dense graphs are almost regular). *There is an absolute constant  $C$  such that the following holds. 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  $0.9d$  and  $1.1d$ .*

*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  $n - 1$  independent  $\text{Ber}(p)$  random variables (the indicators of the edges incident to  $i$ ). Thus we can apply Chernoff's inequality, which yields

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

(We used the version of Chernoff's inequality given in Exercise 2.3.5 here.)

This bound holds for each fixed vertex  $i$ . Next, we can “unfix”  $i$  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, and we have

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

This completes the proof.  $\square$

Sparser graphs, those for which  $d = o(\log n)$ , are no longer almost regular, but there are still useful bounds on their degrees. The following series of exercises makes these claims clear. In all of them, we shall assume that the graph size  $n$  grows to infinity, but we don’t assume the connection probability  $p$  to be constant in  $n$ .

**Exercise 2.4.2** (Bounding the degrees of sparse graphs).  $\clubsuit$  Consider a random graph  $G \sim G(n, p)$  with expected degrees  $d = O(\log n)$ . Show that with high probability (say, 0.9), all vertices of  $G$  have degrees  $O(\log n)$ .

**Hint:** Modify the proof of Proposition 2.4.1.

**Exercise 2.4.3** (Bounding the degrees of very sparse graphs).  $\clubsuit\clubsuit$  Consider a random graph  $G \sim G(n, p)$  with expected degrees  $d = O(1)$ . Show that with high probability (say, 0.9), all vertices of  $G$  have degrees

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

Now we pass to the lower bounds. The next exercise shows that Proposition 2.4.1 does not hold for sparse graphs.

**Exercise 2.4.4** (Sparse graphs are not almost regular).  $\clubsuit\clubsuit\clubsuit$  Consider a random graph  $G \sim G(n, p)$  with expected degrees  $d = o(\log n)$ . Show that with high probability, (say, 0.9),  $G$  has a vertex with degree<sup>3</sup>  $10d$ .

**Hint:** The principal difficulty is that the degrees  $d_i$  are not independent. To fix this, try to replace  $d_i$  by some  $d'_i$  that are independent. (Try to include not all vertices in the counting.) Then use Poisson approximation (2.9).

Moreover, very sparse graphs, those for which  $d = O(1)$ , are even farther from regular. The next exercise gives a lower bound on the degrees that matches the upper bound we gave in Exercise 2.4.3.

**Exercise 2.4.5** (Very sparse graphs are far from being regular).  $\clubsuit\clubsuit$  Consider

<sup>3</sup> We assume here that  $10d$  is an integer. There is nothing particular about the factor 10; it can be replaced by any other constant.

a random graph  $G \sim G(n, p)$  with expected degrees  $d = O(1)$ . Show that with high probability, (say, 0.9),  $G$  has a vertex with degree

$$\Omega\left(\frac{\log n}{\log \log n}\right).$$

## 2.5 Sub-gaussian distributions

So far, we have studied concentration inequalities that apply only for Bernoulli random variables  $X_i$ . It would be useful to extend these results for a wider class of distributions. At the very least, we may expect that the normal distribution belongs to this class, since we think of concentration results as quantitative versions of the central limit theorem.

So let us ask: which random variables  $X_i$  must obey a concentration inequality like Hoeffding's in Theorem 2.2.5, namely

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

If the sum  $\sum_{i=1}^N a_i X_i$  consists of a single term  $X_i$ , this inequality reads as

$$\mathbb{P} \{ |X_i| > t \} \leq 2e^{-ct^2}.$$

This gives us an automatic restriction: if we want Hoeffding's inequality to hold, we must assume that the random variables  $X_i$  have sub-gaussian tails.

This class of such distributions, which we call *sub-gaussian*, deserves special attention. This class is sufficiently wide as it contains Gaussian, Bernoulli and all bounded distributions. And, as we will see shortly, concentration results like Hoeffding's inequality can indeed be proved for all sub-gaussian distributions. This makes the family 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 now explore several equivalent approaches to sub-gaussian distributions, examining 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.

Let  $X \sim N(0, 1)$ . Then using (2.3) and symmetry, we obtain the following tail bound:

$$\mathbb{P} \{ |X| \geq t \} \leq 2e^{-t^2/2} \quad \text{for all } t \geq 0. \quad (2.10)$$

(Deduce this formally!) 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

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

Deduce that

$$\|X\|_{L^p} = O(\sqrt{p}) \quad \text{as } p \rightarrow \infty. \quad (2.11)$$

Finally, a classical formula gives the moment generating function of  $X \sim N(0, 1)$ :

$$\mathbb{E} \exp(\lambda X) = e^{\lambda^2/2} \quad \text{for all } \lambda \in \mathbb{R}. \quad (2.12)$$

### 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 tail decay as in (2.10), the growth of moments as in (2.11), 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.<sup>4</sup>*

(i) *There exists  $K_1 > 0$  such that the tails of  $X$  satisfy*

$$\mathbb{P}\{|X| \geq t\} \leq 2 \exp(-t^2/K_1^2) \quad \text{for all } t \geq 0.$$

(ii) *There exists  $K_2 > 0$  such that the moments of  $X$  satisfy*

$$\|X\|_{L^p} = (\mathbb{E}|X|^p)^{1/p} \leq K_2 \sqrt{p} \quad \text{for all } p \geq 1.$$

(iii) *There exists  $K_3 > 0$  such that the MGF of  $X^2$  satisfies*

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

(iv) *There exists  $K_4 > 0$  such that the MGF of  $X^2$  is bounded at some point, namely*

$$\mathbb{E} \exp(X^2/K_4^2) \leq 2.$$

Moreover, if  $\mathbb{E} X = 0$  then properties i–iv are also equivalent to the following one.

(v) *There exists  $K_5 > 0$  such that the MGF of  $X$  satisfies*

$$\mathbb{E} \exp(\lambda X) \leq \exp(K_5^2 \lambda^2) \quad \text{for all } \lambda \in \mathbb{R}.$$

<sup>4</sup> 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  $i, j = 1, \dots, 5$ .

*Proof* **i**  $\Rightarrow$  **ii**. Assume property **i** holds. By homogeneity, and 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

$$\begin{aligned} \mathbb{E}|X|^p &= \int_0^\infty \mathbb{P}\{|X|^p \geq u\} du \\ &= \int_0^\infty \mathbb{P}\{|X| \geq t\} p t^{p-1} dt \quad (\text{by change of variables } u = t^p) \\ &\leq \int_0^\infty 2e^{-t^2} p t^{p-1} dt \quad (\text{by property i}) \\ &= p\Gamma(p/2) \quad (\text{set } t^2 = s \text{ and use definition of Gamma function}) \\ &\leq 3p(p/2)^{p/2} \quad (\text{since } \Gamma(x) \leq 3x^x \text{ for all } x \geq 1/2). \end{aligned}$$

Taking the  $p$ -th root yields property **ii** with  $K_2 \leq 3$ .

**ii**  $\Rightarrow$  **iii**. Assume property **ii** holds. As before, by homogeneity we may assume that  $K_2 = 1$ . Recalling the Taylor series expansion of the exponential function, 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 **ii** guarantees that  $\mathbb{E}[X^{2p}] \leq (2p)^p$ , while Stirling's approximation yields  $p! \geq (p/e)^p$ . Substituting these two bounds, we get

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

provided that  $2e\lambda^2 < 1$ , in which case the geometric series above converges. To bound this quantity further, we can use the numeric inequality  $1/(1-x) \leq e^{2x}$ , 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 yields property **iii** with  $K_3 = 2\sqrt{e}$ .

**iii**  $\Rightarrow$  **iv** is trivial.

**iv**  $\Rightarrow$  **i**. Assume property **iv** holds. As before, we may assume that  $K_4 = 1$ . Then

$$\begin{aligned} \mathbb{P}\{|X| \geq t\} &= \mathbb{P}\{e^{X^2} \geq e^{t^2}\} \\ &\leq e^{-t^2} \mathbb{E} e^{X^2} \quad (\text{by Markov's inequality, Proposition 1.2.4}) \\ &\leq 2e^{-t^2} \quad (\text{by property iv}). \end{aligned}$$

This proves property **i** with  $K_1 = 1$ .

To prove the second part of the proposition, we show that **iii**  $\Rightarrow$  **v** and **v**  $\Rightarrow$  **i**.

**iii**  $\Rightarrow$  **v**. Assume that property **iii** 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  $x \in \mathbb{R}$ . Then

$$\begin{aligned}\mathbb{E} e^{\lambda X} &\leq \mathbb{E} [\lambda X + e^{\lambda^2 X^2}] \\ &= \mathbb{E} e^{\lambda^2 X^2} \quad (\text{since } \mathbb{E} X = 0 \text{ by assumption}) \\ &\leq e^{\lambda^2} \quad \text{if } |\lambda| \leq 1,\end{aligned}$$

where in the last line we used property **iii**. Thus we have proved property **v** in the range  $|\lambda| \leq 1$ . Now assume that  $|\lambda| \geq 1$ . Here we can use the numeric inequality  $2\lambda x \leq \lambda^2 + x^2$ , which is valid for all  $\lambda$  and  $x$ . It follows that

$$\begin{aligned}\mathbb{E} e^{\lambda X} &\leq e^{\lambda^2/2} \mathbb{E} e^{X^2/2} \leq e^{\lambda^2/2} \cdot \exp(1/2) \quad (\text{by property iii}) \\ &\leq e^{\lambda^2} \quad (\text{since } |\lambda| \geq 1).\end{aligned}$$

This proves property **v** with  $K_5 = 1$ .

**v**  $\Rightarrow$  **i**. Assume property **v** holds; we can assume that  $K_5 = 1$ . 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

$$\begin{aligned}\mathbb{P}\{X \geq t\} &= \mathbb{P}\{e^{\lambda X} \geq e^{\lambda t}\} \\ &\leq e^{-\lambda t} \mathbb{E} e^{\lambda X} \quad (\text{by Markov's inequality}) \\ &\leq e^{-\lambda t} e^{\lambda^2} \quad (\text{by property v}) \\ &= e^{-\lambda t + \lambda^2}.\end{aligned}$$

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 **i** holds with  $K_1 = 2$ . The proposition is proved.  $\square$

**Remark 2.5.3.** The constant 2 that appears in some properties in Proposition 2.5.2 does not have any special meaning; it can be replaced by any other absolute constant that is larger than 1. (Check!)

**Exercise 2.5.4.** ☹☹ Show that the condition  $\mathbb{E} X = 0$  is necessary for property **v** to hold.

**Exercise 2.5.5** (On property **iii** in Proposition 2.5.2). ☹☹

- Show that if  $X \sim N(0, 1)$ , the function  $\lambda \mapsto \mathbb{E} \exp(\lambda^2 X^2)$  is only finite in some bounded neighborhood of zero.
- Suppose that some random variable  $X$  satisfies  $\mathbb{E} \exp(\lambda^2 X^2) \leq \exp(K\lambda^2)$  for all  $\lambda \in \mathbb{R}$  and some constant  $K$ . Show that  $X$  is a bounded random variable, i.e.  $\|X\|_\infty < \infty$ .

### 2.5.2 Definition and examples of sub-gaussian distributions

**Definition 2.5.6** (Sub-gaussian random variables). A random variable  $X$  that satisfies one of the equivalent properties i–iv 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  $K_4$  in property iv. In other words, we define

$$\|X\|_{\psi_2} = \inf \{t > 0 : \mathbb{E} \exp(X^2/t^2) \leq 2\}. \quad (2.13)$$

**Exercise 2.5.7.** ☕☕ Check that  $\|\cdot\|_{\psi_2}$  is indeed a norm on the space of sub-gaussian random variables.

Let us restate Proposition 2.5.2 in terms of the sub-gaussian norm. It states that every sub-gaussian random variable  $X$  satisfies the following bounds:

$$\mathbb{P}\{|X| \geq t\} \leq 2 \exp(-ct^2/\|X\|_{\psi_2}^2) \quad \text{for all } t \geq 0; \quad (2.14)$$

$$\|X\|_{L^p} \leq C\|X\|_{\psi_2} \sqrt{p} \quad \text{for all } p \geq 1; \quad (2.15)$$

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

$$\text{if } \mathbb{E} X = 0 \text{ then } \mathbb{E} \exp(\lambda X) \leq \exp(C\lambda^2\|X\|_{\psi_2}^2) \quad \text{for all } \lambda \in \mathbb{R}. \quad (2.16)$$

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

**Example 2.5.8.** Here are some classical examples of sub-gaussian distributions.

- (a) **(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  $X \sim N(0, \sigma^2)$  then  $X$  is sub-gaussian with

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

(Why?)

- (b) **(Bernoulli):** Let  $X$  be a random variable with symmetric Bernoulli distribution (recall Definition 2.2.1). Since  $|X| = 1$ , it follows that  $X$  is a sub-gaussian random variable with

$$\|X\|_{\psi_2} = \frac{1}{\sqrt{\ln 2}}.$$

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

$$\|X\|_{\psi_2} \leq C\|X\|_{\infty} \quad (2.17)$$

where  $C = 1/\sqrt{\ln 2}$ .

**Exercise 2.5.9.** ☕ Check that Poisson, exponential, Pareto and Cauchy distributions are not sub-gaussian.



**Exercise 2.5.10** (Maximum of sub-gaussians). 🍷🍷🍷 Let  $X_1, X_2, \dots$ , be a sequence of sub-gaussian random variables, which are not necessarily independent. Show that

$$\mathbb{E} \max_i \frac{|X_i|}{\sqrt{1 + \log i}} \leq CK,$$

where  $K = \max_i \|X_i\|_{\psi_2}$ . Deduce that for every  $N \geq 2$  we have

$$\mathbb{E} \max_{i \leq N} |X_i| \leq CK \sqrt{\log N}.$$

**Hint:** Denote  $Y_i := X_i / (CK \sqrt{1 + \log i})$  with absolute constant  $C$  chosen sufficiently large. Use subgaussian tail bound (2.14) and then a union bound to conclude that  $\mathbb{P} \left\{ \exists i : |Y_i| \geq t \right\} \lesssim e^{-t^2}$  for any  $t \geq 1$ . Use the integrated tail formula (Lemma 1.2.1), breaking the integral into two integrals: one over  $[0, 1]$  (whose value should be trivial to bound) and the other over  $[1, \infty)$  (where you can use the tail bound obtained before).

**Exercise 2.5.11** (Lower bound). 🍷🍷 Show that the bound in Exercise 2.5.10 is sharp. Let  $X_1, X_2, \dots, X_N$  be independent  $N(0, 1)$  random variables. Prove that

$$\mathbb{E} \max_{i \leq N} X_i \geq c \sqrt{\log N}.$$

## 2.6 General Hoeffding's and Khintchine'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 *rotation invariance* property of sums of independent sub-gaussians.

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

$$\sum_{i=1}^N X_i \sim N\left(0, \sum_{i=1}^N \sigma_i^2\right). \quad (2.18)$$

This fact is a form of the *rotation invariance* property of the normal distribution, which we recall in Section 3.3.2 in more detail.

The rotation invariance property extends to general sub-gaussian distributions, albeit up to an absolute constant.

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

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

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

$$\begin{aligned} \mathbb{E} \exp \left( \lambda \sum_{i=1}^N X_i \right) &= \prod_{i=1}^N \mathbb{E} \exp(\lambda X_i) \quad (\text{by independence}) \\ &\leq \prod_{i=1}^N \exp(C\lambda^2 \|X_i\|_{\psi_2}^2) \quad (\text{by sub-gaussian property (2.16)}) \\ &= \exp(\lambda^2 K^2) \quad \text{where } K^2 := C \sum_{i=1}^N \|X_i\|_{\psi_2}^2. \end{aligned}$$

To complete the proof, we just need to recall that the bound on MGF we just proved characterizes sub-gaussian distributions. Indeed, the equivalence of properties **v** and **iv** in Proposition 2.5.2 and Definition 2.5.6 imply that the sum  $\sum_{i=1}^N X_i$  is sub-gaussian, and

$$\left\| \sum_{i=1}^N X_i \right\|_{\psi_2} \leq C_1 K$$

where  $C_1$  is an absolute constant. The proposition is proved.  $\square$

The approximate rotation invariance can be restated as a concentration inequality via (2.14):

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


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

To compare this general result with the specific case for Bernoulli distributions (Theorem 2.2.2), let us apply Theorem 2.6.3 for  $a_i X_i$  instead of  $X_i$ . We obtain the following.

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

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

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

**Exercise 2.6.4.**  Deduce Hoeffding's inequality for bounded random variables (Theorem 2.2.6) from Theorem 2.6.3, possibly with some absolute constant instead of 2 in the exponent.

As an application of the general Hoeffding's inequality, we can quickly derive the classical Khintchine's inequality for the  $L^p$ -norms of sums of independent random variables.

**Exercise 2.6.5** (Khintchine's inequality). 🍷🍷 Let  $X_1, \dots, X_N$  be independent sub-gaussian random variables with zero means and unit variances, and let  $a = (a_1, \dots, a_N) \in \mathbb{R}^N$ . Prove that for every  $p \in [2, \infty)$  we have

$$\left( \sum_{i=1}^N a_i^2 \right)^{1/2} \leq \left\| \sum_{i=1}^N a_i X_i \right\|_{L^p} \leq CK \sqrt{p} \left( \sum_{i=1}^N a_i^2 \right)^{1/2}$$

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

**Exercise 2.6.6** (Khintchine's inequality for  $p = 1$ ). 🍷🍷🍷 Show that in the setting of Exercise 2.6.5, we have

$$c(K) \left( \sum_{i=1}^N a_i^2 \right)^{1/2} \leq \left\| \sum_{i=1}^N a_i X_i \right\|_{L^1} \leq \left( \sum_{i=1}^N a_i^2 \right)^{1/2}.$$

Here  $K = \max_i \|X_i\|_{\psi_2}$  and  $c(K) > 0$  is a quantity which may depend only on  $K$ .

**Hint:** Use the following extrapolation trick. Prove the inequality  $\|Z\|_2 \leq \|Z\|_1^{1/4} \|Z\|_3^{3/4}$  and use it for  $Z = \sum a_i X_i$ . Get a bound on  $\|Z\|_3$  from Khintchine's inequality for  $p = 3$ .

**Exercise 2.6.7** (Khintchine's inequality for  $p \in (0, 2)$ ). 🍷🍷 State and prove a version of Khintchine's inequality for  $p \in (0, 2)$ .

**Hint:** Modify the extrapolation trick in Exercise 2.6.6.

### 2.6.1 Centering

In results like Hoeffding's inequality, and in many other results we will encounter later, 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. Let us check that centering does not harm the sub-gaussian property.

First note the following simple centering inequality for the  $L^2$  norm:

$$\|X - \mathbb{E} X\|_{L^2} \leq \|X\|_{L^2}. \quad (2.19)$$

(Check this!) Now let us prove a similar centering inequality for the sub-gaussian norm.

**Lemma 2.6.8** (Centering). *If  $X$  is a sub-gaussian random variable then  $X - \mathbb{E} X$  is sub-gaussian, too, and*

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

where  $C$  is an absolute constant.

*Proof* Recall from Exercise 2.5.7 that  $\|\cdot\|_{\psi_2}$  is a norm. Thus we can use triangle inequality and get

$$\|X - \mathbb{E} X\|_{\psi_2} \leq \|X\|_{\psi_2} + \|\mathbb{E} X\|_{\psi_2}. \quad (2.20)$$

We only have to bound the second term. Note that for any constant random variable  $a$ , we trivially have<sup>5</sup>  $\|a\|_{\psi_2} \lesssim |a|$  (recall 2.17). Using this for  $a = \mathbb{E} X$ , we

<sup>5</sup> In this proof and later, the notation  $a \lesssim b$  means that  $a \leq Cb$  where  $C$  is some absolute constant.

get

$$\begin{aligned}
\|\mathbb{E} X\|_{\psi_2} &\lesssim |\mathbb{E} X| \\
&\leq \mathbb{E} |X| \quad (\text{by Jensen's inequality}) \\
&= \|X\|_1 \\
&\lesssim \|X\|_{\psi_2} \quad (\text{using (2.15) with } p = 1).
\end{aligned}$$

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

**Exercise 2.6.9.** ☕☕☕ Show that, unlike (2.19), the centering inequality in Lemma 2.6.8 does not hold with  $C = 1$ .

## 2.7 Sub-exponential distributions

The class of sub-gaussian distributions is natural and quite large. Nevertheless, it leaves out some important distributions whose tails are heavier than gaussian. Here is one example. Consider a standard normal random vector  $g = (g_1, \dots, g_N)$  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  $g$ , which is

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

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<sup>6</sup>

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

The tails of  $g_i^2$  are like for the exponential distribution, and are strictly heavier than sub-gaussian. This prevents us from using Hoeffding's inequality (Theorem 2.6.2) if we want to study the concentration of  $\|g\|_2$ .

In this section we focus on the class of distributions that have at least an exponential tail decay, and in Section 2.8 we prove an analog of Hoeffding's inequality for them.

Our analysis here will be quite similar to what we did for sub-gaussian distributions in Section 2.5. The following is a version of Proposition 2.5.2 for sub-exponential 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.<sup>7</sup>*

<sup>6</sup> Here we ignored the pre-factor  $1/t$ , which does not make much effect on the exponent.

<sup>7</sup> 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  $i, j = 1, 2, 3, 4$ .

(a) The tails of  $X$  satisfy

$$\mathbb{P}\{|X| \geq t\} \leq 2 \exp(-t/K_1) \quad \text{for all } t \geq 0.$$

(b) The moments of  $X$  satisfy

$$\|X\|_{L^p} = (\mathbb{E}|X|^p)^{1/p} \leq K_2 p \quad \text{for all } p \geq 1.$$

(c) The MGF of  $|X|$  satisfies

$$\mathbb{E} \exp(\lambda|X|) \leq \exp(K_3 \lambda) \quad \text{for all } \lambda \text{ such that } 0 \leq \lambda \leq \frac{1}{K_3}.$$

(d) The MGF of  $|X|$  is bounded at some point, namely

$$\mathbb{E} \exp(|X|/K_4) \leq 2.$$

Moreover, if  $\mathbb{E} X = 0$  then properties **a**–**d** are also equivalent to the following one.

(e) The MGF of  $X$  satisfies

$$\mathbb{E} \exp(\lambda X) \leq \exp(K_5^2 \lambda^2) \quad \text{for all } \lambda \text{ such that } |\lambda| \leq \frac{1}{K_5}.$$

*Proof* We will prove the equivalence of properties **b** and **e** only; you will check the other implications in Exercise 2.7.2.

**b**  $\Rightarrow$  **e**. Without loss of generality we may assume that  $K_2 = 1$ . (Why?) Expanding the exponential function in Taylor series, we obtain

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

where we used the assumption that  $\mathbb{E} X = 0$ . Property **b** guarantees that  $\mathbb{E}[X^p] \leq p^p$ , while Stirling's approximation yields  $p! \geq (p/e)^p$ . Substituting these two bounds, we obtain

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

provided that  $|e\lambda| < 1$ , in which case the geometric series above converges. Moreover, if  $|e\lambda| \leq 1/2$  then we can further bound the quantity above by

$$1 + 2e^2 \lambda^2 \leq \exp(2e^2 \lambda^2).$$

Summarizing, we have shown that

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

This yields property **e** with  $K_5 = 2e$ .

**e**  $\Rightarrow$  **b**. Without loss of generality, we can assume that  $K_5 = 1$ . We will use the numeric inequality

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

which is valid for all  $x \in \mathbb{R}$  and  $p > 0$ . (Check it by dividing both sides by  $p^p$  and taking  $p$ -th roots.) Substituting  $x = X$  and taking expectation, we get

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

Property **e** gives  $\mathbb{E} e^X \leq e$  and  $\mathbb{E} e^{-X} \leq e$ . Thus

$$\mathbb{E} |X|^p \leq 2ep^p.$$

This yields property **b** with  $K_2 = 2e$ .  $\square$

**Exercise 2.7.2.**  $\clubsuit\clubsuit$  Prove the equivalence of properties **a–d** in Proposition 2.7.1 by modifying the proof of Proposition 2.5.2.

**Exercise 2.7.3.**  $\clubsuit\clubsuit\clubsuit$  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.

**Exercise 2.7.4.**  $\clubsuit$  Argue that the bound in property **c** can not be extended for all  $\lambda$  such that  $|\lambda| \leq 1/K_3$ .

**Definition 2.7.5** (Sub-exponential random variables). A random variable  $X$  that satisfies one of the equivalent properties **a–d** 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,

$$\|X\|_{\psi_1} = \inf \{t > 0 : \mathbb{E} \exp(|X|/t) \leq 2\}. \quad (2.21)$$

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 2.7.6** (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  $K = L^2$ .  $\square$

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

**Lemma 2.7.7** (Product of sub-gaussians is sub-exponential). *Let  $X$  and  $Y$  be sub-gaussian random variables. Then  $XY$  is sub-exponential. Moreover,*

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

*Proof* Without loss of generality we may assume that  $\|X\|_{\psi_2} = \|Y\|_{\psi_2} = 1$ . (Why?) The lemma claims that if

$$\mathbb{E} \exp(X^2) \leq 2 \quad \text{and} \quad \mathbb{E} \exp(Y^2) \leq 2 \quad (2.22)$$

then  $\mathbb{E} \exp(|XY|) \leq 2$ . To prove this, let us use the elementary Young's inequality, which states that

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

It yields

$$\begin{aligned} \mathbb{E} \exp(|XY|) &\leq \mathbb{E} \exp\left(\frac{X^2}{2} + \frac{Y^2}{2}\right) \quad (\text{by Young's inequality}) \\ &= \mathbb{E} \left[ \exp\left(\frac{X^2}{2}\right) \exp\left(\frac{Y^2}{2}\right) \right] \\ &\leq \frac{1}{2} \mathbb{E} [\exp(X^2) + \exp(Y^2)] \quad (\text{by Young's inequality}) \\ &= \frac{1}{2}(2 + 2) = 2 \quad (\text{by assumption (2.22)}). \end{aligned}$$

The proof is complete.  $\square$

**Example 2.7.8.** Let us mention a few 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

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

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!)

**Remark 2.7.9** (MGF near the origin). You may be surprised to see the same bound on the MGF near the origin for sub-gaussian and sub-exponential distributions. (Compare property [e](#) in Propositions [2.5.2](#) and [2.7.1](#).) This should not be very surprising though: this kind of local bound is expected from a *general* random variable  $X$  with mean zero and unit variance. To see this, assume for simplicity that  $X$  is bounded. The MGF of  $X$  can be approximated using the first two terms of the Taylor expansion:

$$\mathbb{E} \exp(\lambda X) \approx \mathbb{E} \left[ 1 + \lambda X + \frac{\lambda^2 X^2}{2} + o(\lambda^2 X^2) \right] = 1 + \frac{\lambda^2}{2} \approx e^{\lambda^2/2}$$

as  $\lambda \rightarrow 0$ . For the standard *normal* distribution  $N(0, 1)$ , this approximation

becomes an equality, see (2.12). For *sub-gaussian* distributions, Proposition 2.5.2 says that a bound like this holds for all  $\lambda$ , and this characterizes sub-gaussian distributions. And for *sub-exponential* distributions, Proposition 2.7.1 says that this bound holds for small  $\lambda$ , and this characterizes sub-exponential distributions. For larger  $\lambda$ , no general bound may exist for sub-exponential distributions: indeed, for the *exponential* random variable  $X \sim \text{Exp}(1)$ , the MGF is infinite for  $\lambda \geq 1$ . (Check this!)

**Exercise 2.7.10** (Centering). ♣ Prove an analog of the Centering Lemma 2.6.8 for sub-exponential random variables  $X$ :

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

### 2.7.1 A more general view: Orlicz spaces

Sub-gaussian distributions can be introduced within a more general framework of *Orlicz spaces*. A function  $\psi : [0, \infty) \rightarrow [0, \infty)$  is called an *Orlicz function* if  $\psi$  is convex, increasing, and satisfies

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

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, i.e.

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

**Exercise 2.7.11.** ♣♣ Show that  $\|X\|_{\psi}$  is indeed a norm on the space  $L_{\psi}$ .

It can also be shown that  $L_{\psi}$  is complete and thus a Banach space.

**Example 2.7.12** ( $L^p$  space). Consider the function

$$\psi(x) = x^p,$$

which is obviously an Orlicz function for  $p \geq 1$ . The resulting Orlicz space  $L_{\psi}$  is the classical space  $L^p$ .

**Example 2.7.13** ( $L_{\psi_2}$  space). Consider the function

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

which is obviously an Orlicz function. The resulting Orlicz norm is exactly the sub-gaussian norm  $\|\cdot\|_{\psi_2}$  that we defined in (2.13). The corresponding Orlicz space  $L_{\psi_2}$  consists of all sub-gaussian random variables.



**Remark 2.7.14.** We can easily locate  $L_{\psi_2}$  in the hierarchy of the classical  $L^p$  spaces:

$$L^\infty \subset L_{\psi_2} \subset L^p \quad \text{for every } p \in [1, \infty).$$

The first inclusion follows from Property ii of Proposition 2.5.2, and the second inclusion from bound (2.17). Thus the space of sub-gaussian random variables  $L_{\psi_2}$  is smaller than all of  $L^p$  spaces, but it is still larger than the space of bounded random variables  $L^\infty$ .

## 2.8 Bernstein's inequality

We are ready to state and prove a concentration inequality for sums of independent sub-exponential random variables.

**Theorem 2.8.1** (Bernstein's inequality). *Let  $X_1, \dots, X_N$  be independent, mean zero, sub-exponential random variables. Then, for every  $t \geq 0$ , we have*

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

where  $c > 0$  is an absolute constant.

*Proof* We begin the proof in the same way as we argued about other concentration inequalities for  $S = \sum_{i=1}^N X_i$ , e.g. Theorems 2.2.2 and 2.3.1. Multiply both sides of the inequality  $S \geq t$  by a parameter  $\lambda$ , exponentiate, and then use Markov's inequality and independence. This leads to the bound (2.7), which is

$$\mathbb{P}\{S \geq t\} \leq e^{-\lambda t} \prod_{i=1}^N \mathbb{E} \exp(\lambda X_i). \quad (2.23)$$

To bound the MGF of each term  $X_i$ , we use property e in Proposition 2.7.1. It says that if  $\lambda$  is small enough so that

$$|\lambda| \leq \frac{c}{\max_i \|X_i\|_{\psi_1}}, \quad (2.24)$$

then<sup>8</sup>  $\mathbb{E} \exp(\lambda X_i) \leq \exp(C\lambda^2 \|X_i\|_{\psi_1}^2)$ . Substituting this into (2.23), we obtain

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

Now we minimize this expression in  $\lambda$  subject to the constraint (2.24). The optimal choice is  $\lambda = \min(\frac{t}{2C\sigma^2}, \frac{c}{\max_i \|X_i\|_{\psi_1}})$ , for which we obtain

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

<sup>8</sup> Recall that by Proposition 2.7.1 and definition of the sub-exponential norm, property e holds for a value of  $K_5$  that is within an absolute constant factor of  $\|X\|_{\psi_1}$ .

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

To put Theorem 2.8.1 in a more convenient form, let us apply it for  $a_i X_i$  instead of  $X_i$ .

**Theorem 2.8.2** (Bernstein's inequality). *Let  $X_1, \dots, X_N$  be independent, mean zero, sub-exponential random variables, and  $a = (a_1, \dots, a_N) \in \mathbb{R}^N$ . Then, for every  $t \geq 0$ , we have*

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

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

In the special case where  $a_i = 1/N$ , we obtain a form of Bernstein's inequality for averages:

**Corollary 2.8.3** (Bernstein's inequality). *Let  $X_1, \dots, X_N$  be independent, mean zero, sub-exponential random variables. Then, for every  $t \geq 0$ , we have*

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

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

This result can be considered as a quantitative form of the *law of large numbers* for the averages  $\frac{1}{N} \sum_{i=1}^N X_i$ .

Let us compare Bernstein's inequality (Theorem 2.8.1) 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 2.8.1 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  $\exp(-ct/\|X_i\|_{\psi_1})$ .

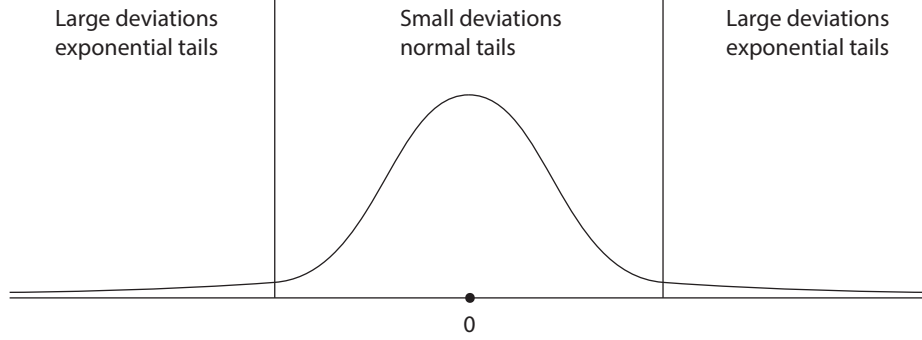
We already saw a similar mixture of two tails, one for small deviations and the other for large deviations, in our analysis of Chernoff's inequality; see Remark 2.3.7. To put Bernstein's inequality in the same perspective, let us normalize the sum as in the central limit theorem and apply Theorem 2.8.2. We obtain<sup>9</sup>

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

Thus, in the *small deviation* regime where  $t \leq C\sqrt{N}$ , we have a sub-gaussian tail bound as if the sum had a *normal distribution* with constant variance. Note

<sup>9</sup> For simplicity, we suppressed here the dependence on  $K$  by allowing the constants  $c, C$  depend on  $K$ .

that this domain widens as  $N$  increases and the central limit theorem becomes more powerful. For *large deviations* where  $t \geq C\sqrt{N}$ , the sum has a heavier, *sub-exponential* tail bound, which can be due to the contribution of a single term  $X_i$ . We illustrate this in Figure 2.3.



**Figure 2.3** Bernstein's inequality for a sum of sub-exponential random variables gives a mixture of two tails: sub-gaussian for small deviations and sub-exponential for large deviations.

Let us mention a strengthening of Bernstein's inequality that is sensitive to the variance of the sum. It holds under the stronger assumption that the random variables  $X_i$  are bounded.

**Theorem 2.8.4** (Bernstein's inequality for bounded distributions). *Let  $X_1, \dots, X_N$  be independent, mean zero random variables, such that  $|X_i| \leq K$  all  $i$ . Then, for every  $t \geq 0$ , we have*

$$\mathbb{P}\left\{\left|\sum_{i=1}^N X_i\right| \geq t\right\} \leq 2 \exp\left(-\frac{t^2/2}{\sigma^2 + Kt/3}\right).$$

Here  $\sigma^2 = \sum_{i=1}^N \mathbb{E} X_i^2$  is the variance of the sum.

We leave the proof of this theorem to the next two exercises.

**Exercise 2.8.5** (A bound on MGF). ☹️☹️ Let  $X$  be a mean-zero random variable such that  $|X| \leq K$ . Prove the following bound on the MGF of  $X$ :

$$\mathbb{E} \exp(\lambda X) \leq \exp(g(\lambda) \mathbb{E} X^2) \quad \text{where} \quad g(\lambda) = \frac{\lambda^2/2}{1 - |\lambda|K/3},$$

provided that  $|\lambda| < 3/K$ .

**Hint:** Check the numeric inequality  $e^z \leq 1 + z + \frac{z^2/2}{1-|z|/3}$  that is valid provided  $|z| < 3$ , apply it for  $z = \lambda X$ , and take expectations on both sides.

**Exercise 2.8.6.** ☹️☹️ Deduce Theorem 2.8.4 from the bound in Exercise 2.8.5.

**Hint:** Follow the proof of Theorem 2.8.1.

## 2.9 Notes

The topic of concentration inequalities is very wide, and we will continue to examine it in Chapter 5. We refer the reader to [8, Appendix A], [152, Chapter 4], [130], [30], [78, Chapter 7], [11, Section 3.5.4], [174, Chapter 1], [14, Chapter 4] for various versions of Hoeffding's, Chernoff's, and Bernstein's inequalities, and related results.

Proposition 2.1.2 on the tails of the normal distribution is borrowed from [72, Theorem 1.4]. The proof of Berry-Esseen's central limit theorem (Theorem 2.1.3) with an extra factor 3 on the right hand side can be found e.g. in [72, Section 2.4.d]; the best currently known factor is  $\approx 0.47$  [120].

It is worthwhile to mention two important concentration inequalities that were omitted in this chapter. One is the *bounded differences inequality*, also called *McDiarmid's inequality*, which works not only for sums but for general functions of independent random variables. It is a generalization of Hoeffding's inequality (Theorem 2.2.6).

**Theorem 2.9.1** (Bounded differences inequality). *Let  $X_1, \dots, X_N$  be independent random variables.<sup>10</sup> Let  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  be a measurable function. Assume that the value of  $f(x)$  can change by at most  $c_i > 0$  under an arbitrary change<sup>11</sup> of a single coordinate of  $x \in \mathbb{R}^n$ . Then, for any  $t > 0$ , we have*

$$\mathbb{P} \{f(X) - \mathbb{E} f(X) \geq t\} \leq \exp \left( -\frac{2t^2}{\sum_{i=1}^N c_i^2} \right)$$

where  $X = (X_1, \dots, X_n)$ .

Another result worth mentioning is *Bennett's inequality*, which can be regarded as a generalization of Chernoff's inequality.

**Theorem 2.9.2** (Bennett's inequality). *Let  $X_1, \dots, X_N$  be independent random variables. Assume that  $|X_i - \mathbb{E} X_i| \leq K$  almost surely for every  $i$ . Then, for any  $t > 0$ , we have*

$$\mathbb{P} \left\{ \sum_{i=1}^N (X_i - \mathbb{E} X_i) \geq t \right\} \leq \exp \left( -\frac{\sigma^2}{K^2} h \left( \frac{Kt}{\sigma^2} \right) \right)$$

where  $\sigma^2 = \sum_{i=1}^N \text{Var}(X_i)$  is the variance of the sum, and  $h(u) = (1+u) \log(1+u) - u$ .

In the small deviation regime, where  $u := Kt/\sigma^2 \ll 1$ , we have asymptotically  $h(u) \approx u^2$  and Bennett's inequality gives approximately the Gaussian tail bound  $\approx \exp(-t^2/\sigma^2)$ . In the large deviations regime, say where  $u \gg Kt/\sigma^2 \geq 2$ , we have  $h(u) \geq \frac{1}{2}u \log u$ , and Bennett's inequality gives a Poisson-like tail  $(\sigma^2/Kt)^{t/2K}$ .

<sup>10</sup> The theorem remains valid if the random variables  $X_i$  take values in an abstract set  $\mathcal{X}$  and  $f : \mathcal{X} \rightarrow \mathbb{R}$ .

<sup>11</sup> This means that for any index  $i$  and any  $x_1, \dots, x_n, x'_i$ , we have  $|f(x_1, \dots, x_{i-1}, x_i, x_{i+1}, \dots, x_n) - f(x_1, \dots, x_{i-1}, x'_i, x_{i+1}, \dots, x_n)| \leq c_i$ .

Both the bounded differences inequality and Bennett's inequality can be proved by the same general method as Hoeffding's inequality (Theorem 2.2.2) and Chernoff's inequality (Theorem 2.3.1), namely by bounding the moment generating function of the sum. This method was pioneered by Sergei Bernstein in the 1920-30's. Our presentation of Chernoff's inequality in Section 2.3 mostly follows [152, Chapter 4].

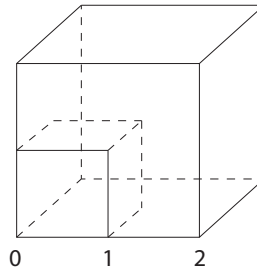
Section 2.4 scratches the surface of the rich theory of *random graphs*. The books [26, 107] offer a comprehensive introduction to the random graph theory.

The presentation in Sections 2.5–2.8 mostly follows [222]; see [78, Chapter 7] for some more elaborate results. For sharp versions of Khintchine's inequalities in Exercises 2.6.5–2.6.7 and related results, see e.g. [195, 95, 118, 155].

## Random vectors in high dimensions

In this chapter we study the distributions of random vectors  $X = (X_1, \dots, X_n) \in \mathbb{R}^n$  where the dimension  $n$  is typically very large. Examples of high-dimensional distributions abound in data science. For instance, computational biologists study the expressions of  $n \sim 10^4$  genes in the human genome, which can be modeled as a random vector  $X = (X_1, \dots, X_n)$  that encodes the gene expressions of a person randomly drawn from a given population.

Life in high dimensions presents new challenges, which stem from the fact that there is *exponentially more room* in higher dimensions than in lower dimensions. For example, in  $\mathbb{R}^n$  the volume of a cube of side 2 is  $2^n$  times larger than the volume of a unit cube, even though the sides of the cubes are just a factor 2 apart (see Figure 3.1). The abundance of room in higher dimensions makes many algorithmic tasks exponentially more difficult, a phenomenon known as the “*curse of dimensionality*”.



**Figure 3.1** The abundance of room in high dimensions: the larger cube has volume exponentially larger than the smaller cube.

Probability in high dimensions offers an array of tools to circumvent these difficulties; some examples will be given in this chapter. We start by examining the Euclidean norm  $\|X\|_2$  of a random vector  $X$  with independent coordinates, and we show in Section 3.1 that the norm concentrates tightly about its mean. Further basic results and examples of high-dimensional distributions (multivariate normal, spherical, Bernoulli, frames, etc.) are covered in Section 3.2, which also discusses principal component analysis, a powerful data exploratory procedure.

In Section 3.5 we give a probabilistic proof of the classical Grothendieck’s inequality, and give an application to semidefinite optimization. We show that

one can sometimes relax hard optimization problems to tractable, semidefinite programs, and use Grothendieck's inequality to analyze the quality of such relaxations. In Section 3.6 we give a remarkable example of a semidefinite relaxation of a hard optimization problem – finding the maximum cut of a given graph. We present there the classical Goemans-Williamson randomized approximation algorithm for the maximum cut problem. In Section 3.7 we give an alternative proof of Grothendieck's inequality (and with almost the best known constant) by introducing the kernel trick, a method that has significant applications in machine learning.

### 3.1 Concentration of the norm

Where in the space  $\mathbb{R}^n$  is a random vector  $X = (X_1, \dots, X_n)$  likely to be located? Assume the coordinates  $X_i$  are independent random variables with zero means and unit variances. What length do we expect  $X$  to have? We have

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

So we should expect the length of  $X$  to be

$$\|X\|_2 \approx \sqrt{n}.$$

We will see now that  $X$  is indeed very close to  $\sqrt{n}$  with high probability.

**Theorem 3.1.1** (Concentration of the norm). *Let  $X = (X_1, \dots, X_n) \in \mathbb{R}^n$  be a random vector with independent, sub-gaussian coordinates  $X_i$  that satisfy  $\mathbb{E} X_i^2 = 1$ . Then*

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

where  $K = \max_i \|X_i\|_{\psi_2}$  and  $C$  is an absolute constant.<sup>1</sup>

*Proof* For simplicity, we 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, mean zero random variables

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

Since the random variable  $X_i$  is sub-gaussian,  $X_i^2 - 1$  is sub-exponential, and more precisely

$$\begin{aligned} \|X_i^2 - 1\|_{\psi_1} &\leq C \|X_i^2\|_{\psi_1} \quad (\text{by centering, see Exercise 2.7.10}) \\ &= C \|X_i\|_{\psi_2}^2 \quad (\text{by Lemma 2.7.6}) \\ &\leq CK^2. \end{aligned}$$

<sup>1</sup> From now on, we will always denote various positive absolute constants by  $C, c, C_1, c_1$  without saying this explicitly.

Applying Bernstein's inequality (Corollary 2.8.3), we obtain for any  $u \geq 0$  that

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

(Here we used that  $K^4 \geq K^2$  since we assumed that  $K \geq 1$ .)

This is a good concentration inequality for  $\|X\|_2^2$ , from which we are going to deduce a concentration inequality for  $\|X\|_2$ . To make the link, we can use the following elementary observation that is valid for all numbers  $z \geq 0$ :

$$|z - 1| \geq \delta \quad \text{implies} \quad |z^2 - 1| \geq \max(\delta, \delta^2). \quad (3.2)$$

(Check it!) We obtain for any  $\delta \geq 0$  that

$$\begin{aligned} \mathbb{P} \left\{ \left| \frac{1}{\sqrt{n}} \|X\|_2 - 1 \right| \geq \delta \right\} &\leq \mathbb{P} \left\{ \left| \frac{1}{n} \|X\|_2^2 - 1 \right| \geq \max(\delta, \delta^2) \right\} \quad (\text{by (3.2)}) \\ &\leq 2 \exp \left( -\frac{cn}{K^4} \cdot \delta^2 \right) \quad (\text{by (3.1) for } u = \max(\delta, \delta^2)). \end{aligned}$$

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

$$\mathbb{P} \left\{ \left| \|X\|_2 - \sqrt{n} \right| \geq t \right\} \leq 2 \exp \left( -\frac{ct^2}{K^4} \right) \quad \text{for all } t \geq 0. \quad (3.3)$$

As we know from Section 2.5.2, this is equivalent to the conclusion of the theorem.  $\square$

**Remark 3.1.2** (Deviation). Theorem 3.1.1 states that with high probability,  $X$  takes values very close to the sphere of radius  $\sqrt{n}$ . In particular, with high probability (say, 0.99),  $X$  even stays within *constant distance* from that sphere. Such small, constant deviations could be surprising at the first sight, so let us explain this intuitively. The square of the norm,  $S_n := \|X\|_2^2$  has mean  $n$  and standard deviation  $O(\sqrt{n})$ . (Why?) Thus  $\|X\|_2 = \sqrt{S_n}$  ought to deviate by  $O(1)$  around  $\sqrt{n}$ . This is because

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

see Figure 3.2 for an illustration.

**Remark 3.1.3** (Anisotropic distributions). After we develop more tools, we will prove a generalization of Theorem 3.1.1 for *anisotropic* random vectors  $X$ ; see Theorem 6.3.2.

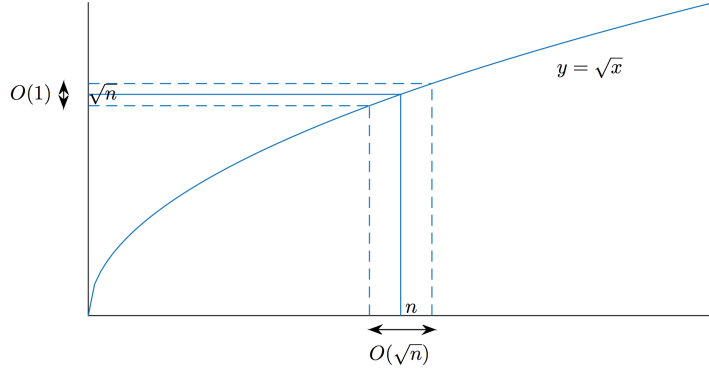
**Exercise 3.1.4** (Expectation of the norm). 🐼🐼🐼

(a) Deduce from Theorem 3.1.1 that

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

(b) Can  $CK^2$  be replaced by  $o(1)$ , a quantity that vanishes as  $n \rightarrow \infty$ ?





**Figure 3.2** Concentration of the norm of a random vector  $X$  in  $\mathbb{R}^n$ . While  $\|X\|_2^2$  deviates by  $O(\sqrt{n})$  around  $n$ ,  $\|X\|_2$  deviates by  $O(1)$  around  $\sqrt{n}$ .

**Exercise 3.1.5** (Variance of the norm). 🐼🐼🐼 Deduce from Theorem 3.1.1 that

$$\text{Var}(\|X\|_2) \leq CK^4.$$

**Hint:** Use Exercise 3.1.4.

The result of the last exercise actually holds not only for sub-gaussian distributions, but for all distributions with bounded fourth moment:

**Exercise 3.1.6** (Variance of the norm under finite moment assumptions). 🐼🐼🐼 Let  $X = (X_1, \dots, X_n) \in \mathbb{R}^n$  be a random vector with independent coordinates  $X_i$  that satisfy  $\mathbb{E} X_i^2 = 1$  and  $\mathbb{E} X_i^4 \leq K^4$ . Show that

$$\text{Var}(\|X\|_2) \leq CK^4.$$

**Hint:** First check that  $\mathbb{E}(\|X\|_2^2 - n)^2 \leq K^4 n$  by expansion. This yields in a simple way that  $\mathbb{E}(\|X\|_2 - \sqrt{n})^2 \leq K^4$ . Finally, replace  $\sqrt{n}$  by  $\mathbb{E}\|X\|_2$  arguing like in Exercise 3.1.4.

**Exercise 3.1.7** (Small ball probabilities). 🐼🐼 Let  $X = (X_1, \dots, X_n) \in \mathbb{R}^n$  be a random vector with independent coordinates  $X_i$  with continuous distributions. Assume that the densities of  $X_i$  are uniformly bounded by 1. Show that, for any  $\varepsilon > 0$ , we have

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

**Hint:** While this inequality does not follow from the result of Exercise 2.2.10 (why?), you can prove it by a similar argument.

### 3.2 Covariance matrices and principal component analysis

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, which the reader may have already seen in basic courses.

The concept of the *mean* of a random variable generalizes in a straightforward way for a random vectors  $X$  taking values in  $\mathbb{R}^n$ . The notion of variance is replaced in high dimensions by the *covariance matrix* of a random vector  $X \in \mathbb{R}^n$ , defined as follows:

$$\text{cov}(X) = \mathbb{E}(X - \mu)(X - \mu)^\top = \mathbb{E} X X^\top - \mu \mu^\top, \quad \text{where } \mu = \mathbb{E} X.$$

Thus  $\text{cov}(X)$  is an  $n \times n$ , symmetric positive semidefinite matrix. The formula for covariance is a direct high-dimensional generalization of the definition of variance for a 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  $\text{cov}(X)$  are the *covariances* of the pairs of coordinates of  $X = (X_1, \dots, X_n)$ :

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

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

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

The second moment matrix is a higher dimensional generalization of the second moment  $\mathbb{E} Z^2$  of a random variable  $Z$ . By translation (replacing  $X$  with  $X - \mu$ ), we can assume in many problems that  $X$  has zero mean, and thus covariance and second moment matrices are equal:

$$\text{cov}(X) = \Sigma(X).$$

This observation allows us to mostly focus on the second moment matrix  $\Sigma = \Sigma(X)$  rather than on the covariance  $\text{cov}(X)$  in the future.

Like the covariance matrix, the second moment matrix  $\Sigma$  is also an  $n \times n$ , symmetric and positive semidefinite matrix. 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

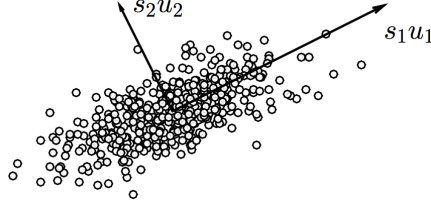
$$\Sigma = \sum_{i=1}^n s_i u_i u_i^\top,$$

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

### 3.2.1 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, for example the genetic data we mentioned on p. 41. The eigenvector  $u_1$  corresponding to the largest eigenvalue  $s_1$  defines the first *principal direction*. This is the direction in

which the distribution is most extended, and it explains most of the variability in the data. The next eigenvector  $u_2$  (corresponding to the next largest eigenvalue  $s_2$ ) defines the next principal direction; it best explains the remaining variations in the data, and so on. This is illustrated in the Figure 3.3.



**Figure 3.3** Illustration of the PCA. 200 sample points are shown from a distribution in  $\mathbb{R}^2$ . The covariance matrix  $\Sigma$  has eigenvalues  $s_i$  and eigenvectors  $u_i$ .

It often happens with real data that only a few eigenvalues  $s_i$  are large and can be considered as informative; the remaining eigenvalues are small and considered as noise. In such situations, a few principal directions can explain most variability in the data. Even though the data is presented in a high-dimensional space  $\mathbb{R}^n$ , such data is essentially *low dimensional*. It clusters near the low-dimensional subspace  $E$  spanned by the first few principal components.

The most basic data analysis algorithm, called the *principal component analysis* (PCA), computes the first few principal components and then projects the data in  $\mathbb{R}^n$  onto the subspace  $E$  spanned by them. This considerably reduces the dimension of the data and simplifies the data analysis. For example, if  $E$  is two- or three-dimensional, PCA allows one to visualize the data.

### 3.2.2 Isotropy

We might remember from a basic probability course how it is often convenient to assume that random variables have zero means and unit variances. This is also true in higher dimensions, where the notion of isotropy generalizes the assumption of unit variance.

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

$$\Sigma(X) = \mathbb{E} X X^\top = I_n$$

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

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

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

The following exercise gives a high-dimensional version of standard score.

**Exercise 3.2.2** (Reduction to isotropy). ☛

- (a) 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 a fixed  $n \times n$  symmetric positive semidefinite matrix. Check that the random vector

$$X := \mu + \Sigma^{1/2}Z$$

has mean  $\mu$  and covariance matrix  $\text{cov}(X) = \Sigma$ .

- (b) Let  $X$  be a random vector with mean  $\mu$  and invertible covariance matrix  $\Sigma = \text{cov}(X)$ . Check that the random vector

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

is an isotropic, mean zero random vector.

This observation will allow us in many future results about random vectors to assume without loss of generality that they have zero means and are isotropic.

### 3.2.3 Properties of isotropic distributions

**Lemma 3.2.3** (Characterization of isotropy). *A random vector  $X$  in  $\mathbb{R}^n$  is isotropic if and only if*

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

*Proof* Recall that two symmetric  $n \times n$  matrices  $A$  and  $B$  are equal if and only if  $x^\top A x = x^\top B x$  for all  $x \in \mathbb{R}^n$ . (Check this!) Thus  $X$  is isotropic if and only if

$$x^\top (\mathbb{E} X X^\top) x = x^\top I_n x \quad \text{for all } x \in \mathbb{R}^n.$$

The left side of this identity equals  $\mathbb{E} \langle X, x \rangle^2$  and the right side is  $\|x\|_2^2$ . This completes the proof.  $\square$

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 a mean-zero random vector  $X$  is isotropic if and only if *all one-dimensional marginals of  $X$  have unit variance*. Informally, this means that an isotropic distribution is extended evenly in all directions.

**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

$$\begin{aligned}
\mathbb{E} \|X\|_2^2 &= \mathbb{E} X^\top X = \mathbb{E} \operatorname{tr}(X^\top X) \quad (\text{viewing } X^\top X \text{ as a } 1 \times 1 \text{ matrix}) \\
&= \mathbb{E} \operatorname{tr}(X X^\top) \quad (\text{by the cyclic property of trace}) \\
&= \operatorname{tr}(\mathbb{E} X X^\top) \quad (\text{by linearity}) \\
&= \operatorname{tr}(I_n) \quad (\text{by isotropy}) \\
&= n.
\end{aligned}$$

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

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

where by  $\mathbb{E}_Y$  we of course denote the expectation with respect to  $Y$ . To compute the inner expectation, we apply Lemma 3.2.3 with  $x = Y$  and conclude that the inner expectation equals  $\|Y\|_2^2$ . Thus

$$\begin{aligned}
\mathbb{E} \langle X, Y \rangle^2 &= \mathbb{E}_Y \|Y\|_2^2 \\
&= n \quad (\text{by the first part of lemma}).
\end{aligned}$$

The proof is complete.  $\square$

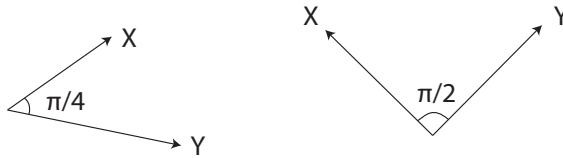
**Remark 3.2.5** (Almost orthogonality of independent vectors). Let us normalize the random vectors  $X$  and  $Y$  in Lemma 3.2.4, setting

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

Lemma 3.2.4 is basically telling us that<sup>2</sup>  $\|X\|_2 \asymp \sqrt{n}$ ,  $\|Y\|_2 \asymp \sqrt{n}$  and  $\langle X, Y \rangle \asymp \sqrt{n}$  with high probability, which implies that

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

Thus, in high-dimensional spaces independent and isotropic random vectors tend to be *almost orthogonal*, see Figure 3.4.



**Figure 3.4** Independent isotropic random vectors tend to be almost orthogonal in high dimensions but not in low dimensions. On the plane, the average angle is  $\pi/4$ , while in high dimensions it is close to  $\pi/2$ .

<sup>2</sup> This argument is not entirely rigorous, since Lemma 3.2.4 is about expectation and not high probability. To make it more rigorous, one can use Theorem 3.1.1 about concentration of the norm.

This may sound surprising since this is not the case in low dimensions. For example, the angle between two random independent and uniformly distributed directions on the plane has mean  $\pi/4$ . (Check!) But in higher dimensions, there is much more room as we mentioned in the beginning of this chapter. This is an intuitive reason why random directions in high-dimensional spaces tend to be very far from each other, i.e. almost orthogonal.

**Exercise 3.2.6** (Distance between independent isotropic vectors). ☛ Let  $X$  and  $Y$  be independent, mean zero, isotropic random vectors in  $\mathbb{R}^n$ . Check that

$$\mathbb{E} \|X - Y\|_2^2 = 2n.$$

### 3.3 Examples of high-dimensional distributions

In this section we give several basic examples of isotropic high-dimensional distributions.

#### 3.3.1 Spherical and Bernoulli distributions

The coordinates of an isotropic random vector are always uncorrelated (why?), but they are not necessarily independent. An example of this situation is the *spherical distribution*, where a random vector  $X$  is uniformly distributed<sup>3</sup> on the Euclidean sphere in  $\mathbb{R}^n$  with center at the origin and radius  $\sqrt{n}$ :

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

**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  $X = (X_1, \dots, X_n)$  is symmetric Bernoulli if the coordinates  $X_i$  are independent, symmetric Bernoulli random variables. Equivalently, we may say that  $X$  is uniformly distributed on the unit discrete cube in  $\mathbb{R}^n$ :

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

The symmetric Bernoulli distribution is isotropic. (Check!)

More generally, we may consider any random vector  $X = (X_1, \dots, X_n)$  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?)

<sup>3</sup> More rigorously, we say that  $X$  is uniformly distributed on  $\sqrt{n} S^{n-1}$  if, for every (Borel) subset  $E \subset S^{n-1}$ , the probability  $\mathbb{P}\{X \in E\}$  equals the ratio of the  $(n-1)$ -dimensional areas of  $E$  and  $S^{n-1}$ .

### 3.3.2 Multivariate normal

One of the most important high-dimensional distributions is Gaussian, or multivariate normal. From a basic probability course we know that a random vector  $g = (g_1, \dots, g_n)$  has the *standard normal distribution* in  $\mathbb{R}^n$ , denoted

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

if the coordinates  $g_i$  are independent standard normal random variables  $N(0, 1)$ . The density of  $Z$  is then the product of the  $n$  standard normal densities (1.6), which is

$$f(x) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} e^{-x_i^2/2} = \frac{1}{(2\pi)^{n/2}} e^{-\|x\|_2^2/2}, \quad x \in \mathbb{R}^n. \quad (3.4)$$

The standard normal distribution is isotropic. (Why?)

Note that the standard normal density (3.4) is *rotation invariant*, since  $f(x)$  depends only on the length but not the direction of  $x$ . We can equivalently express this observation as follows:

**Proposition 3.3.2** (Rotation invariance). *Consider a random vector  $g \sim N(0, I_n)$  and a fixed orthogonal matrix  $U$ . Then*

$$Ug \sim N(0, I_n).$$

**Exercise 3.3.3** (Rotation invariance). 🍷🍷 Deduce the following properties from the rotation invariance of the normal distribution.

- (a) Consider a random vector  $g \sim N(0, I_n)$  and a fixed vector  $u \in \mathbb{R}^n$ . Then

$$\langle g, u \rangle \sim N(0, \|u\|_2^2).$$

- (b) Consider independent random variables  $X_i \sim N(0, \sigma_i^2)$ . Then

$$\sum_{i=1}^n X_i \sim N(0, \sigma^2) \quad \text{where} \quad \sigma^2 = \sum_{i=1}^n \sigma_i^2.$$

- (c) Let  $G$  be an  $m \times n$  Gaussian random matrix, i.e. the entries of  $G$  are independent  $N(0, 1)$  random variables. Let  $u \in \mathbb{R}^n$  be a fixed unit vector. Then

$$Gu \sim N(0, I_m).$$

Let us also recall the notion of the *general* normal distribution  $N(\mu, \Sigma)$ . 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 a general normal distribution in  $\mathbb{R}^n$ , denoted

$$X \sim N(\mu, \Sigma).$$

Summarizing, we have  $X \sim N(\mu, \Sigma)$  if and only if

$$Z := \Sigma^{-1/2}(X - \mu) \sim N(0, I_n).$$

The density of  $X \sim N(\mu, \Sigma)$  can be computed by the change of variables formula, and it equals

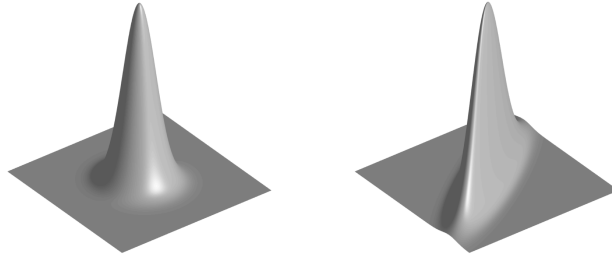
$$f_X(x) = \frac{1}{(2\pi)^{n/2} \det(\Sigma)^{1/2}} e^{-(x-\mu)^\top \Sigma^{-1} (x-\mu)/2}, \quad x \in \mathbb{R}^n. \quad (3.5)$$

Figure 3.5 shows examples of two densities of multivariate normal distributions.

An important observation is that the coordinates of a random vector  $X \sim N(\mu, \Sigma)$  are independent if and only if they are uncorrelated. (In this case  $\Sigma = I_n$ .)

**Exercise 3.3.4** (Characterization of normal distribution). ☹️☹️☹️ Let  $X$  be a random vector in  $\mathbb{R}^n$ . Show that  $X$  has a multivariate normal distribution if and only if every one-dimensional marginal  $\langle X, \theta \rangle$ ,  $\theta \in \mathbb{R}^n$ , has a (univariate) normal distribution.

**Hint:** Utilize a version of Cramér-Wold's theorem, which states that the totality of the distributions of one-dimensional marginals determine the distribution in  $\mathbb{R}^n$  uniquely. More precisely, if  $X$  and  $Y$  are random vectors in  $\mathbb{R}^n$  such that  $\langle X, \theta \rangle$  and  $\langle Y, \theta \rangle$  have the same distribution for each  $\theta \in \mathbb{R}^n$ , then  $X$  and  $Y$  have the same distribution.



**Figure 3.5** The densities of the isotropic distribution  $N(0, I_2)$  and a non-isotropic distribution  $N(0, \Sigma)$ .

**Exercise 3.3.5.** ☹️ Let  $X \sim N(0, I_n)$ .

- (a) Show that, for any fixed vectors  $u, v \in \mathbb{R}^n$ , we have

$$\mathbb{E} \langle X, u \rangle \langle X, v \rangle = \langle u, v \rangle. \quad (3.6)$$

- (b) Given a vector  $u \in \mathbb{R}^n$ , consider the random variable  $X_u := \langle X, u \rangle$ . From Exercise 3.3.3 we know that  $X_u \sim N(0, \|u\|_2^2)$ . Check that

$$\|X_u - X_v\|_{L^2} = \|u - v\|_2$$

for any fixed vectors  $u, v \in \mathbb{R}^n$ . (Here  $\|\cdot\|_{L^2}$  denotes the norm in the Hilbert space  $L^2$  of random variables, which we introduced in (1.1).)

**Exercise 3.3.6.** ☹️ Let  $G$  be an  $m \times n$  Gaussian random matrix, i.e. the entries of  $G$  are independent  $N(0, 1)$  random variables. Let  $u, v \in \mathbb{R}^n$  be unit orthogonal vectors. Prove that  $Gu$  and  $Gv$  are independent  $N(0, I_m)$  random vectors.

**Hint:** Reduce the problem to the case where  $u$  and  $v$  are collinear with canonical basis vectors of  $\mathbb{R}^n$ .




### 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 to 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 (3.3) for the norm of  $g \sim N(0, I_n)$  states that

$$\mathbb{P} \left\{ \left| \|g\|_2 - \sqrt{n} \right| \geq t \right\} \leq 2 \exp(-ct^2) \quad \text{for all } t \geq 0. \quad (3.7)$$

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

**Exercise 3.3.7** (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  $g$ . Prove the following:

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

Concentration inequality (3.7) says that  $r = \|g\|_2 \approx \sqrt{n}$  with high probability, so

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

In other words, the standard normal distribution in high dimensions is close to the uniform distribution on the sphere of radius  $\sqrt{n}$ , i.e.

$$N(0, I_n) \approx \text{Unif} \left( \sqrt{n} S^{n-1} \right). \quad (3.8)$$

Figure 3.6 illustrates this fact that goes against our intuition that has been trained in low dimensions.

### 3.3.4 Frames

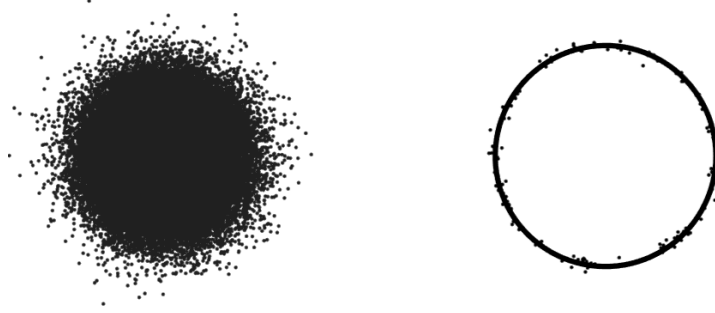
For an example of an extremely discrete distribution, consider a *coordinate random vector*  $X$  uniformly distributed in the set  $\{\sqrt{n} e_i\}_{i=1}^n$  where  $\{e_i\}_{i=1}^n$  is the canonical basis of  $\mathbb{R}^n$ :

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

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 the area of signal processing under the name of *frames*.



**Figure 3.6** A Gaussian point cloud in two dimensions (left) and its intuitive visualization in high dimensions (right). In high dimensions, the standard normal distribution is very close to the uniform distribution on the sphere of radius  $\sqrt{n}$ .

**Definition 3.3.8.** 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  $\{u_i\}_{i=1}^N$  is called a *tight frame*.

**Exercise 3.3.9.** ☕ Show that  $\{u_i\}_{i=1}^N$  is a tight frame in  $\mathbb{R}^n$  with bound  $A$  if and only if

$$\sum_{i=1}^N u_i u_i^\top = AI_n. \quad (3.9)$$

**Hint:** Proceed similarly to the proof of Lemma 3.2.3.

Multiplying both sides of (3.9) by a vector  $x$ , we see that

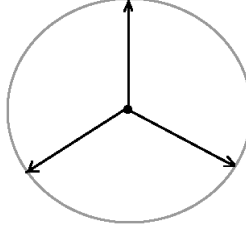
$$\sum_{i=1}^N \langle u_i, x \rangle u_i = Ax \quad \text{for any } x \in \mathbb{R}^n. \quad (3.10)$$

This is a *frame expansion* of a vector  $x$ , and it should look familiar. Indeed, if  $\{u_i\}$  is an orthonormal basis, then (3.10) 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 “Mercedes-Benz frame”, a set of three equidistant points on a circle in  $\mathbb{R}^2$  shown on Figure 3.7.

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

**Lemma 3.3.10** (Tight frames and isotropic distributions). (a) Consider a tight



**Figure 3.7** the Mercedes-Benz frame. A set of equidistant points on the circle form a tight frame in  $\mathbb{R}^2$ .

frame  $\{u_i\}_{i=1}^N$  in  $\mathbb{R}^n$  with frame bounds  $A = B$ . Let  $X$  be a random vector that is uniformly distributed in the set of frame elements, i.e.

$$X \sim \text{Unif} \{u_i : i = 1, \dots, N\}.$$

Then  $(N/A)^{1/2}X$  is an isotropic random vector in  $\mathbb{R}^n$ .

(b) Consider an isotropic random vector  $X$  in  $\mathbb{R}^n$  that takes a finite set of values  $x_i$  with probabilities  $p_i$  each,  $i = 1, \dots, N$ . Then the vectors

$$u_i := \sqrt{p_i} x_i, \quad i = 1, \dots, N,$$

form a tight frame in  $\mathbb{R}^n$  with bounds  $A = B = 1$ .

*Proof* 1. Without loss of generality, we can assume that  $A = N$ . (Why?) The assumptions and (3.9) imply that

$$\sum_{i=1}^N u_i u_i^\top = N I_n.$$

Dividing both sides by  $N$  and interpreting  $\frac{1}{N} \sum_{i=1}^N$  as an expectation, we conclude that  $X$  is isotropic.

2. Isotropy of  $X$  means that

$$\mathbb{E} X X^\top = \sum_{i=1}^N p_i x_i x_i^\top = I_n.$$

Denoting  $u_i := \sqrt{p_i} x_i$ , we obtain (3.9) with  $A = 1$ . □

### 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  $K$ :

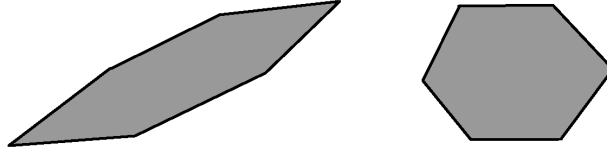
$$X \sim \text{Unif}(K).$$

Assume that  $\mathbb{E}X = 0$  (translate  $K$  appropriately to achieve this) and 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 can think of the isotropic convex body  $TK$  as a *well conditioned* version of  $K$ , with  $T$  playing the role of a pre-conditioner, see Figure 3.8. Algorithms related to convex bodies  $K$  (such as computing the volume of  $K$ ) tend to work better for well-conditioned bodies  $K$ .



**Figure 3.8** 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, which we introduced in Section 2.5, can be extended to higher dimensions. To see how, recall from Exercise 3.3.4 that the multivariate normal distribution can be characterized through its *one-dimensional marginals*, or projections onto lines: a random vector  $X$  has a 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, it is natural to 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*

$X = (X_1, \dots, X_n) \in \mathbb{R}^n$  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 a sum of independent sub-gaussian random variables is sub-gaussian, which we proved in Proposition 2.6.1. Indeed, for a fixed unit vector  $x = (x_1, \dots, x_n) \in S^{n-1}$  we have

$$\begin{aligned} \|\langle X, x \rangle\|_{\psi_2}^2 &= \left\| \sum_{i=1}^n x_i X_i \right\|_{\psi_2}^2 \leq C \sum_{i=1}^n x_i^2 \|X_i\|_{\psi_2}^2 \quad (\text{by Proposition 2.6.1}) \\ &\leq C \max_{i \leq n} \|X_i\|_{\psi_2}^2 \quad (\text{using that } \sum_{i=1}^n x_i^2 = 1). \end{aligned}$$

This completes the proof.  $\square$

**Exercise 3.4.3.** ☕☕ This exercise clarifies the role of independence of coordinates in Lemma 3.4.2.

1. Let  $X = (X_1, \dots, X_n) \in \mathbb{R}^n$  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 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 that we introduced in Section 3.3.1. A random vector  $X$  with this distribution has independent, symmetric Bernoulli coordinates, so 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  $\{\sqrt{n}e_i : i = 1, \dots, n\}$ , where  $e_i$  denotes the 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.

**Exercise 3.4.4.** ☹ Show that

$$\|X\|_{\psi_2} \asymp \sqrt{\frac{n}{\log n}}.$$

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 3.4.5.** ☹☹☹☹ 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 3.3.1. We will show that it is sub-gaussian by reducing it to the Gaussian distribution  $N(0, I_n)$ .

**Theorem 3.4.6** (Uniform distribution on the 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  $\sqrt{n}$ :*

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

*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 Exercise 3.3.7, the direction  $g/\|g\|_2$  is uniformly distributed on the unit sphere  $S^{n-1}$ . Thus, 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 sub-gaussian.

By rotation invariance, we may assume that  $x = (1, 0, \dots, 0)$ , in which case  $\langle X, x \rangle = X_1$ , the first coordinate of  $X$ . We want to bound the tail probability

$$p(t) := \mathbb{P} \{ |X_1| \geq t \} = \mathbb{P} \left\{ \frac{|g_1|}{\|g\|_2} \geq \frac{t}{\sqrt{n}} \right\}.$$

The concentration of norm (Theorem 3.1.1) implies that

$$\|g\|_2 \approx \sqrt{n} \quad \text{with high probability.}$$

This reduces the problem to bounding  $\mathbb{P} \{ |g_1| \geq t \}$ , but as we know from (2.3), this tail is sub-gaussian.

Let us do this argument more carefully. Theorem 3.1.1 implies that

$$\left\| \|g\|_2 - \sqrt{n} \right\|_{\psi_2} \leq C.$$

Thus the event

$$\mathcal{E} := \left\{ \|g\|_2 \geq \frac{\sqrt{n}}{2} \right\}$$

is likely: by (2.14) its complement  $\mathcal{E}^c$  has probability

$$\mathbb{P}(\mathcal{E}^c) \leq 2 \exp(-cn). \quad (3.11)$$


Then the tail probability can be bounded as follows:

$$\begin{aligned} p(t) &\leq \mathbb{P} \left\{ \frac{|g_1|}{\|g\|_2} \geq \frac{t}{\sqrt{n}} \text{ and } \mathcal{E} \right\} + \mathbb{P}(\mathcal{E}^c) \\ &\leq \mathbb{P} \left\{ |g_1| \geq \frac{t}{2} \text{ and } \mathcal{E} \right\} + 2 \exp(-cn) \quad (\text{by definition of } \mathcal{E} \text{ and (3.11)}) \\ &\leq 2 \exp(-t^2/8) + 2 \exp(-cn) \quad (\text{drop } \mathcal{E} \text{ and use (2.3)}). \end{aligned}$$

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 by the characterization of sub-gaussian distributions (recall Proposition 2.5.2 and Remark 2.5.3).  $\square$

**Exercise 3.4.7** (Uniform distribution on the Euclidean ball).  Extend Theorem 3.4.6 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

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

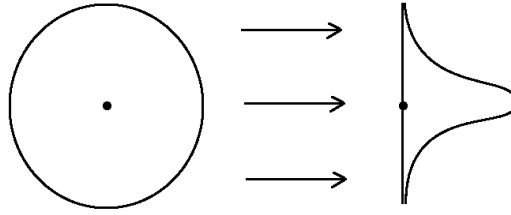
is sub-gaussian, and

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

**Remark 3.4.8** (Projective limit theorem). Theorem 3.4.6 should be compared to the so-called projective central limit theorem. It states that the marginals of the uniform distribution on the sphere become asymptotically normal as  $n$  increases, see Figure 3.9. Precisely, if  $X \sim \text{Unif}(\sqrt{n} S^{n-1})$  then for any fixed unit vector  $x$  we have

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

Thus we can view Theorem 3.4.6 as a concentration version of the Projective Limit Theorem, in the same sense as Hoeffding's inequality in Section 2.2 is a concentration version of the classical central limit theorem.



**Figure 3.9** 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 3.4.7) and the unit cube  $[-1, 1]^n$  (according to Lemma 3.4.2). For some other bodies, this is not true:

**Exercise 3.4.9.** ☕☕☕ Consider a ball of the  $\ell_1$  norm in  $\mathbb{R}^n$ :

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

- Show that the uniform distribution on  $K$  is isotropic for some  $r \asymp n$ .
- Show that the subgaussian norm of this distribution is *not* bounded by an absolute constant as the dimension  $n$  grows.

Nevertheless, it is possible to prove a weaker result 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 [81, Section 2.2.b<sub>3</sub>].

**Exercise 3.4.10.** ☹☹ Show that 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.

### 3.5 Application: Grothendieck's inequality and semidefinite programming

In this and the next section, we use high-dimensional Gaussian distributions to pursue some problems that have seemingly nothing to do with probability. Here we give a probabilistic proof of Grothendieck's inequality, a remarkable result which we will use later in the analysis of some computationally hard problems.

**Theorem 3.5.1** (Grothendieck's inequality). *Consider an  $m \times n$  matrix  $(a_{ij})$  of real numbers. Assume that, for any numbers  $x_i, y_j \in \{-1, 1\}$ , we have*

$$\left| \sum_{i,j} a_{ij} x_i y_j \right| \leq 1.$$

*Then, for any Hilbert space  $H$  and any vectors  $u_i, v_j \in H$  satisfying  $\|u_i\| = \|v_j\| = 1$ , we have*

$$\left| \sum_{i,j} a_{ij} \langle u_i, v_j \rangle \right| \leq K,$$

*where  $K \leq 1.783$  is an absolute constant.*

There is apparently nothing random in the statement of this theorem, but our proof of this result will be probabilistic. We will actually give two proofs of Grothendieck's inequality. The one given in this section will yield a much worse bound on the constant  $K$ , namely  $K \leq 288$ . In Section 3.7, we present an alternative argument that yields the bound  $K \leq 1.783$  as stated in Theorem 3.5.1.

Before we pass to the argument, let us make one simple observation.

**Exercise 3.5.2.** ☹

- (a) Check that the assumption of Grothendieck's inequality can be equivalently stated as follows:

$$\left| \sum_{i,j} a_{ij} x_i y_j \right| \leq \max_i |x_i| \cdot \max_j |y_j|. \quad (3.12)$$

for any real numbers  $x_i$  and  $y_j$ .

- (b) Show that the conclusion of Grothendieck's inequality can be equivalently stated as follows:

$$\left| \sum_{i,j} a_{ij} \langle u_i, v_j \rangle \right| \leq K \max_i \|u_i\| \cdot \max_j \|v_j\| \quad (3.13)$$

for any Hilbert space  $H$  and any vectors  $u_i, v_j \in H$ .

*Proof of Theorem 3.5.1 with  $K \leq 288$ .* **Step 1: Reductions.** Note that Grothendieck's inequality becomes trivial if we allow the value of  $K$  depend on the matrix  $A = (a_{ij})$ . (For example,  $K = \sum_{i,j} |a_{ij}|$  would work – check!) Let us choose  $K = K(A)$  to be the *smallest* number that makes the conclusion (3.13) valid for a given matrix  $A$  and any Hilbert space  $H$  and any vectors  $u_i, v_j \in H$ . Our goal is to show that  $K$  does *not* depend on the matrix  $A$  or the dimensions  $m$  and  $n$ .

Without loss of generality,<sup>4</sup> we may do this for a specific Hilbert space  $H$ , namely for  $\mathbb{R}^N$  equipped with the Euclidean norm  $\|\cdot\|_2$ . Let us fix vectors  $u_i, v_j \in \mathbb{R}^N$  which realize the smallest  $K$ , that is

$$\sum_{i,j} a_{ij} \langle u_i, v_j \rangle = K, \quad \|u_i\|_2 = \|v_j\|_2 = 1.$$

**Step 2: Introducing randomness.** The main idea of the proof is to realize the vectors  $u_i, v_j$  via Gaussian random variables

$$U_i := \langle g, u_i \rangle \quad \text{and} \quad V_j := \langle g, v_j \rangle, \quad \text{where } g \sim N(0, I_N).$$

As we noted in Exercise 3.3.5,  $U_i$  and  $V_j$  are standard normal random variables whose correlations follow exactly the inner products of the vectors  $u_i$  and  $v_j$ :

$$\mathbb{E} U_i V_j = \langle u_i, v_j \rangle.$$

Thus

$$K = \sum_{i,j} a_{ij} \langle u_i, v_j \rangle = \mathbb{E} \sum_{i,j} a_{ij} U_i V_j. \quad (3.14)$$

Assume for a moment that the random variables  $U_i$  and  $V_j$  were bounded almost surely by some constant – say, by  $R$ . Then the assumption (3.12) of Grothendieck's inequality (after rescaling) would yield  $\left| \sum_{i,j} a_{ij} U_i V_j \right| \leq R^2$  almost surely, and (3.14) would then give  $K \leq R^2$ .

**Step 3: Truncation.** Of course, this reasoning is flawed: the random variables  $U_i, V_j \sim N(0, 1)$  are not bounded almost surely. To fix this argument, we can utilize a useful *truncation* trick. Let us fix some level  $R \geq 1$  and decompose the random variables as follows:

$$U_i = U_i^- + U_i^+ \quad \text{where} \quad U_i^- = U_i \mathbf{1}_{\{|U_i| \leq R\}} \quad \text{and} \quad U_i^+ = U_i \mathbf{1}_{\{|U_i| > R\}}.$$

We similarly decompose  $V_j = V_j^- + V_j^+$ . Now  $U_i^-$  and  $V_j^-$  are bounded by  $R$  almost surely as we desired. The remainder terms  $U_i^+$  and  $V_j^+$  are small in the  $L^2$  norm: indeed, the bound in Exercise 2.1.4 gives

$$\|U_i^+\|_{L^2}^2 \leq 2 \left( R + \frac{1}{R} \right) \frac{1}{\sqrt{2\pi}} e^{-R^2/2} < \frac{4}{R^2}, \quad (3.15)$$

<sup>4</sup> To see this, we can first trivially replace  $H$  with the subspace of  $H$  spanned by the vectors  $u_i$  and  $v_j$  (and with the norm inherited from  $H$ ). This subspace has dimension at most  $N := m + n$ . Next, we recall the basic fact that all  $N$ -dimensional Hilbert spaces are isometric with each other, and in particular they are isometric to  $\mathbb{R}^N$  with the norm  $\|\cdot\|_2$ . The isometry can be constructed by identifying orthogonal bases of those spaces.

and similarly for  $V_j^+$ .

**Step 4: Breaking up the sum.** The sum in (3.14) becomes

$$K = \mathbb{E} \sum_{i,j} a_{ij} (U_i^- + U_i^+) (V_j^- + V_j^+).$$

When we expand the product in each term we obtain four sums, which we proceed to bound individually. The first sum,

$$S_1 := \mathbb{E} \sum_{i,j} a_{ij} U_i^- V_j^-,$$

is the best of all. By construction, the random variables  $U_i^-$  and  $V_j^-$  are bounded almost surely by  $R$ . Thus, just like we explained above, we can use the assumption (3.12) of Grothendieck's inequality to get  $S_1 \leq R^2$ .

We are not able to use the same reasoning for the second sum,

$$S_2 := \mathbb{E} \sum_{i,j} a_{ij} U_i^+ V_j^-,$$

since the random variable  $U_i^+$  is unbounded. Instead, we will view the random variables  $U_i^+$  and  $V_j^-$  as elements of the Hilbert space  $L^2$  with the inner product  $\langle X, Y \rangle_{L^2} = \mathbb{E} XY$ . The second sum becomes

$$S_2 = \sum_{i,j} a_{ij} \langle U_i^+, V_j^- \rangle_{L^2}. \quad (3.16)$$

Recall from (3.15) that  $\|U_i^+\|_{L^2} < 2/R$  and  $\|V_j^-\|_{L^2} \leq \|V_j\|_{L^2} = 1$  by construction. Then, applying the conclusion (3.13) of Grothendieck's inequality for the Hilbert space  $H = L^2$ , we find that<sup>5</sup>


$$S_2 \leq K \cdot \frac{2}{R}.$$

The third and fourth sums,  $S_3 := \mathbb{E} \sum_{i,j} a_{ij} U_i^- V_j^+$  and  $S_4 := \mathbb{E} \sum_{i,j} a_{ij} U_i^+ V_j^+$ , can be both bounded just like  $S_2$ . (Check!)

**Step 5: Putting everything together.** Putting the four sums together, we conclude from (3.14) that

$$K \leq R^2 + \frac{6K}{R}.$$

Choosing  $R = 12$  (for example) and solve the resulting inequality, we obtain  $K \leq 288$ . The theorem is proved.  $\square$

**Exercise 3.5.3** (Symmetric matrices,  $x_i = y_i$ ).  Deduce the following version of Grothendieck's inequality for symmetric  $n \times n$  matrices  $A = (a_{ij})$  with

<sup>5</sup> It might seem weird that we are able to apply the inequality that we are trying to prove. Remember, however, that we chose  $K$  in the beginning of the proof as the best number that makes Grothendieck's inequality valid. This is the  $K$  we are using here.

real entries. Suppose that  $A$  is either positive semidefinite or has zero diagonal. Assume that, for any numbers  $x_i \in \{-1, 1\}$ , we have

$$\left| \sum_{i,j} a_{ij} x_i x_j \right| \leq 1.$$

Then, for any Hilbert space  $H$  and any vectors  $u_i, v_j \in H$  satisfying  $\|u_i\| = \|v_j\| = 1$ , we have

$$\left| \sum_{i,j} a_{ij} \langle u_i, v_j \rangle \right| \leq 2K, \quad (3.17)$$

where  $K$  is the absolute constant from Grothendieck's inequality.

**Hint:** Check and use the polarization identity  $\langle Ax, y \rangle = \langle Au, u \rangle - \langle Av, v \rangle$  where  $u = (x + y)/2$  and  $v = (x - y)/2$ .

### 3.5.1 Semidefinite programming

One application area where Grothendieck's inequality can be particularly helpful is the analysis of certain computationally hard problems. A powerful approach to such problems is to try and *relax* them to computationally simpler and more tractable problems. This is often done using semidefinite programming, with Grothendieck's inequality guaranteeing the quality of such relaxations.

**Definition 3.5.4.** A *semidefinite program* is an optimization problem of the following type:

$$\text{maximize } \langle A, X \rangle : \quad X \succeq 0, \quad \langle B_i, X \rangle = b_i \text{ for } i = 1, \dots, m. \quad (3.18)$$

Here  $A$  and  $B_i$  are given  $n \times n$  matrices and  $b_i$  are given real numbers. The running “variable”  $X$  is an  $n \times n$  symmetric positive semidefinite matrix, indicated by the notation  $X \succeq 0$ . The inner product

$$\langle A, X \rangle = \text{tr}(A^T X) = \sum_{i,j=1}^n A_{ij} X_{ij} \quad (3.19)$$

is the canonical inner product on the space of  $n \times n$  matrices.

Note in passing that if we *minimize* instead of maximize in (3.18), we still get a semidefinite program. (To see this, replace  $A$  with  $-A$ .)

Every semidefinite program is a *convex program*, which maximizes a linear function  $\langle A, X \rangle$  over a convex set of matrices. Indeed, the set of positive semidefinite matrices is convex (why?), and so is its intersection with the linear subspace defined by the constraints  $\langle B_i, X \rangle = b_i$ .

This is good news since convex programs are generally algorithmically tractable. There is a variety of computationally efficient solvers available for general convex programs, and for semidefinite programs (3.18) in particular, for example interior point methods.

## Semidefinite relaxations

Semidefinite programs can be designed to provide computationally efficient relaxations of computationally hard problems, such as this one:

$$\text{maximize } \sum_{i,j=1}^n A_{ij}x_i x_j : \quad x_i = \pm 1 \text{ for } i = 1, \dots, n \quad (3.20)$$

where  $A$  is a given  $n \times n$  symmetric matrix. This is an *integer optimization problem*. The feasible set consists of  $2^n$  vectors  $x = (x_i) \in \{-1, 1\}^n$ , so finding the maximum by exhaustive search would take exponential time. Is there a smarter way to solve the problem? This is not likely: the problem (3.20) is known to be computationally hard in general (NP-hard).

Nonetheless, we can “relax” the problem (3.20) to a semidefinite program that can compute the maximum *approximately*, up to a constant factor. To formulate such a relaxation, let us replace in (3.20) the numbers  $x_i = \pm 1$  by their higher-dimensional analogs – unit vectors  $X_i$  in  $\mathbb{R}^n$ . Thus we consider the following optimization problem:

$$\text{maximize } \sum_{i,j=1}^n A_{ij} \langle X_i, X_j \rangle : \quad \|X_i\|_2 = 1 \text{ for } i = 1, \dots, n. \quad (3.21)$$

**Exercise 3.5.5.** ☕☕ Show that the optimization (3.21) is equivalent to the following semidefinite program:

$$\text{maximize } \langle A, X \rangle : \quad X \succeq 0, \quad X_{ii} = 1 \text{ for } i = 1, \dots, n. \quad (3.22)$$

**Hint:** Consider the *Gram matrix* of the vectors  $X_i$ , which is the  $n \times n$  matrix with entries  $\langle X_i, X_j \rangle$ . Do not forget to describe how to translate a solution of (3.22) into a solution of (3.21).

## The guarantee of relaxation

We now see how Grothendieck’s inequality guarantees the accuracy of semidefinite relaxations: the semidefinite program (3.21) approximates the maximum value in the integer optimization problem (3.20) up to an absolute constant factor.

**Theorem 3.5.6.** *Consider an  $n \times n$  symmetric, positive semidefinite matrix  $A$ . Let  $\text{INT}(A)$  denote the maximum in the integer optimization problem (3.20) and  $\text{SDP}(A)$  denote the maximum in the semidefinite problem (3.21). Then*

$$\text{INT}(A) \leq \text{SDP}(A) \leq 2K \cdot \text{INT}(A)$$

where  $K \leq 1.783$  is the constant in Grothendieck’s inequality.

*Proof* The first bound follows with  $X_i = (x_i, 0, 0, \dots, 0)^\top$ . The second bound follows from Grothendieck’s inequality for symmetric matrices in Exercise 3.5.3. (Argue that one can drop absolute values.)  $\square$

Although Theorem 3.5.6 allows us to approximate the maximum value in (3.20), it is not obvious how to compute  $x_i$ ’s that attain this approximate value. Can we translate the vectors  $(X_i)$  that give a solution of the semidefinite program

(3.21) into labels  $x_i = \pm 1$  that approximately solve (3.20)? In the next section, we illustrate this on the example of a remarkable NP-hard problem on graphs – the maximum cut problem.

**Exercise 3.5.7.** ☕☕☕ Let  $A$  be an  $m \times n$  matrix. Consider the optimization problem

$$\text{maximize } \sum_{i,j} A_{ij} \langle X_i, Y_j \rangle : \quad \|X_i\|_2 = \|Y_j\|_2 = 1 \text{ for all } i, j$$

over  $X_i, Y_j \in \mathbb{R}^k$  and  $k \in \mathbb{N}$ . Formulate this problem as a semidefinite program.

**Hint:** First, express the objective function as  $\frac{1}{2} \text{tr}(\tilde{A}ZZ^T)$ , where  $\tilde{A} = \begin{bmatrix} 0 & A \\ A^T & 0 \end{bmatrix}$ ,  $Z = \begin{bmatrix} X \\ Y \end{bmatrix}$  and  $X$  and  $Y$  are the matrices with rows  $X_i^T$  and  $Y_j^T$ , respectively. Then express the set of matrices of the type  $ZZ^T$  with unit rows as the set of symmetric positive semidefinite matrices whose diagonal entries equal 1.

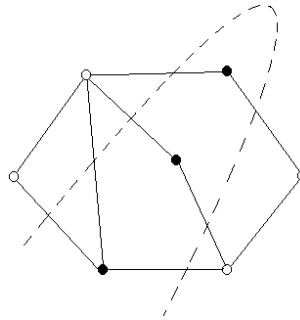
### 3.6 Application: Maximum cut for graphs

We now illustrate the utility of semidefinite relaxations for the problem of finding the *maximum cut* of a graph, which is one of the well known NP-hard problems discussed in the computer science literature.

#### 3.6.1 Graphs and cuts

An undirected *graph*  $G = (V, E)$  is defined as a set  $V$  of vertices together with a set  $E$  of edges; each edge is an unordered pair of vertices. Here we consider finite, *simple* graphs – those with finitely many vertices and with no loops or multiple edges.

**Definition 3.6.1** (Maximum cut). Suppose we partition the set of vertices of a graph  $G$  into two disjoint sets. The *cut* is the number of edges crossing between these two sets. The maximum cut of  $G$ , denoted  $\text{MAX-CUT}(G)$ , is obtained by maximizing the cut over all partitions of vertices; see Figure 3.10 for illustration.



**Figure 3.10** The dashed line illustrates the maximum cut of this graph, obtained by partitioning the vertices into the black and white ones. Here  $\text{MAX-CUT}(G) = 7$ .

Computing the maximum cut of a given graph is known to be a computationally hard problem (NP-hard).

### 3.6.2 A simple 0.5-approximation algorithm

We try to relax the maximum cut problem to a semidefinite program following the method we introduced in Section 3.5.1. To do this, we need to translate the problem into the language of linear algebra.

**Definition 3.6.2** (Adjacency matrix). The *adjacency matrix*  $A$  of a graph  $G$  on  $n$  vertices is a symmetric  $n \times n$  matrix whose entries are defined as  $A_{ij} = 1$  if the vertices  $i$  and  $j$  are connected by an edge and  $A_{ij} = 0$  otherwise.

Let us label the vertices of  $G$  by the integers  $1, \dots, n$ . A partition of the vertices into two sets can be described using a vector of labels

$$x = (x_i) \in \{-1, 1\}^n,$$

the sign of  $x_i$  indicating which subset the vertex  $i$  belongs to. For example, the three black vertices in Figure 3.10 may have labels  $x_i = 1$ , and the four white vertices labels  $x_i = -1$ . The cut of  $G$  corresponding to the partition given by  $x$  is simply the number of edges between the vertices with labels of opposite signs, i.e.

$$\text{CUT}(G, x) = \frac{1}{2} \sum_{i,j: x_i x_j = -1} A_{ij} = \frac{1}{4} \sum_{i,j=1}^n A_{ij} (1 - x_i x_j). \quad (3.23)$$

(The factor  $\frac{1}{2}$  prevents double counting of edges  $(i, j)$  and  $(j, i)$ .) The maximum cut is then obtained by maximizing  $\text{CUT}(G, x)$  over all  $x$ , that is

$$\text{MAX-CUT}(G) = \frac{1}{4} \max \left\{ \sum_{i,j=1}^n A_{ij} (1 - x_i x_j) : x_i = \pm 1 \text{ for all } i \right\}. \quad (3.24)$$

Let us start with a simple 0.5-approximation algorithm for maximum cut – one which finds a cut with at least *half* of the edges of  $G$ .

**Proposition 3.6.3** (0.5-approximation algorithm for maximum cut). *Partition the vertices of  $G$  into two sets at random, uniformly over all  $2^n$  partitions. Then the expectation of the resulting cut equals*

$$0.5|E| \geq 0.5 \text{MAX-CUT}(G),$$

where  $|E|$  denotes the total number of edges of  $G$ .

*Proof* The random cut is generated by a symmetric Bernoulli random vector  $x \sim \text{Unif}(\{-1, 1\}^n)$ , which has independent symmetric Bernoulli coordinates. Then, in (3.23) we have  $\mathbb{E} x_i x_j = 0$  for  $i \neq j$  and  $A_{ij} = 0$  for  $i = j$  (since the graph has no loops). Thus, using linearity of expectation, we get

$$\mathbb{E} \text{CUT}(G, x) = \frac{1}{4} \sum_{i,j=1}^n A_{ij} = \frac{1}{2} |E|.$$

This completes the proof.  $\square$

**Exercise 3.6.4.** ☕☕ For any  $\varepsilon > 0$ , give an  $(0.5 - \varepsilon)$ -approximation algorithm for maximum cut, which is always *guaranteed* to give a suitable cut, but may have a random running time. Give a bound on the expected running time.

**Hint:** Consider cutting  $G$  repeatedly. Give a bound on the expected number of experiments.

### 3.6.3 Semidefinite relaxation

Now we will do much better and give a 0.878-approximation algorithm, which is due to Goemans and Williamson. It is based on a semidefinite relaxation of the NP-hard problem (3.24). It should be easy to guess what such relaxation could be: recalling (3.21), it is natural to consider the semidefinite problem

$$\text{SDP}(G) := \frac{1}{4} \max \left\{ \sum_{i,j=1}^n A_{ij}(1 - \langle X_i, X_j \rangle) : X_i \in \mathbb{R}^n, \|X_i\|_2 = 1 \text{ for all } i \right\}. \quad (3.25)$$

(Again – why is this a semidefinite program?)

As we will see, not only the value  $\text{SDP}(G)$  approximates  $\text{MAX-CUT}(G)$  to within the 0.878 factor, but we can obtain an actual partition of  $G$  (i.e., the labels  $x_i$ ) which attains this value. To do this, we describe how to translate a solution  $(X_i)$  of (3.25) into labels  $x_i = \pm 1$ .

This can be done by the following *randomized rounding* step. Choose a random hyperplane in  $\mathbb{R}^n$  passing through the origin. It cuts the set of vectors  $X_i$  into two parts; let us assign labels  $x_i = 1$  to one part and  $x_i = -1$  to the other part. Equivalently, we may choose a standard normal random vector

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

and define

$$x_i := \text{sign} \langle X_i, g \rangle, \quad i = 1, \dots, n. \quad (3.26)$$

See Figure 3.11 for an illustration.<sup>6</sup>

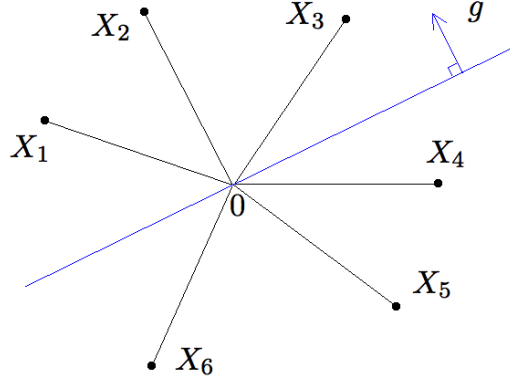
**Theorem 3.6.5** (0.878-approximation algorithm for maximum cut). *Let  $G$  be a graph with adjacency matrix  $A$ . Let  $x = (x_i)$  be the result of a randomized rounding of the solution  $(X_i)$  of the semidefinite program (3.25). Then*

$$\mathbb{E} \text{CUT}(G, x) \geq 0.878 \text{SDP}(G) \geq 0.878 \text{MAX-CUT}(G).$$

The proof of this theorem will be based on the following elementary identity. We can think of it as a more advanced version of the identity (3.6), which we used in the proof of Grothendieck's inequality, Theorem 3.5.1.

<sup>6</sup> In the rounding step, instead of the normal distribution we could use any other rotation invariant distribution in  $\mathbb{R}^n$ , for example the uniform distribution on the sphere  $S^{n-1}$ .





**Figure 3.11** Randomized rounding of vectors  $X_i \in \mathbb{R}^n$  into labels  $x_i = \pm 1$ . For this configuration of points  $X_i$  and a random hyperplane with normal vector  $g$ , we assign  $x_1 = x_2 = x_3 = 1$  and  $x_4 = x_5 = x_6 = -1$ .

**Lemma 3.6.6** (Grothendieck's identity). *Consider a random vector  $g \sim N(0, I_n)$ . Then, for any fixed vectors  $u, v \in S^{n-1}$ , we have*

$$\mathbb{E} \operatorname{sign} \langle g, u \rangle \operatorname{sign} \langle g, v \rangle = \frac{2}{\pi} \arcsin \langle u, v \rangle.$$

**Exercise 3.6.7.** ☕☕ Prove Grothendieck's identity.

**Hint:** It will quickly follow once you show that the probability that  $\langle g, u \rangle$  and  $\langle g, v \rangle$  have opposite signs equals  $\alpha/\pi$ , where  $\alpha \in [0, \pi]$  is the angle between the vectors  $u$  and  $v$ . To check this, use rotation invariance to reduce the problem to  $\mathbb{R}^2$ . Once on the plane, rotation invariance will give the result.

A weak point of Grothendieck's identity is the non-linear function  $\arcsin$ , which would be hard to work with. Let us replace it with a linear function using the numeric inequality

$$1 - \frac{2}{\pi} \arcsin t = \frac{2}{\pi} \arccos t \geq 0.878(1 - t), \quad t \in [-1, 1], \quad (3.27)$$

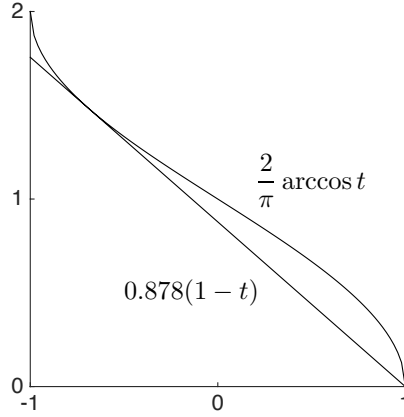
which can be easily verified using software; see Figure 3.12.

*Proof of Theorem 3.6.5* By (3.23) and linearity of expectation, we have

$$\mathbb{E} \operatorname{CUT}(G, x) = \frac{1}{4} \sum_{i,j=1}^n A_{ij} (1 - \mathbb{E} x_i x_j).$$

The definition of labels  $x_i$  in the rounding step (3.26) gives

$$\begin{aligned} 1 - \mathbb{E} x_i x_j &= 1 - \mathbb{E} \operatorname{sign} \langle X_i, g \rangle \operatorname{sign} \langle X_j, g \rangle \\ &= 1 - \frac{2}{\pi} \arcsin \langle X_i, X_j \rangle \quad (\text{by Grothendieck's identity, Lemma 3.6.6}) \\ &\geq 0.878(1 - \langle X_i, X_j \rangle) \quad (\text{by (3.27)}). \end{aligned}$$



**Figure 3.12** The inequality  $\frac{2}{\pi} \arccos t \geq 0.878(1-t)$  holds for all  $t \in [-1, 1]$ .

Therefore

$$\mathbb{E} \text{CUT}(G, x) \geq 0.878 \cdot \frac{1}{4} \sum_{i,j=1}^n A_{ij}(1 - \langle X_i, X_j \rangle) = 0.878 \text{SDP}(G).$$

This proves the first inequality in the theorem. The second inequality is trivial since  $\text{SDP}(G) \geq \text{MAX-CUT}(G)$ . (Why?)  $\square$

### 3.7 Kernel trick, and tightening of Grothendieck's inequality

Our proof of Grothendieck's inequality given in Section 3.5 yields a very loose bound on the absolute constant  $K$ . We now give an alternative proof that gives (almost) the best known constant  $K \leq 1.783$ .

Our new argument will be based on Grothendieck's identity (Lemma (3.6.6)). The main challenge in using this identity arises from the non-linearity of the function  $\arcsin(x)$ . Indeed, suppose there were no such nonlinearity, and we hypothetically had  $\mathbb{E} \text{sign} \langle g, u \rangle \text{sign} \langle g, v \rangle = \frac{2}{\pi} \langle u, v \rangle$ . Then Grothendieck's inequality would easily follow:

$$\frac{2}{\pi} \sum_{i,j} a_{ij} \langle u_i, v_j \rangle = \sum_{i,j} a_{ij} \mathbb{E} \text{sign} \langle g, u_i \rangle \text{sign} \langle g, v_j \rangle \leq 1,$$

where in the last step we swapped the sum and expectation and used the assumption of Grothendieck's inequality with  $x_i = \text{sign} \langle g, u_i \rangle$  and  $y_j = \text{sign} \langle g, v_j \rangle$ . This would give Grothendieck's inequality with  $K \leq \pi/2 \approx 1.57$ .

This argument is of course wrong. To address the non-linear form  $\frac{2}{\pi} \arcsin \langle u, v \rangle$  that appears in Grothendieck's identity, we use the following remarkably powerful trick: represent  $\frac{2}{\pi} \arcsin \langle u, v \rangle$  as the (linear) inner product  $\langle u', v' \rangle$  of some other

vectors  $u', v'$  in some Hilbert space  $H$ . In the literature on machine learning, this method is called the *kernel trick*.

We will explicitly construct the non-linear transformations  $u' = \Phi(u)$ ,  $v' = \Psi(v)$  that will do the job. Our construction is convenient to describe in the language of *tensors*, which are a higher dimensional generalization of the notion of matrices.

**Definition 3.7.1** (Tensors). A tensor can be described as a multidimensional array. Thus, a  $k$ -th order tensor  $(a_{i_1 \dots i_k})$  is a  $k$ -dimensional array of real numbers  $a_{i_1 \dots i_k}$ . The canonical inner product on  $\mathbb{R}^{n_1 \times \dots \times n_k}$  defines the inner product of tensors  $A = (a_{i_1 \dots i_k})$  and  $B = (b_{i_1 \dots i_k})$ :

$$\langle A, B \rangle := \sum_{i_1, \dots, i_k} a_{i_1 \dots i_k} b_{i_1 \dots i_k}. \quad (3.28)$$

**Example 3.7.2.** Scalars, vectors and matrices are examples of tensors. As we noted in (3.19), for  $m \times n$  matrices the inner product of tensors (3.28) specializes to

$$\langle A, B \rangle = \text{tr}(A^T B) = \sum_{i=1}^m \sum_{j=1}^n A_{ij} B_{ij}.$$

**Example 3.7.3** (Rank-one tensors). Every vector  $u \in \mathbb{R}^n$  defines the  $k$ -th order *tensor product*  $u \otimes \dots \otimes u$ , which is the tensor whose entries are the products of all  $k$ -tuples of the entries of  $u$ . In other words,

$$u \otimes \dots \otimes u = u^{\otimes k} := (u_{i_1} \dots u_{i_k}) \in \mathbb{R}^{n \times \dots \times n}.$$

In particular, for  $k = 2$ , the tensor product  $u \otimes u$  is just the  $n \times n$  matrix which is the outer product of  $u$  with itself:

$$u \otimes u = (u_i u_j)_{i,j=1}^n = uu^T.$$

One can similarly define the tensor products  $u \otimes v \otimes \dots \otimes z$  for different vectors  $u, v, \dots, z$ .

**Exercise 3.7.4.** ☞ Show that for any vectors  $u, v \in \mathbb{R}^n$  and  $k \in \mathbb{N}$ , we have

$$\langle u^{\otimes k}, v^{\otimes k} \rangle = \langle u, v \rangle^k.$$

This exercise shows a remarkable fact: we can represent non-linear forms like  $\langle u, v \rangle^k$  as the usual, *linear* inner product in some other space. Formally, there exist a Hilbert space  $H$  and a transformation  $\Phi : \mathbb{R}^n \rightarrow H$  such that

$$\langle \Phi(u), \Phi(v) \rangle = \langle u, v \rangle^k.$$

In this case,  $H$  is the space of  $k$ -th order tensors, and  $\Phi(u) = u^{\otimes k}$ .

In the next two exercises, we extend this observation to more general non-linearities.

**Exercise 3.7.5.** ☞☞

- (a) Show that there exist a Hilbert space  $H$  and a transformation  $\Phi : \mathbb{R}^n \rightarrow H$  such that

$$\langle \Phi(u), \Phi(v) \rangle = 2 \langle u, v \rangle^2 + 5 \langle u, v \rangle^3 \quad \text{for all } u, v \in \mathbb{R}^n.$$

**Hint:** Consider the cartesian product  $H = \mathbb{R}^{n \times n} \oplus \mathbb{R}^{n \times n \times n}$ .


- (b) More generally, consider a polynomial  $f : \mathbb{R} \rightarrow \mathbb{R}$  with non-negative coefficients, and construct  $H$  and  $\Phi$  such that

$$\langle \Phi(u), \Phi(v) \rangle = f(\langle u, v \rangle) \quad \text{for all } u, v \in \mathbb{R}^n.$$

- (c) Show the same for any *real analytic function*  $f : \mathbb{R} \rightarrow \mathbb{R}$  with non-negative coefficients, i.e. for any function that can be represented as a convergent series

$$f(x) = \sum_{k=0}^{\infty} a_k x^k, \quad x \in \mathbb{R}, \quad (3.29)$$

and such that  $a_k \geq 0$  for all  $k$ .

**Exercise 3.7.6.**  Let  $f : \mathbb{R} \rightarrow \mathbb{R}$  be any real analytic function (with possibly negative coefficients in (3.29)). Show that there exist a Hilbert space  $H$  and transformations  $\Phi, \Psi : \mathbb{R}^n \rightarrow H$  such that

$$\langle \Phi(u), \Psi(v) \rangle = f(\langle u, v \rangle) \quad \text{for all } u, v \in \mathbb{R}^n.$$

Moreover, check that

$$\|\Phi(u)\|^2 = \|\Psi(u)\|^2 = \sum_{k=0}^{\infty} |a_k| \|u\|_2^{2k}.$$

**Hint:** Construct  $\Phi$  as in Exercise 3.7.5 with  $\Phi$ , but include the signs of  $a_k$  in the definition of  $\Psi$ .

Let us specialize the kernel trick to the non-linearity  $\frac{2}{\pi} \arcsin \langle u, v \rangle$  that appears in Grothendieck's identity.

**Lemma 3.7.7.** *There exists a Hilbert space  $H$  and transformations<sup>7</sup>  $\Phi, \Psi : S^{n-1} \rightarrow S(H)$  such that*

$$\frac{2}{\pi} \arcsin \langle \Phi(u), \Psi(v) \rangle = \beta \langle u, v \rangle \quad \text{for all } u, v \in S^{n-1}, \quad (3.30)$$

where  $\beta = \frac{2}{\pi} \ln(1 + \sqrt{2})$ .

*Proof* Rewrite the desired identity (3.30) as

$$\langle \Phi(u), \Psi(v) \rangle = \sin \left( \frac{\beta\pi}{2} \langle u, v \rangle \right). \quad (3.31)$$

The result of Exercise 3.7.6 gives us the Hilbert space  $H$  and the maps  $\Phi, \Psi :$

<sup>7</sup> Here  $S(H)$  denotes the unit sphere of the Hilbert space  $H$ .

$\mathbb{R}^n \rightarrow H$  that satisfy (3.31). It only remains to determine the value of  $\beta$  for which  $\Phi$  and  $\Psi$  map unit vectors to unit vectors. To do this, we recall the Taylor series

$$\sin t = t - \frac{t^3}{3!} + \frac{t^5}{5!} - \dots \quad \text{and} \quad \sinh t = t + \frac{t^3}{3!} + \frac{t^5}{5!} + \dots$$

Exercise 3.7.6 then guarantees that for every  $u \in S^{n-1}$ , we have

$$\|\Phi(u)\|^2 = \|\Psi(u)\|^2 = \sinh\left(\frac{\beta\pi}{2}\right).$$

This quantity equals 1 if we set

$$\beta := \frac{2}{\pi} \operatorname{arcsinh}(1) = \frac{2}{\pi} \ln(1 + \sqrt{2}).$$

The lemma is proved.  $\square$

Now we are ready to prove Grothendieck's inequality (Theorem 3.5.1) with constant

$$K \leq \frac{1}{\beta} = \frac{\pi}{2 \ln(1 + \sqrt{2})} \approx 1.783.$$

*Proof of Theorem 3.5.1* We can assume without loss of generality that  $u_i, v_j \in S^{N-1}$  (this is the same reduction as we did in the proof in Section 3.5). Lemma 3.7.7 gives us unit vectors  $u'_i = \Phi(u_i)$  and  $v'_j = \Psi(v_j)$  in some Hilbert space  $H$ , which satisfy

$$\frac{2}{\pi} \arcsin \langle u'_i, v'_j \rangle = \beta \langle u_i, v_j \rangle \quad \text{for all } i, j.$$

We can again assume without loss of generality that  $H = \mathbb{R}^M$  for some  $M$ . (Why?) Then

$$\begin{aligned} \beta \sum_{i,j} a_{ij} \langle u_i, v_j \rangle &= \sum_{i,j} a_{ij} \cdot \frac{2}{\pi} \arcsin \langle u'_i, v'_j \rangle \\ &= \sum_{i,j} a_{ij} \mathbb{E} \operatorname{sign} \langle g, u'_i \rangle \operatorname{sign} \langle g, v'_j \rangle \quad (\text{by Lemma 3.6.6}), \\ &\leq 1, \end{aligned}$$

where in the last step we swapped the sum and expectation and used the assumption of Grothendieck's inequality with  $x_i = \operatorname{sign} \langle g, u'_i \rangle$  and  $y_j = \operatorname{sign} \langle g, v'_j \rangle$ . This yields the conclusion of Grothendieck's inequality for  $K \leq 1/\beta$ .  $\square$

### 3.7.1 Kernels and feature maps

Since the kernel trick was so successful in the proof of Grothendieck's inequality, we may ask – what other non-linearities can be handled with the kernel trick? Let

$$K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$$

be a function of two variables on a set  $\mathcal{X}$ . Under what conditions on  $K$  can we find a Hilbert space  $H$  and a transformation

$$\Phi : \mathcal{X} \rightarrow H$$

so that

$$\langle \Phi(u), \Phi(v) \rangle = K(u, v) \quad \text{for all } u, v \in \mathcal{X} \quad (3.32)$$

The answer to this question is provided by Mercer's and, more precisely, Moore-Aronszajn's theorems. The necessary and sufficient condition is that  $K$  be a *positive semidefinite kernel*, which means that for any finite collection of points  $u_1, \dots, u_N \in \mathcal{X}$ , the matrix

$$\left( K(u_i, u_j) \right)_{i,j=1}^N$$

is symmetric and positive semidefinite. The map  $\Phi$  is called a *feature map*, and the Hilbert space  $H$  can be constructed from the kernel  $K$  as a (unique) *reproducing kernel Hilbert space*.

Examples of positive semidefinite kernels on  $\mathbb{R}^n$  that are common in machine learning include the *Gaussian kernel* (also called the radial basis function kernel)

$$K(u, v) = \exp \left( - \frac{\|u - v\|_2^2}{2\sigma^2} \right), \quad u, v \in \mathbb{R}^n, \sigma > 0$$

and the *polynomial kernel*

$$K(u, v) = \left( \langle u, v \rangle + r \right)^k, \quad u, v \in \mathbb{R}^n, r > 0, k \in \mathbb{N}.$$

The kernel trick (3.32), which represents a general kernel  $K(u, v)$  as an inner product, is very popular in *machine learning*. It allows one to handle non-linear models (determined by kernels  $K$ ) by using methods developed for linear models. In contrast to what we did in this section, in machine learning applications the explicit description of the Hilbert space  $H$  and the feature map  $\Phi : \mathcal{X} \rightarrow H$  is typically not needed. Indeed, to compute the inner product  $\langle \Phi(u), \Phi(v) \rangle$  in  $H$ , one does not need to know  $\Phi$ : the identity (3.32) allows one to compute  $K(u, v)$  instead.

### 3.8 Notes

Theorem 3.1.1 about the concentration of the norm of random vectors is known but difficult to locate in the existing literature. We will later prove a more general result, Theorem 6.3.2, which is valid for anisotropic random vectors. It is unknown if the quadratic dependence on  $K$  in Theorem 3.1.1 is optimal. One may also wonder about concentration of the norm  $\|X\|_2$  of random vectors  $X$  whose coordinates are not necessarily independent. In particular, for a random vector  $X$  that is uniformly distributed in a convex set  $K$ , concentration of the norm is one of the central problems in geometric functional analysis; see [93, Section 2] and [36, Chapter 12].

Exercise 3.3.4 mentions Cramér-Wold’s theorem. It is a straightforward consequence of the uniqueness theorem for characteristic functions, see [23, Section 29].

The concept of frames introduced in Section 3.3.4 is an important extension of the notion of orthogonal bases. One can read more about frames and their applications in signal processing and data compression e.g. in [51, 121].

Sections 3.3.5 and 3.4.4 discuss random vectors uniformly distributed in convex sets. The books [11, 36] study this topic in detail, and the surveys [185, 218] discuss algorithmic aspects of computing the volume of convex sets in high dimensions.

Our discussion of sub-gaussian random vectors in Section 3.4. mostly follows [222]. An alternative geometric proof of Theorem 3.4.6 can be found in [13, Lemma 2.2].

Grothendieck’s inequality (Theorem 3.5.1) was originally proved by A. Grothendieck in 1953 [90] with bound on the constant  $K \leq \sinh(\pi/2) \approx 2.30$ ; a version of this original argument is presented [133, Section 2]. There is a number of alternative proofs of Grothendieck’s inequality with better and worse bounds on  $K$ ; see [35] for the history. The surveys [115, 168] discuss ramifications and applications of Grothendieck’s inequality in various areas of mathematics and computer science. Our first proof of Grothendieck’s inequality, the one given in Section 3.5, is similar to the one in [5, Section 8.1]; it was kindly brought to author’s attention by Mark Rudelson. Our second proof, the one from Section 3.7, is due to J.-L. Krivine [122]; versions of this argument can be found e.g. in [7] and [126]. The bound on the constant  $K \leq \frac{\pi}{2 \ln(1+\sqrt{2})} \approx 1.783$  that follows from Krivine’s argument is currently the best known *explicit* bound on  $K$ . It has been proved, however, that the best possible bound must be strictly smaller than Krivine’s bound, but no explicit number is known [35].

A part of this chapter is about semidefinite relaxations of hard optimization problems. For an introduction to the area of convex optimization, including semidefinite programming, we refer to the books [34, 39, 126, 29]. For the use of Grothendieck’s inequality in analyzing semidefinite relaxations, see [115, 7]. Our presentation of the maximum cut problem in Section 3.6 follows [39, Section 6.6] and [126, Chapter 7]. The semidefinite approach to maximum cut, which we discussed in Section 3.6.3, was pioneered in 1995 by M. Goemans and D. Williamson [83]. The approximation ratio  $\frac{2}{\pi} \min_{0 \leq \theta \leq \pi} \frac{\theta}{1 - \cos(\theta)} \approx 0.878$  guaranteed by the Goemans-Williamson algorithm remains the best known constant for the max-cut problem. If the Unique Games Conjecture is true, this ratio can not be improved, i.e. any better approximation would be NP-hard to compute [114].

In Section 3.7 we give Krivine’s proof of Grothendieck’s inequality [122]. We also briefly discuss kernel methods there. To learn more about kernels, reproducing kernel Hilbert spaces and their applications in machine learning, see e.g. the survey [102].

---

## Random matrices

We begin to study the non-asymptotic theory of random matrices, a study that will be continued in many further chapters. Section 4.1 is a quick reminder about singular values and matrix norms and their relationships. Section 4.2 introduces important geometric concepts – nets, covering and packing numbers, metric entropy, and discusses relations of these quantities with volume and coding. In Sections 4.4 and 4.6, we develop a basic  $\varepsilon$ -net argument and use it for random matrices. We first give a bound on the operator norm (Theorem 4.4.5) and then a stronger, two-sided bound on all singular values (Theorem 4.6.1) of random matrices. Three applications of random matrix theory are discussed in this chapter: a spectral clustering algorithm for recovering clusters, or communities, in complex networks (Section 4.5), covariance estimation (Section 4.7) and a spectral clustering algorithm for data presented as geometric point sets (Section 4.7.1).

### 4.1 Preliminaries on matrices

You should be familiar with the notion of singular value decomposition from a basic course in linear algebra; we recall it nevertheless. We will then introduce two matrix norms – operator and Frobenius, and discuss their relationships.

#### 4.1.1 Singular value decomposition

The main object of our study will be an  $m \times n$  matrix  $A$  with real entries. Recall that  $A$  can be represented using the *singular value decomposition* (SVD), which we can write as

$$A = \sum_{i=1}^r s_i u_i v_i^\top, \quad \text{where } r = \text{rank}(A). \quad (4.1)$$

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  $A$ .

For convenience, we often extend the sequence of singular values by setting  $s_i = 0$  for  $r < i \leq n$ , and we arrange them in a non-increasing order:

$$s_1 \geq s_2 \geq \cdots \geq s_n \geq 0.$$

The left singular vectors  $u_i$  are the orthonormal eigenvectors of  $AA^\top$  and the



right singular vectors  $v_i$  are the orthonormal eigenvectors of  $A^\top A$ . The singular values  $s_i$  are the square roots of the eigenvalues  $\lambda_i$  of both  $AA^\top$  and  $A^\top A$ :

$$s_i(A) = \sqrt{\lambda_i(AA^\top)} = \sqrt{\lambda_i(A^\top A)}.$$

In particular, if  $A$  is a *symmetric* matrix, then 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 eigenvectors of  $A$ .

Courant-Fisher's *min-max theorem* offers the following variational characterization of eigenvalues  $\lambda_i(A)$  of a symmetric matrix  $A$ , assuming they are arranged in a non-increasing order:

$$\lambda_i(A) = \max_{\dim E=i} \min_{x \in S(E)} \langle Ax, x \rangle, \quad (4.2)$$

Here the maximum is over all  $i$ -dimensional subspaces  $E$  of  $\mathbb{R}^n$ , and the minimum is over all unit vectors  $x \in E$ , and  $S(E)$  denotes the unit Euclidean sphere in the subspace  $E$ . For the singular values, the min-max theorem immediately implies that

$$s_i(A) = \max_{\dim E=i} \min_{x \in S(E)} \|Ax\|_2.$$

**Exercise 4.1.1.** 🍷 Suppose  $A$  is an invertible matrix with singular value decomposition

$$A = \sum_{i=1}^n s_i u_i v_i^\top.$$

Check that

$$A^{-1} = \sum_{i=1}^n \frac{1}{s_i} v_i u_i^\top.$$

#### 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 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*, also called the *spectral norm*, is 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  $\langle Ax, y \rangle$  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  $A$  equals 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, if  $A$  has full rank then

$$s_n(A) = \frac{1}{\|A^+\|}$$

where  $A^+$  is the Moore-Penrose pseudoinverse of  $A$ . Its norm  $\|A^+\|$  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 the Frobenius norm is the Euclidean norm on the space of matrices  $\mathbb{R}^{m \times n}$ . In terms of singular values, the Frobenius norm can be computed as

$$\|A\|_F = \left( \sum_{i=1}^r s_i(A)^2 \right)^{1/2}.$$

The canonical inner product on  $\mathbb{R}^{m \times n}$  can be represented in terms of matrices as

$$\langle A, B \rangle = \text{tr}(A^T B) = \sum_{i=1}^m \sum_{j=1}^n A_{ij} B_{ij}. \quad (4.3)$$

Obviously, the canonical inner product generates the canonical Euclidean norm, i.e.

$$\|A\|_F^2 = \langle A, A \rangle.$$

Let us now compare the operator and the Frobenius norm. If we look at the vector  $s = (s_1, \dots, s_r)$  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:

$$\|A\| \leq \|A\|_F \leq \sqrt{r} \|A\|. \quad (4.4)$$

**Exercise 4.1.2.** ☕☕ Prove the following bound on the singular values  $s_i$  of any matrix  $A$ :

$$s_i \leq \frac{1}{\sqrt{i}} \|A\|_F.$$

#### 4.1.4 Low-rank approximation

Suppose we want to approximate a given matrix  $A$  of rank  $r$  by a matrix  $A_k$  that has a given lower rank, say rank  $k < r$ . What is the best choice for  $A_k$ ? In other words, what matrix  $A_k$  of rank  $k$  minimizes the distance to  $A$ ? The distance can be measured by the operator norm or Frobenius norm.

In either case, *Eckart-Young-Mirsky's theorem* gives the answer to the low-rank approximation problem. It states that the minimizer  $A_k$  is obtained by truncating the singular value decomposition of  $A$  at the  $k$ -th term:

$$A_k = \sum_{i=1}^k s_i u_i v_i^\top.$$

In other words, the Eckart-Young-Mirsky theorem states that

$$\|A - A_k\| = \min_{\text{rank}(A') \leq k} \|A - A'\|.$$

A similar statement holds for the Frobenius norm (and, in fact, for any unitary-invariant norm). The matrix  $A_k$  is often called the *best rank  $k$  approximation* of  $A$ .

**Exercise 4.1.3** (Best rank  $k$  approximation). ☕☕ Let  $A_k$  be the best rank  $k$  approximation of a matrix  $A$ . Express  $\|A - A_k\|^2$  and  $\|A - A_k\|_F^2$  in terms of the singular values  $s_i$  of  $A$ .

#### 4.1.5 Approximate isometries

The extreme singular values  $s_1(A)$  and  $s_n(A)$  have an important geometric meaning. They are respectively the smallest number  $M$  and the largest number  $m$  that make the following inequality true:

$$m\|x\|_2 \leq \|Ax\|_2 \leq M\|x\|_2 \quad \text{for all } x \in \mathbb{R}^n. \quad (4.5)$$

(Check!) Applying this inequality for  $x - y$  instead of  $x$  and with the best bounds, we can rewrite it as

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

This means that the matrix  $A$ , acting as an operator from  $\mathbb{R}^n$  to  $\mathbb{R}^m$ , can only change the distance between any points by a factor that lies between  $s_n(A)$  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.

**Exercise 4.1.4** (Isometries). 🍷 Let  $A$  be an  $m \times n$  matrix with  $m \geq n$ . Prove that the following statements are equivalent.

- (a)  $A^T A = I_n$ .
- (b)  $P := AA^T$  is an *orthogonal projection*<sup>1</sup> in  $\mathbb{R}^m$  onto a subspace of dimension  $n$ .
- (c)  $A$  is an *isometry*, or isometric embedding of  $\mathbb{R}^n$  into  $\mathbb{R}^m$ , which means that

$$\|Ax\|_2 = \|x\|_2 \quad \text{for all } x \in \mathbb{R}^n.$$

- (d) All singular values of  $A$  equal 1; equivalently

$$s_n(A) = s_1(A) = 1.$$

Quite often the conditions of Exercise 4.1.4 hold only approximately, in which case we regard  $A$  as an *approximate isometry*.

**Lemma 4.1.5** (Approximate isometries). *Let  $A$  be an  $m \times n$  matrix and  $\delta > 0$ . Suppose that*

$$\|A^T A - I_n\| \leq \max(\delta, \delta^2).$$

*Then*

$$(1 - \delta)\|x\|_2 \leq \|Ax\|_2 \leq (1 + \delta)\|x\|_2 \quad \text{for all } x \in \mathbb{R}^n. \quad (4.6)$$

*Consequently, all singular values of  $A$  are between  $1 - \delta$  and  $1 + \delta$ :*

$$1 - \delta \leq s_n(A) \leq s_1(A) \leq 1 + \delta. \quad (4.7)$$

*Proof* To prove (4.6), we may assume without loss of generality that  $\|x\|_2 = 1$ . (Why?) Then, using the assumption, we get

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

Applying the elementary inequality

$$\max(|z - 1|, |z - 1|^2) \leq |z^2 - 1|, \quad z \geq 0 \quad (4.8)$$

for  $z = \|Ax\|_2$ , we conclude that

$$\left| \|Ax\|_2 - 1 \right| \leq \delta.$$

This proves (4.6), which in turn implies (4.7) as we saw in the beginning of this section.  $\square$

**Exercise 4.1.6** (Approximate isometries). 🍷🍷 Prove the following converse to Lemma 4.1.5: if (4.7) holds, then

$$\|A^T A - I_n\| \leq 3 \max(\delta, \delta^2).$$

<sup>1</sup> Recall that  $P$  is a projection if  $P^2 = P$ , and  $P$  is called orthogonal if the image and kernel of  $P$  are orthogonal subspaces.

**Remark 4.1.7** (Projections vs. isometries). Consider an  $n \times m$  matrix  $Q$ . Then

$$QQ^\top = I_n$$

if and only if

$$P := Q^\top Q$$

is an orthogonal projection in  $\mathbb{R}^m$  onto a subspace of dimension  $n$ . (This can be checked directly or deduced from Exercise 4.1.4 by taking  $A = Q^\top$ .) In case this happens, the matrix  $Q$  itself is often called a *projection* from  $\mathbb{R}^m$  onto  $\mathbb{R}^n$ .

Note that  $A$  is an isometric embedding of  $\mathbb{R}^n$  into  $\mathbb{R}^m$  if and only if  $A^\top$  is a projection from  $\mathbb{R}^m$  onto  $\mathbb{R}^n$ . These remarks can be also made for an approximate isometry  $A$ ; the transpose  $A^\top$  in this case is an *approximate projection*.

**Exercise 4.1.8** (Isometries and projections from unitary matrices). 🍷 Canonical examples of isometries and projections can be constructed from a fixed unitary matrix  $U$ . Check that any sub-matrix of  $U$  obtained by selecting a subset of columns is an isometry, and any sub-matrix obtained by selecting a subset of rows is a projection.

## 4.2 Nets, covering numbers and packing numbers

We are going to develop a simple but powerful method – an  $\varepsilon$ -net argument – and illustrate its usefulness for the analysis of random matrices. In this section, we recall the concept of an  $\varepsilon$ -net, which you may have seen in a course in real analysis, and we relate it to some other basic notions – covering, packing, entropy, volume, and coding.

**Definition 4.2.1** ( $\varepsilon$ -net). Let  $(T, d)$  be a metric space. Consider a subset  $K \subset T$  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.

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

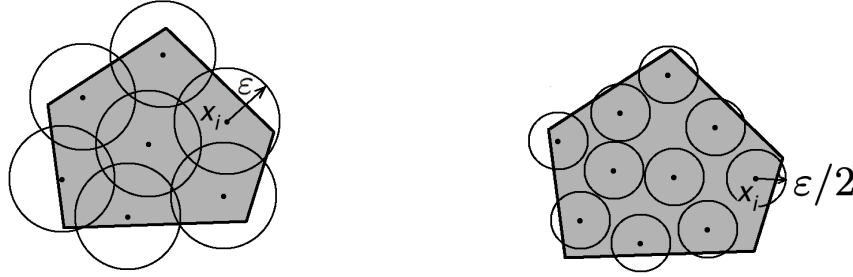
Equivalently,  $\mathcal{N}$  is an  $\varepsilon$ -net of  $K$  if and only if  $K$  can be covered by balls with centers in  $\mathcal{N}$  and radii  $\varepsilon$ , see Figure 4.1a.

If you ever feel confused by too much generality, it might be helpful to keep in mind an important example. Let  $T = \mathbb{R}^n$  with  $d$  being the Euclidean distance, i.e.

$$d(x, y) = \|x - y\|_2, \quad x, y \in \mathbb{R}^n. \quad (4.9)$$

In this case, we cover a subset  $K \subset \mathbb{R}^n$  by *round balls*, as shown in Figure 4.1a. We already saw an example of such covering in Corollary 0.0.4 where  $K$  was a polytope.

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



(a) This covering of a pentagon  $K$  by seven  $\varepsilon$ -balls shows that  $\mathcal{N}(K, \varepsilon) \leq 7$ .

(b) This packing of a pentagon  $K$  by ten  $\varepsilon/2$ -balls shows that  $\mathcal{P}(K, \varepsilon) \geq 10$ .

**Figure 4.1** Packing and covering

**Remark 4.2.3** (Compactness). An important result in real analysis states that a subset  $K$  of a complete metric space  $(T, d)$  is *precompact* (i.e. the closure of  $K$  is compact) if and only if

$$\mathcal{N}(K, d, \varepsilon) < \infty \quad \text{for every } \varepsilon > 0.$$

Thus we can think about the magnitude  $\mathcal{N}(K, d, \varepsilon)$  as a quantitative measure of compactness of  $K$ .

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

**Definition 4.2.4** (Packing numbers). A subset  $\mathcal{N}$  of a metric space  $(T, d)$  is  $\varepsilon$ -*separated* if  $d(x, y) > \varepsilon$  for all distinct points  $x, y \in \mathcal{N}$ . The largest possible cardinality of an  $\varepsilon$ -separated subset of a given set  $K \subset T$  is called the *packing number* of  $K$  and is denoted  $\mathcal{P}(K, d, \varepsilon)$ .

**Exercise 4.2.5** (Packing the balls into  $K$ ). ☹☹

- Suppose  $T$  is a normed space. Prove that  $\mathcal{P}(K, d, \varepsilon)$  is the largest number of closed disjoint balls with centers in  $K$  and radii  $\varepsilon/2$ . See Figure 4.1b for an illustration.
- Show by example that the previous statement may be false for a general metric space  $T$ .

**Lemma 4.2.6** (Nets from separated sets). Let  $\mathcal{N}$  be a maximal<sup>2</sup>  $\varepsilon$ -separated subset of  $K$ . Then  $\mathcal{N}$  is an  $\varepsilon$ -net of  $K$ .

*Proof* Let  $x \in K$ ; we want to show that there exists  $x_0 \in \mathcal{N}$  such that  $d(x, x_0) \leq \varepsilon$ . If  $x \in \mathcal{N}$ , the conclusion is trivial by choosing  $x_0 = x$ . Suppose now  $x \notin \mathcal{N}$ . The maximality assumption implies that  $\mathcal{N} \cup \{x\}$  is not  $\varepsilon$ -separated. But this means precisely that  $d(x, x_0) \leq \varepsilon$  for some  $x_0 \in \mathcal{N}$ .  $\square$

**Remark 4.2.7** (Constructing a net). Lemma 4.2.6 leads to the following simple algorithm for constructing an  $\varepsilon$ -net of a given set  $K$ . Choose a point  $x_1 \in K$

<sup>2</sup> Here by “maximal” we mean that adding any new point to  $\mathcal{N}$  destroys the separation property.

arbitrarily, choose a point  $x_2 \in K$  which is farther than  $\varepsilon$  from  $x_1$ , choose  $x_3$  so that it is farther than  $\varepsilon$  from both  $x_1$  and  $x_2$ , and so on. If  $K$  is compact, the algorithm terminates in finite time (why?) and gives an  $\varepsilon$ -net of  $K$ .


The covering and packing numbers are essentially equivalent:

**Lemma 4.2.8** (Equivalence of covering and packing numbers). *For any set  $K \subset T$  and any  $\varepsilon > 0$ , we have*


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

*Proof* The upper bound follows from Lemma 4.2.6. (How?)

To prove the lower bound, choose an  $2\varepsilon$ -separated subset  $\mathcal{P} = \{x_i\}$  in  $K$  and an  $\varepsilon$ -net  $\mathcal{N} = \{y_j\}$  of  $K$ . By the definition of a net, each point  $x_i$  belongs to a closed  $\varepsilon$ -ball centered at some point  $y_j$ . Moreover, since any closed  $\varepsilon$ -ball can not contain a pair of  $2\varepsilon$ -separated points, each  $\varepsilon$ -ball centered at  $y_j$  may contain at most one point  $x_i$ . The pigeonhole principle then yields  $|\mathcal{P}| \leq |\mathcal{N}|$ . Since this happens for arbitrary packing  $\mathcal{P}$  and covering  $\mathcal{N}$ , the lower bound in the lemma is proved.  $\square$

**Exercise 4.2.9** (Allowing the centers to be outside  $K$ ).  In our definition of the covering numbers of  $K$ , we required that the centers  $x_i$  of the balls  $B(x_i, \varepsilon)$  that form a covering lie in  $K$ . Relaxing this condition, define the *exterior covering number*  $\mathcal{N}^{\text{ext}}(K, d, \varepsilon)$  similarly but without requiring that  $x_i \in K$ . Prove that

$$\mathcal{N}^{\text{ext}}(K, d, \varepsilon) \leq \mathcal{N}(K, d, \varepsilon) \leq \mathcal{N}^{\text{ext}}(K, d, \varepsilon/2).$$

**Exercise 4.2.10** (Monotonicity).  Give a counterexample to the following monotonicity property:

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

Prove an approximate version of monotonicity:

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

#### 4.2.1 Covering numbers and volume

Let us now specialize our study of covering numbers to the most important example where  $T = \mathbb{R}^n$  with the Euclidean metric

$$d(x, y) = \|x - y\|_2$$

as in (4.9). To ease the notation, we often omit the metric when it is understood, thus writing

$$\mathcal{N}(K, \varepsilon) = \mathcal{N}(K, d, \varepsilon).$$

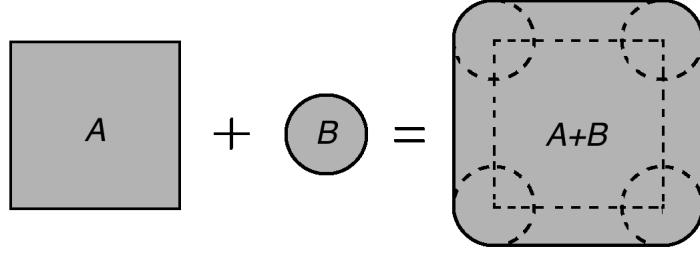
If the covering numbers measure the size of  $K$ , how are they related to the most classical measure of size, the volume of  $K$  in  $\mathbb{R}^n$ ? There could not be a full equivalence between these two quantities, since “flat” sets have zero volume but non-zero covering numbers.

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

**Definition 4.2.11** (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 4.2 shows an example of Minkowski sum of two sets on the plane.



**Figure 4.2** Minkowski sum of a square and a circle is a square with rounded corners.

**Proposition 4.2.12** (Covering numbers and volume). Let  $K$  be a subset of  $\mathbb{R}^n$  and  $\varepsilon > 0$ . Then

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

Here  $|\cdot|$  denotes the volume in  $\mathbb{R}^n$ ,  $B_2^n$  denotes the unit Euclidean ball<sup>3</sup> in  $\mathbb{R}^n$ , so  $\varepsilon B_2^n$  is a Euclidean ball with radius  $\varepsilon$ .

*Proof* The middle inequality follows from Lemma 4.2.8, so all we need to prove is the left and right bounds.

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

$$|K| \leq N \cdot |\varepsilon B_2^n|,$$

Dividing both sides by  $|\varepsilon B_2^n|$  yields the lower bound.

**(Upper bound)** Let  $N := \mathcal{P}(K, \varepsilon)$ . Then one can construct  $N$  closed disjoint balls  $B(x_i, \varepsilon/2)$  with centers  $x_i \in K$  and radii  $\varepsilon/2$  (see Exercise 4.2.5). While these balls may not need to fit entirely in  $K$  (see Figure 4.1b), they do fit in a slightly inflated set, namely  $K + (\varepsilon/2)B_2^n$ . (Why?) Comparing the volumes, we obtain

$$N \cdot |(\varepsilon/2)B_2^n| \leq |K + (\varepsilon/2)B_2^n|.$$

which leads to the upper bound in the proposition.  $\square$

<sup>3</sup> Thus  $B_2^n = \{x \in \mathbb{R}^n : \|x\|_2 \leq 1\}$ .



An important consequence of the volumetric bound (4.10) is that the covering (and thus packing) numbers of the Euclidean ball, as well as many other sets, are *exponential* in the dimension  $n$ . Let us check this.

**Corollary 4.2.13** (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 4.2.12, since the volume in  $\mathbb{R}^n$  scales as

$$|\varepsilon B_2^n| = \varepsilon^n |B_2^n|.$$

The upper bound follows from Proposition 4.2.12, too:

$$\mathcal{N}(B_2^n, \varepsilon) \leq \frac{|(1 + \varepsilon/2)B_2^n|}{|(\varepsilon/2)B_2^n|} = \frac{(1 + \varepsilon/2)^n}{(\varepsilon/2)^n} = \left(\frac{2}{\varepsilon} + 1\right)^n.$$

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

To simplify the bound a bit, note that in the non-trivial range  $\varepsilon \in (0, 1]$  we have

$$\left(\frac{1}{\varepsilon}\right)^n \leq \mathcal{N}(B_2^n, \varepsilon) \leq \left(\frac{3}{\varepsilon}\right)^n. \quad (4.10)$$

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

The volumetric argument we just gave works well in many other situations. Let us give an important example.

**Definition 4.2.14** (Hamming cube). The Hamming cube  $\{0, 1\}^n$  consists of all binary strings of length  $n$ . The *Hamming distance*  $d_H(x, y)$  between two binary strings is defined as the number of bits where  $x$  and  $y$  disagree, i.e.

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

Endowed with this metric, the Hamming cube is a metric space  $(\{0, 1\}^n, d_H)$ , which is sometimes called the *Hamming space*.

**Exercise 4.2.15.** ☞ Check that  $d_H$  is indeed a metric.

**Exercise 4.2.16** (Covering and packing numbers of the Hamming cube). ☞☞☞ Let  $K = \{0, 1\}^n$ . Prove that for every integer  $m \in [0, n]$ , we have

$$\frac{2^n}{\sum_{k=0}^m \binom{n}{k}} \leq \mathcal{N}(K, d_H, m) \leq \mathcal{P}(K, d_H, m) \leq \frac{2^n}{\sum_{k=0}^{\lfloor m/2 \rfloor} \binom{n}{k}}$$

**Hint:** Adapt the volumetric argument by replacing volume by cardinality.

To make these bounds easier to compute, one can use the bounds for binomial sums from Exercise 0.0.5.

### 4.3 Application: error correcting codes

Covering and packing arguments frequently appear in applications to *coding theory*. Here we give two examples that relate covering and packing numbers to complexity and error correction.

#### 4.3.1 Metric entropy and complexity

Intuitively, the covering and packing numbers measure the *complexity* of a set  $K$ . The logarithm of the covering numbers  $\log_2 \mathcal{N}(K, \varepsilon)$  is often called the *metric entropy* of  $K$ . As we will see now, the metric entropy is equivalent to the number of bits needed to encode points in  $K$ .

**Proposition 4.3.1** (Metric entropy and coding). *Let  $(T, d)$  be a metric space, and consider a subset  $K \subset T$ . Let  $\mathcal{C}(K, d, \varepsilon)$  denote the smallest number of bits sufficient to specify every point  $x \in K$  with accuracy  $\varepsilon$  in the metric  $d$ . Then*

$$\log_2 \mathcal{N}(K, d, \varepsilon) \leq \mathcal{C}(K, d, \varepsilon) \leq \lceil \log_2 \mathcal{N}(K, d, \varepsilon/2) \rceil.$$

*Proof* **(Lower bound)** Assume  $\mathcal{C}(K, d, \varepsilon) \leq N$ . This means that there exists a transformation (“encoding”) of points  $x \in K$  into bit strings of length  $N$ , which specifies every point with accuracy  $\varepsilon$ . Such a transformation induces a partition of  $K$  into at most  $2^N$  subsets, which are obtained by grouping the points represented by the same bit string; see Figure 4.3 for an illustration. Each subset must have diameter<sup>4</sup> at most  $\varepsilon$ , and thus it can be covered by a ball centered in  $K$  and with radius  $\varepsilon$ . (Why?) Thus  $K$  can be covered by at most  $2^N$  balls with radii  $\varepsilon$ . This implies that  $\mathcal{N}(K, d, \varepsilon) \leq 2^N$ . Taking logarithms on both sides, we obtain the lower bound in the proposition.

**(Upper bound)** Assume  $\log_2 \mathcal{N}(K, d, \varepsilon/2) \leq N$  for some integer  $N$ . This means that 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  $x_0 \in \mathcal{N}$  that is closest to  $x$ . Since there are at most  $2^N$  such points,  $N$  bits are sufficient to specify the point  $x_0$ . It remains to note that 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, by triangle inequality,

$$d(x, y) \leq d(x, x_0) + d(y, x_0) \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon$$

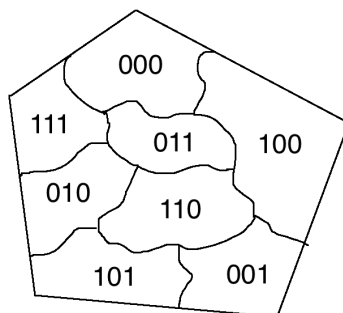
This shows that  $\mathcal{C}(K, d, \varepsilon) \leq N$ . This completes the proof.  $\square$

#### 4.3.2 Error correcting codes

Suppose Alice wants to send Bob a message that consists of  $k$  letters, such as

$$x := \text{“fill the glass”}.$$

<sup>4</sup> If  $(T, d)$  is a metric space and  $K \subset T$ , the diameter of the set  $K$  is defined as  $\text{diam}(K) := \sup\{d(x, y) : x, y \in K\}$ .



**Figure 4.3** Encoding points in  $K$  as  $N$ -bit strings induces a partition of  $K$  into at most  $2^N$  subsets.

Suppose further that an adversary may corrupt Alice's message by changing at most  $r$  letters in it. For example, Bob may receive

$$y := \text{"bill the class"}$$

if  $r = 2$ . Is there a way to protect the communication channel between Alice and Bob, a method that can correct adversarial errors?

A common approach relies on using *redundancy*. Alice would encode her  $k$ -letter message into a longer,  $n$ -letter, message for some  $n > k$ , hoping that the extra information would help Bob get her message right despite any  $r$  errors.

**Example 4.3.2** (Repetition code). Alice may just repeat her message several times, thus sending to Bob

$$E(x) := \text{"fill the glass fill the glass fill the glass fill the glass fill the glass"}.$$

Bob could then use the *majority decoding*: to determine the value of any particular letter, he would look at the received copies of it in  $E(x)$  and choose the value that occurs most frequently. If the original message  $x$  is repeated  $2r + 1$  times, then the majority decoding recovers  $x$  exactly even when  $r$  letters of  $E(x)$  are corrupted. (Why?)

The problem with majority decoding is that it is very inefficient: it uses

$$n = (2r + 1)k \tag{4.11}$$

letters to encode a  $k$ -letter message. As we will see shortly, there exist error correcting codes with much smaller  $n$ .

But first let us formalize the notion of an error correcting code – an encoding of  $k$ -letter strings into  $n$ -letter strings that can correct  $r$  errors. For convenience, instead of using the English alphabet we shall work with the binary alphabet consisting of two letters 0 and 1.

**Definition 4.3.3** (Error correcting code). Fix integers  $k$ ,  $n$  and  $r$ . Two maps

$$E : \{0, 1\}^k \rightarrow \{0, 1\}^n \quad \text{and} \quad D : \{0, 1\}^n \rightarrow \{0, 1\}^k$$

are called *encoding* and *decoding* maps that can correct  $r$  errors if we have

$$D(y) = x$$

for every word  $x \in \{0, 1\}^k$  and every string  $y \in \{0, 1\}^n$  that differs from  $E(x)$  in at most  $r$  bits. The encoding map  $E$  is called an *error correcting code*; its image  $E(\{0, 1\}^k)$  is called a *codebook* (and very often the image itself is called the *error correcting code*); the elements  $E(x)$  of the image are called *codewords*.

We now relate error correction to packing numbers of the Hamming cube  $(\{0, 1\}^n, d_H)$  where  $d_H$  is the Hamming metric we introduced in Definition 4.2.14.

**Lemma 4.3.4** (Error correction and packing). *Assume that positive integers  $k$ ,  $n$  and  $r$  are such that*

$$\log_2 \mathcal{P}(\{0, 1\}^n, d_H, 2r) \geq k.$$

*Then there exists an error correcting code that encodes  $k$ -bit strings into  $n$ -bit strings and can correct  $r$  errors.*

*Proof* By assumption, there exists a subset  $\mathcal{N} \subset \{0, 1\}^n$  with cardinality  $|\mathcal{N}| = 2^k$  and such that the closed balls centered at the points in  $\mathcal{N}$  and with radii  $r$  are disjoint. (Why?) We then define the encoding and decoding maps as follows: choose  $E : \{0, 1\}^k \rightarrow \mathcal{N}$  to be an arbitrary one-to-one map and  $D : \{0, 1\}^n \rightarrow \{0, 1\}^k$  to be a nearest neighbor decoder.<sup>5</sup>

Now, if  $y \in \{0, 1\}^n$  differs from  $E(x)$  in at most  $r$  bits,  $y$  lies in the closed ball centered at  $E(x)$  and with radius  $r$ . Since such balls are disjoint by construction,  $y$  must be strictly closer to  $E(x)$  than to any other codeword  $E(x')$  in  $\mathcal{N}$ . Thus the nearest-neighbor decoding decodes  $y$  correctly, i.e.  $D(y) = x$ . This completes the proof.  $\square$

Let us substitute into Lemma 4.3.4 the bounds on the packing numbers of the Hamming cube from Exercise 4.2.16.

**Theorem 4.3.5** (Guarantees for an error correcting code). *Assume that positive integers  $k$ ,  $n$  and  $r$  are such that*

$$n \geq k + 2r \log_2 \left( \frac{en}{2r} \right).$$

*Then there exists an error correcting code that encodes  $k$ -bit strings into  $n$ -bit strings and can correct  $r$  errors.*

*Proof* Passing from packing to covering numbers using Lemma 4.2.8 and then using the bounds on the covering numbers from Exercises 4.2.16 (and simplifying using Exercise 0.0.5), we get

$$\mathcal{P}(\{0, 1\}^n, d_H, 2r) \geq \mathcal{N}(\{0, 1\}^n, d_H, 2r) \geq 2^n \left( \frac{2r}{en} \right)^{2r}.$$

<sup>5</sup> Formally, we set  $D(y) = x_0$  where  $E(x_0)$  is the closest codeword in  $\mathcal{N}$  to  $y$ ; break ties arbitrarily.

By assumption, this quantity is further bounded below by  $2^k$ . An application of Lemma 4.3.4 completes the proof.  $\square$

Informally, Theorem 4.3.5 shows that we can correct  $r$  errors if we make the information overhead  $n - k$  almost linear in  $r$ :

$$n - k \asymp r \log \left( \frac{n}{r} \right).$$

This overhead is much smaller than for the repetition code (4.11). For example, to correct two errors in Alice's twelve-letter message "*fill the glass*", encoding it into a 30-letter codeword would suffice.

**Remark 4.3.6** (Rate). The guarantees of a given error correcting code are traditionally expressed in terms of the tradeoff between the *rate* and *fraction of errors*, defined as

$$R := \frac{k}{n} \quad \text{and} \quad \delta := \frac{r}{n}.$$

Theorem 4.3.5 states that there exist error correcting codes with rate as high as

$$R \geq 1 - f(2\delta)$$

where  $f(t) = t \log_2(e/t)$ .

**Exercise 4.3.7** (Optimality). ☕☕☕

- (a) Prove the converse to the statement of Lemma 4.3.4.
- (b) Deduce a converse to Theorem 4.3.5. Conclude that for any error correcting code that encodes  $k$ -bit strings into  $n$ -bit strings and can correct  $r$  errors, the rate must be

$$R \leq 1 - f(\delta)$$

where  $f(t) = t \log_2(1/t)$  as before.

#### 4.4 Upper bounds on random sub-gaussian matrices

We are now ready to begin to study the non-asymptotic theory of random matrices. Random matrix theory is concerned with  $m \times n$  matrices  $A$  with random entries. The central questions of this theory are about the distributions of singular values, eigenvalues (if  $A$  is symmetric) and eigenvectors of  $A$ .

Theorem 4.4.5 will give a first bound on the operator norm (equivalently, on the largest singular value) of a random matrix with independent sub-gaussian entries. It is neither the sharpest nor the most general result; it will be sharpened and extended in Sections 4.6 and 6.5.

But before we do this, let us pause to learn how  $\varepsilon$ -nets can help us compute the operator norm of a matrix.

### 4.4.1 Computing the norm on a net

The notion of  $\varepsilon$ -nets can help us to simplify various problems involving high-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

$$\|A\| = \max_{x \in S^{n-1}} \|Ax\|_2.$$

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 gain control just over an  $\varepsilon$ -net of the sphere (in the Euclidean metric).

**Lemma 4.4.1** (Computing the operator norm on a net). *Let  $A$  be an  $m \times n$  matrix and  $\varepsilon \in [0, 1)$ . Then, for any  $\varepsilon$ -net  $\mathcal{N}$  of the sphere  $S^{n-1}$ , we have*

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

*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

$$\|x - x_0\|_2 \leq \varepsilon.$$


By the definition of the operator norm, this implies

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

Using triangle inequality, we find that

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

Dividing both sides of this inequality by  $1 - \varepsilon$ , we complete the proof.  $\square$



**Exercise 4.4.2.**  Let  $x \in \mathbb{R}^n$  and  $\mathcal{N}$  be an  $\varepsilon$ -net of the sphere  $S^{n-1}$ . Show that

$$\sup_{y \in \mathcal{N}} \langle x, y \rangle \leq \|x\|_2 \leq \frac{1}{1 - \varepsilon} \sup_{y \in \mathcal{N}} \langle x, y \rangle.$$

Recall from Section 4.1.2 that the operator norm of  $A$  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  $x = y$  in this formula. The following exercise shows that instead of controlling the quadratic form on the spheres, it suffices to have control just over the  $\varepsilon$ -nets.

**Exercise 4.4.3** (Quadratic form on a net).   Let  $A$  be an  $m \times n$  matrix and  $\varepsilon \in [0, 1/2)$ .

- (a) Show that for any  $\varepsilon$ -net  $\mathcal{N}$  of the sphere  $S^{n-1}$  and any  $\varepsilon$ -net  $\mathcal{M}$  of the sphere  $S^{m-1}$ , we have

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

- (b) Moreover, if  $m = n$  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 4.4.1 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 4.4.4** (Deviation of the norm on a net). ☹☹☹ 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

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

**Hint:** Assume without loss of generality that  $\mu = 1$ . Represent  $\|Ax\|_2^2 - 1$  as a quadratic form  $\langle Rx, x \rangle$  where  $R = A^\top A - I_n$ . Use Exercise 4.4.3 to compute the maximum of this quadratic form on a net.

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

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

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

with high probability.

**Theorem 4.4.5** (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<sup>6</sup>*

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

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

*Proof* This proof is an 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 4.2.13, 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

$$|\mathcal{N}| \leq 9^n \quad \text{and} \quad |\mathcal{M}| \leq 9^m. \quad (4.12)$$

<sup>6</sup> In results like this,  $C$  and  $c$  will always denote some positive absolute constants.

By Exercise 4.4.3, the operator norm of  $A$  can be bounded using these nets as follows:

$$\|A\| \leq 2 \max_{x \in \mathcal{N}, y \in \mathcal{M}} \langle Ax, y \rangle. \quad (4.13)$$

**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

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

Recalling (2.14), we can restate this as the tail bound

$$\mathbb{P} \{ \langle Ax, y \rangle \geq u \} \leq 2 \exp(-cu^2/K^2), \quad u \geq 0. \quad (4.14)$$

**Step 3: Union bound.** Next, we 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

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

Using the tail bound (4.14) and the estimate (4.12) on the sizes of  $\mathcal{N}$  and  $\mathcal{M}$ , we bound the probability above by

$$9^{n+m} \cdot 2 \exp(-cu^2/K^2). \quad (4.15)$$

Choose

$$u = CK(\sqrt{n} + \sqrt{m} + t). \quad (4.16)$$

Then  $u^2 \geq C^2 K^2 (n + m + t^2)$ , and if the constant  $C$  is chosen sufficiently large, the exponent in (4.15) is large enough, say  $cu^2/K^2 \geq 3(n + m) + t^2$ . Thus

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

Finally, combining this with (4.13), we conclude that

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

Recalling our choice of  $u$  in (4.16), we complete the proof.  $\square$

**Exercise 4.4.6** (Expected norm).  $\clubsuit$  Deduce from Theorem 4.4.5 that

$$\mathbb{E} \|A\| \leq CK(\sqrt{m} + \sqrt{n}).$$



**Exercise 4.4.7** (Optimality). ☹️☹️ Suppose that in Theorem 4.4.5 the entries  $A_{ij}$  have unit variances. Prove that for sufficiently large  $n$  and  $m$  one has

$$\mathbb{E} \|A\| \geq \frac{1}{4} (\sqrt{m} + \sqrt{n}).$$

**Hint:** Bound the operator norm of  $A$  below by the Euclidean norm of the first column and first row; use the concentration of the norm (Exercise 3.1.4) to complete the proof.

Theorem 4.4.5 can be easily extended for symmetric matrices, and the bound for them is

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

with high probability.

**Corollary 4.4.8** (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 the 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 4.4.5 applies for each part  $A^+$  and  $A^-$  separately. By a 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 the triangle inequality  $\|A\| \leq \|A^+\| + \|A^-\|$ , the proof is complete.  $\square$

## 4.5 Application: community detection in networks

Results of random matrix theory are useful in many applications. Here we give an illustration in the analysis of networks.

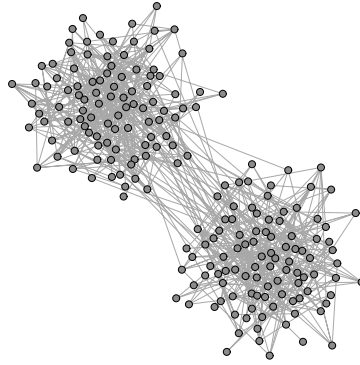
Real-world networks tend to have *communities* – clusters of tightly connected vertices. Finding the communities accurately and efficiently is one of the main problems in network analysis, known as the *community detection problem*.

### 4.5.1 Stochastic Block Model

We will try to solve the community detection 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 4.5.1** (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*<sup>7</sup> 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 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.



**Figure 4.4** A random graph generated according to the stochastic block model  $G(n, p, q)$  with  $n = 200$ ,  $p = 1/20$  and  $q = 1/200$ .

#### 4.5.2 Expected adjacency matrix

It is convenient to identify a graph  $G$  with its adjacency matrix  $A$  which we introduced in Definition 3.6.2. For a random graph  $G \sim G(n, p, q)$ , the adjacency matrix  $A$  is a *random matrix*, and we will examine  $A$  using the tools we developed earlier in this chapter.

It is enlightening to split  $A$  into deterministic and random parts,

$$A = D + R,$$

where  $D$  is the expectation of  $A$ . We may think about  $D$  as an informative part (the “signal”) and  $R$  as “noise”.

To see why  $D$  is informative, let us compute its eigenstructure. The entries  $A_{ij}$  have a Bernoulli distribution; they are either  $\text{Ber}(p)$  or  $\text{Ber}(q)$  depending on the 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

<sup>7</sup> The term *stochastic block model* can also refer to a more general model of random graphs with multiple communities of variable sizes.

like this:

$$D = \mathbb{E} A = \left[ \begin{array}{cc|cc} p & p & q & q \\ p & p & q & q \\ \hline q & q & p & p \\ q & q & p & p \end{array} \right].$$

**Exercise 4.5.2.** ☹️☹️ Check that the matrix  $D$  has rank 2, and the non-zero eigenvalues  $\lambda_i$  and the corresponding eigenvectors  $u_i$  are

$$\lambda_1 = \left(\frac{p+q}{2}\right)n, \quad u_1 = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}; \quad \lambda_2 = \left(\frac{p-q}{2}\right)n, \quad u_2 = \begin{bmatrix} 1 \\ \vdots \\ -1 \end{bmatrix}. \quad (4.17)$$

The important object here is the second eigenvector  $u_2$ . It contains all information about the 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  $D = \mathbb{E} A$  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  $D$  is

$$\|D\| = \lambda_1 \asymp n$$

while the level of the noise  $R$  can be estimated using Corollary 4.4.8:

$$\|R\| \leq C\sqrt{n} \quad \text{with probability at least } 1 - 4e^{-n}. \quad (4.18)$$

Thus, for large  $n$ , the noise  $R$  is much smaller than the signal  $D$ . In other words,  $A$  is close to  $D$ , and thus we should be able to use  $A$  instead of  $D$  to extract the community information. This can be justified using the classical perturbation theory for matrices.

### 4.5.3 Perturbation theory

Perturbation theory describes how the eigenvalues and eigenvectors change under matrix perturbations. For the eigenvalues, we have

**Theorem 4.5.3** (Weyl's inequality). *For any symmetric matrices  $S$  and  $T$  with the same dimensions, we have*

$$\max_i |\lambda_i(S) - \lambda_i(T)| \leq \|S - T\|.$$

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

**Exercise 4.5.4.** ☹️☹️ Deduce Weyl's inequality from the Courant-Fisher's min-max characterization of eigenvalues (4.2).

A similar result holds for eigenvectors, but 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 the wrong eigenvectors. To prevent this from happening, we can assume that the eigenvalues of  $S$  are well separated.

**Theorem 4.5.5** (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:*

$$\min_{j:j \neq i} |\lambda_i(S) - \lambda_j(S)| = \delta > 0.$$

*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*

$$\sin \angle (v_i(S), v_i(T)) \leq \frac{2\|S - T\|}{\delta}.$$

We do not prove the Davis-Kahan theorem here.

The conclusion of the Davis-Kahan theorem implies that the *unit* eigenvectors  $v_i(S)$  and  $v_i(T)$  are close to each other up to a sign, namely

$$\exists \theta \in \{-1, 1\} : \|v_i(S) - \theta v_i(T)\|_2 \leq \frac{2^{3/2}\|S - T\|}{\delta}. \quad (4.19)$$

(Check!)

#### 4.5.4 Spectral Clustering

Returning to the community detection problem, let us apply the 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 (4.18) on  $R = T - S$  and applying (4.19), we can bound the distance between the unit eigenvectors of  $D$  and  $A$ . It follows that there exists a sign  $\theta \in \{-1, 1\}$  such that

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

with probability at least  $1 - 4e^{-n}$ . We already computed the eigenvectors  $u_i(D)$  of  $D$  in (4.17), 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 that the *signs* of most coefficients of  $\theta v_2(A)$  and  $v_2(D)$  must agree. Indeed, we know that

$$\sum_{j=1}^n |u_2(D)_j - \theta u_2(A)_j|^2 \leq \frac{C}{\mu^2}. \quad (4.20)$$

and we also know from (4.17) 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 (4.20). Thus the number of disagreeing signs must be bounded by

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

Summarizing, we can use the vector  $v_2(A)$  to accurately estimate the vector  $v_2 = v_2(D)$  in (4.17), whose signs identify the two communities. This method for community detection is usually called *spectral clustering*. Let us explicitly state this method and the guarantees that we just obtained.

---

### Spectral Clustering Algorithm

---

**Input:** graph  $G$

**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.)
- 

**Theorem 4.5.6** (Spectral clustering for 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 algorithm correctly classifies all except 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 ( $p - q \geq \text{const}$ ).

### 4.6 Two-sided bounds on sub-gaussian matrices

Let us return to Theorem 4.4.5, which gives an upper bound on the spectrum of an  $m \times n$  matrix  $A$  with independent sub-gaussian entries. It essentially states that

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

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

First, we are going to prove sharper and *two-sided* bounds on the entire spectrum of  $A$ :

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

In other words, we will show that a tall random matrix (with  $m \gg n$ ) is an *approximate isometry* in the sense of Section 4.1.5.

Second, the independence of entries is going to be relaxed to just *independence of rows*. Thus we assume that the rows of  $A$  are sub-gaussian random vectors. (We studied such vectors in Section 3.4). This relaxation of independence is important in some applications to data science, where the rows of  $A$  could be 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 usually not independent.

**Theorem 4.6.1** (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*

$$\sqrt{m} - CK^2(\sqrt{n} + t) \leq s_n(A) \leq s_1(A) \leq \sqrt{m} + CK^2(\sqrt{n} + t) \quad (4.21)$$

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

We will prove a slightly stronger conclusion than (4.21), namely that

$$\left\| \frac{1}{m} A^\top A - I_n \right\| \leq K^2 \max(\delta, \delta^2) \quad \text{where} \quad \delta = C \left( \sqrt{\frac{n}{m}} + \frac{t}{\sqrt{m}} \right). \quad (4.22)$$

Using Lemma 4.1.5, one can quickly check that (4.22) indeed implies (4.21). (Do this!)

*Proof* We will prove (4.22) using an  $\varepsilon$ -net argument. This will be similar to the proof of Theorem 4.4.5, but we now use Bernstein’s concentration inequality instead of Hoeffding’s.

**Step 1: Approximation.** Using Corollary 4.2.13, 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 the result of Exercise 4.4.3, we can evaluate the operator norm in (4.22) on the  $\mathcal{N}$ :

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

To complete the proof of (4.22) it suffices to show that, with the required probability,

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

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

$$\|Ax\|_2^2 = \sum_{i=1}^m \langle A_i, x \rangle^2 =: \sum_{i=1}^m X_i^2 \quad (4.23)$$

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 2.8.3) and obtain

$$\begin{aligned} \mathbb{P} \left\{ \left| \frac{1}{m} \|Ax\|_2^2 - 1 \right| \geq \frac{\varepsilon}{2} \right\} &= \mathbb{P} \left\{ \left| \frac{1}{m} \sum_{i=1}^m X_i^2 - 1 \right| \geq \frac{\varepsilon}{2} \right\} \\ &\leq 2 \exp \left[ -c_1 \min \left( \frac{\varepsilon^2}{K^4}, \frac{\varepsilon}{K^2} \right) m \right] \\ &= 2 \exp \left[ -c_1 \delta^2 m \right] \quad (\text{since } \frac{\varepsilon}{K^2} = \max(\delta, \delta^2)) \\ &\leq 2 \exp \left[ -c_1 C^2 (n + t^2) \right]. \end{aligned}$$

The last bound follows from the definition of  $\delta$  in (4.22) 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

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

if we chose the absolute constant  $C$  in (4.22) large enough. As we noted in Step 1, this completes the proof of the theorem.  $\square$


**Exercise 4.6.2.**  Deduce from (4.22) that

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

**Hint:** Use the integral identity from Lemma 1.2.1.

**Exercise 4.6.3.**  Deduce from Theorem 4.6.1 the following bounds on the expectation:

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

**Exercise 4.6.4.**  Give a simpler proof of Theorem 4.6.1, using Theorem 3.1.1 to obtain a concentration bound for  $\|Ax\|_2$  and Exercise 4.4.4 to reduce to a union bound over a net.

### 4.7 Application: covariance estimation and clustering

Suppose we are analyzing some high-dimensional data, which is represented as points  $X_1, \dots, X_m$  sampled from an unknown distribution in  $\mathbb{R}^n$ . One of the most basic data exploration tools is principal component analysis (PCA), which we discussed briefly in Section 3.2.1.

Since we do not have access to the full distribution but only to the finite sample  $\{X_1, \dots, X_m\}$ , we can only expect to compute the covariance matrix of the underlying distribution approximately. If we can do so, the Davis-Kahan theorem 4.5.5 would allow us to estimate the principal components of the underlying distribution, which are the eigenvectors of the covariance matrix.

So, how can we estimate the covariance matrix from the data? Let  $X$  denote the random vector drawn from the (unknown) distribution. Assume for simplicity that  $X$  has zero mean, and let us denote its covariance matrix by

$$\Sigma = \mathbb{E} X X^\top.$$

(Actually, our analysis will not require zero mean, in which case  $\Sigma$  is simply the second moment matrix of  $X$ , as we explained in Section 3.2.)

To estimate  $\Sigma$ , we can use the *sample covariance* matrix  $\Sigma_m$  that is computed from the sample  $X_1, \dots, X_m$  as follows:

$$\Sigma_m = \frac{1}{m} \sum_{i=1}^m X_i X_i^\top.$$

In other words, to compute  $\Sigma$  we replace the expectation over the entire distribution (“population expectation”) by the average over the sample (“sample expectation”).

Since  $X_i$  and  $X$  are identically distributed, our estimate is unbiased, that is

$$\mathbb{E} \Sigma_m = \Sigma.$$

Then the law of large numbers (Theorem 1.3.1) applied to each entry of  $\Sigma$  yields

$$\Sigma_m \rightarrow \Sigma \quad \text{almost surely}$$

as the sample size  $m$  increases to infinity. This leads to the quantitative question: how large must the sample size  $m$  be 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 now show that  $m \asymp n$  sample points suffice.

**Theorem 4.7.1** (Covariance estimation). *Let  $X$  be a sub-gaussian random vector in  $\mathbb{R}^n$ . More precisely, assume that there exists  $K \geq 1$  such that<sup>8</sup>*

$$\|\langle X, x \rangle\|_{\psi_2} \leq K \|\langle X, x \rangle\|_{L^2} \quad \text{for any } x \in \mathbb{R}^n. \quad (4.24)$$

<sup>8</sup> Here we used the notation for the  $L^2$  norm of random variables from Section 1.1, namely  $\|\langle X, x \rangle\|_{L^2}^2 = \mathbb{E} \langle X, x \rangle^2 = \langle \Sigma x, x \rangle$ .



Then, for every positive integer  $m$ , we have

$$\mathbb{E} \|\Sigma_m - \Sigma\| \leq CK^2 \left( \sqrt{\frac{n}{m}} + \frac{n}{m} \right) \|\Sigma\|.$$

*Proof* Let us first bring the random vectors  $X, X_1, \dots, X_m$  to the isotropic position. (This is only possible if  $\Sigma$  is invertible; think how to modify the argument in the general case.) There exist independent isotropic random vectors  $Z, Z_1, \dots, Z_m$  such that

$$X = \Sigma^{1/2}Z \quad \text{and} \quad X_i = \Sigma^{1/2}Z_i.$$

(We checked this in Exercise 3.2.2.) The assumption (4.24) then implies that

$$\|Z\|_{\psi_2} \leq K \quad \text{and} \quad \|Z_i\|_{\psi_2} \leq K.$$

(Check!) Then

$$\|\Sigma_m - \Sigma\| = \|\Sigma^{1/2}R_m\Sigma^{1/2}\| \leq \|R_m\| \|\Sigma\| \quad \text{where} \quad R_m := \frac{1}{m} \sum_{i=1}^m Z_i Z_i^\top - I_n. \quad (4.25)$$

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

$$\frac{1}{m} A^\top A - I_n = \frac{1}{m} \sum_{i=1}^m Z_i Z_i^\top - I_n = R_m.$$

We can apply Theorem 4.6.1 for  $A$  and get

$$\mathbb{E} \|R_m\| \leq CK^2 \left( \sqrt{\frac{n}{m}} + \frac{n}{m} \right).$$

(See Exercise 4.6.2.) Substituting this into (4.25), we complete the proof.  $\square$

**Remark 4.7.2** (Sample complexity). Theorem 4.7.1 implies that for any  $\varepsilon \in (0, 1)$ , we are guaranteed to have covariance estimation with a good relative error,

$$\mathbb{E} \|\Sigma_m - \Sigma\| \leq \varepsilon \|\Sigma\|,$$

if we take a sample of size

$$m \asymp \varepsilon^{-2}n.$$

In other words, the covariance matrix can be estimated accurately by the sample covariance matrix *if the sample size  $m$  is proportional to the dimension  $n$ .*

**Exercise 4.7.3** (Tail bound).  $\clubsuit$  Our argument also implies the following high-probability guarantee. Check that for any  $u \geq 0$ , we have

$$\|\Sigma_m - \Sigma\| \leq CK^2 \left( \sqrt{\frac{n+u}{m}} + \frac{n+u}{m} \right) \|\Sigma\|$$

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

### 4.7.1 Application: clustering of point sets

We are going to illustrate Theorem 4.7.1 with an application to clustering. Like in Section 4.5, we try to identify clusters in the data. 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 a given set of points into few clusters. What exactly constitutes cluster is not well defined in data science. But common sense suggests that the points in the same cluster should tend to be closer to each other than the points taken from different clusters.

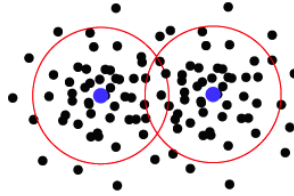
Just like we did for networks, we will design a basic probabilistic model of point sets in  $\mathbb{R}^n$  with two communities, and we will study the clustering problem for that model.

**Definition 4.7.4** (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  $N(-\mu, I_n)$ . This distribution of points is called the 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  $X_1, \dots, X_m$  of independent random vectors identically distributed with  $X$ . Then the sample is distributed according to the Gaussian mixture model; see Figure 4.5 for illustration.



**Figure 4.5** A simulation of points generated according to the Gaussian mixture model, which has two clusters with different means.

Suppose we are given a sample of  $m$  points drawn according to the Gaussian mixture model. Our goal is to identify which points belong to which cluster. To this end, we can use a variant of the *spectral clustering* algorithm that we introduced for networks in Section 3.2.1.

To see why a spectral method has a chance to work here, note that the distribution of  $X$  is not isotropic, but rather stretched in the direction of  $\mu$ . (This is the horizontal direction in Figure 4.5.) Thus, we can approximately compute  $\mu$  by computing the first principal component of the data. Next, we can project the data points onto the line spanned by  $\mu$ , and thus classify them – just look at which side of the origin the projections lie. This leads to the following algorithm.

---

**Spectral Clustering Algorithm**


---

**Input:** points  $X_1, \dots, X_m$  in  $\mathbb{R}^n$

**Output:** a partition of the points into two clusters

- 1: Compute the sample covariance matrix  $\Sigma_m = \frac{1}{m} \sum_{i=1}^m X_i X_i^\top$ .
  - 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 the first community, otherwise in the second.)
- 

**Theorem 4.7.5** (Guarantees of spectral clustering of the Gaussian mixture model). *Let  $X_1, \dots, X_m$  be points in  $\mathbb{R}^n$  drawn from the Gaussian mixture model as above, i.e. there are two communities with means  $\mu$  and  $-\mu$ . Let  $\varepsilon > 0$  be such that  $\|\mu\|_2 \geq C\sqrt{\log(1/\varepsilon)}$ . Suppose the sample size satisfies*

$$m \geq \left( \frac{n}{\|\mu\|_2} \right)^c$$

*where  $c > 0$  is an appropriate absolute constant. Then, with probability at least  $1 - 4e^{-n}$ , the Spectral Clustering Algorithm identifies the communities correctly up to  $\varepsilon m$  misclassified points.*

**Exercise 4.7.6** (Spectral clustering of the Gaussian mixture model). 🍷🍷🍷 Prove Theorem 4.7.5 for the spectral clustering algorithm applied for the Gaussian mixture model. Proceed as follows.

- (a) Compute the covariance matrix  $\Sigma$  of  $X$ ; note that the eigenvector corresponding to the largest eigenvalue is parallel to  $\mu$ .
- (b) Use results about covariance estimation to show that the sample covariance matrix  $\Sigma_m$  is close to  $\Sigma$ , if the sample size  $m$  is relatively large.
- (c) Use the Davis-Kahan Theorem 4.5.5 to deduce that the first eigenvector  $v = v_1(\Sigma_m)$  is close to the direction of  $\mu$ .
- (d) Conclude that the signs of  $\langle \mu, X_i \rangle$  predict well which community  $X_i$  belongs to.
- (e) Since  $v \approx \mu$ , conclude the same for  $v$ .

## 4.8 Notes

The notions of covering and packing numbers and metric entropy introduced in Section 4.2 are thoroughly studied in asymptotic geometric analysis. Most of the material we covered in that section can be found in standard sources such as [11, Chapter 4] and [168].

In Section 4.3.2 we gave some basic results about error correcting codes. The book [216] offers a more systematic introduction to coding theory. Theorem 4.3.5 is a simplified version of the landmark *Gilbert-Varshamov bound* on the rate of

error correcting codes. Our proof of this result relies on a bound on the binomial sum from Exercise 0.0.5. A slight tightening of the binomial sum bound leads to the following improved bound on the rate in Remark 4.3.6: there exist codes with rate

$$R \geq 1 - h(2\delta) - o(1),$$

where

$$h(x) = -x \log_2(x) + (1 - x) \log_2(1 - x)$$

is the *binary entropy function*. This result is known as the *Gilbert-Varshamov bound*. One can tighten up the result of Exercise 4.3.7 similarly and prove that for any error correcting code, the rate is bounded as

$$R \leq 1 - h(\delta).$$

This result is known as the *Hamming bound*.

Our introduction to non-asymptotic random matrix theory in Sections 4.4 and 4.6 mostly follows [222].

In Section 4.5 we gave an application of random matrix theory to networks. For a comprehensive introduction into the interdisciplinary area of network analysis, see e.g. the book [158]. Stochastic block models (Definition 4.5.1) were introduced in [103]. The community detection problem in stochastic block models has attracted a lot of attention: see the book [158], the survey [77], papers including [141, 230, 157, 96, 1, 27, 55, 128, 94, 108] and the references therein.

Davis-Kahan's Theorem 4.5.5, originally proved in [60], has become an invaluable tool in numerical analysis and statistics. There are numerous extensions, variants, and alternative proofs of this theorem, see in particular [226, 229, 225], [21, Section VII.3], [188, Chapter V].

In Section 4.7 we discussed covariance estimation following [222]; more general results will appear in Section 9.2.3. The covariance estimation problem has been studied extensively in high-dimensional statistics, see e.g. [222, 174, 119, 43, 131, 53] and the references therein.

In Section 4.7.1 we gave an application to the clustering of Gaussian mixture models. This problem has been well studied in statistics and computer science communities; see e.g. [153, Chapter 6] and [112, 154, 19, 104, 10, 89].