High Dimensional Probability

for Mathematicians and Data Scientists

Roman Vershynin[1]

[1]University of Michigan. Webpage: www.umich.edu/~romanv

Preface

Who is this book for? This is a textbook in probability in high dimensions with a view toward applications in data sciences. It will be useful for doctoral and advanced masters students in mathematics, statistics, electrical engineering and other related areas, who are looking to expand their knowledge of powerful theoretical methods used in modern data sciences.

To accommodate readers from various disciplines, a good previous masters level course in probability and an excellent command of linear algebra should be enough preparation. Familiarity with measure theory would be helpful but is not essential.

Contents

Chapter 1

Preliminaries on random variables

1.1 Basic quantities associated with random variables

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

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

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

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

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

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

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

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

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

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

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

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

For $p \in [1, \infty]$ the quantity $\|X\|_p$ is a norm and L_p is a *Banach space*. This fact follows from Minkowski's inequality, which we will recall in (1.2). For $p < 1$, the triangle inequality fails and $\|X\|_p$ is not a norm.

The exponent $p = 2$ is special: L_2 is a *Hilbert space*. The inner product on L_2 is given by

$$\mathbb{E}\,\langle X, Y\rangle = \mathbb{E}\,XY.$$

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

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

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

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

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

1.2 Some classical inequalities

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

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

$$\varphi(\mathbb{E}\,X) \le \mathbb{E}\,\varphi(X).$$

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

As a simple consequence of Jensen's inequality, we may note that $\|X\|_p$ is an *increasing function in* p, that is

$$\|X\|_p \le \|X\|_q \quad \text{for any } 0 \le p \le q = \infty. \tag{1.1}$$

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

Minkowski's inequality states that for any $p \in [1, \infty]$ and any random variables $X, Y \in L_p$, we have

$$\|X + Y\|_p \le \|X\|_p + \|Y\|_p. \tag{1.2}$$

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

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

$$\mathbb{E}\, XY \le \|X\|_2 \, \|Y\|_2.$$

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

$$\mathbb{E}\, XY \le \|X\|_p \, \|Y\|_q.$$

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

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

$$F_X(t) = \mathbb{P}\{X \le t\} \qquad \text{for } t \in \mathbb{R}.$$

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

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

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

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

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

In particular, for a non-negative random variable X, we have

$$\mathbb{E}\, X = \int_0^\infty \mathbb{P}\{X > t\}\, dt.$$

Exercise 1.2.2. *Prove the Expectation Integral Identity.*

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

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

$$\mathbb{P}\{X \geq t\} \leq \frac{\mathbb{E}\, X}{t}.$$

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

Corollary 1.2.4 (Chebyshev's inequality)**.** *Let X be a random variable with mean μ and variance σ^2. Then, for any $t > 0$, we have*

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

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

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

1.3 The Law of Large Numbers and the Central Limit Theorem

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

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

$$S_N = X_1 + \cdots X_N.$$

Then, as $N \to \infty$,

$$\frac{S_N}{N} \to \mu \quad \textit{almost surely.}$$

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

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad x \in \mathbb{R}. \tag{1.3}$$

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

$$S_N = X_1 + \cdots X_N$$

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

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

Then, as $N \to \infty$,

$$Z_N \to N(0,1) \quad \textit{in distribution.}$$

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

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

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

Chapter 2

Concentration of sums of independent random variables

2.1 Why concentration inequalities?

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

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

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

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

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

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

Applying Chebyshev's inequality, we can bound the probability of at

least $\frac{3}{4}N$ heads as follows:

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

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

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

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

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

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

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

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

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

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

$$\mathbb{P}\{g \geq t\} \leq \frac{1}{\sqrt{2\pi}} e^{-t^2/2}. \tag{2.2}$$

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

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

a quantity that decays to zero *exponentially* fast in N.

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

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

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

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

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

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

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

the last estimate can be obtained using Stirling's approximation. (Do it!) On the other hand, since the normal distribution is continuous, we have $\mathbb{P}\left\{ g = \frac{N}{2} \right\} = 0$. 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 normal distribution. The normal distribution is especially nice since it has very light, exponentially decaying tails. At the same time, the error of approximation in Central Limit Theorem decays too slow, even slower than linear. This big error ruins desired concentration properties for S_N, which should guarantee light, exponentially decaying tails for S_N.

In order to resolve this issue, we need different ways to obtain concentration inequalities for S_N. We will now develop a direct approach to concentration, which completely bypasses the Central Limit Theorem.

2.2 Hoeffding's inequality

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

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

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

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

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

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

Proof. By homogeneity, we can assume without loss of generality that $\|a\|_2 = 1$.

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

$$\begin{aligned}\mathbb{P}\left\{\sum_{i=1}^N a_i X_i \geq t\right\} &= \mathbb{P}\left\{\exp\left(\lambda \sum_{i=1}^N a_i X_i\right) \geq \exp(\lambda t)\right\} \\ &\leq e^{-\lambda t}\, \mathbb{E} \exp\left(\lambda \sum_{i=1}^N a_i X_i\right). \qquad (2.4)\end{aligned}$$

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

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

$$\mathbb{E} \exp\left(\lambda \sum_{i=1}^N a_i X_i\right) = \prod_{i=1}^N \mathbb{E} \exp(\lambda a_i X_i). \qquad (2.5)$$

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

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

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

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

(Compare the Taylor's expansions of both sides.)

This bound shows that

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

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

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

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

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

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

This completes the proof of Hoeffding's inequality. □

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

Armed with Hoeffding's inequality, we can now return to Question 2.1.1 of bounding the probability of at least $\frac{3}{4}N$ heads in N tosses of a fair coin.

After rescaling from Bernoulli to symmetric Bernoulli, we obtain that this probability is *exponentially small* in N, namely

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

(Check this.)

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 $N \to \infty$. The larger N, the stronger inequality becomes. As we will see later, the non-asymptotic nature of concentration inequalities like Hoeffding makes them attractive in applications in data sciences, where N corresponds to *sample size*.

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

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

Thus the bound doubles, and we obtain:

Theorem 2.2.4 (Hoeffding's inequality, two-sided)**.** *Let $X_1, \dots, X_N$ be independent symmetric Bernoulli random variables, and $a = (a_1, \dots, a_N) \in \mathbb{R}$. Then, for any $t > 0$, we have*

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

Our proof of Hoeffdin'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 2.2.5 (Hoeffding's inequality for general bounded random variables)**.** *Let $X_1, \dots, X_N$ be independent random variables. Assume that $X_i \in [m_i, M_i]$ almost surely for every i. Then, for any $t > 0$, we have*

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

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

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

Write it

Exercise 2.2.7 (Small ball probabilities)**.**

2.3 Chernoff's inequality and Poisson tails

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

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

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

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

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

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

$$\mathbb{P}\{S_N \geq t\} \leq e^{-t}. \tag{2.6}$$

Proof. We will use the same method – based on moment generating function – as we did in the proof of Hoeffding's inequality, Theorem 2.2.2. We repeat the first steps of that argument, leading to 2.4 and (2.5) – multiply both

sides of the inequality $S_N \geq t$ by a parameter λ, exponentiate, and then use Markov's inequality and independence. This yields

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

It remains to bound the MGF of each Bernoulli random variable X_i separately. Since X_i takes value 1 with probability p_i and 1 with probability 0, we have

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

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

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

Substituting this into (2.7), we obtain

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

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

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

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

2.3.1 Poisson tails

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

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

$$\mathbb{P}\{X = k\} = e^{-\lambda} \frac{\lambda^k}{k!}, \quad k = 0, 1, 2, \ldots \tag{2.8}$$

Theorem 2.3.3 (Poisson Limit Theorem)**.** *Consider a sequence of independent random variables $X_i \sim \mathrm{Ber}(p_i)$ and let $S_N = \sum_{i=1}^N X_i$. Assume that, as $N \to \infty$,*

$$\max_{i \le N} p_i \to 0 \quad \text{and} \quad \mathbb{E}\, S_N \to \lambda.$$

Then, as $N \to \infty$,

$$S_N \to \mathrm{Pois}(\lambda) \quad \text{in distribution.}$$

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

Corollary 2.3.4 (Possion tails)**.** *Let $X \sim \mathrm{Pois}(\lambda)$. Then, for any $t > \lambda$ we have*

$$\mathbb{P}\{X \ge t\} \le e^{-\lambda} \Big(\frac{e\lambda}{t} \Big)^t.$$

In particular, for any $t \ge e^2 \lambda$ we have

$$\mathbb{P}\{X \ge t\} \le e^{-t}.$$

How good are these bounds on Poisson tails? Let us compare them with (2.8). Using Stirling's approximation $k! \sim \sqrt{2\pi k}(k/e)^k$, we can approximate the Poisson probability mass function as

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

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

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

2.3.2 Small deviations

It is sometimes useful to distinguish two regimes for concentration inequalities like $\mathbb{P}\{|X - \mu| \ge t\}$. In the *large deviation* regime, t is large, usually larger than the mean μ. In the *small deviation* regime, t is small. The exponential bound (2.6) is an example of a large deviation inequality. In the small deviation regime, Chernoff's inequality is also very useful.

Exercise 2.3.6. [Difficulty=7] *Deduce from Chernoff's inequalities (Theorem 2.3.1 and Exercise 2.3.2) that, for $\delta \in (0, 1]$ we have*

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

(Apply those results with $t = (1 + \delta)\mu$ and $t = (1 - \delta)\mu$, respectively, and analyze the bounds for small δ.)

c absolute const

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

Corollary 2.3.7 (Chernoff's inequality: small deviations)**.** *In the setting of Theorem 2.3.1, for any $t \in (0, \mu]$ we have*

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

In particular, by the Poisson Limit Theorem 2.3.3, a similar bound holds for $X \sim \text{Pois}(\lambda)$. For any $t \in (0, \lambda]$, we have

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

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

Theorem 2.3.8 (Normal approximation to Poisson)**.** *Let $X \sim \text{Pois}(\lambda)$. Then, as $\lambda \to \infty$, we have*

$$\frac{X - \lambda}{\sqrt{\lambda}} \to N(0, 1) \quad \textit{in probability.}$$

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

Exercise 2.3.9. *Give a formal proof of Theorem 2.3.8.*

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

2.3.3 Summary

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

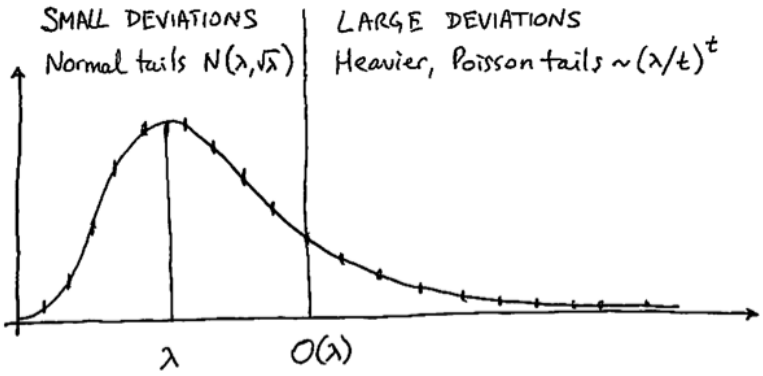


Figure 2.1: This is a sketch of the probability mass function of Poisson distribution Pois(λ). In the small deviation regime, within $O(\lambda)$ from the mean λ, the tails of Pois(λ) are like for the normal distribution $N(\lambda, \sqrt{\lambda})$. In the large deviation regime, the tails are heavier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

2.4 Application: degrees of random graphs

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

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

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

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

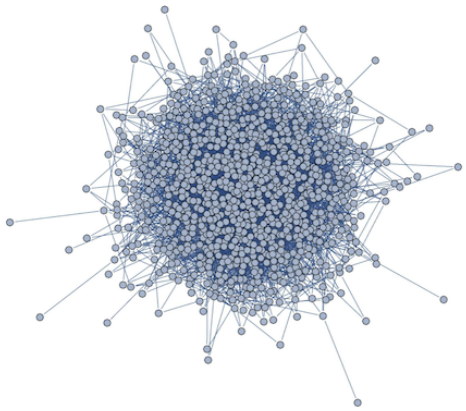


Figure 2.2: A random graph from Erdös-Rényi model $G(n, p)$ with $n = 1000$, $p = 0.00095$.

Let us show that for relatively *dense graphs* where $d \gtrsim \log n$, all degrees tend to concentrate near d.

Proposition 2.4.1 (Degrees of dense graphs concentrate)**.** *Consider a random graph $G \sim G(n, p)$ with expected degree satisfying $d \geq C \log n$. Then, with high probability (for example, 0.9), the following occurs: all vertices of G have degrees between $0.9d$ 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 independent random $\mathrm{Ber}(p)$ random variables. So we can control this degree using Chernoff's inequality for small deviations, which we obtained in Exercise 2.3.6. It follows that

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

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

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

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

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

This completes the proof. □

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

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

1. Show that with high probability (say, 0.9), G has at least one isolated vertex – a vertex with zero degree.

2. Show that with high probability, (say, 0.9), G has a vertex with degree larger than $10d$. (Here 10 can be replaced by any other constant).

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

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

$$O\Big(\frac{\log n}{\log\log n}\Big).$$

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

2.5 Sub-gaussian distributions

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

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

Setting a single coefficient a_i to 1 and the other coefficients to zero, we find that X_i must have sub-gaussian tails:

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

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

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

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

$$\mathbb{P}\{|X| \ge t\} \le 2e^{-t^2/2} \quad \text{for all } t \ge 0. \tag{2.10}$$

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

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

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

Deduce that

$$\|X\|_p = O(\sqrt{p}) \quad \text{as } p \to \infty. \tag{2.11}$$

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

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

2.5.1 Sub-gaussian properties

Now let X be a general random variable. The following proposition states that that the properties we just considered are equivalent – a sub-gaussian

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

Proposition 2.5.2 (Sub-gaussian properties)**.** *Let X be a random variable. Then the following properties are equivalent; the parameters $K_i > 0$ appearing in these properties differ from each other by at most an absolute constant factor.*[1]

1. *The tails of X satisfy*
$$\mathbb{P}\{|X| \geq t\} \leq 2\exp(-t^2/K_1^2) \quad \text{for all } t \geq 0.$$
2. *The moments of X satisfy*
$$\|X\|_p = (\mathbb{E}\,|X|^p)^{1/p} \leq K_2\sqrt{p} \quad \text{for all } p \geq 1.$$
3. *The MGF of X^2 is finite at some point, namely*
$$\mathbb{E}\exp(X^2/K_3^2) \leq 2.$$

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

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

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

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

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

Taking the p-th root yields property 2 with $K_2 \le 2$.

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

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

Expanding the exponential function in Taylor series, we obtain

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

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

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

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

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

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

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

$$\begin{aligned} \mathbb{P}\{|X| > t\} &= \mathbb{P}\{e^{X^2} \ge e^{t^2}\} \\ &\le e^{-t^2}\, \mathbb{E}\, e^{X^2} \quad \text{(by Markov's inequality, Proposition 1.2.3)} \\ &\le 2e^{-t^2} \quad \text{(by property 3).} \end{aligned}$$

This proves property 1 with $K_1 = 1$.

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

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

$$\mathbb{E}\, e^{\lambda X} \le \mathbb{E}\left[\lambda X + e^{\lambda^2 X^2}\right] = \mathbb{E}\, e^{\lambda^2 X^2},$$

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

$$\mathbb{E}\, e^{\lambda^2 X^2} \le e^{\lambda^2} \quad \text{if } |\lambda| \le 1.$$

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

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

$$\begin{aligned} \mathbb{E}\, e^{\lambda X} &\le e^{\lambda^2}\, \mathbb{E}\, e^{X^2} \le 2e^{\lambda^2} \quad \text{(by property 3)} \\ &\le e^{2\lambda^2} \quad \text{(since } |\lambda| \ge 1\text{)}. \end{aligned}$$

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

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

$$\begin{aligned} \mathbb{P}\{X \ge t\} &= \mathbb{P}\{e^{\lambda X} \ge e^{\lambda t}\} \\ &\le e^{-\lambda t}\, \mathbb{E}\, e^{\lambda X} \quad \text{(by Markov's inequality)} \\ &\le e^{-\lambda t} e^{\lambda^2} \quad \text{(by property 4)} \\ &= e^{-\lambda t + \lambda t^2}. \end{aligned}$$

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

$$\mathbb{P}\{X \ge t\} \le e^{-t^2/4}.$$

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

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

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

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

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

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

2.5.2 Definition and examples of sub-gaussian distributions

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

$$\|X\|_{\psi_2} = \inf\left\{t > 0 : \ \mathbb{E}\exp(X^2/t^2) \le 2\right\}. \tag{2.14}$$

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

$$\mathbb{P}\{|X| \ge t\} \le 2\exp(-ct^2/\|X\|_{\psi_2}^2) \quad \text{for all } t \ge 0; \tag{2.15}$$

$$\|X\|_p \le C\|X\|_{\psi_2}\sqrt{p} \quad \text{for all } p \ge 1; \tag{2.16}$$

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

$$\text{if } \mathbb{E}\,X = 0 \text{ then } \ \mathbb{E}\exp(\lambda X) \le \exp(C\lambda^2\|X\|_{\psi_2}^2) \quad \text{for all } \lambda \in \mathbb{R}. \tag{2.17}$$

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

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

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

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

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

$$\|X\|_{\psi_2} \le C\|X\|_\infty, \tag{2.18}$$

where $C = 1/\ln 2$.

Exercise 2.5.7. *Show that Poisson, exponential, Pareto and Cauchy distributions are not sub-gaussian.*

Exercise 2.5.8 (Maximum of sub-gaussians)**.** [Difficulty=6] *Let $X_1, X_2, \dots,$ be a sequence of sub-gaussian random variables, not necessarily independent. Show that*

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

where $K = \max_i \|X_i\|_{\psi_2}$. In particular, for every $N \ge 2$ we have

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

Exercise 2.5.9 (Maximum of gaussians)**.** *Let $X_1, X_2, \dots, X_N$ be independent $N(0,1)$ random variables. Show that*

$$\mathbb{E} \max_{i \le N} g_i \ge c\sqrt{\log N}.$$

2.5.3 A more general view via Orlicz spaces

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

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

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

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

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

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

Exercise 2.5.10. *Show that $\|X\|_\psi$ is indeed a norm on the space L_ψ.*

Thus L_ψ is a normed space; it can also be shown that it is a Banach space.

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

Another example is for the Orlicz function

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

The resulting Orlicz norm is exactly the sub-gaussian norm $\|\cdot\|_{\psi_2}$ that we defined in (2.14), and Orlicz space L_{ψ_2} consists of all sub-gaussian random variables..

Property 2 of Proposition 2.5.2 and (2.18) determine the place of the space L_{ψ_2} of sub-gaussian random variables in the hierarchy of L_p spaces:

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

Thus the space of sub-gaussian random variables L_{ψ_2} is smaller than all L_p spaces, but it is still larger than their limit, the space of bounded random variables L_∞.

2.6 General Hoeffding's and Khinchine's inequalities

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

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

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

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

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

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

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

$$\begin{aligned}
\mathbb{E} \exp\big(\lambda \sum_{i=1}^N X_i\big) &= \prod_{i=1}^N \mathbb{E} \exp(\lambda X_i) \quad \text{(by independence)} \\
&\le \prod_{i=1}^N \exp(C\lambda^2 \|X_i\|_{\psi_2}^2) \quad \text{(by sub-gaussian property (2.17))} \\
&= \exp(\lambda^2 K^2) \quad \text{where } K^2 := C \sum_{i=1}^N \|X_i\|_{\psi_2}^2.
\end{aligned}$$

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

$$\Big\| \sum_{i=1}^N X_i \Big\|_{\psi_2} \le C_1 K$$

where C_1 is an absolute constant. The proof is complete. □

We can restate this result as a concentration inequality.

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 $t \ge 0$, we have*

$$\mathbb{P}\Big\{ \Big| \sum_{i=1}^N X_i \Big| \ge t \Big\} \le 2 \exp\Big(- \frac{ct^2}{\sum_{i=1}^N \|X_i\|_{\psi_2}^2} \Big).$$

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

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

$$\mathbb{P}\Big\{ \Big| \sum_{i=1}^N a_i X_i \Big| \ge t \Big\} \le 2 \exp\Big(- \frac{ct^2}{K^2 \|a\|_2^2} \Big)$$

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

2.6.1 Khinchine's inequality

Consider making section an exercise

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

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

$$\Big(\sum_{i=1}^N a_i^2\Big)^{1/2} \leq \Big\|\sum_{i=1}^N a_i X_i\Big\|_p \leq CK\sqrt{p}\,\Big(\sum_{i=1}^N a_i^2\Big)^{1/2}$$

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

Proof. Hoeffding's inequality in (**??**) yields that the sum is sub-gaussian. Then (2.16) bounds the growth of the moments of the sum. This proves the upper bound on the L_p norm.

To obtain the lower bound, we bound

$$\begin{aligned}
\Big\|\sum_{i=1}^N a_i X_i\Big\|_p &\geq \Big\|\sum_{i=1}^N a_i X_i\Big\|_2 \quad (\text{since } p \geq 2) \\
&= \Big[\operatorname{Var}\Big(\sum_{i=1}^N a_i X_i\Big)\Big]^{1/2} \quad (\text{since } \mathbb{E}\, X_i = 0) \\
&= \Big[\sum_{i=1}^N \operatorname{Var}(a_i X_i)\Big]^{1/2} \quad (\text{by independence}) \\
&= \Big(\sum_{i=1}^N a_i^2\Big)^{1/2} \quad (\text{since } \operatorname{Var}(X_i) = 1).
\end{aligned}$$

This completes the proof. □

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

Provide a hint.

2.6.2 Centering

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

One can quickly check the centering inequality for the L_2 norm, namely

$$\|X - \mathbb{E}\, X\|_2 \le \|X\|_2.$$

(Do this!) The next lemma provides a similar centering inequality for the sub-gaussian norm.

Lemma 2.6.6 (Centering)**.** *If X is a sub-gaussian random variable then $X - \mathbb{E}\, X$ is sub-gaussian, too, and*

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

Proof. Recall that $\|\cdot\|_{\psi_2}$ is a norm, so triangle inequality yields

$$\|X - \mathbb{E}\, X\|_{\psi_2} \le \|X\|_{\psi_2} + \|\mathbb{E}\, X\|_{\psi_2}. \tag{2.20}$$

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

$$\begin{aligned} \|\mathbb{E}\, X\|_{\psi_2} &= C_1|\mathbb{E}\, X| \\ &\le C_1\, \mathbb{E}\,|X| \quad \text{(by Jensen's inequality)} \\ &= C_1\|X\|_1 \\ &\le C_2\|X\|_{\psi_2} \quad \text{(by sub-gaussian moment property (2.16)).} \end{aligned}$$

Substituting this into (2.20), we complete the proof. □

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

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

2.7 Sub-exponential distributions

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

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

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

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

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

In this section we will focus on the class of distributions that have at least an exponential tail decay; we will call sub-exponential distributions. Our analysis here will be quite similar to what we did for sub-gaussian distributions in Section 2.5, and we will leave some details to the reader. In particular, it is not difficult to prove a version of Proposition 2.5.2 for sub-exponential (and more general) distributions.

Proposition 2.7.1 (Sub-exponential properties)**.** *Let X be a random variable. Then the following properties are equivalent; the parameters $K_i > 0$ appearing in these properties differ from each other by at most an absolute constant factor.*[3]

1. *The tails of X satisfy*

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

[2] Here we ignored the pre-factor $1/t$, which does not make much effect on the exponent.

[3] 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 $i, j = 1, 2, 3, 4$.

2. *The moments of X satisfy*

$$\|X\|_p = (\mathbb{E}\,|X|^p)^{1/p} \le K_2 p \quad \text{for all } p \ge 1.$$

3. *The MGF of $|X|$ is finite at some point, namely*

$$\mathbb{E}\exp(|X|/K_3) \le 2.$$

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

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

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

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

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

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

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

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

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

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

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

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

It yields

$$\begin{aligned}
\mathbb{E}\exp\Big(\frac{|XY|}{KL}\Big) &\le \mathbb{E}\exp\Big(\frac{X^2}{2K^2} + \frac{Y^2}{2L^2}\Big) \\
&= \mathbb{E}\left[\exp\Big(\frac{X^2}{2K^2}\Big)\exp\Big(\frac{Y^2}{2L^2}\Big)\right] \\
&\le \frac{1}{2}\,\mathbb{E}\left[\exp\Big(\frac{X^2}{K^2}\Big) + \exp\Big(\frac{Y^2}{L^2}\Big)\right] \quad \text{(by Young's inequality)} \\
&= \frac{1}{2}(2+2) = 2 \quad \text{(by definition of } K \text{ and } L\text{)}.
\end{aligned}$$

The proof is complete. □

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

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

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

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

(Check this!)

2.8 Bernstein's inequality

Our next goal is to prove a concentration inequality for sums of sub-exponential random variables. Just like in the proof of the previous concentration inequalities – Hoeffding's and Chernoff's – our argument will be based on the moment generating function. We will have to be a little more careful now,

since the tails of sub-exponential distributions may not be light enough to make the MGF finite everywhere.

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

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

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

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

$$\mathbb{E}\exp(\lambda X) \leq \exp(C\lambda^2 \|X\|_{\psi_1}^2).$$

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

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

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

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

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

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

This completes the proof. □

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

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

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

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

$$\mathbb{P}\left\{S_N \geq t\right\} \leq e^{-\lambda t}\prod_{i=1}^N \mathbb{E}\exp(\lambda X_i). \tag{2.22}$$

Lemma 2.8.1 can bound the MGF of each term X_i. If we choose λ small enough so that

$$|\lambda| \leq \frac{c}{\max_i \|X_i\|_{\psi_1}}, \tag{2.23}$$

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

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

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

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

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. □

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

Theorem 2.8.3 (Bernstein's inequality)**.** *Let $X_1, \dots, X_N$ 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*

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

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

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

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

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

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

2.8.1 Summary

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

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

$$\mathbb{P}\Big\{\Big|\frac{1}{\sqrt{N}}\sum_{i=1}^{N} X_i\Big| \geq t\Big\} \leq \begin{cases} 2\exp\Big(-\frac{ct^2}{K^2}\Big), & t \leq K\sqrt{N} \\ 2\exp\Big(-\frac{t\sqrt{N}}{K}\Big), & t \geq K\sqrt{N} \end{cases} \tag{2.24}$$

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

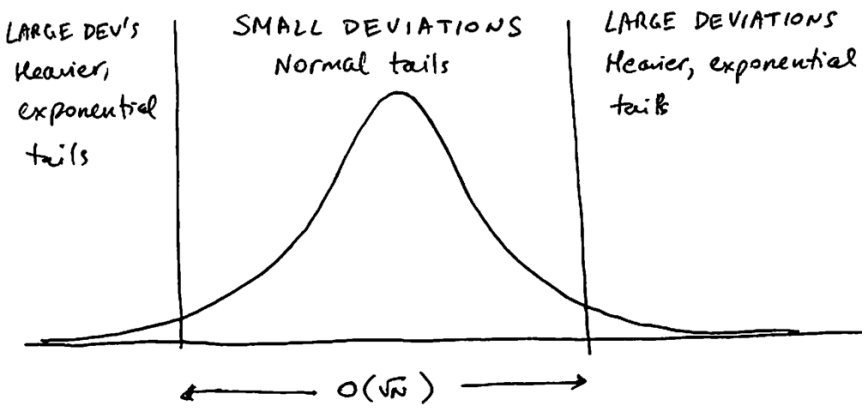


Figure 2.3: Bernstein's inequality for a sum of sub-exponential random variables $N^{-1/2}\sum_{i=1}^N X_i$ is a mixture of two tails, sub-gaussian for small deviations and sub-exponential for large deviations. The sub-gaussian tail emerges in the $O(\sqrt{N})$ neighborhood of zero, and it can be explained by Central Limit Theorem. The heavier, sub-exponential tail, is produced by a single term in the sum.

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

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

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

Theorem 2.8.6 (Bernstein's inequality for bounded distributions)**.** *Let $X_1, \dots, X_N$ be independent, mean zero random variables, such that $|X_i| \le K$ almost surely for all i. Then, for every $t \ge 0$, we have*

$$\mathbb{P}\Big\{\Big|\sum_{i=1}^N X_i\Big| \ge t\Big\} \le 2\exp\Big(-\frac{-t^2/2}{\sigma^2 + CKt}\Big).$$

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

Think behow to present this better. We are reproving the same result in the section of matrix Bernstein's inequality.

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

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

Chapter 3

Random vectors in high dimensions

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

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

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

3.1 Concentration of the norm

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

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

So we should expect that the length of X is

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

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

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*

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

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

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

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

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

$$\begin{aligned} \|X_i^2 - 1\|_{\psi_1} &\le C\|X_i^2\|_{\psi_1} \quad \text{(by Centering Lemma 2.6.6)} \\ &= C\|X_i\|_{\psi_2}^2 \quad \text{(by Lemma 2.7.4)} \\ &\le CK^2. \end{aligned}$$

Applying Bernstein's inequality (Corollary 2.8.4), we obtain for any $u \ge 0$ that

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

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

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

$$|z - 1| \ge \delta \quad \text{implies} \quad |z^2 - 1| \ge \max(\delta, \delta^2) \quad \text{for } z \ge 0.$$

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

$$\begin{aligned} \mathbb{P}\left\{ \left| \frac{1}{\sqrt{n}}\|X\|_2 - 1 \right| \ge \delta \right\} &\le \mathbb{P}\left\{ \left| \frac{1}{n}\|X\|_2^2 - 1 \right| \ge \max(\delta, \delta^2) \right\} \\ &\le 2\exp\left(-\frac{cn}{K^4} \cdot \delta^2 \right). \end{aligned}$$

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

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

The proof is complete. □

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

$$\mathbb{P}\left\{\left|\|X\|_2 - \sqrt{n}\right| \ge t\right\} \le 2\exp\Big(-\frac{ct^2}{K^4}\Big). \tag{3.2}$$

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

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

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

see Figure 3.1 for illustration.

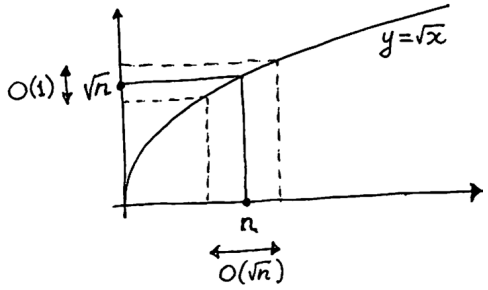


Figure 3.1: Concentration of the norm of a random vector X in $\mathbb{R}^n$. When $\|X\|_2^2$ deviates by $O(\sqrt{n})$ around n, the square root of this quantity, $\|X\|_2$, deviates by $O(1)$ around $\sqrt{n}$.

Question 3.1.2. *Is the quadratic dependence on K in Theorem 3.1.1 optimal? Can it be improved to the linear dependence?*

Theorem 3.1.1 and Centering Lemma 2.6.6 imply the following concentration inequality of a random vector *about its mean*. Let $X = (X_1, \dots, X_n) \in \mathbb{R}^n$ be a random vector with independent, sub-gaussian coordinates X_i that satisfy $\mathrm{Var}(X_i^2) = 1$. Then

$$\Big\| \|X - \mathbb{E}\, X\|_2 - \sqrt{n} \Big\|_{\psi_2} \le CK^2$$

where $K = \max_i \|X_i\|_{\psi_2}$. To see this, apply Theorem 3.1.1 for $X - \mathbb{E}\, X$ instead of X, and use that $\|X_i - \mathbb{E}\, X_i\|_{\psi_2} \le CK$ by centering.

Exercise 3.1.3 (Expected norm)**.** *Let X be a random vector as in Theorem 3.1.1. Show that*

$$\sqrt{n} - C \le \mathbb{E}\, \|X\|_2 \le \sqrt{n} + C.$$

Exercise 3.1.4. [Difficulty=5] *Let X be a random vector as in Theorem 3.1.1. Show that*

$$\mathrm{Var}(\|X\|_2) = O(1).$$

Exercise 3.1.5 (Sub-gaussian concentration squared)**.** [Difficulty=7] *Let X be a random variable with sub-gaussian concentration around its mean μ, say*

$$\|X - \mu\|_{\psi_2} \le 1.$$

What kind of concentration does X^2 have around μ^2? Give a tail bound.

3.2 Covariance matrices and isotropic distributions

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

The concept of the *mean* of random variables generalizes in a straightforward way for random vectors. The notion of variance is replaced in high dimensions by the *covariance matrix* of a random vector X, defined as follows:

$$\mathrm{cov}(X) = \mathbb{E}(X - \mu)(X - \mu)^{\mathsf{T}} = \mathbb{E}\, XX^{\mathsf{T}} - \mu\mu^{\mathsf{T}}$$

where $\mu = \mathbb{E}\, X$. Thus $\operatorname{cov}(X)$ is an $n \times n$ matrix. Note that the formula for covariance is a direct high-dimensional generalization of the definition of variance for random variables Z, which is

$$\operatorname{Var}(Z) = \mathbb{E}(Z - \mu)^2 = \mathbb{E}\, Z^2 - \mu^2, \quad \text{where } \mu = \mathbb{E}\, Z.$$

The entries of Σ are the covariances of the coordinates of $X = (X_1, \dots, X_n)$:

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

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

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

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

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

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

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

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

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

3.2.1 The Principal Component Analysis

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

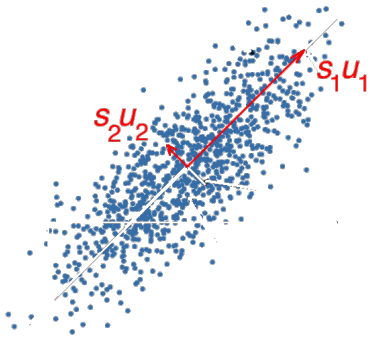


Figure 3.2: Illustration of the PCA. A few dozen samples are shown from a distribution in $\mathbb{R}^2$. The covariance matrix Σ has eigenvalues s_i and eigenvectors u_i. The vectors s_1u_1 and s_2u_2 are shown in red.

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

3.2.2 Isotropy

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

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

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

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

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

dilation, namely

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

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

Exercise 3.2.2 (Reduction to isotropy)**.** [Difficulty=3] *1. Let Z be a mean zero, isotropic random vector in $\mathbb{R}^n$. Let $\mu \in \mathbb{R}^n$ be a fixed vector and Σ be an $n \times n$ positive-semidefinite matrix. Then the random vector $X := \mu + \Sigma^{1/2} Z$ has mean μ and covariance matrix $\mathrm{cov}(X) = \Sigma$.*

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

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

3.2.3 Properties of isotropic distributions

Lemma 3.2.3 (Characterization of isotropy)**.** *A random vector X is isotropic if and only if*

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

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

$$x^{\mathsf{T}} \left(\mathbb{E}\, X X^{\mathsf{T}} \right) x = x^{\mathsf{T}} I_n x \quad \text{for all } x \in \mathbb{R}^n.$$

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

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

Lemma 3.2.4. *Let X be an isotropic random vector in $\mathbb{R}^n$. Then*

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

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

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

Proof. To prove the first part, we have

$$\begin{aligned}
\mathbb{E}\,\|X\|_2^2 &= \mathbb{E}\,X^{\mathsf{T}}X = \mathbb{E}\,\mathrm{tr}(X^{\mathsf{T}}X) \quad \text{(viewing } X^{\mathsf{T}}X \text{ as a } 1\times 1 \text{ matrix)} \\
&= \mathbb{E}\,\mathrm{tr}(XX^{\mathsf{T}}) \quad \text{(by invariance of trace under cyclic permutations)} \\
&= \mathrm{tr}(\mathbb{E}\,XX^{\mathsf{T}}) \quad \text{(by linearity)} \\
&= \mathrm{tr}(I_n) \quad \text{(by isotropy)} \\
&= n.
\end{aligned}$$

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

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

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

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

The proof is complete. □

3.2.4 Almost orthogonality of independent random vectors

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

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

Then we should expect from Lemma 3.2.4 that

$$\left|\langle \overline{X}, \overline{Y}\rangle\right| \sim \frac{1}{\sqrt{n}}$$

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

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

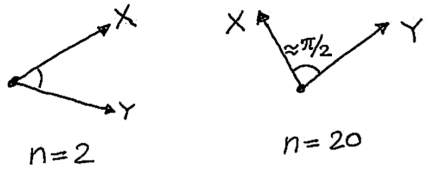


Figure 3.3: Independent isotropic random vectors tend to be almost orthogonal in high dimensions n but not in low dimensions.

3.3 Examples of high dimensional distributions

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

3.3.1 Spherical and Bernoulli distributions

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

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

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

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

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

The symmetric Bernoulli distribution is isotropic. (Check!)

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

3.3.2 Multivariate normal

The most important high dimensional distribution is arguably the Gaussian, or multivariate normal. From the basic probability course we recall that a

random vector $Z = (Z_1, \dots, Z_n)$ has *standard normal distribution* in $\mathbb{R}^n$, denoted

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

if the coordinates Z_i are independent standard normal random variables. The density of Z is then the product of the n standard normal densities (1.3), which is

$$f_Z(x) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-x_i^2/2} = \frac{1}{(2\pi)^{n/2}} e^{-\|x\|_2^2/2}, \quad x \in \mathbb{R}^n. \tag{3.3}$$

The standard normal distribution is *isotropic.* (Why?)

Note that the standard normal density is *rotation invariant.* In particular, for any fixed unitary matrix U,

$$Z \sim N(0, I_n) \quad \text{implies} \quad UZ \sim N(0, I_n).$$

Exercise 3.3.2 (Sum of independent gaussians)**.** [Difficulty=5] *Let $X_i \sim N(0, \sigma_i^2)$ be independent random variables. Deduce from rotation invariance that the following property we mentioned in (2.19):*

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

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

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

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

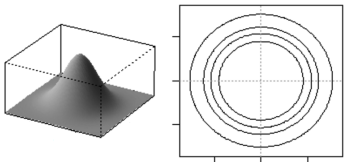
Summarizing, we have

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

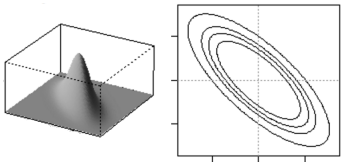
The density of $X \sim N(\mu, \Sigma)$ can be computed by change of variables formula, and it is

$$f_X(x) = \frac{1}{(2\pi)^{n/2} \det(\Sigma)^{1/2}} e^{-(x-\mu)^{\mathsf{T}} \Sigma^{-1} (x-\mu)/2}, \quad x \in \mathbb{R}^n. \tag{3.4}$$

Figure 3.4 shows examples of the densities of multivariate normal distributions.



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



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

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

3.3.3 Similarity of normal and spherical distributions

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

$$\mathbb{P}\left\{\left|\|X\|_2 - \sqrt{n}\right| \geq t\right\} \leq 2\exp(-ct^2) \quad \text{for all } t \geq 0. \tag{3.5}$$

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

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

$$g = r\theta$$

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

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

Concentration inequality (3.5) for the length says that $r \approx \sqrt{n}$ with high probability, so

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

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

$$N(0, I_n) \approx \text{Unif}\left(\sqrt{n} S^{n-1}\right). \tag{3.6}$$

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

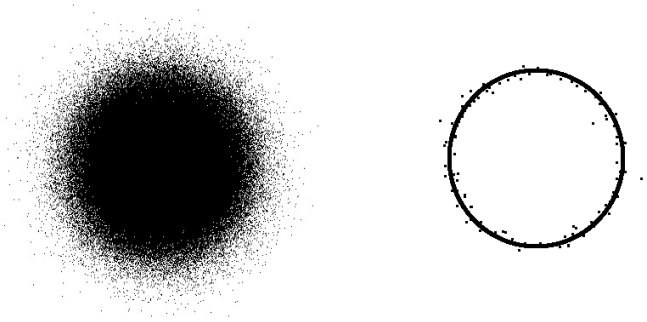


Figure 3.5: A Gaussian point cloud in two dimensions (left) and in high dimensions (right). In high dimensions, gaussian distribution is very close to the uniform distribution on the sphere of radius $\sqrt{n}$.

3.3.4 Frames

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

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

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 signal processing under the name of *frames*.

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

frame bounds *such that*

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

If $A = B$ the set is called a tight frame.

Exercise 3.3.5. *Show that $\{u_i\}_{i=1}^N$ is a tight frame in $\mathbb{R}^n$ with bound A if*

$$\sum_{i=1}^{N} u_i u_i^{\mathsf{T}} = AI_n. \tag{3.7}$$

Hint: Proceed similarly to the proof of Lemma 3.2.3.

Multiplying both sides of (3.7) by a vector x, we see that

$$\sum_{i=1}^{N} \langle u_i, x\rangle u_i = Ax \quad \text{for any } x \in \mathbb{R}^n. \tag{3.8}$$

This is a *frame expansion* of a vector x, and it should look familiar. Indeed, if $\{u_i\}$ is an orthonormal basis, then (3.8) is just a classical basis expansion of x, and it holds with $A = 1$.

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

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

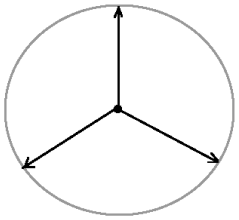


Figure 3.6: the Mercedes-Benz frame. A set of equidistant points on the circle form a tight frame in $\mathbb{R}^2$.

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

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

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

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

2. Consider an isotropic random vector X in $\mathbb{R}^n$ that takes a finite set of values x_i with probabilities p_i each, $i = 1, \dots, N$. 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. The assumptions and (3.7) imply that

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

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

2. Isotropy of X means that

$$\mathbb{E}\, XX^{\mathsf{T}} = \sum_{i=1}^N p_i x_i x_i^{\mathsf{T}} = I_n.$$

Denoting $u_i := \sqrt{p_i}\, x_i$, we obtain (3.7) with $A = I_n$. □

3.3.5 Isotropic convex sets

Our last example of a high dimensional distribution comes from convex geometry. Consider a bounded convex set K in $\mathbb{R}^n$ with non-empty interior; such sets are called *convex bodies*. Let X be a random vector uniformly distributed in K, according to the probability measure given by normalized volume in K:

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

Denote the covariance matrix of X by Σ. Then by Exercise 3.2.2, the random vector $Z := \Sigma^{-1/2} X$ is isotropic. Note that Z is uniformly distributed in the linearly transformed copy of K:

$$Z \sim \text{Unif}(\Sigma^{-1/2} K).$$

(Why?) Summarizing, we found a linear transformation $T := \Sigma^{-1/2}$ which makes the uniform distribution on TK isotropic. The body TK is sometimes called isotropic itself.

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

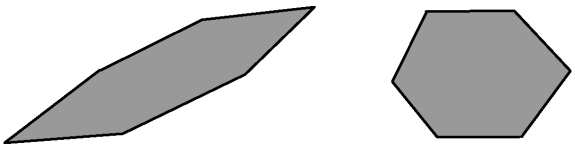


Figure 3.7: A convex body K on the left is transformed into an isotropic convex body TK on the right. The pre-conditioner T is computed from the covariance matrix Σ of K as $T = \Sigma^{-1/2}$.

3.4 Sub-gaussian distributions in higher dimensions

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

Can anyone suggest a reference for this characterization?

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*

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

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

$$\| \langle X, x \rangle \|_{\psi_2}^2 = \Big\| \sum_{i=1}^n x_i X_i \Big\|_{\psi_2}^2 \le C \sum_{i=1}^n x_i^2 \|X_i\|_{\psi_2}^2 \quad \text{(by Proposition 2.6.1)}$$

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

This completes the proof. □

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

1. Let $X = (X_1, \dots, X_n) \in \mathbb{R}^n$ 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

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

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

3.4.1 Gaussian and Bernoulli distributions

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

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

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

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

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

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:

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

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

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

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

$$|T| \geq e^{cn}.$$

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

3.4.3 Uniform distribution on the sphere

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

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

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

Then X is sub-gaussian, and

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

Proof. Consider a standard normal random vector $g \sim N(0, I_n)$. As we noted in Section 3.3.3, the direction $g/\|g\|_2$ is uniformly distributed on the unit sphere S^{n-1}. By rescaling, we can represent a random vector $X \sim \text{Unif}\left(\sqrt{n}\, S^{n-1}\right)$ 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 $x = e_1$. So it is enough to analyze $\langle X, x\rangle = X_1$ the first coordinate of X. Thus we want to bound the tail probability

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

Heuristically, 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 the tail $\mathbb{P}\left\{|g_1| \geq t\right\}$, but as we know from (2.2), this tail is sub-gaussian.

Let us do this argument more carefully. The concentration of norm, Theorem 3.1.1, implies that

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

Thus the event

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

is likely: by (2.15) its complement $\mathcal{E}^c$ has probability

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

Then the tail probability can be bounded as follows:

$$\begin{aligned} p(t) &\leq \mathbb{P}\left\{\frac{|g_1|}{\|g\|_2} \geq \frac{t}{\sqrt{n}} \text{ and } \mathcal{E}\right\} + \mathbb{P}(\mathcal{E}^c) \\ &\leq \mathbb{P}\left\{|g_1| \geq \frac{t}{2}\right\} + 2\exp(-cn) \\ &\leq 2\exp(-t^2/8) + 2\exp(-cn) \quad \text{(using (2.2)).} \end{aligned}$$

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.

□

Exercise 3.4.6 (Uniform distribution on the Euclidean ball)**.** [Difficulty=5] *Extend Theorem 3.4.5 for the uniform distribution on the Euclidean ball $B(0, \sqrt{n})$ in $\mathbb{R}^n$ centered at the origin and with radius $\sqrt{n}$. Namely, show that a random vector*

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

is sub-gaussian, and

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

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

Projective Limit Theorem

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

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

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

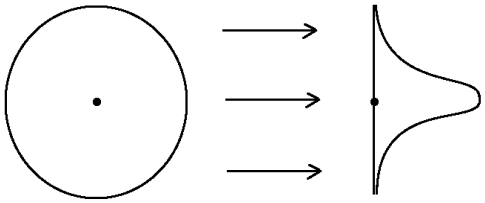


Figure 3.8: The Projective Central Limit Theorem. The projection of the uniform distribution on the sphere of radius $\sqrt{n}$ onto a line converges to the normal distribution $N(0,1)$ as $n \to \infty$.

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 is a convex body and

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

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

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

Exercise 3.4.8. [Difficulty=7] *Consider a ball of the ℓ_1 norm in $\mathbb{R}^n$:*

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

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

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

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

for all unit vectors x. This result follows from C. Borell's lemma, which itself is a consequence of Brunn-Minkowski inequality; see [**?**, Section 2.2.b$_3$].

Borrow reference from my tutorial on non-asymptotic RMT

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

Chapter 4

Sub-gaussian random matrices

4.1 Preliminaries on matrices

4.1.1 Singular value decomposition

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

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

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

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

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

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

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

In particular, if A is a *symmetric* matrix, the singular values of A are the absolute values of the eigenvalues λ_i of A:

$$s_i(A) = |\lambda_i(A)|,$$

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

4.1.2 Operator norm and the extreme singular values

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

When we think of the space $\mathbb{R}^m$ along with the Euclidean norm $\|\cdot\|_2$ on it, we denote this Hilbert space ℓ_2^m. The matrix A acts as a linear operator from $\ell_2^n \to \ell_2^m$. Its *operator norm* of A, also called the *spectral norm*, is then defined as

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

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

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

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

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

(Check!)

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

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

where $A^\dagger$ is the pseudo-inverse of A. Its norm $\|A^\dagger\|$ is the norm of the operator A^{-1} restricted to the image of A.

4.1.3 Frobenius norm

The *Frobenius norm*, also called *Hilbert-Schmidt* norm of a matrix A with entries A_{ij} is defined as

$$\|A\|_F = \Big(\sum_{i=1}^{m} \sum_{j=1}^{n} |A_{ij}|^2 \Big)^{1/2}.$$

Thus Frobenius norm is the Euclidean norm on the space of matrices $\mathbb{R}^{m\times n}$. In terms of spectrum, the Frobenius norm can be computed as

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

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

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

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 Frobenius norm. If we look at the vector $s = (s_1, \ldots, s_r)$ of singular values of A, these norms become the ℓ_∞ and ℓ_2 norms, respectively:

$$\|A\| = \|s\|_\infty, \quad \|A\|_F = \|s\|_2.$$

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

$$\|A\| \le \|A\|_F \le \sqrt{r}\, \|A\|. \tag{4.1}$$

4.1.4 Approximate isometries

The extreme singular values $s_1(A)$ and $s_r(A)$ have an important geometric meaning. They are respectively the smallest number M and the largest number m that make the following inequality true:

$$m\|x\|_2 \le \|Ax\|_2 \le M\|x\|_2 \quad \text{for all } x \in \mathbb{R}^n. \tag{4.2}$$

(Check!) Applying this inequality for $x - y$ instead of x and with the best bounds, we can rewrite is as

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

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

The best possible matrices in this sense, which preserve distances exactly, are called *isometries*. Let us recall their characterization, which can be proved using elementary linear algebra. (Do it!)

Lemma 4.1.1 (Isometries)**.** *Let A be an $m \times n$ matrix with $m \ge n$. Then the following are equivalent.*

1. A is an isometry*, or isometric embedding of $\mathbb{R}^n$ into $\mathbb{R}^m$. This means that*

$$\|Ax\|_2 = \|x\|_2 \quad \text{for all } x \in \mathbb{R}^n.$$

2. $A^{\mathsf{T}}A = I_n$.

3. All singular values of A equal 1; equivalently

$$s_n(A) = s_1(A) = 1.$$

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

Lemma 4.1.2 (Approximate isometries)**.** *Let A be an $m \times n$ matrix and $\delta > 0$. Suppose that*

$$\|A^{\mathsf{T}}A - I_n\| \le \max(\delta, \delta^2).$$

Then

$$(1 - \delta)\|x\|_2 \le \|Ax\|_2 \le (1 + \delta)\|x\|_2 \quad \text{for all } x \in \mathbb{R}^n. \tag{4.3}$$

Equivalently, all singular values of A are between $1 - \delta$ and $1 + \delta$:

$$1 - \delta \le s_n(A) \le s_1(A) \le 1 + \delta. \tag{4.4}$$

Proof. Without loss of generality, we may assume that $\|x\|_2 = 1$. (Why?) By assumption, we have

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

Applying the elementary inequality

$$\max(|z-1|, |z-1|^2) \le |z^2-1|, \quad z \ge 0 \tag{4.5}$$

for $z = \|Ax\|_2$, we conclude that

$$|\|Ax\|_2 - 1| \le \delta.$$

This proves (4.3), which in turn implies (4.4) as we saw before. □

Exercise 4.1.3 (Approximate isometries)**.** [Difficulty=3] *Prove the following converse to Lemma 4.1.2. If A is an approximate isometry, i.e.* (4.4) *holds, then*

$$\|A^{\mathsf{T}}A - I_n\| \le 2\max(\delta, \delta^2).$$

Suppose A is a *fat* $m \times n$ matrix, that is $m \le n$, and

$$s_1(A) \approx s_m(A) \approx 1.$$

Then A can be viewed as an *approximate projection* from $\mathbb{R}^m$ into $\mathbb{R}^n$. Thus A is an approximate isometry if and only if A^{T} is an approximate projection.

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

4.2 Nets, covering numbers and packing numbers

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

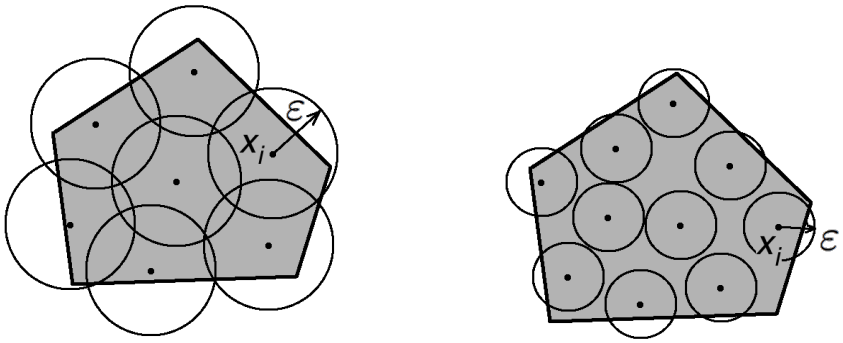
Definition 4.2.1 (ε-net)**.** *Consider a subset K of $\mathbb{R}^n$ and let $\varepsilon > 0$. A subset $\mathcal{N} \subseteq K$ is called an ε-net of K if every point in K is within distance ε of some point of $\mathcal{N}$, i.e.*

$$\forall x \in K \ \exists x_0 \in \mathcal{N} : \ \|x - x_0\|_2 \le \varepsilon.$$

Equivalently, $\mathcal{N} \subseteq K$ is an ε-net of K if and only if K can be covered by balls with centers in $\mathcal{N}$ and radii ε, see Figure 4.1a.

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

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



(a) This covering of a pentagon K by seven ε-balls shows that $\mathcal{N}(K, \varepsilon) \leq 7$.

(b) This packing of a pentagon K by ten ε-balls shows that $\mathcal{P}(K, \varepsilon) \geq 10$.

Figure 4.1: Packing and covering

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

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

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

Exercise 4.2.4. [Difficulty=3] *A collection of points $x_i \in K$ are centers of balls that form a packing in K if and only if*

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

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

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

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

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

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

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

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

This proves the lower bound.

Upper bound. Let $\mathcal{P} = \{x_i\}$ be a maximal packing of K by $\varepsilon/2$-balls. Here by "maximal" we mean that an addition of any $\varepsilon/2$-ball to $\mathcal{P}$ will always destroy the packing property; Figure 4.1b shows an example of a maximal packing. Equivalently, by Exercise 4.2.4, $\mathcal{P}$ is a maximal ε-separated set of points $x_i \in K$:

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

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

Exercise 4.2.6 (Allowing the centers to be outside K)**.** [Difficulty=5] *In our definition of covering numbers of K, we required that the centers x_i of the balls $B(x_i, \varepsilon)$ that form covering lie in K. Relaxing this condition, define* external covering number $\mathcal{N}^{\text{ext}}(K, \varepsilon)$ *similarly but without requiring that $x_i \in K$. Prove that*

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

Exercise 4.2.7 (Monotonicity of covering numbers)**.** [Difficulty=6] *Give a counterexample to the monotonicity property*

$$L \subset K \quad \textit{implies} \quad \mathcal{N}(L, \varepsilon) \leq \mathcal{N}(K, \varepsilon).$$

Prove an approximate version of monotonicity:

$$L \subset K \quad \textit{implies} \quad \mathcal{N}(L, \varepsilon) \leq \mathcal{N}(K, \varepsilon/2).$$

4.2.1 Metric entropy and coding

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

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

$$\log_2 \mathcal{N}(K, \varepsilon) \le \mathcal{C}(K, \varepsilon) \le \log_2 \mathcal{N}(K, \varepsilon/2).$$

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

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

4.2.2 Covering numbers and volume

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

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

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

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

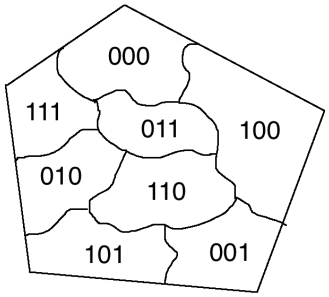


Figure 4.2: Encoding points in K as bit strings of length N induces a partition of K into at most 2^N subsets.

Figure 4.3 shows an example of Minkowski sum of two sets on the plane.

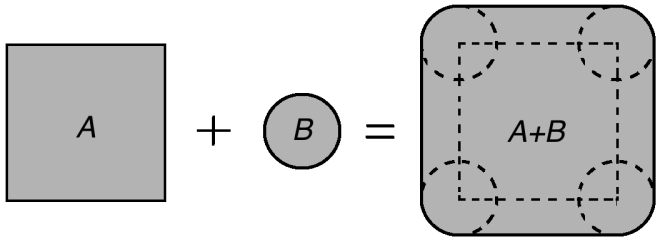


Figure 4.3: Minkowski sum of a square and a circle is a square with rounded corners.

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

$$\frac{\mathrm{Vol}(K)}{\mathrm{Vol}(\varepsilon B_2^n)} \leq \mathcal{N}(K, \varepsilon) \leq \frac{\mathrm{Vol}\left(K + (\varepsilon/2)B_2^n\right)}{\mathrm{Vol}((\varepsilon/2)B_2^n)}.$$

Here as usual B_2^n denotes the unit Euclidean ball in $\mathbb{R}^n$, so εB_2^n is the Euclidean ball with radius ε.

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

$$\mathrm{Vol}(K) \leq N \cdot \mathrm{Vol}(\varepsilon B_2^n),$$

which proves the lower bound in the proposition.

(Upper bound) By Lemma 4.2.5, it is enough to prove the same upper bound but for the packing number $\mathcal{P}(K, \varepsilon/2) =: P$. Consider P open disjoint balls $B(x_i, \varepsilon/2)$ with centers $x_i \in K$ and radii $\varepsilon/2$. While these balls do not

need to fit entirely in K (see Figure 4.1b), they fit in a slightly inflated set, namely $K + (\varepsilon/2)B_2^n$. (Why?) Comparing the volumes, we obtain

$$P \cdot \mathrm{Vol}((\varepsilon/2)B_2^n) \le \mathrm{Vol}\left(K + (\varepsilon/2)B_2^n\right),$$

which leads to the upper bound in the proposition. □

Let us give some examples for the volumetric bound.

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

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

The same upper bound is true for the unit Euclidean sphere S^{n-1}.

Proof. The lower bound follows immediately from Proposition 4.2.10, since the volume scales as $\mathrm{Vol}(\varepsilon B_2^n) = \varepsilon^n \cdot \mathrm{Vol}(B_2^n)$. The upper bound follows from Proposition 4.2.10, too:

$$\mathcal{N}(K, \varepsilon) \le \frac{\mathrm{Vol}\left((1 + \varepsilon/2)B_2^n\right)}{\mathrm{Vol}((\varepsilon/2)B_2^n)} = \frac{(1 + \varepsilon/2)^n}{(\varepsilon/2)^n} = \left(\frac{2}{\varepsilon} + 1\right)^n.$$

The upper bound for the sphere can be proved in the same way. □

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

$$\left(\frac{1}{\varepsilon}\right)^n \le \mathcal{N}(B_2^n, \varepsilon) \le \left(\frac{3}{\varepsilon}\right)^n. \tag{4.6}$$

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

The important message of (4.6) is that the covering numbers are *exponential in the dimension n*. This should not be surprising if we recall the coding theory perspective we discussed in Section 4.2.1. Indeed, to encode a vector in dimension n, one should be prepared to spend at least one bit per coefficient, so n bits total. This makes the metric entropy linear in n, and the covering numbers exponential in n.

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

We need to define covering numbers for general metric spaces somewhere. Either here or in Section 7.5 where we use it for the first time.

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

Write this exercise more clearly

4.3 Upper bounds on sub-gaussian matrices

4.3.1 Computing the norm on a net

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

$$\|A\| = \sup_{x \in S^{n-1}} \|Ax\|_2.$$

Thus, to evaluate $\|A\|$ one needs to control $\|Ax\|$ uniformly over the sphere S^{n-1}. We will show that instead of the entire sphere, it is enough to have a control just over an ε-net of the sphere.

Lemma 4.3.1 (Computing the operator norm on a net)**.** *Let A be an $m \times n$ matrix and $\varepsilon \in [0, 1)$. Then, for any ε-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 triangle inequality, this implies

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

Using the triangle inequality again, 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. □

Exercise 4.3.2. *Let $x \in \mathbb{R}^n$ and $\mathcal{N}$ be an ε-net of the sphere S^{n-1}. Show that*

$$\sup_{y \in \mathcal{N}} \langle x, y \rangle \leq \|x\|_2 \leq \frac{1}{1-\varepsilon} \sup_{y \in \mathcal{N}} \langle x, y \rangle .$$

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

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

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

Exercise 4.3.3. [Difficulty=4] *1. Let A be an $m \times n$ matrix and $\varepsilon \in [0, 1/2)$. Show that for any ε-net $\mathcal{N}$ of the sphere S^{n-1} and any ε-net $\mathcal{M}$ of the sphere S^{m-1}, we have*

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

2. Moreover, if $n = m$ and A is symmetric, show that

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

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

Exercise 4.3.4 (Bounding the norm deviation on a net)**.** [Difficulty=7] *1. Let A be an $m \times n$ matrix, $\mu \in \mathbb{R}$ and $\varepsilon \in [0, 1/2)$. Show that for any ε-net $\mathcal{N}$ of the sphere S^{n-1}, we have*

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

Hint: Assume that $\mu = 1$ without loss of generality. Represent $\|Ax\|_2^2 - 1$ as a quadratic form $\langle Rx, x \rangle$ where $R = A^{\mathsf{T}} A - I_n$. Use Exercise 4.3.3 to compute the maximum of this quadratic form on a net.

4.3.2 The norms of sub-gaussian random matrices

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

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

with high probability.

Theorem 4.3.5 (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*

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

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

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

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

$$|\mathcal{N}| \le 9^n \quad \text{and} \quad |\mathcal{M}| \le 9^m. \tag{4.7}$$

By Exercise 4.3.3, the operator norm of A can be bounded using these nets as follows:

$$\|A\| \le 2 \max_{x \in \mathcal{N},\, y \in \mathcal{M}} \langle Ax, y\rangle . \tag{4.8}$$

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

$$\begin{aligned} \|\langle Ax, y\rangle\|_{\psi_2}^2 &\le C \sum_{i=1}^{n} \sum_{j=1}^{m} \|A_{ij} x_i y_j\|_{\psi_2}^2 \le CK^2 \sum_{i=1}^{n} \sum_{j=1}^{m} x_i^2 y_i^2 \\ &= CK^2 \Big(\sum_{i=1}^{n} x_i^2 \Big) \Big(\sum_{j=1}^{m} y_i^2 \Big) = CK^2. \end{aligned}$$

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

$$\mathbb{P}\left\{\langle Ax, y\rangle \ge u\right\} \le 2\exp(-cu^2/K^2), \quad u \ge 0. \tag{4.9}$$

Step 3: Union bound. Next, we will unfix x and y using a union bound. Suppose the event $\max_{x \in \mathcal{N},\, y \in \mathcal{M}} \langle Ax, y\rangle \ge u$ 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

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

Using the tail bound (4.9) and the estimate (4.7) on the sizes of $\mathcal{N}$ and $\mathcal{M}$, we bound the probability above by

$$9^{n+m} \cdot 2\exp(-cu^2/K^2). \tag{4.10}$$

Choose

$$u = CK(\sqrt{n} + \sqrt{m} + t). \tag{4.11}$$

Then $u^2 \geq C^2K^2(n + m + t)$, and if constant C is chosen sufficiently large, the exponent in (4.10) is large enough, say $cu^2/K^2 \geq 3(n + m) + t^2$. Thus

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

Finally, combining this with (4.8), we conclude that

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

Recalling our choice of u in (4.11), we complete the proof. □

Optimality

Theorem 4.3.5 states that

$$\|A\| \lesssim \sqrt{m} + \sqrt{n} \tag{4.12}$$

with high probability. Is this bound optimal?

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

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

so the upper bound (4.12) has the optimal form.

Symmetric matrices

Theorem 4.3.5 can be easily extended for symmetric matrices, and the bound for them is

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

with high probability.

Corollary 4.3.6 (Norm of symmetric matrices with sub-gaussian entries)**.** *Let A be an $n \times n$ symmetric random matrix whose entries A_{ij} on and above diagonal are independent, mean zero, sub-gaussian random variables. Then, for any $t > 0$ we have*

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

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

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

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

Theorem 4.3.5 applies for each part A^+ and A^- separately. By union bound, we have simultaneously

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

with probability at least $1 - 4\exp(-t^2)$. Since by triangle inequality $\|A\| \le \|A^+\| + \|A^-\|$, the proof is complete. □

4.4 Application: community detection in networks

We are going to illustrate Corollary 4.3.6 with an application to the analysis of networks.

Real-world networks tend to have *communities*, or clusters, of tightly connected vertices. Finding the communities accurately and efficiently is one of the main problems in network analysis.

4.4.1 Stochastic Block Model

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

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

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

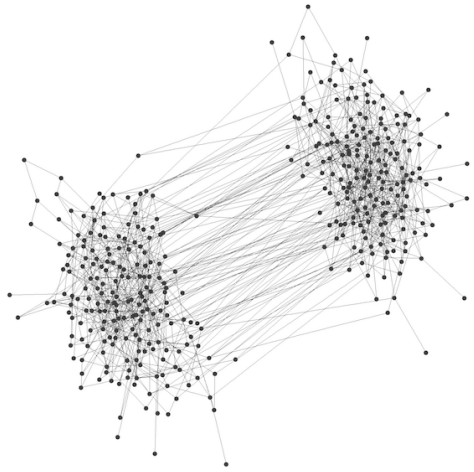


Figure 4.4: A random graph generated according to the stochastic block model $G(n, p, q)$.

4.4.2 Expected adjacency matrix

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

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

$$A = D + R,$$

[1] The term *stochastic block model* can also refer a more general model of random graphs with multiple communities of variable sizes.

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

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

$$D = \mathbb{E}\, A = \left[\begin{array}{cc|cc} p & p & q & q \\ p & p & q & q \\ \hline q & q & p & p \\ q & q & p & p \end{array}\right]$$

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

$$\lambda_1 = \Big(\frac{p+q}{2}\Big)n, \quad u_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}; \quad \lambda_2 = \Big(\frac{p-q}{2}\Big)n, \quad u_2 = \begin{bmatrix} 1 \\ 1 \\ -1 \\ -1 \end{bmatrix}. \tag{4.13}$$

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

But we do not know $D = \mathbb{E}\, A$ 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

$$\|D\| = \lambda_1 \sim n$$

while the level of the noise R can be estimated using Corollary 4.3.6:

$$\|R\| \le C\sqrt{n} \quad \text{with probability at least } 1 - 4e^{-n}. \tag{4.14}$$

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

4.4.3 Perturbation theory

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

$$\max_i |\lambda_i(S) - \lambda_i(T)| \le \|S - T\|.$$

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

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

Check!

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

$$\min\left(|\lambda_i(S) - \lambda_j(S)| : \ j \neq i\right) = \delta > 0.$$

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

$$\sin \angle\left(v_i(S),\, v_i(T)\right) \leq \frac{C\|S - T\|}{\delta}.$$

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

$$\exists \theta \in \{-1, 1\} : \quad \|v_i(S) - \theta v_i(T)\|_2 \leq \frac{C\|S - T\|}{\delta}. \tag{4.15}$$

4.4.4 Spectral Clustering

Let us apply Davis-Kahan Theorem for $S = D$ and $T = A = D + R$, and for the second largest eigenvalue. We need to check that λ_2 is well separated from the rest of the spectrum of D, that is from 0 and λ_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 (4.14) on $R = T - S$ and applying (4.15), we can bound the distance between the normalized eigenvectors of D and A. There exists a sign $\theta \in \{-1, 1\}$ such that

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

with probability at least $1 - 4e^{-n}$. We computed the eigenvectors of D in (4.13), but there they had norm $\sqrt{n}$. So, multiplying both sides by $\sqrt{n}$, we

obtain in this normalization that

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

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

$$\sum_{j=1}^{n} |u_2(D)_j - \theta u_2(A)_j|^2 \leq \frac{C}{\mu^2}. \tag{4.16}$$

and we also know from (4.13) that the coefficients $u_2(D)_j$ are all ± 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 (4.16). Thus the number of disagreeing signs must be bounded by

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

Summarizing, we can use the vector $v_2(A)$ that can be computed from the network to accurately estimate the vector $v_2 = v_2(D)$ in (4.13), whose signs identify the two communities. This method for community detection is usually called em *spectral clustering*. Let us explicitly state this method and the guarantees that we just obtained.

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

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

4.5 Two-sided bounds on sub-gaussian matrices

Let us return to Theorem 4.3.5, which gives an upper bound on the spectrum of an $n \times m$ matrix A with independent sub-gaussian entries:

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

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

First, we are going to prove sharper, *two-sided bounds* on the entire spectrum of A:

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

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

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

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

$$\sqrt{m} - CK^2(\sqrt{n} + t) \le s_n(A) \le s_1(A) \le \sqrt{m} + CK^2(\sqrt{n} + t) \qquad (4.17)$$

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

Before we prove this theorem, let us note that by Lemma 4.1.2, we can equivalently restate the conclusion (4.17) in the following form:

$$\Big\| \frac{1}{m} A^{\mathsf{T}} A - I_n \Big\| \le \max(\delta, \delta^2) \quad \text{where} \quad \delta = CK^2\Big(\sqrt{\frac{n}{m}} + \frac{t}{\sqrt{m}}\Big). \qquad (4.18)$$

Proof. We will prove (4.18) using an *ε-net argument*. This will be similar to the proof of Theorem 4.3.5, but we will now use Bernstein's concentration inequality instead of Hoeffding's.

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

$$|\mathcal{N}| \le 9^n.$$

Using Lemma 4.3.1, we can evaluate the operator norm in (4.18) on the $\mathcal{N}$:

$$\Big\| \frac{1}{m} A^*A - I_m \Big\| \le 2 \max_{x \in \mathcal{N}} \Big| \Big\langle \Big(\frac{1}{m} A^*A - I \Big) x, x \Big\rangle \Big| = 2 \max_{x \in \mathcal{N}} \Big| \frac{1}{m} \|Ax\|_2^2 - 1 \Big|.$$

To complete the proof of (4.18) it suffices to show that, with the required probability,

$$\max_{x \in \mathcal{N}} \Big| \frac{1}{m} \|Ax\|_2^2 - 1 \Big| \le \frac{\varepsilon}{2} \quad \text{where} \quad \varepsilon := \max(\delta, \delta^2).$$

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

$$\|Ax\|_2^2 = \sum_{i=1}^m \langle A_i, x \rangle^2 =: \sum_{i=1}^m X_i^2 \tag{4.19}$$

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

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

(Check this; we did a similar computation in the proof of Theorem 3.1.1.) Thus we can use Bernstein's inequality (Corollary 2.8.4) and obtain

$$\begin{aligned}
\mathbb{P} \left\{ \Big| \frac{1}{m} \|Ax\|_2^2 - 1 \Big| \ge \frac{\varepsilon}{2} \right\} &= \mathbb{P} \left\{ \Big| \frac{1}{m} \sum_{i=1}^m X_i^2 - 1 \Big| \ge \frac{\varepsilon}{2} \right\} \\
&\le 2 \exp \Big[- \frac{c_1}{K^4} \min(\varepsilon^2, \varepsilon) m \Big] \\
&= 2 \exp \Big[- \frac{c_1}{K^4} \delta^2 m \Big] \quad (\text{since } \varepsilon = \max(\delta, \delta^2)) \\
&\le 2 \exp \big[- c_1 C^2 (n + t^2) \big].
\end{aligned}$$

The last bound follows from the definition of δ in (4.18) and using the inequality $(a+b)^2 \ge a^2 + b^2$ for $a, b \ge 0$.

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

$$\mathbb{P}\Big\{ \max_{x \in \mathcal{N}} \Big| \frac{1}{N} \|Ax\|_2^2 - 1 \Big| \geq \frac{\varepsilon}{2} \Big\} \leq 9^n \cdot 2 \exp\big[- c_1 C^2 (n + t^2) \big] \leq 2 \exp(-t^2)$$

if we chose absolute constant C in (4.18) large enough. As we noted in Step 1, this completes the proof of the theorem. □

Exercise 4.5.2. [Difficulty=6] *Give a simpler proof of Theorem 4.5.1, using Theorem 3.1.1 to obtain a concentration bound for $\|Ax\|_2$ and Exercise 4.3.4 to reduce to a union bound over a net.*

Let us emphasize the alternative form (4.18) of Theorem 4.5.1, which is important on its own. Replacing there t with $t\sqrt{n}$ and assuming that $t \geq 1$, we can restate it as follows:

$$\Big\| \frac{1}{m} A^{\mathsf{T}} A - I_n \Big\| \leq \max(\delta, \delta^2) \quad \text{where} \quad \delta = CK^2 t \sqrt{\frac{n}{m}} \tag{4.20}$$

with probability at least $1 - 2\exp(-t^2 n)$.

Exercise 4.5.3 (Non-isotropic distributions)**.** [Difficulty=7] *Prove the following version of (4.20) for non-isotropic distributions. Let A be an $m \times n$ matrix whose rows A_i are independent, mean zero, sub-gaussian random vectors in $\mathbb{R}^n$ with the same covariance matrix*

$$\Sigma = \mathbb{E}\, A_i A_i^{\mathsf{T}}.$$

Then for any $t \geq 0$ we have

$$\Big\| \frac{1}{m} A^{\mathsf{T}} A - \Sigma \Big\| \leq \max(\delta, \delta^2) \quad \textit{where} \quad \delta = CLt \sqrt{\frac{n}{m}}$$

with probability at least $1 - 2\exp(-t^2 n)$. Here $L = \max(K, K^2)$ and $K = \max_i \|A_i\|_{\psi_2}$.

Check!

4.6 Application: covariance estimation and clustering

Suppose we are analyzing high dimensional data, which is represented as points $X_1, \dots, X_m$ sampled from an unknown distribution in $\mathbb{R}^n$. The most basic method is Principal Component Analysis (PCA), which we discussed

briefly in Section 3.2.1. The goal of PCA is to identify the *principal components* the eigenvectors of the covariance matrix of the distribution.

Since we do not have access to the full distribution but only to the finite sample $\{X_1, \dots, X_m\}$, we can only expect to compute the covariance matrix and its eigenvectors approximately. How can we do this? Let X denote the random vector drawn from the (unknown) distribution. Assume for simplicity that X have zero mean, and let us denote the covariance matrix

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

To estimate Σ, we can use the *sample covariance* matrix Σ_m that is computed from the sample $X_1, \dots, X_m$ as follows:

$$\Sigma_m = \frac{1}{m} \sum_{i=1}^{m} X_i X_i^{\mathsf{T}}.$$

(Basically, we replace the expectation over the entire distribution by expectation over the sample.)

Since X_i and X are identically distributed, our estimate is unbiased, that is

$$\mathbb{E}\, \Sigma_m = \Sigma.$$

Moreover, by the Law of Large Numbers (Theorem 1.3.1),

$$\Sigma_m \to \Sigma \quad \text{almost surely}$$

as the sample size m increases to infinity. (Justify that we can apply the Law of Larde Numbers to matrices.)

This leads to the quantitative question: how many sample points m are needed to guarantee that

$$\Sigma_m \approx \Sigma$$

with high probability? For dimension reasons, we need at least $m \gtrsim n$ sample points. (Why?) And we will now show that $m \sim n$ sample points suffice.

Theorem 4.6.1 (Covariance estimation)**.** *Consider a sub-gaussian random vector X in $\mathbb{R}^n$ with zero mean and covariance matrix Σ, and let $\varepsilon \in (0, 1)$ and $t \geq 1$. Suppose the sample size satisfies*

$$m \geq CK^4 (t/\varepsilon)^2 n.$$

Then the sample covariance matrix Σ_m satisfies

$$\|\Sigma_m - \Sigma\| \leq \varepsilon$$

with probability at least $1 - 2\exp(-t^2 n)$. Here $K = \|X\|_{\psi_2}$.

Proof. Consider the $m \times n$ matrix A whose rows are the sample points X_i^T. Then the sample covariance matrix Σ can be represented as

$$\Sigma_m = \sum_{i=1}^{m} X_i X_i^\mathsf{T} = \frac{1}{m} A^\mathsf{T} A.$$

So we can apply the non-isotropic form of Theorem 4.5.1 stated in Exercise 4.5.3 to bound the error $\|\Sigma_m - \Sigma\|$. Setting the error bound $\max(\delta, \delta^2)$ to ε and solving for m, we complete the proof. □

Let us emphasize the meaning of this result. *The covariance matrix can be estimated accurately by the sample covariance matrix, if the size of the sample m is proportional to the dimension n.*

4.6.1 Application: clustering of point sets

We are going to illustrate Theorem 4.6.1 with an application to clustering. The problem here will be similar to the community detection problem we studied in Section 4.4. Like before, we will try to partition data into clusters, but the nature of data will be different. Instead of networks, we will now be working with point sets in $\mathbb{R}^n$. The general goal is to partition the points into few subsets ("clusters"). What exactly constitutes cluster is not well defined in data sciences. But the common sense suggests that the points in the same cluster should tend to be closer to each other than points taken from different clusters.

Just like we did for networks, we will design a basic probabilistic model of point sets in $\mathbb{R}^n$ with two communities, and we will study the clustering problem for that model.

Definition 4.6.2 (Gaussian mixture model)**.** *Generate m random points in $\mathbb{R}^n$ as follows. Flip a fair coin. If we get heads, draw a point from $N(\mu, I_n)$, and if we get tails, from $N(-\mu, I_n)$. We call this a Gaussian mixture model with means μ and $-\mu$.*

Equivalently, we may consider a random vector

$$X = \theta\mu + g$$

where θ is a symmetric Bernoulli random variable, $g \in N(0, I_n)$, and θ 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

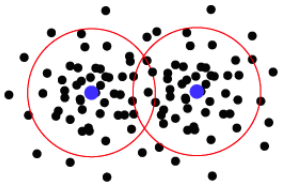


Figure 4.5: A simulation of points generated according to the Gaussian mixture model, which has two clusters with different means.

the Gaussian mixture model. Figure 4.5 illustrates a simulation of Gaussian mixture model.

A basic clustering method, called *spectral clustering*, is based on the Principal Component Analysis (PCA) of the data, which we outlined in Section 3.2.1. The distribution of X is not isotropic; it is stretched in the direction of μ. The first principal component of the data should also be close to μ, and thus one should be able classify the data points by projecting them onto that first principal component. This is formalized in the following basic clustering algorithm.

Spectral Clustering Algorithm

Input: points $X_1, \dots, X_m$ in $\mathbb{R}^n$

Output: a partition of the points into two clusters

1: Compute the sample covariance matrix $\Sigma_m = \frac{1}{m} \sum_{i=1}^m X_i X_i^{\mathsf{T}}$.

2: Compute the eigenvector $v = v_1(\Sigma_m)$ corresponding to the largest eigenvalue of Σ_m.

3: Partition the vertices into two communities based on the signs of the inner product of v with the data points. (To be specific, if $\langle v, X_i \rangle > 0$ put point X_i into first community, otherwise in the second.)

Theorem 4.6.3 (Spectral clustering of the Gaussian mixture model)**.** *Let $X_1, \dots, X_m$ be points in $\mathbb{R}^n$ drawn from the Gaussian mixture model as above, i.e. there are two communities with means μ and $-\mu$, and let $\varepsilon, t > 0$. Suppose the sample size satisfies*

$$m \geq \text{poly}\Big(n, \frac{1}{\varepsilon}, \frac{1}{\|\mu\|_2}\Big).$$

Then, with probability at least $1 - 4e^{-n}$, the Spectral Clustering Algorithm identifies the communities correctly up to εm misclassified vertices.

Correct the statement: ε should depend on μ.

Exercise 4.6.4. [Difficulty=8] *Prove Theorem 4.6.3 along the following lines.*

1. Compute the covariance matrix Σ of X and note that the eigenvector corresponding to the largest eigenvalue is parallel to μ.

2. Use results about covariance estimation (such as Exercise 4.5.3) to show that the sample covariance matrix Σ_m is close to Σ.

3. Use Davis-Kahan Theorem 4.4.3 to deduce that the eigenvector $v = v_1(\Sigma_m)$ is close to μ.

4. Conclude that the signs of $\langle \mu, X_i \rangle$ predict well which community X_i belongs to.

5. Since $v \approx \mu$, conclude the same for v.

Chapter 5

Concentration without independence

Our approach to concentration was crucially based on independence of random variables. This was clearly the case for sums of independent random variables we studied in Chapter 2; our later results were based on concentration for these sums. We will now develop alternative approaches to concentration that are not based on independence.

5.1 Concentration of Lipschitz functions on the sphere

Consider a Gaussian random vector $X \sim N(0, I_n)$ and a function $f : \mathbb{R}^n \to \mathbb{R}$. When does the random vector $f(X)$ concentrate about its mean, i.e.

$$f(X) \approx \mathbb{E} f(X) \quad \text{with high probability?}$$

This question is easy for *linear functions* f. Indeed, in this case $f(X)$ has normal distribution, and it concentrates around its mean well.

In this section, we will study concentration of *non-linear functions* $f(X)$ of random vectors $X \sim \text{Unif}(S^{n-1})$ and $X \sim N(0, I_n)$. While we can not expect concentration for completely arbitrary f (why?), the Lipschitz requirement for f will be enough.

5.1.1 Lipschitz functions

Definition 5.1.1 (Lipschitz functions)**.** *Let (X, d_X) and (Y, d_Y) be metric spaces. A function $f : X \to Y$ is called* Lipschitz *if there exists $L \in \mathbb{R}$ such*

that

$$d_X(f(u), f(v)) \le d_Y(u, v) \quad \text{for every } u, v \in X.$$

The infimum of all L in this definition is called the Lipschitz norm *of f and is denoted $\|f\|_{\mathrm{Lip}}$.*

Lipschitz functions with $\|f\|_{\mathrm{Lip}} \le 1$ are usually called *contractions*.

Exercise 5.1.2. *1. $f(x) = |x|$ is a Lipschitz a function on $\mathbb{R}$, while $f(x) = \sqrt{x}$ and $f(x) = x^2$ are not.*

2. $f(x) = \|x\|_2$ is a Lipschitz function on $\mathbb{R}^n$, and $\|f\|_{\mathrm{Lip}} = 1$.

3. Every differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is Lipschitz, and

$$\|f\|_{\mathrm{Lip}} \le \|\nabla f\|_\infty.$$

4. For a fixed $\theta \in \mathbb{R}^n$, the linear functional $f(x) = \langle x, \theta \rangle$ is a Lipschitz function on $\mathbb{R}^n$, and $\|f\|_{\mathrm{Lip}} = \|\theta\|_2$.

5. More generally, an $m \times n$ matrix A acting as a linear operator between $A : (\mathbb{R}^n, \|\cdot\|_2) \to (\mathbb{R}^m, \|\cdot\|_2)$ is Lipschitz, and

$$\|A\|_{\mathrm{Lip}} = \|A\|.$$

6. Any norm $f(x) = \|x\|$ on $\mathbb{R}^n$ is a Lipschitz function. The Lipschitz norm of f is the smallest L such that

$$\|x\| \le L\|x\|_2 \quad \text{for all } x \in \mathbb{R}^n.$$

5.1.2 Concentration via isoperimetric inequalities

The main result of this section is that any Lipschitz function on the sphere concentrates well.

Theorem 5.1.3 (Concentration of Lipschitz functions on the sphere)**.** *Consider a random vector $X \sim \mathrm{Unif}(\sqrt{n}S^{n-1})$ and a Lipschitz function[1] $f : \sqrt{n}S^{n-1} \to \mathbb{R}$. Then*

$$\|f(X) - \mathbb{E}\, f(X)\|_{\psi_2} \le C\|f\|_{\mathrm{Lip}}.$$

[1] This theorem is valid for both the geodesic metric on the sphere (where $d(x, y)$ is the length of the shortest arc connecting x and y) and the Euclidean metric $d(x, y) = \|x - y\|_2$. We will prove the theorem for the Euclidean metric; Exercise **??** extends it to geodesic metric.

Equivalently, Theorem 5.1.3 states that for every $t \geq 0$, we have

$$\mathbb{P}\left\{|f(X) - \mathbb{E}\, f(X)| \geq t\right\} \leq 2\exp\Big(-\frac{ct^2}{\|f\|_{\text{Lip}}^2}\Big)$$

We already know that Theorem 5.1.3 holds for linear functions. Indeed, Theorem 3.4.5 states that $X \sim \text{Unif}(\sqrt{n}S^{n-1})$ is a sub-gaussian random vector. By definition, this means that any linear function of X is a sub-gaussian random variable.

To prove Theorem 5.1.3 in full generality, we will argue that non-linear functions must concentrate at least as good as linear functions. To compare non-linear to linear functions, it is enough to compare their sub-level sets: arbitrary sets on the sphere and the spherical caps. Such comparison will be based on a geometric principle called *isoperimetric inequality*.

The classical isoperimetric inequality in $\mathbb{R}^3$ (and also in $\mathbb{R}^n$) states that among all sets with given volume, the area is minimal for the Euclidean balls. A similar isoperimetric inequality holds on the sphere, and the minimizers are the spherical caps. To state it carefully, we will denote the normalized area ($n-1$-dimensional Lebesgue measure) on the sphere S^{n-1} by σ_{n-1}. An ε-neighborhood of a set $A \subset S^{n-1}$ is defined as

$$A_\varepsilon := \left\{x \in S^{n-1} :\ \exists y \in A,\ \|x-y\|_2 \leq \varepsilon\right\} = (A + \varepsilon B_2^n) \cap S^{n-1}. \qquad (5.1)$$

Here we used the notation for Minkowski sum introduced in Definintion 4.2.9. Figure 5.1 illustrates the ε-neighborhood of A. The perimeter of A is the

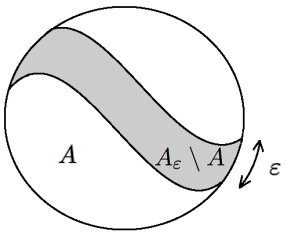


Figure 5.1: The points that are within Euclidean distance ε from a given set A on the sphere S^{n-1} form the ε-neighborhood A_ε.

$(n-2)$-dimensional area of boundary ∂A, and it can be defined as

$$\text{Area}(\partial A) := \lim_{\varepsilon \to 0} \frac{\sigma_{n-1}(A_\varepsilon) - \sigma_{n-1}(A)}{\varepsilon}.$$

Theorem 5.1.4 (Isoperimetric inequality on the sphere)**.** *1. Let $\varepsilon > 0$. Then, among all sets $A \subset S^{n-1}$ with fixed area $\sigma_{n-1}(A)$, the spherical caps minimize the area of the neighborhood $\sigma_{n-1}(A_\varepsilon)$.*

2. Among all sets $A \subset S^{n-1}$ with fixed area $\sigma_{n-1}(A)$, the spherical caps minimize the perimeter $\mathrm{Area}(\partial A)$.

We will not prove this theorem but just note that Part 2 follows from part 1 by letting $\varepsilon \to 0$. (Check!)

Refer

5.1.3 Blow-up of sets on the sphere

Now we will deduce form isoperimetric a statement that may sound counter-intuitive. If A makes up at least half of the sphere (in terms of volume) then A_ε will make up most of the sphere. This fact nevertheless is simple to check for a hemisphere, and then extend to general sets using isoperimetric inequality. In view of Theorem 5.1.3, it will be convenient to work with the sphere that is scaled by the factor $\sqrt{n}$.

Lemma 5.1.5 (Blow-up of neighborhoods on the sphere)**.** *Let A be a subset of the sphere $\sqrt{n}S^{n-1}$, and let σ denote the normalized area on that sphere. If $\sigma(A) \geq 1/2$ then, for every $t \geq 0$,*

$$\sigma(A_t) \geq 1 - 2\exp(-ct^2).$$

Proof. Consider the hemisphere

$$H := \left\{x \in \sqrt{n}S^{n-1} : \ x_1 \leq 0\right\}.$$

Then $\sigma(A) \geq \sigma(H) = 1/2$, and the isoperimetric inequality (Theorem 5.1.4) implies that

$$\sigma(A_t) \geq \sigma(H_t). \tag{5.2}$$

The set H_t is a spherical cap, and it should be easy to compute its area. It is even easier to use Theorem 3.4.5 instead, which states a random vector

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

is sub-gaussian, and $\|X\|_{\psi_2} \leq C$. Since σ is the uniform probability measure on the sphere, it follows that

$$\sigma(H_t) = \mathbb{P}\left\{X \in H_t\right\}.$$

Now, the definition of the neighborhood (5.1) implies that

$$H_t \supset \left\{x \in \sqrt{n}S^{n-1} : \ x_1 \leq \sqrt{2}t\right\}.$$

(Check this!) Thus

$$\sigma(H_t) \geq \mathbb{P}\left\{X_1 \leq \sqrt{2}t\right\} \geq 1 - 2\exp(-ct^2).$$

The last inequality holds because $\|X_1\|_{\psi_2} \leq \|X\|_{\psi_2} \leq C$. In view of (5.2), the lemma is proved. □

5.1.4 Proof of Theorem 5.1.3

Without loss of generality, we can assume that $\|f\|_{\text{Lip}} = 1$. (Why?) Let M denote a median of $f(X)$, which by definition is a number satisfying[2]

$$\mathbb{P}\{f(X) \leq M\} \geq \frac{1}{2} \quad \text{and} \quad \mathbb{P}\{f(X) \geq M\} \geq \frac{1}{2}.$$

Consider the sub-level set

$$A := \left\{x \in \sqrt{n}S^{n-1} : \ f(x) \leq M\right\}.$$

Since $\mathbb{P}\{X \in A\} \geq 1/2$, Lemma 5.1.5 implies that

$$\mathbb{P}\{X \in A_t\} \geq 1 - 2\exp(-ct^2). \tag{5.3}$$

On the other hand, we claim that

$$\mathbb{P}\{X \in A_t\} \leq \mathbb{P}\{f(X) \leq M + t\}. \tag{5.4}$$

Indeed, if $X \in A_t$ then $\|X - y\|_2 \leq t$ for some point $y \in A$. By definition, $f(y) \leq M$. Since f Lipschitz with $\|f\|_{\text{Lip}} = 1$, it follows that

$$f(X) \leq f(y) + \|X - y\|_2 \leq M + t.$$

This proves our claim (5.4).

Combining (5.3) and (5.4), we conclude that

$$\mathbb{P}\{f(X) \leq M + t\} \geq 1 - 2\exp(-ct^2).$$

Repeating the argument for $-f$, we obtain a similar bound for the probability that $f(X) \geq M - t$. Combining the two, we obtain a similar bound for the probability that $|f(X) - M| \leq t$, showing that

$$\|f(X) - M\|_{\psi_2} \leq C.$$

It remains to replace the median M by the expectation $\mathbb{E}f$. This can be done automatically by applying the Centering Lemma 2.6.6. (Do this!) The proof of Theorem 5.1.3 is now complete. □

[2]The median may not be unique. However, for continuous and one-to-one functions f, the median is unique. (Check!)

Exercise 5.1.6 (Geodesic metric)**.** [Difficulty=4] *We proved Theorem 5.1.3 for functions f that are Lipschitz with respect to the Euclidean metric $\|x - y\|_2$ on the sphere. Argue that the same result holds for the geodesic metric, which is the length of the shortest arc connecting x and y.*

Exercise 5.1.7 (Concentration on the unit sphere)**.** [Difficulty=7] *We stated Theorem 5.1.3 for the scaled sphere $\sqrt{n}S^{n-1}$. Deduce that a Lipschitz function f on the* unit *sphere S^{n-1} satisfies*

$$\|f(X) - \mathbb{E}\, f(X)\|_{\psi_2} \le \frac{C\|f\|_{\mathrm{Lip}}}{\sqrt{n}}. \tag{5.5}$$

where $X \sim \mathrm{Unif}(S^{n-1})$. Equivalently, for every $t \ge 0$, we have

$$\mathbb{P}\left\{|f(X) - \mathbb{E}\, f(X)| \ge t\right\} \le 2\exp\Big(-\frac{cnt^2}{\|f\|_{\mathrm{Lip}}^2}\Big) \tag{5.6}$$

Exercise 5.1.8 (Exponential set of mutually almost orthogonal points)**.** *Fix $\varepsilon \in (0,1)$. Show that there exists a set $\{x_1, \dots, x_N\}$ of unit vectors in $\mathbb{R}^n$ which are mutually almost orthogonal:*

$$|\langle x_i, x_j\rangle| \le \varepsilon \quad \text{for all } i \ne j,$$

and the set is exponentially large in n:

$$N \ge \exp\left(c(\varepsilon)n\right).$$

Hint: Construct the points $x_i \in S^{n-1}$ one at a time. Note that the set of points on the sphere that are almost orthogonal with a given point x_0 form a spherical cap. Show that the normalized area of that cap is exponentially small.

5.2 Concentration on other metric measure spaces

In this section, we will extend the concentration for the sphere to other spaces. To do this, note that our proof of Theorem 5.1.3. was based on two main ingredients:

(a) an isoperimetric inequality;

(b) a blow-up of the minimizers for the isoperimetric inequality.

The sphere is not the only space where these two ingredients are in place. In this section, we will briefly survey several other metric measure spaces where (a) and (b) can be shown, and thus concentration of Lipschitz functions almost automatically follows.

5.2.1 Gauss space

Theorem 5.1.3 can be proved for the normal random vector $X \sim N(0, I_n)$.

Theorem 5.2.1 (Concentration on the Gauss space)**.** *Consider a random vector $X \sim N(0, I_n)$ and a Lipschitz function $f : \mathbb{R}^n \to \mathbb{R}$ (with respect to the Euclidean metric). Then*

$$\|f(X) - \mathbb{E}\, f(X)\|_{\psi_2} \le C\|f\|_{\text{Lip}}. \tag{5.7}$$

The proof is similar to Theorem 5.1.3, and it is based on the following isoperimetric inequality in the Gauss space[3] $(\mathbb{R}^n, \|\cdot\|_2, \gamma_n)$.

Theorem 5.2.2 (Isoperimetric inequality on the Gauss space)**.** *Let $\varepsilon > 0$. Then, among all sets $A \subset \mathbb{R}^n$ with fixed Gaussian measure $\gamma_n(A)$, the half spaces minimize the area of the neighborhood $\gamma_n(A_\varepsilon)$.*

Exercise 5.2.3. [Difficulty=4] *Deduce Theorem 5.2.1. Hint: The ε-neighborhood of a half-space is still a half-space, and its Gaussian measure should be easy to compute.*

Remark 5.2.4. We came across two partial cases of Theorem 5.2.1 before.

1. For *linear functions* f, concentration follows from the fact the normal distribution $N(0, I_n)$ is sub-gaussian.

2. For the *Euclidean norm* $f(x) = \|x\|_2$, concentration follows from Theorem 3.1.1.

Exercise 5.2.5 (Replacing expectation by L_p norm)**.** [Difficulty=5] *Prove that in the concentration results for sphere and Gauss space (Theorems 5.1.3 and 5.2.1), the expectation $\mathbb{E}\, f(X)$ can be replaced by the L_p norm $(\mathbb{E}\, f^p)^{1/p}$ for any $p > 0$ and for any non-negative function f. The constants may depend on p.*

5.2.2 Discrete cube

A similar method based on isoperimetry yields concentration on many other metric measure spaces. One of them is the discrete cube

$$(\{0,1\}^n, d, \mathbb{P})\,.$$

Here $d(x, y)$ is the *Hamming distance* , which is defined for binary strings $x, y \in \{0,1\}^n$ as the fraction of the digits where x and y disagree:

$$d(x, y) = \frac{1}{n}\,|\{i : x_i \ne y_i\}|\,.$$

[3] Here the measure γ_n has the standard normal density (3.3).

The measure $\mathbb{P}$ is the uniform probability measure on the discrete cube. So, the coordinates of the random vector

$$X \sim \text{Unif}\left(\{0,1\}^n\right)$$

are independent Ber(1/2) random variables.

The concentration inequality for the discrete cube states that

$$\|f(X) - \mathbb{E}\, f(X)\|_{\psi_2} \le \frac{C\|f\|_{\text{Lip}}}{\sqrt{n}} \tag{5.8}$$

for any function $f : \{0,1\}^n \to \mathbb{R}$, see [4, Section 6.2]. This result can be deduced from the isoperimetric inequality on the cube, whose minimizers are known to be the *Hamming cubes* – the neighborhoods of single points with respect to the Hamming distance.

5.2.3 Symmetric group

The permutation group S_n consists of permutations of n symbols, which we choose to be $\{1, \dots, n\}$ to be specific. We can view the symmetric group as a metric measure space

$$(S_n, d, \mathbb{P}).$$

Here $d(\pi, \rho)$ is the Hamming distance – the fraction of the symbols where the permutations π and ρ disagree:

$$d(\pi, \rho) = \frac{1}{n} \left|\{i : \pi(i) \neq \rho(i)\}\right|.$$

The measure $\mathbb{P}$ is the uniform probability measure on S_n.

Then concentration inequality (5.8) holds for any function $f : S_n \to \mathbb{R}$, see [4, Section 6.3].

5.2.4 Riemanian manifolds with strictly positive curvature

In addition of arguments based on isoperimetry, there are several other methods to prove concentration of Lipschitz functions. For a thorough treatment of this topic, we refer the reader to the books [2, 1, 4, 3]; here we will briefly survey some of these results.

A very general class of examples is covered by the notion of a *Riemannian manifold*. We refer the reader to ... for necessary background in differential geometry, and here we will just mention a relevant concentration result.

Refer

Let (M, g) be a compact connected smooth Riemannian manifold. The canonical distance $d(x, y)$ on M is defined as the the arclength (with respect

to the Riemmanian tensor g) of a minimizing geodesic connecting x and y. The Riemannian manifold can be viewed as a metric measure space

$$(M, d, \mathbb{P})$$

where $\mathbb{P} = \frac{dv}{V}$ is the probability measure on M obtained from the Riemann volume element dv by normalization (here V is the total volume of M).

Let $c(M)$ denote the infimum of the Ricci curvature tensor over all tangent vectors. Assuming that $c(M) > 0$, it can be proved that

$$\|f(X) - \mathbb{E}\, f(X)\|_{\psi_2} \leq \frac{C\|f\|_{\text{Lip}}}{\sqrt{c(M)}} \tag{5.9}$$

for any Lipschitz function $f : M \to \mathbb{R}$. This concentration inequality can be proved by semigroup tools, see [2, Section 2.3].

To give an example, it is known that $c(S^{n-1}) = n - 1$. Thus (5.9) gives an alternative approach to proving the concentration inequality (5.5) for the sphere S^{n-1}. We will give several other examples next.

5.2.5 Special orthogonal group

The special orthogonal group $SO(n)$ consists of all distance preserving linear transformations on $\mathbb{R}^n$. Equivalently, the elements of $SO(n)$ are $n \times n$ orthogonal matrices whose determinant equals 1. We will view the special orthogonal group as a metric measure space

$$(SO(n), \|\cdot\|_F, \mathbb{P})\,,$$

where the distance is the Frobenius norm $\|A - B\|_F$ (see Section 4.1.3) and $\mathbb{P}$ is the uniform probability measure on $SO(n)$.

Technically, $\mathbb{P}$ is the *Haar measure* on $SO(n)$ – the unique probability measure that is invariant under the action on the group;[4] see ... . This measure allows us to talk about *random orthogonal matrices* Cite

$$X \sim \text{Unif}(SO(n))$$

and discuss concentration inequalities for $f(X)$ where f is a Lipschitz function on $SO(n)$.

Alternatively, a random orthogonal matrix $U \sim \text{Unif}(SO(n))$ (and thus the Haar measure on the special orthogonal group) can be constructed by

[4] A measure μ on $SO(n)$ is rotation invariant if for any measurable set $E \subset SO(n)$ and any $T \in SO(n)$, one has $\mu(E) = \mu(T(E))$.

computing the singular value decomposition $G = UGV^\mathsf{T}$ of a random Gaussian matrix G with i.i.d. $N(0,1)$ entries. (Why? Check rotation invariance.)

The concentration inequality (5.8) holds for for any Lipschitz function $f : SO(n) \to \mathbb{R}$, see [4, Section 6.5.1]. This result follows from concentration on general Riemannian manifolds discussed in Section 5.2.4.

5.2.6 Grassmann manifold

The Grassmann manifold $G_{n,m}$ consists of all m-dimensional subspaces of $\mathbb{R}^n$. In the special case where $m = 1$, the Grassman manifold can be identified with the sphere S^{n-1} (how?), so the result below will generalize concentration on the sphere.

We can view the Grassmann manifold as a metric measure space

$$(G_{n,m}, d, \mathbb{P}).$$

The distance between subspaces E and F can be defined as the operator norm

$$d(E, F) = \|P_E - P_F\|$$

where P_E and P_F are the orthogonal projections onto E and F, respectively.

The probability P is, like before, the uniform (Haar) probability measure on $G_{n,m}$. This measure allows us to talk about *random m-dimensional subspaces of $\mathbb{R}^n$*

$$E \sim \text{Unif}(G_{n,m}),$$

and discuss concentration inequalities for $f(E)$ where f is a Lipschitz function on $G_{n,m}$.

Alternatively, a random subspace $E \sim \text{Unif}(G_{n,m})$, and thus the Haar measure on the Grassmann manifold, can be constructed by computing the column span (i.e. the image) of a random $n \times m$ Gaussian random matrix G with i.i.d. $N(0,1)$ entries. (Why? Check rotation invariance.)

The concentration inequality (5.8) holds for for any Lipschitz function $f : G_{n,m} \to \mathbb{R}$, see [4, Section 6.7.2]. This result can be deduced from concentration on the special orthogonal group discussed in Section 5.2.5. Indeed, one expresses Grassmann manifold as a quotient $G_{n,k} = SO(n)/(SO_m \times SO_{n-m})$ and note that concentration passes on to quotients, see [4, Section 6.6].

Exercise 5.2.6. *State and prove a concentration inequality for Lipschitz functions on the set of all $n \times m$ matrices with orthonormal columns.*

5.2.7 Euclidean ball

The same concentration inequality (5.5) that holds for the unit Euclidean sphere S^{n-1} also holds for the unit Euclidean ball

$$(B_2^n, \|\cdot\|_2, \mathbb{P})$$

equipped with the uniform probability measure. This can be deduced from concentration in Gauss space, see [2, Proposition 2.9]. Exercise?

5.2.8 Continuous cube

Consider the continuous cube

$$([0,1]^n, \|\cdot\|_2, \mathbb{P})$$

as a metric measure space, equipped with the Euclidean distance and the uniform probability measure. Thus the coordinates of a random vector

$$X \sim \text{Unif}\,([0,1]^n)$$

are independent random variables uniformly distributed in the interval $[0,1]$.

Then concentration inequality (5.7) holds for any Lipschitz function $f : [0,1]^n \to \mathbb{R}$. This result can be deduced from concentration in Gauss space, see [2, Proposition 2.8]. Exercise?

Other than for Euclidean balls and cubes, sub-gaussian concentration inequalities hold for many other convex bodies, but not all of them. (A unit ball in ℓ_1 norm is a counterexample.) A weaker sub-exponential concentration can be proved for general convex bodies using C. Borell's inequality, see [4, Section III.3].

5.2.9 Densities $e^{-U(x)}$

Concentration inequality for Gaussian distribution we proved in Theorem 5.2.1 can be extended for distributions more general densities. Let X be a random vector in $\mathbb{R}^n$ whose density has the form

$$f(x) = e^{-U(x)}, \quad x \in \mathbb{R}^n,$$

for some function $U : \mathbb{R}^n \to \mathbb{R}$. As an example, if $X \sim N(0, I_n)$ then the normal density (3.3) gives

$$U(x) = \|x\|_2^2 + c$$

where c is a constant (that depends on n but not x).

Suppose U has curvature like $\|x\|_2^2$ or better. More rigorously, we require that the Hessian of U be lower bounded on all of the space. So, suppose there exists $\kappa > 0$ such that

$$\operatorname{Hess} U(x) \succeq \kappa I_n \quad \text{for all } x \in \mathbb{R}^n.$$

Then the concentration inequality

$$\|f(X) - \mathbb{E}\, f(X)\|_{\psi_2} \le \frac{C\|f\|_{\mathrm{Lip}}}{\sqrt{\kappa}}$$

for any Lipschitz function $f : \mathbb{R}^n \to \mathbb{R}$.

Note a similarity of this result with the concentration inequality (5.9) for Riemannian manifolds. Both of them can be proved using semigroup tools [2, Proposition 2.18].

5.2.10 Random vectors with independent bounded coordinates

In Section 5.2.8, we mentioned a concentration inequality for random vectors $X = (X_1, \dots, X_n)$ whose coordinates are independent random variables uniformly distributed in $[0, 1]$. We may wonder if this can be extended from uniform to more general distributions.

This indeed can be done. Suppose the coordinates of $X = (X_1, \dots, X_n)$ are independent bounded random variables; to be specific, assume that

$$|X_i| \le 1 \quad \text{almost surely for every } i.$$

Then concentration inequality (5.7) holds for any *convex* Lipschitz function $f : [0, 1]^n \to \mathbb{R}$. In particular, this holds for any *norm* on $\mathbb{R}^n$. This result is due to M. Talagrand; see [2, Corollary 4.10].

5.2.11 Bounded differences inequality?

Write

5.3 Application: Johnson-Lindenstrauss Lemma

Suppose we have N data points in $\mathbb{R}^n$ where n is very large. We would like to reduce dimension of the data without sacrificing too much of its geometry. The simplest form of *dimension reduction* is to project the data points onto a low-dimensional subspace

$$E \subset \mathbb{R}^n, \quad \dim(E) := m \ll n,$$

see Figure **??** for illustration. How shall we choose the subspace E, and how small its dimension m can be?

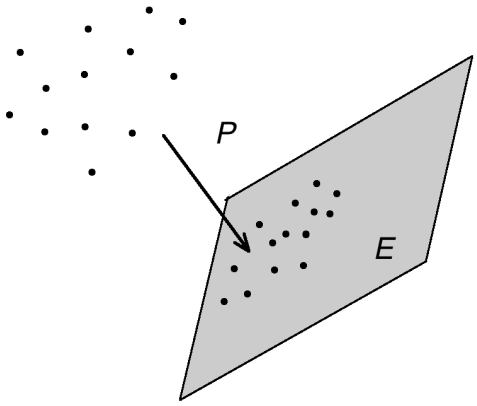


Figure 5.2: In Johnson-Lindenstrauss Lemma, the dimension of the data is reduced by projection onto a random low-dimensional subspace.

Johnson-Lindenstrauss Lemma states that the geometry of data is well preserved if we choose E to be a *random subspace* of dimension

$$m \sim \log N.$$

We already came across the notion of a random subspace in Section **??**.. Let us recall it here. The *Grassmann manifold* $G_{n,m}$ is the set of all m-dimensional subspaces in $\mathbb{R}^n$. It is equipped with Haar measure , which is the unique rotation-invariant[5] probability measure on $G_{n,m}$. This measure allows us to talk about random m-dimensional subspaces uniformly distributed in the Grassman manifold

$$E \sim \mathrm{Unif}(G_{n,m}).$$

Theorem 5.3.1 (Johnson-Lindenstrauss Lemma)**.** *Let $\mathcal{X}$ be a set of N points in $\mathbb{R}^n$ and $\varepsilon > 0$. Assume that*

$$m \geq (C/\varepsilon^2) \log N.$$

Consider a random m-dimensional subspace E in $\mathbb{R}^n$ uniformly distributed in $G_{n,m}$. Denote the orthogonal projection onto E by P. Then, with probability

[5]Rotation invariance means that for any measurable set $\mathcal{E} \in G_{n,m}$ and any orthogonal matrix $U \in O(n)$, the set $U(\mathcal{E})$ has the same measure as $\mathcal{E}$.

at least $1 - 2\exp(-c\varepsilon^2 m)$*, the scaled projection*

$$Q := \sqrt{\frac{n}{m}}\, P$$

is an approximate isometry on $\mathcal{X}$*:*

$$(1-\varepsilon)\|x-y\|_2 \le \|Qx - Qy\|_2 \le (1+\varepsilon)\|x-y\|_2 \quad \text{for all } x, y \in \mathcal{X}. \tag{5.10}$$

The proof of Johnson-Lindenstrauss Lemma is based on concentration of Lipschitz functions on the sphere, which we studied in Section 5.1. We will use it to first examine the action of the random projection P on a fixed vector $x - y$, and then take union bound over all N^2 vectors $x - y$.

Lemma 5.3.2 (Random projection)**.** *Let* P *be a projection in* $\mathbb{R}^n$ *onto a random* m*-dimensional subspace uniformly distributed in* $G_{n,m}$*. Let* $x \in \mathbb{R}^n$ *be a (fixed) point and* $\varepsilon > 0$*. Then:*

1. $\big(\mathbb{E}\,\|Pz\|_2^2\big)^{1/2} = \sqrt{\frac{m}{n}}\,\|z\|_2$.
2. *With probability at least* $1 - 2\exp(-c\varepsilon^2 m)$*, we have*

$$(1-\varepsilon)\sqrt{\frac{m}{n}}\,\|z\|_2 \le \|Pz\|_2 \le (1-\varepsilon)\sqrt{\frac{m}{n}}\,\|z\|_2.$$

Proof. Without loss of generality, we may assume that $\|z\|_2 = 1$. (Why?) Next, we will change the model: instead of a random projection P acting on a fixed vector z, we will consider a fixed projection P acting on a random vector z. Specifically, the distribution of $\|Pz\|_2$ will not change if we let P be fixed and

$$z \sim \mathrm{Unif}(S^{n-1}).$$

(Show this using rotation invariance.)

Using rotation invariance again, we may assume without loss of generality that P is the coordinate projection onto $\mathbb{R}^m$ viewed as a subspace of $\mathbb{R}^n$. Thus

$$\mathbb{E}\,\|Pz\|_2^2 = \mathbb{E}\sum_{i=1}^m z_i^2 = \sum_{i=1}^m \mathbb{E}\, z_i^2 = m\,\mathbb{E}\, z_1^2 \tag{5.11}$$

since the coordinates z_i of the random vector $z \sim \mathrm{Unif}(S^{n-1})$ are identically distributed. To compute $\mathbb{E}\, z_1^2$, note that $1 = \|z\|_2^2 = \sum_{i=1}^m z_i^2$, Taking expectations of both sides, we obtain

$$1 = \sum_{i=1}^m \mathbb{E}\, z_i^2 = m\,\mathbb{E}\, z_1^2,$$

which yields

$$\mathbb{E}\, z_1^2 = \frac{1}{n}.$$

Putting this into (5.11), we get

$$\mathbb{E}\, \|Pz\|_2^2 = \frac{m}{n}.$$

This proves the first part of the lemma.

The second part follows from concentration of Lipschitz functions on the sphere. Indeed,

$$f(x) := \|Px\|_2$$

is a Lipschitz function on S^{n-1}, and $\|f\|_{\text{Lip}} = 1$. (Why?) Then concentration inequality (5.6) yields

$$\mathbb{P}\left\{ \left| \|Px\|_2 - \sqrt{\frac{m}{n}} \right| \geq t \right\} \leq 2\exp(-cnt^2).$$

(Here we also used Exercise 5.2.5 to replace $\mathbb{E}\, \|x\|_2$ by the $(\mathbb{E}\, \|x\|_2^2)^{1/2}$ in the concentration inequality.) Choosing $t := \varepsilon\sqrt{m/n}$, we complete the proof of the lemma. □

Proof of Johnson-Lindenstrauss Lemma. Consider the difference set

$$\mathcal{X} - \mathcal{X} = \{x - y : \ x, y \in \mathcal{X}\}.$$

We would like to show that, with required probability, the inequality

$$(1-\varepsilon)\|z\|_2 \leq \|Qz\|_2 \leq (1+\varepsilon)\|z\|_2$$

holds for all $z \in \mathcal{X} - \mathcal{X}$ Since $Q = \sqrt{n/m}\, P$, this is inequality is equivalent to

$$(1-\varepsilon)\sqrt{\frac{m}{n}}\, \|z\|_2 \leq \|Pz\|_2 \leq (1+\varepsilon)\sqrt{\frac{m}{n}}\, \|z\|_2. \tag{5.12}$$

For any fixed z, Lemma 5.3.2 states that (5.12) holds with probability at least $1 - 2\exp(-c\varepsilon^2 m)$. It remains to take a union bound over $z \in \mathcal{X} - \mathcal{X}$. By doing this, we make the inequality (5.12) hold with probability at least

$$1 - |\mathcal{X} - \mathcal{X}| \cdot 2\exp(-c\varepsilon^2 m) \geq 1 - N^2 \cdot 2\exp(-c\varepsilon^2 m).$$

If we choose $m \geq (C/\varepsilon^2)\log N$ then this probability is at least $1 - 3\exp(-c\varepsilon^2 m/2)$, as claimed. Johnson-Lindenstrauss Lemma is proved. □

A remarkable feature of Johnson-Lindenstrauss lemma is dimension reduction map A is *non-adaptive*, it does not depend on the data. It is also interesting that the ambient dimension n of the data plays no role in this result.

Johnson-Lindenstrauss Lemma uses a random projection as a means of dimension reduction. Other linear and non-linear maps are possible to use, too:

Exercise 5.3.3 (Johnson-Lindenstrauss with sub-gaussian matrices)**.** [Difficulty=6] *Let G be an $m \times n$ random matrix whose rows are independent, mean zero, sub-gaussian isotropic random vectors in $\mathbb{R}^n$. Show that the conclusion of Johnson-Lindenstrauss lemma holds for $Q = \frac{1}{\sqrt{m}} G$.*

TODO: Interpret Q as a *sub-gaussian projection*. Discuss this concept here or, better, before. Make a coupling between random orthogonal projections and random sub-gaussian projections, using the two-sided bounds on the spectrum.

5.4 Matrix Bernstein's inequality

Concentration inequalities for sums of independent random variables $\sum X_i$ can be generalized for sums of independent *random matrices*. In this section, we will prove a matrix version of Bernstein's inequality (Theorem 2.8.6). It is exactly the same inequality, but where random variables X_i are replaced by random matrices, and the absolute value $|\cdot|$ is replaced by the operator norm $\|\cdot\|$.

A remarkable feature of matrix Bernstein's inequality is that it will not require any independence of entries, rows, or columns within each random matrix X_i.

Theorem 5.4.1 (Matrix Bernstein's inequality)**.** *Let $X_1, \dots, X_N$ be independent, mean zero, $n \times n$ symmetric random matrices, such that $\|X_i\| \le K$ almost surely for all i. Then, for every $t \ge 0$, we have*

$$\mathbb{P}\Big\{ \Big\| \sum_{i=1}^N X_i \Big\| \ge t \Big\} \le 2n \exp\Big(- \frac{-t^2/2}{\sigma^2 + Kt/3} \Big).$$

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

In particular, we can express this bound as the mixture of sub-gaussian and sub-exponential tail, just like in the scalar Bernstein's inequality:

$$\mathbb{P}\Big\{ \Big\| \sum_{i=1}^N X_i \Big\| \ge t \Big\} \le 2n \exp\Big[- c \cdot \min\Big(- \frac{t^2}{\sigma^2}, \frac{t}{K} \Big) \Big].$$

The proof of matrix Bernstein's inequality will be based on the following naïve idea. Can we repeat the classical argument based on moment generating functions (see Section 2.8), replacing scalars by matrices at each occurrence? This can indeed be done. In most of the places, scalars can be replaced by matrices without any problem, but one inequality will be non-trivial. To prepare for this, we will now develop the basics of matrix calculus, which basically allows us to treat matrices as scalars.

5.4.1 Matrix calculus

Throughout this section, we will work with symmetric $n \times n$ matrices. As we know, the operation of addition $A+B$ generalizes painlessly from scalars to matrices. We need to be more careful with multiplication, since it is not commutative for matrices: in general, $AB \neq BA$. For this reason, matrix Bernstein's inequality is sometimes called *non-commutative* Bernstein's inequality. Functions of matrices are defined as follows.

Definition 5.4.2 (Functions of matrices)**.** *Consider a function $f : \mathbb{R} \to \mathbb{R}$ and an $n \times n$ symmetric matrix X. Express X through its spectral decomposition:*

$$X = \sum_{i=1}^{n} \lambda_i u_i u_i^{\mathsf{T}}.$$

Then define

$$f(X) := \sum_{i=1}^{n} f(\lambda_i) u_i u_i^{\mathsf{T}}.$$

In other words, to obtain the matrix $f(X)$ from X, we do not change the eigenvectors and apply f to the eigenvalues.

Note that this definition agrees with addition and multiplication of matrices:

Exercise 5.4.3 (Matrix polynomials and power series)**.** *1. Consider a polynomial*

$$f(x) = a_0 + a_1 x + \cdots + a_p x^p.$$

Check that for matrices X, we have

$$f(X) = a_0 I + a_1 X + \cdots + a_p X^p.$$

In the right side, we use the standard operations of matrix addition and multiplication, so in particular $X^p = X \cdots X$ (p times) there.

2. Consider a convergent power series expansion of f about x_0:

$$f(x) = \sum_{k=1}^{\infty} a_k (x - x_0)^k.$$

Check that the series of matrix terms converges, and

$$f(X) = \sum_{k=1}^{\infty} a_k (X - X_0)^k.$$

As an example, for each $n \times n$ symmetric matrix X we have

$$e^X = I + X + \frac{X^2}{2!} + \frac{X^3}{3!} + \cdots$$

Just like scalars, matrices can be compared to each other. To do this, we define a *partial order* on the set of $n \times n$ symmetric matrices as follows. First, we say that

$$X \succeq 0 \quad \text{if} \quad X \text{ is positive semi-definite.}$$

Equivalently, $X \succeq 0$ if all eigenvalues of X satisfy $\lambda_i(X) \geq 0$. Next, we set

$$X \succeq Y \quad \text{if} \quad X - Y \succeq 0.$$

Finally, we obviously set $Y \preceq X$ if $X \succeq Y$.

Note that $\succeq$ is a partial, as opposed to total, order, since there are matrices for which neither $X \succeq Y$ nor $Y \succeq X$ hold. (Give an example!)

Exercise 5.4.4. *Prove the following properties.*

1. *$\|X\| \leq t$ if and only if $-tI \preceq X \preceq tI$.*

2. *Let $f : \mathbb{R} \to \mathbb{R}$ be an increasing function and X, Y are commuting matrices. Then $X \preceq Y$ implies $f(X) \preceq f(Y)$.*

3. *Let $f, g : \mathbb{R} \to \mathbb{R}$ be two functions. If $f(x) \leq g(x)$ for all $x \in \mathbb{R}$ satisfying $|x| \leq K$, then $f(X) \leq g(X)$ for all X satisfying $\|X\| \leq K$.*

5.4.2 Trace inequalities

So far, generalization from scalars to matrices was smooth. But the non-commutativity of the matrix product ($AB \neq BA$) causes some important identities to fail for matrices. One of such identities is $e^{x+y} = e^x e^y$, which holds for scalars but fails for matrices:

Exercise 5.4.5. [Difficulty=7] *Let X and Y be $n \times n$ symmetric matrices.*
1. Show that if the matrices commute, i.e. $XY = YX$, then

$$e^{X+Y} = e^X e^Y.$$

2. Find an example of matrices X and Y such that

$$e^{X+Y} \neq e^X e^Y.$$

This is unfortunate for us, because the identity $e^{x+y} = e^x e^y$ was crucially used in our approach to concentration of sums of random variables. Indeed, this approach was based on computing the moment generating function $\mathbb{E} \exp(\lambda S)$ of the sum, and then breaking it into the product of exponentials using this identity.

Nevertheless, there exists some useful substitutes for the missing identity $e^{X+Y} = e^X e^Y$. We will state two of them here without proof; they belong to the rich family of *trace inequalities* for matrices.

Theorem 5.4.6 (Golden-Thompson inequality)**.** *For any $n \times n$ symmetric matrices A and B, we have*

$$\operatorname{tr}(e^{A+B}) \leq \operatorname{tr}(e^A e^B).$$

Theorem 5.4.7 (Lieb's inequality)**.** *Let H be an $n \times n$ symmetric matrix. Consider the function*

$$f(X) := \operatorname{tr} \exp(H + \log X).$$

Then f is concave on the space on $n \times n$ symmetric matrices.[6]

Note that in the scalar case where $n = 1$, the function f is linear and Lieb's inequality holds trivially.

A proof of matrix Bernstein's inequality can be based on either Golden-Thompson or Lieb's inequalities. We will use Lieb's inequality, which we

[6] Concavity means that the inequality $f(\lambda X + (1-\lambda)Y) \geq \lambda f(X) + (1-\lambda) f(Y)$ holds for matrices X and Y, and for $\lambda \in [0, 1]$.

will now restate for random matrices. If X is a random matrix, then Lieb's and Jensen's inequalities imply that

$$\mathbb{E}\, f(X) \le f(\mathbb{E}\, X).$$

Applying this with $X = e^Z$, we obtain the following.

Lemma 5.4.8 (Lieb's inequality for random matrices)**.** *Let H be a fixed $n \times n$ symmetric matrix, and Z be a random $n \times n$ symmetric matrix. Then*

$$\mathbb{E}\, \mathrm{tr} \exp(H + Z) \le \mathrm{tr} \exp(H + \log \mathbb{E}\, e^Z).$$

5.4.3 Proof of matrix Bernstein's inequality

In this section we prove matrix Bernstein's inequality, Theorem 5.4.1, using Lieb's inequality.

Step 1: Reduction to MGF. To bound the norm of the sum

$$S := \sum_{i=1}^N X_i,$$

we need to control the largest and smallest eigenvalues of S. We can do this separately. To put this formally, consider the largest eigenvalue

$$\lambda_{\max}(S) := \max_i \lambda_i(S)$$

and note that

$$\|S\| = \max_i |\lambda_i(S)| = \max\left(\lambda_{\max}(S),\, \lambda_{\max}(-S)\right). \tag{5.13}$$

To bound $\lambda_{\max(S)}$, we will proceed with the method based on computing the moment generating function as in the scalar case. To this end, fix $\lambda \ge 0$ and use Markov's inequality to obtain

$$\mathbb{P}\left\{\lambda_{\max}(S) \ge t\right\} = \mathbb{P}\left\{e^{\lambda \cdot \lambda_{\max}(S)} \ge e^{\lambda t}\right\} \le e^{-\lambda t}\, \mathbb{E}\, e^{\lambda \cdot \lambda_{\max}(S)}. \tag{5.14}$$

Since the eigenvalues of $e^{\lambda S}$ are $e^{\lambda \cdot \lambda_i(S)}$, we have

$$E := \mathbb{E}\, e^{\lambda \cdot \lambda_{\max}(S)} = \mathbb{E}\, \lambda_{\max}(e^{\lambda S}).$$

The eigenvalues of $e^{\lambda S}$ are positive. Then we can bound the maximal eigenvalue of $e^{\lambda S}$ by the sum of all eigenvalues, which is the trace of $e^{\lambda S}$, leading to

$$E \le \mathbb{E}\, \mathrm{tr}\, e^{\lambda S}.$$

Step 2: Application of Lieb's inequality. To prepare this quantity for an application of Lieb's inequality (Lemma 5.4.8), we separate one term from the sum S:

$$E \le \mathbb{E} \operatorname{tr} \exp\Big[\sum_{i=1}^{N-1} \lambda X_i + \lambda X_N\Big].$$

Let us condition on $(X_i)_{i=1}^{N-1}$ and apply Lemma 5.4.8 for the fixed matrix $H := \sum_{i=1}^{N-1} \lambda X_i$ and the random matrix $Z := \lambda X_N$. We obtain

$$E \le \mathbb{E} \operatorname{tr} \exp\Big[\sum_{i=1}^{N-1} \lambda X_i + \log \mathbb{E}\, e^{\lambda X_N}\Big].$$

(To be more specific here, we first apply Lemma 5.4.8 for the conditional expectation with respect to X_N. Then we take expectation of both sides with respect to $X_1, \dots, X_N$ and use the law of total expectation.)

Next we separate the term λX_{N-1} from the sum $\sum_{i=1}^{N-1} \lambda X_i$, and apply Lemma 5.4.8 again for $Z = \lambda X_{N-1}$. Repeating these steps N times, we obtain

$$E \le \operatorname{tr} \exp\Big[\sum_{i=1}^{N} \log \mathbb{E}\, e^{\lambda X_i}\Big]. \tag{5.15}$$

Step 3: MGF of the individual terms. It remains to bound the matrix-valued moment generating function $\mathbb{E}\, e^{\lambda X_i}$ for each term X_i. This is a standard task, and the argument will be similar to the scalar case.

Lemma 5.4.9 (Moment generating function)**.** *Let X be an $n \times n$ symmetric random matrix such that $\|X\| \le K$ almost surely. Then*

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

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

Proof. First, note that we can bound the (scalar) exponential function by the first few terms of its Taylor's expansion as follows:

$$e^z \le 1 + z + \frac{1}{1 - |z|/3} \cdot \frac{z^2}{2}, \quad \text{if } |z| < 3.$$

(To get this inequality, write $e^z = 1 + z + z^2 \cdot \sum_{p=2}^{\infty} z^p/p!$ and use the bound $p! \ge 2 \cdot 3^{p-2}$.) Next, apply this inequality for $z = \lambda x$. If $|x| \le K$ and $|\lambda| < 3/K$ then we obtain

$$e^{\lambda x} \le 1 + \lambda x + g(\lambda) x^2,$$

where $g(\lambda)$ is the function in the statement of the lemma.

Finally, we can transfer this inequality from scalars to matrices using part 3 of Exercise 5.4.4. We obtain that if $\|X\| \le K$ and $|\lambda| < 3/K$, then

$$e^{\lambda X} \preceq I + \lambda X + g(\lambda) X^2.$$

Take expectation of both sides and use the assumption that $\mathbb{E}\, X = 0$ to obtain

$$\mathbb{E}\, e^{\lambda X} \preceq I + g(\lambda)\, \mathbb{E}\, X^2.$$

To bound the right hand side, we may use the inequality $1 + z \le e^z$ which holds for all scalars z. Thus the inequality $I + Z \preceq e^Z$ holds for all matrices Z, and in particular for $Z = g(\lambda)\, \mathbb{E}\, X^2$. (Here we again refer to part 3 of Exercise 5.4.4.) This yields the conclusion of the lemma. □

Step 4: Completion of the proof. Let us return to bounding the quantity in (5.15). Using Lemma 5.4.9, we obtain

$$\exp\Big[\sum_{i=1}^{N-1} \log \mathbb{E}\, e^{\lambda X_i}\Big] \preceq \exp\Big[g(\lambda) \sum_{i=1}^{N-1} \mathbb{E}\, X_i^2\Big].$$

(Here we also used that part 2 of Exercise 5.4.4 for the logarithmic and the exponential function.)

Since the matrices on both sides of the inequality in (**??**) have positive eigenvalues, the trace of the left side is bounded by the trace of the right side:

Revise the previous few lines, think how to make them clear, more concise

$$E \le \operatorname{tr} \exp\left[g(\lambda) Z\right], \quad \text{where} \quad Z := \sum_{i=1}^{N-1} \mathbb{E}\, X_i^2.$$

Since the trace here is a sum of n positive eigenvalues, it is bounded by n times the maximum eigenvalue, so

$$\begin{aligned} E &\le n \cdot \lambda_{\max}\left(\exp[g(\lambda) Z]\right) = n \cdot \exp\left[g(\lambda) \lambda_{\max}(Z)\right] \qquad \text{(Why?)} \\ &= n \cdot \exp\left[g(\lambda) \|Z\|\right] \qquad \text{(since } Z \succeq 0\text{)} \\ &= n \cdot \exp\left[g(\lambda) \sigma^2\right] \qquad \text{(by definition of } \sigma \text{ in the theorem).} \end{aligned}$$

Plugging this bound for $E = \mathbb{E}\, e^{\lambda \cdot \lambda_{\max}(S)}$ into (5.14), we obtain

$$\mathbb{P}\left\{\lambda_{\max}(S) \ge t\right\} \le n \cdot \exp\left[-\lambda t + g(\lambda) \sigma^2\right].$$

Now we optimize the right hand side in λ. The bound is minimized for $\lambda = t/(\sigma^2 + Kt/3)$. (Check!) Substituting this value for λ, we conclude that

$$\mathbb{P}\left\{\lambda_{\max}(S) \ge t\right\} \le n \cdot \exp\Big(-\frac{-t^2/2}{\sigma^2 + Kt/3}\Big).$$

Repeating the argument for $-S$ and combining the two bounds via (5.13), we complete the proof of Theorem 5.4.1. (Do this!) □

5.4.4 Discussion of matrix Bernstein's inequality

Matrix Bernstein's inequality, Theorem 5.4.1 is a direct generalization of the scalar Bernstein's inequality (Theorem 2.8.6). We will now pause to note its remarkable strength and sharpness.

As an immediate consequence of Theorem 5.4.1, we can bound the expected deviation of the sum as follows.

This is not quite immediate. Make this corollary an exercise?

Corollary 5.4.10 (Expected deviation of sum of random matrices). *Let $X_1, \dots, X_N$ be independent, mean zero, $n \times n$ symmetric random matrices, such that $\|X_i\| \le K$ almost surely for all i. Then*

$$\mathbb{E} \Big\| \sum_{i=1}^N X_i \Big\| \lesssim \Big\| \sum_{i=1}^N \mathbb{E}\, X_i^2 \Big\|^{1/2} \sqrt{\log n} + K \log n.$$

In the scalar case where $n = 1$, such a bound on the expected deviation is trivial. Indeed,

$$\mathbb{E} \Big| \sum_{i=1}^N X_i \Big| \le \Big(\mathbb{E} \Big| \sum_{i=1}^N X_i \Big|^2 \Big)^{1/2} = \Big(\sum_{i=1}^N \mathbb{E}\, X_i^2 \Big)^{1/2}.$$

where we used that the variance of a sum of independent random variables equals the sum of variances.

This simple argument fails for matrices (why?). Bounding the expected deviation for a sum of independent random matrices is a non-trivial problem, which matrix Bernstein's inequality successfully solves.

The price of going from scalar to matrices is the pre-factor n in the probability bound in Theorem 5.4.1. It is a light price, considering that this factor becomes *logarithmic in dimension n* in Corollary 5.4.10.

The following example shows that a logarithmic factor is needed in general; we will give another example in ....

Covariance estimation

Exercise 5.4.11 (Sharpness of matrix Bernstein's inequality). *Let X be an $n \times n$ random matrix that takes values $e_k e_k^\mathsf{T}$, $k = 1, \dots, n$, with probability $1/n$ each. (Here as usual (e_k) denotes the standard basis in $\mathbb{R}^n$.) Let $X_1, \dots, X_N$ be independent copies of X. Consider the sum*

$$S := \sum_{i=1}^N X_i.$$

Then S is a diagonal matrix.

1. Show that the entry S_{ii} has the same distribution as the number of balls in i-th bin when N balls are thrown into n bins independently.

2. Relating this to the coupon collector's problem, show that if $N \asymp n$ then

$$\mathbb{E}\,\|S\| \asymp \frac{\log n}{\log\log n}.$$

Deduce that Corollary 5.4.10 would fail if the logarithmic factors were removed from its bound.

5.5 Application: community detection in sparse networks

In Section 4.4, we analyzed a basic method for community detection in networks – the Spectral Clustering Algorithm. We examined the performance of spectral clustering for the stochastic block model $G(n, p, q)$ with two communities (see Definition 4.4.1). We found that the communities are identified with high accuracy and high probability (recall Theorem 4.4.4).

We will now re-examine the same method but using matrix Bernstein's inequality. We will find that spectral clustering works for *much sparser* networks than we were able to analyze before. This will be done in a series of exercises.

As in Section 4.4, we denote by A the adjacency matrix of a random graph from $G(n, p, q)$. We express A as

$$A = D + R$$

where $D = \mathbb{E}\,A$ is a deterministic matrix ("signal") and R is random ("noise"). The success of the method hinges on showing that the noise $\|R\|$ is small with high probability (see (4.14)).

Exercise 5.5.1. [Difficulty=6] *Represent the adjacency matrix A as a sum of independent random matrices*

$$A = \sum_{i,j=1}^{n} Z_{ij}.$$

Make it so that each Z_{ij} encode the contribution of an edge between vertices i and j. Thus the only non-zero entries of Z_{ij} should be (ij) and (ji), and they should be the same as in A.

Apply matrix Bernstein's inequality to find that

$$\mathbb{E}\,\|R\| \lesssim \sqrt{d \log n} + \log n,$$

where $d = \frac{1}{2}(p+q)n$. *Argue that* d *is the* average expected degree *of the graph.*

Exercise 5.5.2. [Difficulty=6] *Similarly to Section 4.4, conclude that spectral clustering works. Identify the conditions on* p *and* q. *In particular, argue that spectral clustering works for* sparse networks, *as long as the average expected degrees satisfy*

$$d \gg \log n.$$

5.6 Random matrices with general independent rows

We will illustrate the matrix Bernstein inequality with an application to random matrices. In this section, we will prove a remarkably general version of Theorem 4.5.1 which holds for random matrices with arbitrary, not necessarily sub-gaussian distributions of rows.

Such generality is important because, as we noted in Section 3.4.2, discrete distributions are poorly sub-gaussian. So Theorem 4.5.1 does not usually give a satisfactory result for random matrices whose rows have discrete distributions. The next result will be more useful.

Theorem 5.6.1 (Random matrices with general rows)**.** *Let* A *be an* $m \times n$ *matrix whose rows* A_i *are independent isotropic random vectors in* $\mathbb{R}^n$. *Assume that for some* $L \geq 0$,

$$\|A_i\|_2 \leq L\sqrt{n} \quad \text{almost surely for every } i. \tag{5.16}$$

Then for every $t \geq 0$, *one has*

$$\sqrt{m} - tL\sqrt{n} \leq s_n(A) \leq s_1(A) \leq \sqrt{m} + tL\sqrt{n} \tag{5.17}$$

with probability at least $1 - 2n \cdot \exp(-ct^2)$.

Before we prove this theorem, let us pause to make two remarks. First, why does the boundedness assumption (5.16) have this form? The isotropy of random vectors A_i implies by Lemma 3.2.4 that

$$\mathbb{E}\,\|A_i\|_2^2 = n.$$

By Markov's inequality, it follows that with high probability (let's say, 0.99), we have

$$\|A_i\|_2 = O(\sqrt{n}).$$

So in applications, we expect that the boundedness assumption (5.16) hold with

$$L = O(1).$$

Next, let us note that by Lemma 4.1.2, we can equivalently restate the conclusion (5.17) in the following form:

$$\left\| \frac{1}{m} A^{\mathsf{T}} A - I_n \right\| \le \max(\delta, \delta^2) \quad \text{where} \quad \delta = tL\sqrt{\frac{n}{m}}. \tag{5.18}$$

Proof. Our proof of (5.18) will be based on matrix Bernstein's inequality, Theorem 5.4.1. To use this result, we express the matrix in question a sum of independent random matrices:

$$\frac{1}{m} A^{\mathsf{T}} A - I_n = \frac{1}{m} \sum_{i=1}^{m} A_i A_i^{\mathsf{T}} - I_n = \sum_{i=1}^{N} X_i \quad \text{where } X_i := \frac{1}{m}(A_i A_i^{\mathsf{T}} - I_n).$$

Note that X_i are independent symmetric $n \times n$ random matrices, and they have zero means by the isotropy assumption. So, in order to apply Theorem 5.4.1, it remains to bound the range $\|X_i\|$ and the norm of the matrix variance $\left\| \sum_{i=1}^m \mathbb{E}\, X_i^2 \right\|$.

To bound the range, we can use boundedness assumption (5.16), where we necessarily have $L \ge 1$ by isotropy. (Why?) By triangle inequality, we have

$$\begin{aligned} \|X_i\| &\le \frac{1}{m}(\|A_i A_i^{\mathsf{T}}\| + 1) = \frac{1}{m}(\|A_i\|_2^2 + 1) \\ &\le \frac{1}{m}(L^2 n + 1) \qquad \text{(by the boundedness assumption)} \\ &\le \frac{2L^2 n}{m} \qquad \text{(since } L \ge 1\text{)} \\ &=: K. \end{aligned} \tag{5.19}$$

To estimate the norm of the matrix variance $\left\| \sum_{i=1}^m \mathbb{E}\, X_i^2 \right\|$, we first compute

$$X_i^2 = \frac{1}{m^2}\left[(A_i A_i^{\mathsf{T}})^2 - 2(A_i A_i^{\mathsf{T}}) + I_n \right].$$

Taking expectations of both sides and using isotropy of A_i, we obtain

$$\mathbb{E}\, X_i^2 = \frac{1}{m^2}\left[\mathbb{E}(A_i A_i^{\mathsf{T}})^2 - I_n \right].$$

r to an exercise?

Now, using boundedness assumption (5.16), we have

$$\begin{aligned}\mathbb{E}(A_i A_i^{\mathsf{T}})^2 &= \mathbb{E}\,\|A_i\|_2^2\, A_i A_i^{\mathsf{T}} \\ &\preceq \mathbb{E}\left[L^2 n \cdot A_i A_i^{\mathsf{T}}\right] \qquad \text{(by the boundedness assumption– check!)} \\ &= L^2 n \cdot I_n \qquad \text{(by isotropy).}\end{aligned}$$

Thus

$$\mathbb{E}\, X_i^2 \preceq \frac{L^2 n}{m^2} \cdot I_n.$$

Summing up, we obtain a bound on the matrix variance:

$$\sum_{i=1}^m \mathbb{E}\, X_i^2 \preceq \frac{L^2 n}{m} \cdot I_n.$$

Since the matrix variance is a positive semidefinite matrix, it follows that

$$\Big\| \sum_{i=1}^m \mathbb{E}\, X_i^2 \Big\| \le \frac{L^2 n}{m} := \sigma^2. \tag{5.20}$$

Now we are ready to apply the matrix Bernstein inequality (**??**).

$$\begin{aligned}\mathbb{P}\left\{ \Big\| \frac{1}{m} A^{\mathsf{T}} A - I_n \Big\| \ge \varepsilon \right\} &= \mathbb{P}\left\{ \Big\| \sum_{i=1}^m X_i \Big\| \ge \varepsilon \right\} \\ &\le 2n \cdot \exp\Big[- c \min\Big(\frac{\varepsilon^2}{\sigma^2}, \frac{\varepsilon}{K} \Big) \Big].\end{aligned}$$

Substituting σ and K from (5.20) and (5.19), we bound this further by

$$2n \cdot \exp\Big[- c \min(\varepsilon^2, \varepsilon) \cdot \frac{m}{2L^2 n} \Big]. \tag{5.21}$$

Returning to (5.18), we set $\varepsilon := \max(\delta, \delta^2)$ and $\delta = tL\sqrt{n/m}$. Substituting these values into (5.21), we bound the probability that (5.18) fails by

$$2n \cdot \exp\Big(- \frac{c\delta^2 m}{2L^2 n} \Big) = 2n \cdot \exp(-ct^2/2).$$

This completes the proof. □

Notice the pre-factor n in the probability bound $1 - 2n \cdot \exp(-ct^2)$ in Theorem 5.6.1. This theorem is non-trivial when the probability is positive,

and this happens when $t \geq C\sqrt{\log n}$. So we can restate the conclusion of Theorem 5.6.1 as follows. For every $s \geq 1$, one has

$$\sqrt{m} - sL\sqrt{n \log n} \leq s_n(A) \leq s_1(A) \leq \sqrt{m} + sL\sqrt{n \log n} \tag{5.22}$$

with probability at least $1 - 2n^{-cs^2}$. (To obtain this, set $t = s\sqrt{\log n}$.)

Summarizing, we extended Theorem 4.5.1 from sub-gaussian to general distributions, and we paid just a logarithmic factor for that.

Exercise 5.6.2 (Non-isotropic distributions)**.** [Difficulty=7] *Prove the following version of* (5.18) *for non-isotropic distributions. Let A be an $m \times n$ matrix which satisfies all assumptions of Theorem 5.6.1 except the isotropy. Assume that the rows have the same covariance matrix*

$$\Sigma = \mathbb{E}\, A_i A_i^{\mathsf{T}}.$$

Then, for every $t \geq 0$, the following inequality holds with probability at least $1 - 2n \cdot \exp(-ct^2)$:

$$\left\| \frac{1}{m} A^{\mathsf{T}} A - \Sigma \right\| \leq \max(\|\Sigma\|^{1/2}\delta, \delta^2) \quad \textit{where} \quad \delta = tL\sqrt{\frac{n}{m}}. \tag{5.23}$$

5.7 Application: covariance estimation for general distributions

Rewrite the results of this section as in my PCMI lectures

In Section 4.6, we showed the covariance matrix of a sub-gaussian distribution in $\mathbb{R}^n$ can be accurately estimated using $O(n)$ samples. In this section, we will remove the sub-gaussian requirement thus making covariance estimation possible for very general, in particular discrete, distributions in $\mathbb{R}^n$. The price we will pay is the logarithmic oversampling factor. The following results shows that $O(n \log n)$ samples suffice for covariance estimation of general distributions in $\mathbb{R}^n$.

Theorem 5.7.1 (General covariance estimation)**.** *Consider a random vector X in $\mathbb{R}^n$ with zero mean and covariance matrix Σ. Assume that for some $K > 0$,*

$$\|X\|_2 \leq K\sqrt{\|\Sigma\| n} \quad \textit{almost surely.} \tag{5.24}$$

Let $\varepsilon \in (0, 1)$ and $t \geq 1$. Suppose the sample size satisfies

$$m \geq C(Kt/\varepsilon)^2 n \log n. \tag{5.25}$$

Then the sample covariance matrix Σ_m satisfies

$$\|\Sigma_m - \Sigma\| \le \varepsilon \|\Sigma\|$$

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

Proof. Consider the $m \times n$ matrix A whose rows are the sample points X_i^T. Then the sample covariance matrix Σ can be represented as

$$\Sigma_m = \sum_{i=1}^m X_i X_i^\mathsf{T} = \frac{1}{m} A^\mathsf{T} A.$$

So, to bound the error $\|\Sigma_m - \Sigma\|$ we can apply the non-isotropic form of Theorem 4.5.1 stated in Exercise 5.6.2. It states that with probability at least $1 - 2n \cdot \exp(-cs^2)$, we have

$$\|\Sigma_m - \Sigma\| \le \max(\|\Sigma\|^{1/2}\delta, \delta^2) \quad \text{where} \quad \delta = sK\sqrt{\frac{\|\Sigma\| n}{m}}.$$

After simplifying, this becomes

$$\|\Sigma_m - \Sigma\| \le \max\left(sK\sqrt{\frac{n}{m}}, \frac{s^2K^2 n}{m}\right)\|\Sigma\|.$$

So, if $m \ge (Ks/\varepsilon)^2 n$ the right hand side is bounded by $\varepsilon\|\Sigma\|$ as required. It remains to choose $s = Ct\sqrt{\log n}$ to complete the proof. □

Let us clarify the form of the boundedness assumption (5.24).

Move this exercise to the covariance section?

Exercise 5.7.2. *Let X be a random vector in $\mathbb{R}^n$ with $\mathbb{E}\, XX^\mathsf{T} = \Sigma$. Show that*

$$\mathbb{E}\,\|X\|_2^2 = \operatorname{tr}(\Sigma).$$

Since we always have

$$\operatorname{tr}(\Sigma) \le \|\Sigma\| n$$

Markov's inequality implies that with high probability (let's say, 0.99), we have

$$\|X\|_2 = O(\sqrt{\|\Sigma\| n}).$$

So in applications, we expect that the boundedness assumption (5.24) hold with

$$K = O(1).$$

Such bounded assumption indeed holds in many applications. When it fails, one may consider enforcing it by truncation, thus rejecting a small fraction of samples with the largest norm.

Exercise 5.7.3. [Difficulty=6] *Show that if the boundedness assumption* (5.24) *is removed from Theorem 5.7.1, the result may in general fail.*

Exercise 5.7.4 (Sampling from frames)**.** [Difficulty=4] *Consider a tight frame* $(u_i)_{i=1}^N$ *in* $\mathbb{R}^n$ *(recall Section 3.3.4). State and prove a result that shows that a random sample of*

$$m \gtrsim n \log n$$

elements of (u_i) *forms a frame with good frame bounds (as close to tight as one wants). The quality of the result should not depend on the frame size* N*.*

5.7.1 Logarithmic oversampling

We will now argue that the factor $\log n$ can not be removed from (5.25). In other words, logarithmic oversampling is in general needed for covariance estimation.

To give an example, consider a random vector X with the most discrete isotropic distribution in $\mathbb{R}^n$, namely the *coordinate distribution* introduced in Section 3.3.4. Thus

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

where $\{e_i\}_{i=1}^n$ is the canonical basis of $\mathbb{R}^n$.

The distribution of X is isotropic, so $\Sigma = I$. In order to have a non-trivial estimation of the form

$$\|\Sigma_m - \Sigma\| < \varepsilon \|\Sigma\|$$

with any $\varepsilon < 1$, the sample covariance matrix Σ_m must have full rank n. (Why?) But recalling that

$$\Sigma_m = \frac{1}{m} \sum_{i=1}^m X_i X_i^{\mathsf{T}},$$

we see that for this to happen, the sample $\{X_1, \dots, X_m\}$ must contain all basis vectors $\sqrt{n}\, e_1, \dots, \sqrt{n}\, e_n$. (Why?)

Rethinking this condition in terms of the classical *coupon collector's problem*, we see that each of the n "coupons" $\sqrt{n}\, e_1, \dots, \sqrt{n}\, e_n$ must be picked at least once in m independent trials performed by $X_1, \dots, X_m$.

Refer

By the known result on the coupon collector's problem, one must make at least

$$m \gtrsim n \log n$$

trials to pick each of the n coupons with high (and even constant) probability. This shows that the logarithmic factor $\log n$ is necessary for any non-trivial covariance estimation for the coordinate distribution.

We can track the source of this logarithmic factor to matrix Bernstein inequality (Theorem 5.4.1). As we saw, its only weakness compared with the scalar Bernstein's inequality is the pre-factor n in the probability bound. Our analysis then shows that this pre-factor is needed, and is optimal.

Chapter 6

Quadratic forms, symmetrization and contraction

6.1 Decoupling

In the beginning of this book, we thoroughly studied sums of independent random variables of the type

$$\sum_{i=1}^{n} a_i X_i \tag{6.1}$$

where $X_1, \dots, X_n$ are independent random variables and a_i are fixed coefficients. In this section, we will study quadratic forms of the type

$$\sum_{i,j=1}^{n} a_{ij} X_i X_j = X^{\mathsf{T}} A X = \langle AX, X \rangle \tag{6.2}$$

where $A = (a_{ij})$ is an $n \times n$ matrix of coefficients, and $X = (X_1, \dots, X_n)$ is a random vector with independent coordinates. Such a quadratic form is called a *chaos* in probability theory.

Computing the expectation of a chaos is easy. For simplicity, assume that X_i have zero means and unit variances. Then

$$\mathbb{E}\, X^{\mathsf{T}} A X = \sum_{i=1}^{n} a_{ii} = \operatorname{tr} A.$$

(Check!)

It is less trivial to establish concentration of a chaos. The main difficulty is that the terms of the chaos are not independent. To overcome this problem, we will now introduce the technique of *decoupling*. This method will allow us to replace X with a random vector X' that is independent of X yet has the same distribution as X. We call such X' an *independent copy* of X. So we will seek to replace the quadratic form (6.2) with

$$\sum_{i,j=1}^{n} a_{ij} X_i X'_j = X^{\mathsf{T}} A X' = \langle AX, X' \rangle .$$

The usefulness of this new, decoupled form is that it is *linear* in X, and this makes it easy to analyze. Indeed, we may condition on X' and treat the decoupled form as a sum of independent random variables

$$\sum_{i=1}^{n} \Big(\sum_{j=1}^{n} a_{ij} X'_j \Big) X_i = \sum_{i=1}^{n} c_i X_i$$

with fixed coefficients c_i, much like we treated the sums (6.1) before.

Theorem 6.1.1 (Decoupling)**.** *Let A be an $n \times n$, diagonal-free matrix. Let $X = (X_1, \dots, X_n)$ be a random vector with independent mean zero coefficients. Then, for every convex function F, one has*

$$\mathbb{E}\, F(X^{\mathsf{T}} A X) \le \mathbb{E}\, F(4 X^{\mathsf{T}} A X') \tag{6.3}$$

where X' is an independent copy of X.

We will actually prove a slightly stronger version of decoupling, where A needs not to be diagonal-free. Thus, for every matrix A we will show that

$$\mathbb{E}\, F\Big(\sum_{i,j:\, i \ne j} a_{ij} X_i X_j \Big) \le \mathbb{E}\, F\Big(4 \sum_{i,j} a_{ij} X_i X'_j \Big) \tag{6.4}$$

where $X' = (X'_1, \dots, X'_n)$.

The proof will be based on the following observation.

Lemma 6.1.2. *Let Y and Z be independent random variables such that $\mathbb{E}\, Z = 0$. Then, for every convex function F, one has*

$$\mathbb{E}\, F(Y) \le \mathbb{E}\, F(Y + Z).$$

Proof. This is a simple consequence of Jensen's inequality. Denote $\mathbb{E}_Y$ and $\mathbb{E}_Z$ the conditional expectations with respect to Y and Z respectively. Condition on Y. Since $\mathbb{E}_Z\, Z = 0$, we have

$$F(Y) = F(Y + \mathbb{E}_Z\, Z) \le \mathbb{E}_Z\, F(Y + Z).$$

Taking expectation of both sides with respect to Y completes the proof. □

Proof of Decoupling Theorem 6.1.1. Here is what our proof of (6.4) will look like, in a nutshell. First, we will replace the chaos $\sum_{i\neq j} a_{ij}X_iX_j$ by the "partial chaos"

$$\sum_{(i,j)\in I\times I^c} a_{ij}X_iX_j$$

where the subset of indices $I \subset \{1,\dots,n\}$ will be chosen by random sampling. The advantage of partial chaos is that the summation is done over disjoint sets for i and j. Thus one can automatically replace X_j by X'_j without changing the distribution. Finally, we will complete the partial chaos $\sum_{(i,j)\in I\times I^c} a_{ij}X_iX'_j$ to the full sum using Lemma 6.1.2.

Let us pass to a detailed proof. To randomly select a subset of indices I, consider *selectors* $\delta_1,\dots,\delta_n \in \{0,1\}$ – independent Bernoulli random variables with $\mathbb{P}\{\delta_i = 0\} = \mathbb{P}\{\delta_i = 1\} = 1/2$. We define

$$I := \{i :\ \delta_i = 1\}.$$

We shall denote the conditional expectation with respect to the selectors (or, equivalently, with respect to the random subset I) by $\mathbb{E}_\delta$ and $\mathbb{E}_I$, and the conditional expectations with respect to X and X' by $\mathbb{E}_X$ and $\mathbb{E}_{X'}$ respectively. Since

$$\mathbb{E}\,\delta_i(1-\delta_i) = \frac{1}{2}\cdot\frac{1}{2} = \frac{1}{4} \quad \text{for all } i,$$

we may express our chaos as

$$\sum_{i,j=1}^n a_{ij}X_iX_j = 4\,\mathbb{E}_\delta \sum_{i,j=1}^n \delta_i(1-\delta_j)a_{ij}X_iX_j = 4\,\mathbb{E}_I \sum_{(i,j)\in I\times I^c} a_{ij}X_iX_j.$$

Then, using Jensen's and Fubini inequalities, we obtain

$$\mathbb{E}_X\, F(X^{\mathsf{T}}AX) \le \mathbb{E}_I\,\mathbb{E}_X\, F\Big(4\sum_{(i,j)\in I\times I^c} a_{ij}X_iX_j\Big).$$

It follows that there exists a realization of a random subset I such that

$$\mathbb{E}_X F(X^{\mathsf{T}}AX) \le \mathbb{E}_X F\Big(4 \sum_{(i,j)\in I\times I^c} a_{ij}X_iX_j\Big).$$

Fix such realization. Since the random variables $(X_i)_{i\in I}$ are independent from $(X_j)_{j\in I^c}$, the distribution of the sum in the right side will not change if we replace X_j by X'_j. So we obtained

$$\mathbb{E}\, F(X^{\mathsf{T}}AX) \le \mathbb{E}\, F\Big(4 \sum_{(i,j)\in I\times I^c} a_{ij}X_iX'_j\Big).$$

It remains to complete the sum in the right side to the sum over all pairs of indices. In other words, we want to show that

$$\mathbb{E}\, F\Big(4 \sum_{(i,j)\in I\times I^c} a_{ij}X_iX'_j\Big) \le \mathbb{E}\, F\Big(4 \sum_{(i,j)\in [n]\times[n]} a_{ij}X_iX'_j\Big), \tag{6.5}$$

where we use the notation $[n] = \{1,\dots,n\}$. To do this, we decompose the chaos in the right side as

$$\sum_{(i,j)\in[n]\times[n]} a_{ij}X_iX'_j = Y + Z_1 + Z_2$$

where

$$Y = \sum_{(i,j)\in I\times I^c} a_{ij}X_iX'_j, \quad Z_1 = \sum_{(i,j)\in I\times I} a_{ij}X_iX'_j, \quad Z_2 = \sum_{(i,j)\in I^c\times[n]} a_{ij}X_iX'_j.$$

Condition on all random variables except $(X'_j)_{j\in I}$ and $(X_i)_{i\in I^c}$. This fixes Y, while Z_1 and Z_2 are random variables with zero conditional expectations. Use Lemma 6.1.2 to conclude that the conditional expectation satisfies

$$F(4Y) \le \mathbb{E}\, F(4Y + 4Z_1 + 4Z_2).$$

Finally, we take the expectation of both sides with respect to all other random variables, and conclude that

$$\mathbb{E}\, F(4Y) \le \mathbb{E}\, F(4Y + 4Z_1 + 4Z_2).$$

This proves (6.5) and finishes the proof. □

Remark 6.1.3. The proof of Decoupling Theorem 6.1.1 generalizes without change to higher dimensions. If X_i are vectors in Hilbert space, then

$$\mathbb{E}\, F\Big(\sum_{ij} a_{ij} \langle X_i, X_j \rangle\Big) \le \mathbb{E}\, F\Big(4 \sum_{ij} a_{ij} \langle X_i, X_j' \rangle\Big).$$

(Check!) If X_i are vectors in $\mathbb{R}^m$ then this can be written in a matrix form similarly to (6.3). If X is an $m \times n$ random matrix with independent, mean zero columns, then

$$\mathbb{E}\, F(X^{\mathsf{T}} A X) \le \mathbb{E}\, F(4 X^{\mathsf{T}} A X').$$

(Check!)

Theorem 6.1.4 (Decoupling for vectors)**.** *Prove the following version of decoupling in normed spaces. Let $(u_{ij})_{i,j=1}^n$ be vectors in a normed space such that $u_{ii} = 0$ for all i. Let $X = (X_1, \dots, X_n)$ be a random vector with independent mean zero coefficients. Then, for every convex function F, one has*

$$\mathbb{E}\, F\Big(\Big\| \sum_{i,j} X_i X_j u_{ij} \Big\|\Big) \le F\Big(4\, \mathbb{E} \Big\| \sum_{i,j} X_i X_j' u_{ij} \Big\|\Big).$$

where X' is an independent copy of X.

Consider including a multivariate-subgaussian exercise, see my unpublished folder.

6.2 Hanson-Wright Inequality

We will now prove a general concentration inequality for a chaos. It can be viewed a as a chaos version of Berstein's inequalty.

Theorem 6.2.1 (Hanson-Wright inequality)**.** *Let $X = (X_1, \dots, X_n) \in \mathbb{R}^n$ be a random vector with independent, mean zero, sub-gaussian coordinates. Let A be an $n \times n$ matrix. Then, for every $t \ge 0$, we have*

$$\mathbb{P}\Big\{ |X^{\mathsf{T}} A X - \mathbb{E}\, X^{\mathsf{T}} A X| \ge t \Big\} \le 2 \exp\Big[- c \min\Big(\frac{t^2}{K^4 \|A\|_F^2}, \frac{t}{K^2 \|A\|} \Big)\Big],$$

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

We will first prove this result for Gaussian random variables X_i; then we will extend it to general distributions by an replacement trick.

Proof of Theorem 6.2.1 for normal distribution. Let us assume that $X \sim N(0, I_n)$. As usual, it is enough to bound the one-sided tail

$$p := \mathbb{P}\left\{X^{\mathsf{T}}AX - \mathbb{E}\, X^{\mathsf{T}}AX \geq t\right\}.$$

Indeed, once we have a bound on this upper tail, a similar bound will hold for the lower tail as well (since one can replace A with $-A$). By combining the two tails, we will complete the proof.

In terms of the entries of $A = (a_{ij})_{i,j=1}^n$, we have

$$X^{\mathsf{T}}AX = \sum_{i,j} a_{ij}X_iX_j \quad \text{and} \quad \mathbb{E}\, X^{\mathsf{T}}AX = \sum_i a_{ii}\,\mathbb{E}\, X_i^2,$$

where we used that $\mathbb{E}\, X_iX_j = 0$ by mean zero assumption and independence. So we can express the deviation as

$$X^{\mathsf{T}}AX - \mathbb{E}\, X^{\mathsf{T}}AX = \sum_i a_{ii}(X_i^2 - \mathbb{E}\, X_i^2) + \sum_{i,j:\, i\neq j} a_{ij}X_iX_j.$$

The problem reduces to estimating the diagonal and off-diagonal sums:

$$p \leq \mathbb{P}\left\{\sum_i a_{ii}(X_i^2 - \mathbb{E}\, X_i^2) \geq t/2\right\} + \mathbb{P}\left\{\sum_{i,j:\, i\neq j} a_{ij}X_iX_j \geq t/2\right\} =: p_1 + p_2.$$

Step 1: diagonal sum. Since X_i are independent, sub-gaussian random variables, $X_i^2 - \mathbb{E}\, X_i^2$ are independent, mean-zero, sub-exponential random variables, and

$$\|X_i^2 - \mathbb{E}\, X_i^2\|_{\psi_1} \lesssim \|X_i^2\|_{\psi_1} \lesssim \|X_i\|_{\psi_2}^2 \lesssim 1.$$

(This follows from the Centering Exercise 2.8.5 and Lemma 2.7.4 that states that a sub-gaussian random variable squared is sub-exponential.)

Then Bernstein's inequality (Theorem 2.8.3) yields

$$p_1 \leq \Big[-c\min\Big(\frac{t^2}{\sum_i a_{ii}^2}, \frac{t}{\max_i |a_{ii}|}\Big)\Big] \leq \exp\Big[-c\min\Big(\frac{t^2}{\|A\|_F^2}, \frac{t}{\|A\|}\Big)\Big].$$

Step 2: decoupling. It remains to bound the off-diagonal sum

$$S := \sum_{i,j:\, i\neq j} a_{ij}X_iX_j.$$

The argument will be based on estimating the moment generating function of S by decoupling and then using rotation invariance of the normal distribution.

Let $\lambda > 0$ be a parameter whose value we will determine later. By Chebyshev's inequality, we have

$$p_2 = \mathbb{P}\{S \geq t/2\} = \mathbb{P}\{\lambda S \geq \lambda t/2\} \leq \exp(-\lambda t/2)\, \mathbb{E} \exp(\lambda S). \tag{6.6}$$

By using decoupling (Theorem 6.1.1) , we can relate the moment generating function of S to that of

$$T := \sum_{i,j} a_{ij} X_i X'_j = X^{\mathsf{T}} A X'$$

where X' is an independent copy of X. We have

$$\mathbb{E} \exp(\lambda S) \leq \mathbb{E} \exp(4\lambda T). \tag{6.7}$$

(Note that we do not to remove the diagonal terms a_{ii} from T by Exercise **??**.)

Step 3: rotation invariance. Express A through its singular value decomposition

$$A = U \Sigma V^{\mathsf{T}}.$$

Then

$$T = X^{\mathsf{T}} A X' = \sum_i s_i \langle u_i, X \rangle \langle v_i, X' \rangle .$$

By rotation invariance of the normal distribution (see Exercise 3.3.2), $g := (\langle u_i, X \rangle)_{i=1}^n$ and $g' := (\langle v_i, X' \rangle)_{i=1}^n$ are independent standard normal random vectors in $\mathbb{R}^n$. Thus we can represent T as

$$T = \sum_i s_i g_i g'_i$$

where $g, g' \sim N(0, I_n)$ are independent and s_i are the singular values of A.

In this representation, T becomes a sum of independent random variables. This allows us to easily bound its moment generating function. Indeed, by independence we have

$$\mathbb{E} \exp(4\lambda T) = \prod_i \mathbb{E} \exp(4\lambda s_i g_i g'_i).$$

Now, $g_i g_i'$ are mean zero, sub-exponential random variables with

$$\|g_i g_i'\|_{\psi_1} \lesssim 1.$$

(Here we may use Lemma 2.7.5 which states that a product of two sub-gaussian random variables is sub-exponential.) We have already seen a bound on the moment generating function of sub-exponential random variables in Lemma 2.8.1. It implies that

$$\mathbb{E} \exp(4\lambda s_i g_i g_i') \le \exp(C\lambda^2 s_i^2) \quad \text{provided that } \lambda^2 s_i^2 \le c.$$

It follows that

$$\mathbb{E} \exp(4\lambda T) \le \exp\Big[C\lambda^2 \sum_i s_i^2\Big] \quad \text{provided that } |\lambda| \le \frac{c}{\max_i s_i}.$$

Recall that s_i are the singular values of A, so $\sum_i s_i^2 = \|A\|_F^2$ and $\max_i s_i = \|A\|$. Substituting this into the previous inequality and then into (6.7), we can bound the moment generating function of S as follows:

$$\mathbb{E} \exp(\lambda S) \le \exp(C\lambda^2 \|A\|_F^2) \quad \text{provided that } |\lambda| \le \frac{c}{\|A\|}.$$

Step 5: conclusion. Putting this bound into the exponential Chebyshev's inequality (6.6), we obtain

$$p_2 \le \exp\big(-\lambda t/2 + C\lambda^2 \|A\|_F^2\big) \quad \text{provided that } |\lambda| \le \frac{c}{\|A\|}.$$

Optimizing over λ, we conclude that

$$p_2 \le \exp\Big[-c \min\Big(\frac{t^2}{\|A\|_F^2}, \frac{t}{\|A\|}\Big)\Big].$$

Summarizing, we obtained the desired bounds for the probabilities of diagonal deviation p_1 and off-diagonal deviation p_2. Putting them together, we complete the proof. □

Our proof of Theorem 6.2.1 for general distributions will be based on a *replacement trick*, where we seek to replace general distributions by standard normal. The following comparison of moment generating functions will make this possible.

Lemma 6.2.2 (Comparison for MGF)**.** *Let X be a sub-gaussian random vector in $\mathbb{R}^n$ with $\|X\|_{\psi_2} \le K$ and $g \sim N(0, I_n)$. Then, for every fixed $\theta \in \mathbb{R}^n$ and every $\lambda \in \mathbb{R}$, we have*

$$\mathbb{E} \exp(\lambda \langle X, \theta \rangle) \le \mathbb{E} \exp(CK\lambda \langle g, \theta \rangle).$$

Proof. Without loss of generality, we may assume that $\|\theta\|_2 = 1$. (Why?) Then the bound $\|X\|_{\psi_2} \le K$ implies by definition that $\| \langle X, \theta \rangle \|_{\psi_2} \le K$. By (2.17), the moment generating function can be bounded as follows:

$$\mathbb{E} \exp(\lambda \langle X, \theta \rangle) \le \exp(CK^2\lambda^2) \quad \text{for all } \lambda \in \mathbb{R}.$$

On the other hand, $\langle g, \theta \rangle \sim N(0, 1)$. So, the formula (2.12) for the moment generating function of the normal distribution implies that

$$\mathbb{E} \exp(\lambda \langle g, \theta \rangle) = \exp(\lambda^2/2) \quad \text{for all } \lambda \in \mathbb{R}.$$

Comparing the two expressions, we complete the proof. □

Proof of Theorem 6.2.1 for general distributions. Now we assume that the random vector X may have a general distribution as in Theorem 6.2.1. Without loss of generality, we may assume that $K = 1$. (Why?)

Let us examine the previous argument. We had not used normal distribution until Step 3 which relied on the rotation invariance. In particular, we can control the diagonal sum can be controlled exactly as in Step 1 and obtain the same good bound for p_1.

To bound the contribution of the off-diagonal sum, the decoupling method works as before, and gives (6.7). The problem is now to bound the moment generating function

$$\mathbb{E} \exp(4\lambda T) \quad \text{for} \quad T = X^{\mathsf{T}} A X'.$$

We will do it by a replacement trick: we will replace X and X' by independent random vectors

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

using Lemma 6.2.2. (We can apply this lemma since X and X' are indeed sub-gaussian random vectors with $\|X\| \le C$ and $\|X'\| \le C$ by Lemma 3.4.2.)

Let us write $\mathbb{E}_X$ when we integrate with respect to X, and similarly for other random vectors. We have

$$\begin{aligned}
\mathbb{E}\exp(4\lambda T) &= \mathbb{E}_{X'}\,\mathbb{E}_X \exp(4\lambda \langle X, AX' \rangle) \\
&\le \mathbb{E}_{X'}\,\mathbb{E}_g \exp(C_1\lambda \langle g, AX' \rangle) && \text{(by Lemma 6.2.2 for } X\text{)} \\
&= \mathbb{E}_g\,\mathbb{E}_{X'} \exp(C_1\lambda \left\langle X', A^{\mathsf{T}} g \right\rangle) && \text{(by Fubini Theorem)} \\
&\le \mathbb{E}_g\,\mathbb{E}_{g'} \exp(C_2\lambda \left\langle g', A^{\mathsf{T}} g \right\rangle) && \text{(by Lemma 6.2.2 for } X'\text{)} \\
&= \mathbb{E}\exp(C_2\lambda g^{\mathsf{T}} A g').
\end{aligned}$$

Summarizing, we showed that X and X' in the definition of T can be replaced by standard normal random variables; the moment generating function of T will remain the same (except the constant 4 will change to some other absolute constant).

This means that from this point on, we can continue the argument as in the former proof for normal distributions. Theorem 6.2.1 is proved. □

Exercise 6.2.3. [Difficulty=7] *Give an alternative proof of Hanson-Wright inequality for normal distributions, without separating the diagonal part or decoupling. Use the singular value decomposition for A and rotation invariance of X to simplify and control the quadratic form $X^{\mathsf{T}}AX$.*

Exercise 6.2.4 (Comparison)**.** [Difficulty=6] *1. Let B be an $m \times n$ matrix, X be a sub-gaussian random vector in $\mathbb{R}^n$ with $\|X\|_{\psi_2} \le K$, and $g \sim N(0, I_n)$. Prove that for every $\lambda \in \mathbb{R}$, we have*

$$\mathbb{E}\exp(\lambda^2 \|BX\|_2^2) \le \mathbb{E}\exp(CK^2\lambda^2 \|Bg\|_2^2).$$

Hint: Check and use the identity $\mathbb{E}\exp(\lambda \langle g, x \rangle) = \exp(\lambda^2 \|x\|_2^2/2)$, which is valid for any fixed $x \in \mathbb{R}^n$ and $\lambda \ge 0$.

2. Check that for any λ such that $|\lambda| \le 1/\|B\|$, we have

$$\mathbb{E}\exp(\lambda^2 \|Bg\|_2^2) \le \exp(C\lambda^2 \|B\|_F^2).$$

Hint: Use rotation invariance to pass to a diagonal matrix.

6.2.1 Concentration of anisotropic random vectors

Change A to B in this section; we are applying HW with $A = B^{\mathsf{T}}B$.

As a consequence of Hanson-Wright inequality, we will now obtain concentration for *anisotropic* random vectors, which have the form AX, where A is a fixed matrix and X is an isotropic random vector.

Exercise 6.2.5. *Let A is an $m \times n$ matrix and X is an isotropic random vector in $\mathbb{R}^n$. Check that*

$$\mathbb{E}\,\|AX\|_2^2 = \|A\|_F^2.$$

Theorem 6.2.6 (Concentration of random vectors)**.** *Let A be an $m \times n$ matrix, and let $X = (X_1, \dots, X_n) \in \mathbb{R}^n$ be a random vector with independent, mean zero, unit variance, sub-gaussian coordinates. Then*

$$\Big\| \|AX\|_2 - \|A\|_F \Big\|_{\psi_2} \le CK^2 \|A\|,$$

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

An important partial case of this theorem when $A = I_n$. In this case, we have

$$\Big\| \|X\|_2 - \sqrt{n} \Big\|_{\psi_2} \le CK^2.$$

We proved this concentration result in Theorem 3.1.1 using Bernstein's inequality. And now we will prove a more general result, Theorem 6.2.6, using Hanson-Wright inequality.

Before we start the proof, note that the conclusion of Corollary 6.2.6 can be stated as a tail bound: for every $t \ge 0$, we have

$$\mathbb{P}\left\{ \big| \|AX\|_2 - \|A\|_F \big| > t \right\} \le 2 \exp\Big(- \frac{ct^2}{K^4 \|A\|^2} \Big). \tag{6.8}$$

Proof. Let us apply Hanson-Wright inequality, Theorem 6.2.1, for the matrix $Q = A^{\mathsf{T}} A$ instead of A. Then

$$X^{\mathsf{T}} Q X = \|AX\|_2^2, \qquad \mathbb{E}\, X^{\mathsf{T}} Q X = \|A\|_F^2 \quad \text{and } \|Q\| = \|A\|^2.$$

Thus we can state the conclusion of Hanson-Wright inequality as follows: for every $u \ge 0$, we have

$$\mathbb{P}\left\{ \big| \|AX\|_2^2 - \|A\|_F^2 \big| \ge u \right\} \le 2 \exp\Big[- \frac{c}{K^4} \min\Big(\frac{u^2}{\|A^{\mathsf{T}} A\|_F^2}, \frac{u}{\|A\|^2} \Big) \Big].$$

(Here we used that $K \ge c$. Why?)

To simplify this bound, note that

$$\|A^{\mathsf{T}} A\|_F^2 \le \|A^{\mathsf{T}}\|^2 \|A\|_F^2 = \|A\|^2 \|A\|_F^2.$$

(Check the inequality!) Using this and substituting the value $u = \varepsilon \|A\|_F^2$ for $\varepsilon \geq 0$, we obtain

$$\mathbb{P}\left\{ \left| \|AX\|_2^2 - \|A\|_F^2 \right| \geq \varepsilon \|A\|_F^2 \right\} \leq 2 \exp\Big[- c \min(\varepsilon^2, \varepsilon) \frac{\|A\|_F^2}{K^4 \|A\|^2} \Big].$$

It remains to remove the squares in the inequality in the left hand side. Denote $\delta^2 = \min(\varepsilon^2, \varepsilon)$, or equivalently set $\varepsilon = \max(\delta, \delta^2)$. Observe that that the following implication holds:

$$\text{If } \left| \|AX\|_2 - \|A\|_F \right| \geq \delta \|A\|_F \quad \text{then } \left| \|AX\|_2^2 - \|A\|_F^2 \right| \geq \varepsilon \|A\|_F^2.$$

(Check it!) Then we conclude that

$$\mathbb{P}\left\{ \left| \|AX\|_2 - \|A\|_F \right| \geq \delta \|A\|_F \right\} \leq 2 \exp\Big(- c\delta^2 \frac{\|A\|_F^2}{K^4 \|A\|^2} \Big).$$

Setting $\delta = t / \|A\|_F$, we obtain the desired inequality (6.8). □

A version of Theorem 6.2.6 holds for general sub-gaussian random vectors X in $\mathbb{R}^n$, without assuming independence of the coordinates.

Theorem 6.2.7 (Tails for sub-gaussian random vectors)**.** *Let B be an $m \times n$ matrix, and let X be a sub-gaussian random vector in $\mathbb{R}^n$ with $\|X\|_{\psi_2} \leq K$. Then for any $t \geq 0$, we have*

$$\mathbb{P}\left\{ \|BX\|_2 \geq CK \|B\|_F + t \right\} \leq \exp\Big(- \frac{ct^2}{K^2 \|B\|^2} \Big).$$

Exercise 6.2.8. [Difficulty=5] *1. Prove Theorem 6.2.7 by using the results in Exercise 6.2.4.*

2. Argue that there one can not have any non-trivial bound on the probability of the lower tail $\mathbb{P}\left\{ \|BX\|_2 \leq cK \|B\|_F - t \right\}$, even if X is isotropic.

Exercise 6.2.9 (Distance to a subspace)**.** [Difficulty=5] *Let E be a subspace of $\mathbb{R}^n$ of dimension d. Consider a random vector $X = (X_1, \dots, X_n) \in \mathbb{R}^n$ with independent, mean zero, unit variance, sub-gaussian coordinates.*

1. Check that

$$\big(\mathbb{E} \operatorname{dist}(X, E)^2 \big)^{1/2} = \sqrt{n - d}.$$

2. Prove that for any $t \geq 0$, the distance nicely concentrates:

$$\mathbb{P}\left\{ \left| d(X, E) - \sqrt{n - d} \right| > t \right\} \leq 2 \exp(-ct^2 / K^4),$$

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

6.3 Symmetrization

We say that a random variable X is *symmetric* if X and $-X$ are identically distributed random variables. A simplest example of a symmetric random variable is *symmetric Bernoulli*, which takes values -1 and 1 with probabilities $1/2$ each:

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

Normal, mean zero distribution $N(0, \sigma^2)$ is also symmetric, while Poisson or exponential distributions are not.

Exercise 6.3.1 (Constructing symmetric distributions)**.** *Let X be a random variable with zero mean, and ξ be an independent symmetric Bernoulli random variable.*

1. Check that ξX and $\xi|X|$ are symmetric random variables, and they have the same distribution.

2. If X is symmetric, show that the distribution of ξX and $\xi|X|$ is the same as of X.

3. Let X' be an independent copy of X. Check that $X - X'$ is symmetric.

In this section we will develop the simple and useful technique of *symmetrization*. It allows one to reduce problems about arbitrary distributions to symmetric distributions, and in some cases even to the symmetric Bernoulli distribution.

Throughout this section, we will denote by

$$\varepsilon_1, \varepsilon_2, \varepsilon_3, \dots$$

a sequence of independent symmetric Bernoulli random variables. We will assume that they are independent not only of each other, but also of any other random variables in question.

Lemma 6.3.2 (Symmetrization)**.** *Let $X_1, \dots, X_N$ be independent, mean zero random vectors in a normed space. Then*

$$\frac{1}{2}\,\mathbb{E}\left\|\sum_{i=1}^N \varepsilon_i X_i\right\| \le \mathbb{E}\left\|\sum_{i=1}^N X_i\right\| \le 2\,\mathbb{E}\left\|\sum_{i=1}^N \varepsilon_i X_i\right\|.$$

This lemma allows one to replace general random variables X_i by the symmetric random variables $\varepsilon_i X_i$.

Proof. **Upper bound.** Let (X_i') be an independent copy of the random vectors (X_i). Since $\sum_i X_i'$ has zero mean, we have

$$p := \mathbb{E} \left\| \sum_i X_i \right\| \le \mathbb{E} \left\| \sum_i X_i - \sum_i X_i' \right\| = \mathbb{E} \left\| \sum_i (X_i - X_i') \right\|.$$

To see this, use the following version of Lemma 6.1.2 for indepdendent random vectors Y and Z:

$$\text{if } \mathbb{E}\, Z = 0 \quad \text{then } \mathbb{E}\, \|Y\| \le \mathbb{E}\, \|Y + Z\|. \tag{6.9}$$

(Check it!)

Next, since $(X_i - X_i')$ are symmetric random vectors, they have the same distribution as $\varepsilon_i(X_i - X_i')$ (see Exercise 6.3.1). Then

$$\begin{aligned} p &\le \mathbb{E} \left\| \sum_i \varepsilon_i (X_i - X_i') \right\| \\ &\le \mathbb{E} \left\| \sum_i \varepsilon_i X_i \right\| + \mathbb{E} \left\| \sum_i \varepsilon_i X_i' \right\| \quad \text{(by triangle inequality)} \\ &= 2\, \mathbb{E} \left\| \sum_i \varepsilon_i X_i \right\| \quad \text{(since the two terms are identically distributed).} \end{aligned}$$

Lower bound. The argument here is similar:

$$\begin{aligned} \mathbb{E} \left\| \sum_i \varepsilon_i X_i \right\| &\le \mathbb{E} \left\| \sum_i \varepsilon_i (X_i - X_i') \right\| \quad \text{(using (6.9))} \\ &= \mathbb{E} \left\| \sum_i (X_i - X_i') \right\| \quad \text{(the distribution is the same)} \\ &= \mathbb{E} \left\| \sum_i X_i \right\| + \mathbb{E} \left\| \sum_i X_i' \right\| \quad \text{(by triangle inequality)} \\ &\le 2\, \mathbb{E} \left\| \sum_i X_i \right\| \quad \text{(by identical distribution).} \end{aligned}$$

This completes the proof of Symmetrization Lemma. □

Exercise 6.3.3. *Where in this argument did we use the independence of the random variables X_i? Is mean zero assumption needed for both upper and lower bounds?*

Exercise 6.3.4 (Removing the mean zero assumption)**.** [Difficulty=4] *1. Prove the following generalization of Symmetrization Lemma 6.3.2 for random vectors X_i that do not necessarily have zero means:*

$$\mathbb{E}\Big\|\sum_{i=1}^N X_i - \sum_{i=1}^N \mathbb{E}\, X_i\Big\| \le 2\,\mathbb{E}\Big\|\sum_{i=1}^N \varepsilon_i X_i\Big\|.$$

2. Argue that there can not be any non-trivial reverse inequality.

Exercise 6.3.5. *Prove the following generalization of Symmetrization Lemma 6.3.2. Let $F : \mathbb{R}_+ \to \mathbb{R}$ be an increasing, convex function. Show that the same inequalities in Lemma 6.3.2 hold if the norm $\|\cdot\|$ is replaced with $F(\|\cdot\|)$, namely*

$$\mathbb{E}\, F\Big(\frac{1}{2}\Big\|\sum_{i=1}^N \varepsilon_i X_i\Big\|\Big) \le \mathbb{E}\, F\Big(\Big\|\sum_{i=1}^N X_i\Big\|\Big) \le \mathbb{E}\, F\Big(2\Big\|\sum_{i=1}^N \varepsilon_i X_i\Big\|\Big).$$

Such generalization can be used to derive tail bounds *on the sums. In such situations, to bound the moment generating function one can use the function $F(x) = \exp(\lambda x)$.* Make an exercise?

Exercise 6.3.6. *Let $X_1, \dots, X_N$ be independent random variables. Show that their sum $\sum_i X_i$ is sub-gaussian if and only if $\sum_i \varepsilon_i X_i$ is sub-gaussian, and*

$$c\Big\|\sum_{i=1}^N \varepsilon_i X_i\Big\|_{\psi_2} \le \Big\|\sum_{i=1}^N X_i\Big\|_{\psi_2} \le C\Big\|\sum_{i=1}^N \varepsilon_i X_i\Big\|_{\psi_2}.$$

6.4 Random matrices with non-i.i.d. entries

A typical application of Symmetrization Lemma 6.3.2 has two steps. First, general random variables X_i are replaced by symmetric random variables $\varepsilon_i X_i$. Next, one conditions on X_i and leaves the entire randomness within ε_i. This reduces the problem to *symmetric Bernoulli* random variables ε_i, which are often simpler to deal with.

To illustrate this technique, we will prove a general bound on the norms of random matrices with independent but not identically distributed entries.

Theorem 6.4.1 (Norms of random matrices with non-i.i.d. entries)**.** *Let A be an $n \times n$ symmetric random matrix whose entries on and above the diagonal are independent, mean zero random variables. Then*

$$\mathbb{E}\,\|A\| \le C \log n \cdot \mathbb{E} \max_i \|A_i\|_2,$$

where A_i denote the rows of A.

This result is sharp up to the logarithmic factor. Indeed, since the operator norm of any matrix is bounded below by the Euclidean norms of the rows (why?), we trivially have

$$\mathbb{E}\,\|A\| \geq \mathbb{E}\max_i \|A_i\|_2.$$

Note also that unlike all results we have seen before, Theorem 6.4.1 does not require *any moment assumptions* on the entries of A.

Proof. The proof of Theorem 6.4.1 will be based on a combination of symmetrization with matrix Bernstein's inequality.

First, we decompose A into a sum of independent, mean zero, symmetric random matrices X_{ij}, each of which contains a pair of symmetric entries of A (or one diagonal entry). Previsely,

$$A = \sum_{i\leq j} Z_{ij}, \quad Z_{ij} := \begin{cases} A_{ij}(e_i e_j^{\mathsf{T}} + e_j e_i^{\mathsf{T}}), & i < j \\ A_{ii} e_i e_i^{\mathsf{T}}, & i = j \end{cases}$$

where (e_i) denotes the canonical basis of $\mathbb{R}^n$.

Apply Symmetrization Lemma 6.3.2. This gives

$$\mathbb{E}\,\|A\| = \mathbb{E}\,\Big\|\sum_{i\leq j} Z_{ij}\Big\| \leq 2\,\mathbb{E}\,\Big\|\sum_{i\leq j} \varepsilon_{ij} Z_{ij}\Big\|, \tag{6.10}$$

where (ε_{ij}) are independent symmetric Bernoulli random variables.

Now we condition on A. This fixes random variables (Z_{ij}), leaving all randomness within (ε_{ij}). Apply matrix Bernstein's inequality (Corollary 5.4.10) for

$$X_{ij} := \varepsilon_{ij} Z_{ij}.$$

Then the conditional expectation is bounded as follows:

$$\mathbb{E}_\varepsilon \Big\|\sum_{i\leq j} X_{ij}\Big\| \lesssim \sigma\sqrt{\log n} + K\log n, \tag{6.11}$$

where

$$\sigma^2 = \Big\|\sum_{i\leq j} \mathbb{E}\,X_{ij}^2\Big\| \quad \text{and} \quad K = \max_{ij}\|X_{ij}\|.$$

Now, a quick check verifies that

$$\mathbb{E}\,X_{ij}^2 = Z_{ij}^2 = \begin{cases} A_{ij}^2(e_i e_i^{\mathsf{T}} + e_j e_j^{\mathsf{T}}), & i < j \\ A_{ii}^2 e_i e_i^{\mathsf{T}}, & i = j. \end{cases}$$

The sum of these is the diagonal matrix

$$\sum_{i\le j}\mathbb{E}\,X_{ij}^2=\sum_{i<j}A_{ij}^2(e_ie_i^{\mathsf{T}}+e_je_j^{\mathsf{T}})+\sum_i A_{ii}^2e_ie_i^{\mathsf{T}}$$
$$\preceq 2\sum_{i=1}^n\Big(\sum_{j=1}^n A_{ij}^2\Big)e_ie_i^{\mathsf{T}}=2\sum_{i=1}^n\|A_i\|_2^2e_ie_i^{\mathsf{T}}.$$

(Check the matrix inequality!) Thus

$$\sigma=\Big\|\sum_{i\le j}\mathbb{E}\,X_{ij}^2\Big\|^{1/2}\le\sqrt{2}\max_i\|A_i\|_2$$

and, similarly,

$$K=\max_{i\le j}\|X_{ij}\|=\max_{i\le j}\|Z_{ij}\|\le 2\max_{i\le j}|A_{ij}|\le 2\max_i\|A_i\|_2.$$

Substituting the bounds for σ and K into matrix Bernstein's inequality (6.11), we get

$$\mathbb{E}_\varepsilon\Big\|\sum_{i\le j}X_{ij}\Big\|\lesssim\log n\cdot\max_i\|A_i\|_2.$$

Finally, we unfix A by taking expectation of both sides with respect to A (equivalently, with respect to (Z_{ij})). Using (6.10), we complete the proof. □

By using the so-called "Hermitization trick", we can obtain a version of Theorem 6.4.1 for non-symmetric matrices.

Corollary 6.4.2 (Non-symmetric matrices)**.** *Let A be an $m\times n$ random matrix whose entries are independent, mean zero random variables. Then*

$$\mathbb{E}\,\|A\|\le C\log(m+n)\cdot\big(\mathbb{E}\max_i\|A_i\|_2+\mathbb{E}\max_i\|A^j\|_2\big)$$

where A_i and A^j denote the rows and columns of A, respectively.

Proof. It is enough to apply Theorem 6.4.1 for the $(m+n)\times(m+n)$ symmetric random matrix

$$\begin{bmatrix}0 & A\\ A^{\mathsf{T}} & 0\end{bmatrix}.$$

(Write down the details!) □

Again, note that Corollary 6.4.2 is sharp up to the logarithmic factor. Indeed, since the operator norm of any matrix is bounded below by the Euclidean norms of the rows and columns, we trivially have

$$\mathbb{E}\,\|A\|\ge\mathbb{E}\max_{i,j}\big(\|A_i\|_2,\,\|A^j\|_2\big)\ge\frac{1}{2}\big(\mathbb{E}\max_i\|A_i\|_2+\mathbb{E}\max_i\|A^j\|_2\big).$$

6.5 Application: matrix completion

A remarkable application of the methods we have studied is to the problem of matrix completion. Suppose we are shown a few entries of a matrix; can we guess the other entries? We obviously can not, unless we know something else about the matrix. We will show that if the matrix has *low rank* then matrix completion is possible.

To describe the problem mathematically, consider a fixed $n \times n$ matrix X with

$$\operatorname{rank}(X) = r$$

where $r \ll n$. Suppose we are shown a few *randomly chosen entries* of X. Each entry X_{ij} is revealed to us independently with some probability $p \in (0, 1)$ and is hidden from us with probability $1 - p$. In other words, assume that we are shown the $n \times n$ matrix Y whose entries are

$$Y_{ij} := \delta_{ij} X_{ij} \quad \text{where} \quad \delta_{ij} \sim \text{Ber}(p) \text{ are independent.}$$

These δ_{ij} are *selectors* – Bernoulli random variables that indicate whether an entry is revealed to us or not (in the latter case, it is replaced with zero). If

$$p = \frac{m}{n^2} \tag{6.12}$$

then we are shown m entries of X on average.

How can we infer X from Y? Although X has small rank r by assumption, Y may not have small rank. (Why?) It is thus natural enforce small rank by choosing a *best rank r approximation* to Y.[1] The result, properly scaled, will be a good approximation to X:

Theorem 6.5.1 (Matrix completion)**.** *Let $\hat{X}$ be a best rank r approximation to $p^{-1}Y$. Then*

$$\mathbb{E}\, \frac{1}{n} \|\hat{X} - X\|_F \le C \log(n) \sqrt{\frac{rn}{m}}\, \|X\|_\infty, \tag{6.13}$$

as long as $m \ge n \log n$. Here $\|X\|_\infty = \max_{i,j} |X_{ij}|$ is the maximum magnitude of the entries of X.

[1] A *best rank k approximation* to a matrix A is obtained by minimizing $\|B - A\|$ over all rank k matrices B. The minimizer can be computed by truncating the singular value decomposition $A = \sum_i s_i u_i v_i^{\mathsf{T}}$ at k-th term, thus giving $B = \sum_{i=1}^{k} s_i u_i v_i^{\mathsf{T}}$. According to Eckart-Young-Minsky's theorem , the same holds not only for the operator norm but for general unitary-invariant norm, e.g. Frobienus.

Let us pause quickly to better understand the bound in this theorem. The assumption $m \gtrsim n \log n$ guarantees that we will see at least one entry in each row and column of X. (Why?) In the left side of (6.13), the recovery error can be expressed as

$$\frac{1}{n}\|\hat{X} - X\|_F = \Big(\frac{1}{n^2}\sum_{i,j=1}^{n} |\hat{X}_{ij} - X_{ij}|^2\Big)^{1/2}.$$

This is simply the average error per entry (in the L_2 sense). To make the right side small, assume that

$$m = Crn \log^2 n.$$

If $C > 0$ is a large constant, Theorem 6.5.1 gives an error that is much smaller than $\|X\|_\infty$. So, if the rank r of the original matrix X is small, the number of observed entries m needed for matrix completion is also small – much smaller than the total number of entries n^2.

To summarize, *matrix completion is possible if the number of observed entries exceeds rn by a logarithmic margin.* In this case, the expected average error per entry is much smaller than the maximal magnitude of an entry. Thus, for low rank matrices, matrix completion is possible with few observed entries.

Proof. We will first bound the recovery error in the operator norm, and then pass to the Frobenius norm using the low rank assumption.

Step 1: bounding the error in the operator norm. Using triangle inequality, let us split the error as follows:

$$\|\hat{X} - X\| \le \|\hat{X} - p^{-1}Y\| + \|p^{-1}Y - X\|.$$

Since we have chosen $\hat{X}$ as a best approximation to $p^{-1}Y$, the second summand dominates, i.e. $\|\hat{X} - p^{-1}Y\| \le \|p^{-1}Y - X\|$, so we have

$$\|\hat{X} - X\| \le 2\|p^{-1}Y - X\| = \frac{2}{p}\|Y - pX\|. \tag{6.14}$$

Note that the matrix $\hat{X}$, which would be hard to handle, has disappeared from the bound. Instead, $Y - pX$ is a matrix that is easy to understand. Its entries

$$(Y - pX)_{ij} = (\delta_{ij} - p)X_{ij}$$

are independent and mean zero random variables. So we can apply Corollary 6.4.2, which gives

$$\mathbb{E}\,\|Y-pX\| \le C\log n\cdot\big(\mathbb{E}\max_{i\in[n]}\|(Y-pX)_i\|_2+\mathbb{E}\max_{i\in[n]}\|(Y-pX)^j\|_2\big). \quad (6.15)$$

To bound the norms of the rows and columns of $Y-pX$, we can express them as

$$\|(Y-pX)_i\|_2^2=\sum_{j=1}^n(\delta_{ij}-p)^2X_{ij}^2\le\sum_{j=1}^n(\delta_{ij}-p)^2\cdot\|X\|_\infty^2,$$

and similarly for columns. These sums of independent random variables can be easily bounded using Bernstein's (or Chernoff's) inequality; this gives

$$\mathbb{E}\max_{i\in[n]}\sum_{j=1}^n(\delta_{ij}-p)^2\le Cpn.$$

(We will do this calculation in Exercise 6.5.2.) Combining with a similar bound for the columns and substituting into (6.15), we obtain

$$\mathbb{E}\,\|Y-pX\|\lesssim\log(n)\sqrt{pn}\,\|X\|_\infty.$$

Then, by (6.14), we get

$$\mathbb{E}\,\|\hat{X}-X\|\lesssim\log(n)\sqrt{\frac{n}{p}}\,\|X\|_\infty. \quad (6.16)$$

Step 2: passing to Frobenius norm. We have not used the low rank assumption yet, and this is what we will do now. Since $\operatorname{rank}(X)\le r$ by assumption and $\operatorname{rank}(\hat{X})\le r$ by construction, we have $\operatorname{rank}(\hat{X}-X)\le 2r$. The simple relationship (4.1) between the operator and Frobenius norms thus gives

$$\|\hat{X}-X\|_F\le\sqrt{2r}\|\hat{X}-X\|.$$

Taking expectations and using the bound on the error in the operator norm (6.16), we get

$$\mathbb{E}\,\|\hat{X}-X\|_F\le\sqrt{2r}\,\mathbb{E}\,\|\hat{X}-X\|\lesssim\log(n)\sqrt{\frac{rn}{p}}\,\|X\|_\infty.$$

Dividing both sides by n, we can rewrite this bound as

$$\mathbb{E}\,\frac{1}{n}\|\hat{X}-X\|_F\lesssim\log(n)\sqrt{\frac{rn}{pn^2}}\,\|X\|_\infty.$$

Recalling the definition (6.12) of p we have $pn^2=m$, so the desired bound (6.13) is proved. □

Exercise 6.5.2 (Bounding rows of random matrices). *Consider i.i.d. random variables $\delta_{ij} \sim \mathrm{Ber}(p)$, where $i, j = 1, \dots, n$. Assuming that $pn \geq \log n$, show that*

$$\mathbb{E} \max_{i \in [n]} \sum_{j=1}^{n} (\delta_{ij} - p)^2 \leq Cpn.$$

Hint: Fix i and use Bernstein's inequality (Corollary 2.8.4) to get a tail bound for $\sum_{j=1}^{n} (\delta_{ij} - p)^2$. Conclude by taking a union bound over $i \in [n]$.

Write this based on Seginer's theorem as in my tutorial. However, this theorem applies for iid entries, which we don't have here. Fix? Check and write down.

Remark 6.5.3 (Removal of logarithmic factor).

Remark 6.5.4 (Exact matrix completion).

Exercise 6.5.5 (Rectangular and symmetric matrices). *(a) State and prove a version of Matrix Completion Theorem 6.5.1 for general rectangular $n_1 \times n_2$ matrices X.*

(b) State and prove a version of Matrix Completion Theorem 6.5.1 for symmetric matrices X.

Exercise 6.5.6 (Noisy observations). *Extend Matrix Completion Theorem 6.5.1 to noisy observations, where we are shown noisy versions $X_{ij} + \nu_{ij}$ of some entries of X. Here ν_{ij} are independent and mean zero random variables representing noise.*

6.6 Application: covariance estimation for unbounded distributions

Let us give one more application of the symmetrization technique. We will prove a version of matrix Bernstein inequality for unbounded random variables, and then apply it for covariance estimation for unbounded distributions.

Theorem 6.6.1 (Matrix Bernstein for unbounded distributions). *Let $Z_1, \dots, Z_N$ be independent $n \times n$ positive-semidefinite random matrices. Consider the sum*

$$S := \sum_{i=1}^{N} Z_i.$$

Then

$$\mathbb{E} \|S - \mathbb{E} S\| \leq C\left(\sqrt{\|\mathbb{E} S\| \cdot L} + L\right) \tag{6.17}$$

where $L = \log n \cdot \mathbb{E} \max_i \|Z_i\|$.

Proof. Using symmetrization (see Exercise 6.3.4), we obtain

$$\mathbb{E}\,\|S - \mathbb{E}\,S\| \le 2\Big\|\sum_i \varepsilon_i Z_i\Big\|. \tag{6.18}$$

Condition on (Z_i) and apply matrix Bernstein inequality (Corollary 5.4.10) for the bounded random matrices $X_i = \varepsilon_i Z_i$. Afterwards, take expectation with respect to (Z_i). This gives

$$\mathbb{E}\,\|S - \mathbb{E}\,S\| \lesssim \mathbb{E}\left[\Big\|\sum_i Z_i^2\Big\|^{1/2}\sqrt{\log n} + \max_i \|Z_i\|\log n\right]$$

To simplify this bound, observe that

$$0 \preceq \sum_i Z_i^2 \preceq \max_i \|Z_i\| \cdot \sum_i Z_i = \max_i \|Z_i\| \cdot S.$$

(Check this!) This implies that

$$\Big\|\sum_i Z_i^2\Big\|^{1/2} \le \big(\max_i \|Z_i\|\big)^{1/2}\|S\|^{1/2}.$$

Taking expectation and using Cauchy-Schwarz inequality, we obtain

$$\mathbb{E}\Big\|\sum_i Z_i^2\Big\|^{1/2} \le \big(\mathbb{E}\max_i \|Z_i\| \cdot \mathbb{E}\,\|S\|\big)^{1/2}.$$

Substituting into (6.18) and denoting $L = \log(n)\,\mathbb{E}\max_i \|Z_i\|$, we get

$$\mathbb{E}\,\|S - \mathbb{E}\,S\| \lesssim \sqrt{\mathbb{E}\,\|S\| \cdot L} + L.$$

We almost obtained the desired conclusion (6.17), but not quite: we will need to replace $\mathbb{E}\,\|S\|$ with the smaller quantity $\|\mathbb{E}\,S\|$. By triangle inequality, we bound

$$\mathbb{E}\,\|S\| \le \mathbb{E}\,\|S - \mathbb{E}\,S\| + \|\mathbb{E}\,S\|.$$

Thus, denoting $x := \mathbb{E}\,\|S - \mathbb{E}\,S\|$, we have

$$x \lesssim \sqrt{(x + \|\mathbb{E}\,S\|) \cdot L} + L.$$

Solving this inequality and simplifying the solution, we get

$$x \lesssim \sqrt{\|\mathbb{E}\,S\| \cdot L} + L.$$

(Do this computation!) The proof is complete. □

Let us illustrate this theorem with an application to covariance estimation. In Section 5.7, we showed that

$$m \sim n \log n$$

samples are enough to estimate the covariance matrix of a general bounded distribution in $\mathbb{R}^n$. We will now relax the boundedness assumption (5.24), and will prove the following version of Theorem 5.7.1. (For simplicity, we prove only an expectation version here.)

Exercise 6.6.2 (Covariance estimation for unbounded distributions)**.** [Difficulty=4] *Consider a random vector X in $\mathbb{R}^n$ with zero mean and covariance matrix Σ. Let $\varepsilon \in (0,1)$ and $K > 0$. Suppose the sample size satisfies*

$$m \geq C(K/\varepsilon)^2 n \log n.$$

Assume also that

$$\big(\mathbb{E} \max_{i \leq m} \|X\|_2^2\big)^{1/2} \leq K\sqrt{\|\Sigma\| n}. \tag{6.19}$$

Then

$$\mathbb{E}\,\|\Sigma_m - \Sigma\| \leq \varepsilon \|\Sigma\|.$$

Hint: Apply Theorem 6.6.1 for $Z_i = \frac{1}{m} X_i X_i^{\mathsf{T}}$.

6.7 Contraction Principle

We conclude this chapter with one more useful comparison inequality. Here will will keep denoting by $\varepsilon_1, \varepsilon_2, \varepsilon_3, \ldots$ a sequence of independent symmetric Bernoulli random variables. (which is also independent of any other random variables in question).

Theorem 6.7.1 (Contraction principle)**.** *Let $x_1, \ldots, x_N$ be (deterministic) vectors in a normed space, and let $a = (a_1, \ldots, a_n) \in \mathbb{R}^n$ be a coefficient vector. Then*

$$\mathbb{E}\Big\| \sum_{i=1}^N a_i \varepsilon_i x_i \Big\| \leq \|a\|_\infty \cdot \mathbb{E}\Big\| \sum_{i=1}^N \varepsilon_i x_i \Big\|.$$

Proof. Without loss of generality, we may assume that $\|a\|_\infty \leq 1$. (Why?) Define the function

$$f(a) := \mathbb{E}\Big\| \sum_{i=1}^N a_i \varepsilon_i x_i \Big\|.$$

Then $f : \mathbb{R}^N \to \mathbb{R}^N$ is a convex function. (Check!)

We would like find an upper bound for f on the set of points a satisfying $\|a\|_\infty \le 1$, i.e. on the unit cube $[-1, 1]^n$. Recall that every convex function defined on a closed, compact set attains its maximum on an extreme point of the set. (Check this!) Then the maximum of f is attained at a vertex of the unit cube, i.e. at a point a whose coefficients are all $a_i = \pm 1$.

For this point a, the random variables $(\varepsilon_i a_i)$ have the same distribution as (ε_i) due to symmetry. Thus

$$\mathbb{E} \Big\| \sum_{i=1}^N a_i \varepsilon_i x_i \Big\| = \mathbb{E} \Big\| \sum_{i=1}^N \varepsilon_i x_i \Big\|.$$

The proof is complete. □

Using symmetrization, we can immediately extend the contraction principle for general distributions.

Theorem 6.7.2 (Contraction principle for general distributions)**.** *Let $X_1, \dots, X_N$ be independent, mean zero random vectors in a normed space, and let $a = (a_1, \dots, a_n) \in \mathbb{R}^n$ be a coefficient vector. Then*

$$\mathbb{E} \Big\| \sum_{i=1}^N a_i X_i \Big\| \le 4 \|a\|_\infty \cdot \mathbb{E} \Big\| \sum_{i=1}^N X_i \Big\|.$$

Proof. It is enough to apply symmetrization (Lemma 6.3.2), then use contraction principle (Theorem 6.7.2) conditioned on (X_i), and finish by applying symmetrization again. (Write down the details!) □

As an application, we will now show how symmetrization can be done using *Gaussian* random variables $g_i \sim N(0, 1)$ instead of symmetric Bernoulli random variables ε_i.

Lemma 6.7.3 (Symmetrization with Gaussians)**.** *Let $X_1, \dots, X_N$ be independent, mean zero random vectors in a normed space. Let $g_1, \dots, g_N \in \mathbb{N}(0, 1)$ be independent Gaussian random variables, which are also independent of X_i. Then*

$$\frac{c}{\sqrt{\log N}} \mathbb{E} \Big\| \sum_{i=1}^N g_i X_i \Big\| \le \mathbb{E} \Big\| \sum_{i=1}^N X_i \Big\| \le 3 \mathbb{E} \Big\| \sum_{i=1}^N g_i X_i \Big\|.$$

Proof. This is a standard argument, which combines symmetrization (Lemma 6.3.2) and contraction (Theorem 6.7.2). **Upper bound.** We have

$$\mathbb{E}\Big\|\sum_{i=1}^N X_i\Big\| \le 2\,\mathbb{E}\Big\|\sum_{i=1}^N \varepsilon_i X_i\Big\| \quad \text{(by symmetrization)}.$$

To interject Gaussian random variables, recall that $\mathbb{E}\,|g_i| = \sqrt{2/\pi}$. Thus we can continue our bound as follows:

$$\begin{aligned}
&\le 2\sqrt{\frac{\pi}{2}}\;\mathbb{E}_X\Big\|\sum_{i=1}^N \varepsilon_i\,\mathbb{E}_g\,|g_i|X_i\Big\| \\
&\le 2\sqrt{\frac{\pi}{2}}\;\mathbb{E}\Big\|\sum_{i=1}^N \varepsilon_i|g_i|X_i\Big\| \quad \text{(by Jensen's inequality)} \\
&= 2\sqrt{\frac{\pi}{2}}\;\mathbb{E}\Big\|\sum_{i=1}^N g_i X_i\Big\|.
\end{aligned}$$

The last equality follows by symmetry of Gaussian distribution, which implies that the random variables $\varepsilon_i|g_i|$ have the same distribution as g_i (recall Exercise 6.3.1).

Lower bound. Condition on the random vector $g = (g_i)_{i=1}^N$ and apply the contraction principle (Theorem 6.7.2). This gives

$$\begin{aligned}
\mathbb{E}\Big\|\sum_{i=1}^N g_i X_i\Big\| &\le \mathbb{E}_g\Big(\|g\|_\infty \cdot \mathbb{E}_X\Big\|\sum_{i=1}^N X_i\Big\|\Big) \\
&\le \Big(\mathbb{E}\,\|g\|_\infty\Big)\Big(\mathbb{E}\Big\|\sum_{i=1}^N X_i\Big\|\Big) \quad \text{(by independence)}.
\end{aligned}$$

It remains to recall from Exercise 2.5.8 that

$$\mathbb{E}\,\|g\|_\infty \le C\sqrt{\log N}.$$

The proof is complete. □

Exercise 6.7.4. *Show that the factor $\sqrt{\log N}$ in Lemma 6.7.3 is needed in general, and is optimal. (This is where symmetrization with Gaussian random variables is weaker than symmetrization with symmetric Bernoullis.)*

Exercise 6.7.5 (Symmetrization and contraction for functions of norms)**.** *Let $F : \mathbb{R}_+ \to \mathbb{R}$ be a convex increasing function. Generalize the symmetrization and contraction results of this and previous section by replacing the norm $\|\cdot\|$ with $F(\|\cdot\|)$ throughout.*

This generalization is useful for the functions of the form $F(z) = \exp(\lambda z)$, since it allows to bound the moment generating functions. Such bounds, as we know, are instrumental in proving tail bounds *rather than bounds on expectation.*

Chapter 7

Random processes

7.1 Basic concepts and examples

Definition 7.1.1 (Random process)**.** *A* random process *is a collection of random variables $(X_t)_{t\in T}$ on the same probability space, and indexed by elements t of a set T. As a standing assumption, we will always suppose that*

$$\mathbb{E}\, X_t = 0 \quad \textit{for all } t \in T,$$

that is the process has zero drift.

The variable t is classically thought of as *time*, and in this case T is a subset of $\mathbb{R}$. But we will primarily study processes in high-dimensional settings, where T is a subset of $\mathbb{R}^n$ and where the analogy with time will be loose.

Example 7.1.2 (Discrete time). If $T = \{1, \dots, n\}$ then the random process

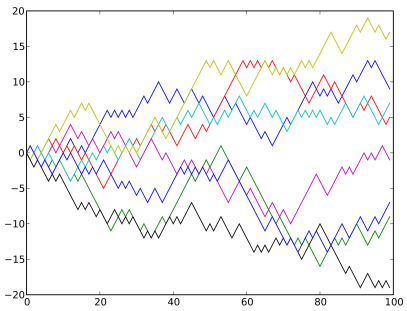
$$(X_1, \dots, X_n)$$

can be identifies with a *random vector* in $\mathbb{R}^n$.

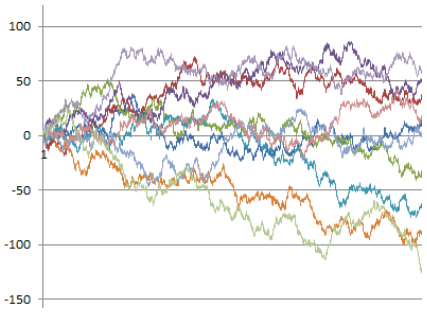
Example 7.1.3 (Random walks). If $T = \mathbb{N}$, a discrete-time random process $(X_n)_{n\in\mathbb{N}}$ is simply a *sequence* of random variables. An important example is a *random walk* defined as

$$X_n := \sum_{i=1}^{n} Z_i,$$

where the increments Z_i are independent, mean zero random variables. See Figure 7.1 for illustration.



(a) A few trials of a random walk with symmetric Bernoulli Z_i.



(b) A few trials of the standard Brownian motion.

Figure 7.1: Random walks and Brownian motion in $\mathbb{R}$.

Example 7.1.4 (Brownian motion). The most classical continuous-time random process is the standard *Brownian motion* $(X_t)_{t\geq 0}$, also called the *Wiener process*. It can be characterized as follows:

- The process has continuos sample paths, i.e. the random function $t \mapsto X_t$ is almost surely continuous.
- The increments satisfy $X_t - X_s \sim N(0, t-s)$ for all $t \geq s$.

Figure 7.1 illustrates a few trials of the standard Brownian motion.

Example 7.1.5 (Random fields). When the index set T is a subset of $\mathbb{R}^n$, a random process $(X_t)_{t\in T}$ are sometimes called a spacial random process, or a *random field*. For example, if X_t denotes the water temperature at the location on Earth that is parametrized by t, it can be modeled as a spacial random process.

7.1.1 Covariance and increments

Similarly to covariance matrices for random vectors, we may define the *covariance function* for a random process $(X_t)_{t\in T}$ as

$$\Sigma(t,s) := \operatorname{cov}(X_t, X_s) = \mathbb{E}\, X_t X_s, \quad t,s \in T.$$

We may also study a random process $(X_t)_{t\in T}$ through the behavior of its *increments*

$$\|X_t - X_s\|_2 = \big(\mathbb{E}(X_t - X_s)^2\big)^{1/2}, \quad t,s \in T.$$

Example 7.1.6. The increments of the standard Brownian motion satisfy

$$\|X_t - X_s\|_2 = \sqrt{t-s}, \quad t \geq s$$

by definition. The increments of a random walk of Example 7.1.3 with $\mathbb{E}_i Z_i = 1$ behave similarly:

$$\|X_n - X_m\|_2 = \sqrt{n-m}, \quad n \geq m.$$

(Check!)

In principle, the index set T may be an abstract set without any structure. But the increments define a *metric* on T,

$$d(t,s) := \|X_t - X_s\|_2, \quad t, s \in T, \tag{7.1}$$

thus turning T into a *metric space*. The examples above show, however, that this metric may not agree with the canonical metric on $\mathbb{R}$.

Exercise 7.1.7 (Covariance and increments)**.** [Difficulty=3] *Let $(X_t)_{t\in T}$ be a random process.*

1. Express the increments $\|X_t - X_s\|_2$ in terms of the covariance function $\Sigma(t,s)$.

2. Assuming that the zero random variable 0 belongs to the process, express the covariance function $\Sigma(t,s)$ in terms of the increments $\|X_t - X_s\|_2$. This shows that the distribution of a Gaussian random process containing the zero random variable is completely determined by the increments $\|X_t - X_s\|_2$, $t, s \in T$.

7.2 Gaussian processes

Definition 7.2.1 (Gaussian process)**.** *A random process $(X_t)_{t\in T}$ is called a* Gaussian process *if, for any finite subset $T_0 \subset T$, the random vector*

$$(X_t)_{t\in T_0}$$

has normal distribution. Equivalently, $(X_t)_{t\in T}$ is Gaussian if every finite linear combination

$$\sum_{t\in T_0} a_t X_t$$

is a normal random variable.

A classical example of a Gaussian process is the standard Brownian motion.

The notion of Gaussian processes generalizes that of Gaussian random vectors in $\mathbb{R}^n$.

From the formula (3.4) for multivariate normal density we may recall that the distribution of a Gaussian random vector X in $\mathbb{R}^n$ is completely determined by its covariance matrix. Then, by definition, the distribution of a Gaussian process $(X_t)_{t\in T}$ is also completely determined[1] by its covariance function $\Sigma(t,s)$.

7.2.1 Canonical Gaussian processes

We will now consider a wide class of examples of a Gaussian processes indexed by higher-dimensional sets $T \subset \mathbb{R}^n$. Consider the standard normal random vector $g \sim N(0,1)$ and define the random process

$$X_t := \langle g, t \rangle, \quad t \in T. \tag{7.2}$$

Then $(X_t)_{t\in T}$ is clearly a Gaussian process, and we call it a *canonical Gaussian process*. The increments of this process define the Euclidean distance

$$\|X_t - X_s\|_2 = \|t - s\|_2, \quad t, s \in T.$$

(Check!)

One can essentially realize any Gaussian process as the canonical process (7.2). This follows from a simple observation about Gaussian vectors.

Lemma 7.2.2 (Gaussian random vectors)**.** *Let Y be a mean zero Gaussian random vector in $\mathbb{R}^n$. Then there exist points $t_1, \dots, t_n \in \mathbb{R}^n$ such that*

$$Y \equiv (\langle g, t_i \rangle)_{i=1}^n, \quad \text{where } g \sim N(0, I_n).$$

Here "$\equiv$" means that the distributions of the two random vectors are the same.

Proof. Let Σ denote the covariance matrix of Y. Then we may realize

$$Y \equiv \Sigma^{1/2} g \quad \text{where } g \sim N(0, I_n)$$

(recall Section 3.3.2). Next, the coordinates of the vector $\Sigma^{1/2} g$ are $\langle t_i, g \rangle$ where t_i denote the rows of the matrix $\Sigma^{1/2}$. This completes the proof. □

[1] To avoid measurability issues, we do not formally define the distribution of a random process here. So the statement above should be understood as the fact that the covariance function determines the distribution of all marginals $(X_t)_{t\in T_0}$ with finite $T_0 \subset T$.

It follows that for any Gaussian process $(Y_s)_{s\in S}$, all finite-dimensional marginals $(Y_s)_{s\in S_0}$, $|S_0| = n$ can be represented as the canonical Gaussian process (7.2) indexed in a certain subset $T_0 \subset \mathbb{R}^n$.

Exercise 7.2.3. *Realize an N-step random walk of Example 7.1.3 with $Z_i \sim N(0,1)$ as a canonical Gaussian process (7.2) with $T \subset \mathbb{R}^N$. Hint: It might be simpler to think about increments $\|X_t - X_s\|_2$ instead of the covariance matrix.*

7.3 Slepian's inequality

In many applications, it is useful to have a uniform control on a random process $(X_t)_{t\in T}$, that is to have a bound on[2]

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

For the standard Brownian motion, the answer is known exactly. As a consequence of the so-called *reflection principle*, we have

$$\mathbb{E}\sup_{t\le t_0} X_t = \sqrt{\frac{2t_0}{\pi}} \quad \text{for every } t_0 \ge 0.$$

For general random processes, even Gaussian, this problem is very non-trivial.

The first general bound we will prove is Slepian's comparison inequality for Gaussian processes. It basically states that the faster the process grows (in terms of the magnitude of increments), the farther it gets.

Theorem 7.3.1 (Slepian's inequality)**.** *Let $(X_t)_{t\in T}$ and $(Y_t)_{t\in T}$ be two Gaussian processes. Assume that for all $t, s \in T$, we have*

$$\mathbb{E}\, X_t^2 = \mathbb{E}\, Y_t^2 \quad \text{and} \quad \mathbb{E}(X_t - X_s)^2 \le \mathbb{E}(Y_t - Y_s)^2. \tag{7.3}$$

Then for every $\tau \ge 0$ we have

$$\mathbb{P}\left\{\sup_{t\in T} X_t \ge \tau\right\} \le \mathbb{P}\left\{\sup_{t\in T} Y_t \ge \tau\right\}. \tag{7.4}$$

Consequently,

$$\mathbb{E}\sup_{t\in T} X_t \le \mathbb{E}\sup_{t\in T} Y_t. \tag{7.5}$$

Whenever the tail comparison inequality (7.4) holds, we say that the random variable X is *stochastically dominated* by the random variable Y.

We will prove Slepian's inequality now.

[2] To avoid measurability issues, we will study random processes through their finite-dimensional marginals as before. Thus we interpret $\mathbb{E}\sup_{t\in T} X_t$ more formally as $\sup_{T_0\subset T} \mathbb{E}\max_{t\in T_0} X_t$ where the supremum is over all finite subsets $T_0 \subset T$.

7.3.1 Gaussian interpolation

The proof of Slepian's inequality that we are about to give will be based on the useful technique of *Gaussian interpolation*. Let us describe it briefly. Assume that T is finite; then $X = (X_t)_{t\in T}$ and $Y = (Y_t)_{t\in T}$ are Gaussian random vectors in $\mathbb{R}^n$. We may also assume that X and Y are independent. (Why?) Define the Gaussian random vector $Z(u)$ that continuously interpolates between $Z(0) = Y$ and $Z(1) = X$:

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

Then the covariance matrix of $Z(u)$ interpolates linearly between the covariance matrices of Y and X:

$$\Sigma(Z(u)) = u\,\Sigma(X) + (1-u)\,\Sigma(Y).$$

(Check this!)

We will study how the quantity

$$\mathbb{E}\, f(Z(u))$$

changes as u increases from 0 to 1, where $f : \mathbb{R}^n \to \mathbb{R}$ is a general function. Of specific interest to us is f that is a smooth approximation to the indicator function $\mathbf{1}_{\{\max_i x_i < u\}}$. We will be able to show that in this case, $\mathbb{E}\, f(Z(u))$ increases in u. This would imply the conclusion of Slepian's inequality at once, since then

$$\mathbb{E}\, f(Z(1)) \geq \mathbb{E}\, f(Z(0)), \quad \text{thus} \quad \mathbb{P}\left\{\max_i X_i < \tau\right\} \geq \mathbb{P}\left\{\max_i Y_i < \tau\right\}$$

as claimed.

We will now pass to a detailed argument. To develop Gaussian interpolation, let us start with the following useful identity.

Lemma 7.3.2 (Gaussian integration by parts)**.** *Let $X \sim N(0,1)$. Then for any differentiable function $f : \mathbb{R} \to \mathbb{R}$ we have*

$$\mathbb{E}\, f'(X) = \mathbb{E}\, X f(X).$$

Proof. Assume first that f has bounded support. Denoting the Gaussian density of X by

$$p(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2},$$

we can express the expectation as an integral, and integrate it by parts:

$$\mathbb{E}\, f'(X) = \int_{\mathbb{R}} f'(x)p(x)\,dx = -\int_{\mathbb{R}} f(x)p'(x)\,dx. \tag{7.6}$$

Now, a direct check gives

$$p'(x) = -xp(x),$$

so the integral in (7.6) equals

$$\int_{\mathbb{R}} f(x)p(x)x\,dx = \mathbb{E}\, Xf(X),$$

as claimed. The identity can be extended to general functions by an approximation argument. The lemma is proved. □

Exercise 7.3.3. *If $X \sim N(0, \sigma^2)$, show that*

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

Hint: Represent $X = \sigma Z$ for $Z \sim N(0, 1)$, and apply Gaussian integration by parts.

Gaussian integration by parts generalizes nicely to high dimenions.

Lemma 7.3.4 (Multivariate Gaussian integration by parts)**.** *Let $X \sim N(0, \Sigma)$. Then for any differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ we have*

$$\mathbb{E}\, Xf(X) = \Sigma \cdot \mathbb{E}\, \nabla f(X).$$

Exercise 7.3.5. [Difficulty=4] *Prove Lemma 7.3.4. According to the matrix-by-vector multiplication, note that the conclusion of the lemma is equivalent to*

$$\mathbb{E}\, X_i f(X) = \sum_{j=1}^{n} \Sigma_{ij}\, \mathbb{E}\, \frac{\partial f}{\partial x_j}(X), \quad i = 1, \ldots, n. \tag{7.7}$$

Hint: Represent $X = \Sigma^{1/2} Z$ for $Z \sim N(0, I_n)$. Then

$$X_i = \sum_{k=1}^{n} (\Sigma^{1/2})_{ik} Z_k \quad \text{and} \quad \mathbb{E}\, X_i f(X) = \sum_{k=1}^{n} (\Sigma^{1/2})_{ik}\, \mathbb{E}\, Z_k f(\Sigma^{1/2} Z).$$

Apply univariate Gaussian integration by parts (Lemma 7.3.2) for $\mathbb{E}\, Z_k f(\Sigma^{1/2} Z)$ as a function of $Z_k \sim N(0, 1)$, and simplify.

Lemma 7.3.6 (Gaussian interpolation)**.** *Consider two independent Gaussian random vectors $X \sim N(0, \Sigma^X)$ and $Y \sim N(0, \Sigma^Y)$. Define the interpolation Gaussian vector*

$$Z(u) := \sqrt{u}\, X + \sqrt{1-u}\, Y, \quad u \in [0,1]. \tag{7.8}$$

Then for any twice-differentiable function $f : \mathbb{R}^n \to \mathbb{R}$, we have

$$\frac{d}{du}\, \mathbb{E}\, f(Z(u)) = \frac{1}{2} \sum_{i,j=1}^{n} (\Sigma_{ij}^X - \Sigma_{ij}^Y)\, \mathbb{E}\left[\frac{\partial^2 f}{\partial x_i\, \partial x_j}(Z(u))\right]. \tag{7.9}$$

Proof. Using the chain rule,[3] we have

$$\begin{aligned} \frac{d}{du}\, \mathbb{E}\, f(Z(u)) &= \sum_{i=1}^{n} \mathbb{E}\, \frac{\partial f}{\partial x_i}(Z(u)) \frac{dZ_i(t)}{du} \\ &= \frac{1}{2} \sum_{i=1}^{n} \mathbb{E}\, \frac{\partial f}{\partial x_i}(Z(u)) \Big(\frac{X_i}{\sqrt{u}} - \frac{Y_i}{\sqrt{1-u}} \Big) \quad \text{(by (7.8))}. \end{aligned} \tag{7.10}$$

Let us break this sum into two, and first compute the contribution of the terms containing X_i. To this end, we condition on Y and express

$$\sum_{i=1}^{n} \frac{1}{\sqrt{u}}\, \mathbb{E}\, X_i \frac{\partial f}{\partial x_j}(Z(u)) = \sum_{i=1}^{n} \frac{1}{\sqrt{u}}\, \mathbb{E}\, X_i g_i(X), \tag{7.11}$$

where

$$g_i(X) = \frac{\partial f}{\partial x_i}(\sqrt{u}\, X + \sqrt{1-u}\, Y).$$

Apply the multivariate Gaussian integration by parts (Lemma 7.3.4). According to (7.7), we have

$$\begin{aligned} \mathbb{E}\, X_i g_i(X) &= \sum_{j=1}^{n} \Sigma_{ij}^X\, \mathbb{E}\, \frac{\partial g_i}{\partial x_j}(X) \\ &= \sum_{j=1}^{n} \Sigma_{ij}^X\, \mathbb{E}\, \frac{\partial^2 f}{\partial x_i\, \partial x_j}(\sqrt{u}\, X + \sqrt{1-u}\, Y) \cdot \sqrt{u}. \end{aligned}$$

Substitute this into (7.11) to get

$$\sum_{i=1}^{n} \frac{1}{\sqrt{u}}\, \mathbb{E}\, X_i \frac{\partial f}{\partial x_j}(Z(u)) = \sum_{i,j=1}^{n} \Sigma_{ij}^X\, \mathbb{E}\, \frac{\partial^2 f}{\partial x_i\, \partial x_j}(Z(u)).$$

[3] Here we use the multivariate chain rule to differentiate a function $f(g_1(u), \dots, g_n(u))$ as follows: $\frac{df}{du} = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} \frac{dg_i}{du}$.

Taking expectation of both sides with respect to Y, we lift the conditioning on Y.

We can simiarly evaluate the other sum in (7.10), the one containing the terms Y_i. Combining the two sums we complete the proof. □

7.3.2 Proof of Slepian's inequality

We are ready to establish a preliminary, functional form Slepian's inequality.

Lemma 7.3.7 (Slepian's inequality, functional form)**.** *Consider two mean zero Gaussian random vectors X and Y in $\mathbb{R}^n$. Assume that for all $i, j = 1, \dots, n$, we have*

$$\mathbb{E}\, X_i^2 = \mathbb{E}\, Y_i^2 \quad \text{and} \quad \mathbb{E}(X_i - X_j)^2 \le \mathbb{E}(Y_i - Y_j)^2.$$

Consider a twice-differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ such that

$$\frac{\partial^2 f}{\partial x_i\, \partial x_j} \ge 0 \quad \text{for all } i \ne j.$$

Then

$$\mathbb{E}\, f(X) \ge \mathbb{E}\, f(Y).$$

Proof. The assumptions imply that the entries of the covariance matrices Σ^X and Σ^Y of X and Y satisfy

$$\Sigma^X_{ii} = \Sigma^Y_{ii} \quad \text{and} \quad \Sigma^X_{ij} \ge \Sigma^Y_{ij}.$$

for all $i, j = 1, \dots, n$. We can assume that X and Y are independent. (Why?) Apply Lemma 7.3.6 and using our assumptions, we conclude that

$$\frac{d}{du}\, \mathbb{E}\, f(Z(u)) \ge 0,$$

so $\mathbb{E}\, f(Z(u))$ increases in u. Then $\mathbb{E}\, f(Z(1)) = \mathbb{E}\, f(X)$ is at least as large as $\mathbb{E}\, f(Z(0)) = \mathbb{E}\, f(Y)$. This completes the proof. □

Now we are ready to prove Slepian's inequality, Theorem 7.3.1. Let us state and prove it in the equivalent form for Gaussian random vectors.

Theorem 7.3.8 (Slepian's inequality)**.** *Let X and Y be Gaussian random vectors as in Lemma 7.3.7. Then for every $u \ge 0$ we have*

$$\mathbb{P}\left\{\max_{i \le n} X_i \ge \tau\right\} \le \mathbb{P}\left\{\max_{i \le n} Y_i \ge \tau\right\}.$$

Consequently,

$$\mathbb{E} \max_{i \le n} X_i \le \mathbb{E} \max_{i \le n} Y_i.$$

Proof. Let $h : \mathbb{R} \to [0,1]$ be a twice-differentiable, non-increasing approximation to the indicator function of the interval $(-\infty, \tau)$:

$$h(x) \approx \mathbf{1}_{(-\infty,\tau)},$$

see Figure 7.2. Define the function $f : \mathbb{R}^n \to \mathbb{R}$ by

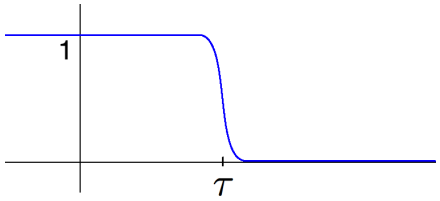


Figure 7.2: The function $h(x)$ is a smooth, non-increasing approximation to the indicator function $\mathbf{1}_{(-\infty,\tau)}$.

$$f(x) = h(x_1) \cdots h(x_n).$$

Then $f(x)$ is an approximation to the indicator function

$$f(x) \approx \mathbf{1}_{\{\max_i x_i < \tau\}}.$$

We are looking to apply the functional form of Slepian's inequality, Lemma 7.3.7, for $f(x)$. To check the assumptions of this result, note that for $i \neq j$ we have

$$\frac{\partial^2 f}{\partial x_i \, \partial x_j} = h'(x_i)h'(x_j) \cdot \prod_{k \notin \{i,j\}} h(x_k).$$

The first two factors are negative and the others are positive by the assumption. Thus second derivative is positive, as required.

It follows that

$$\mathbb{E}\, f(X) \geq \mathbb{E}\, f(Y).$$

By approximation, this implies

$$\mathbb{P}\left\{\max_{i \leq n} X_i < \tau\right\} \geq \mathbb{P}\left\{\max_{i \leq n} Y_i < \tau\right\}.$$

This proves the first part of the conclusion. The second part follows using the integral identity in Lemma 1.2.1, see Exercise 7.3.9. □

Exercise 7.3.9. *Using the integral identity in Lemma 1.2.1, deduce the second part of Slepian's inequality (comparison of expectations).*

7.3.3 Sudakov-Fernique's and Gordon's inequalities

Slepian's inequality has two assumptions on the processes (X_t) and (Y_t) in (7.3): the equality of variances and the dominance of increments. We will now remove the assumption on the equality of variances, and still be able to obtain (7.5). This more practically useful result is due to Sudakov and Fernique.

Theorem 7.3.10 (Sudakov-Fernique's inequality)**.** *Let $(X_t)_{t\in T}$ and $(Y_t)_{t\in T}$ be two Gaussian processes. Assume that for all $t, s \in T$, we have*

$$\mathbb{E}(X_t - X_s)^2 \le \mathbb{E}(Y_t - Y_s)^2.$$

Then

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

Proof. It is enough to prove this theorem for Gaussian random vectors X and Y in $\mathbb{R}^n$, just like we did for Slepian's inequality in Theorem 7.3.8. We will again deduced the result from Gaussian Interpolation Lemma 7.3.6. But this time, instead of choosing $f(x)$ that approximates the indicator function of $\{\max_i x_i < \tau\}$, we want $f(x)$ to approximate $\max_i x_i$.

To this end, let $\beta > 0$ be a parameter and define the function[4]

$$f(x) := \frac{1}{\beta}\log\sum_{i=1}^{n} e^{\beta x_i}. \tag{7.12}$$

A quick check shows that

$$f(x) \to \max_{i\le n} x_i \quad \text{as } \beta \to \infty.$$

Substituting $f(x)$ into the Gaussian interpolation formula (7.9) and simplifying the expression shows that $\frac{d}{du}\mathbb{E}\, f(Z(u)) \le 0$ for all u (see the exercise below). The proof can then be completed just like the proof of Slepian's inequality. □

Exercise 7.3.11. [Difficulty=6] *Show that $\frac{d}{du}\mathbb{E}\, f(Z(u)) \le 0$ in Sudakov-Fernique's Theorem 7.3.10. Hint: Differentiate f and check that*

$$\frac{\partial f}{\partial x_i} = \frac{e^{\beta x_i}}{\sum_k e^{\beta x_k}} =: p_i(x) \quad \text{and} \quad \frac{\partial^2 f}{\partial x_i \partial x_j} = \beta\left(\delta_{ij} p_i(x) - p_i(x)p_j(x)\right)$$

[4]The motivation for considering this form of $f(x)$ comes from statistical mechanics, where the right side of (7.12) can be interpreted as a *log-partition function* and β as the *inverse temperature*.

where δ_{ij} is the Kronecker delta, which equals 1 is $i = j$ and 0 otherwise. Next, check the following numeric identity:

$$\text{If } \sum_{i=1}^n p_i = 1 \quad \text{then} \quad \sum_{i,j=1}^n \sigma_{ij}(\delta_{ij} p_i - p_i p_j) = \frac{1}{2} \sum_{i \neq j} (\sigma_{ii} + \sigma_{jj} - 2\sigma_{ij}) p_i p_j.$$

Use Gaussian interpolation formula 7.3.6. Simplify the expression using the identity above with $\sigma_{ij} = \Sigma^X_{ij} - \Sigma^Y_{ij}$ and $p_i = p_i(Z(u))$. Deduce that

$$\frac{d}{du} \mathbb{E}\, f(Z(u)) = \frac{\beta}{4} \sum_{i \neq j} \left[\mathbb{E}(X_i - X_j)^2 - \mathbb{E}(Y_i - Y_j)^2 \right] \mathbb{E}\, p_i(Z(u))\, p_j(Z(u)).$$

By the assumptions, this expression is non-positive.

Exercise 7.3.12 (Gordon's inequality)**.** [Difficulty=6] *Prove the following extension of Slepian's inequality due to Y. Gordon. Let $(X_{ut})_{u \in U,\, t \in T}$ and $Y = (Y_{ut})_{u \in U,\, t \in T}$ be two Gaussian processes indexed by pairs of points (u, t) in a product set $U \times T$. Assume that we have*

$$\mathbb{E}\, X_{ut}^2 = \mathbb{E}\, Y_{ut}^2, \quad \mathbb{E}(X_{ut} - X_{us})^2 \leq \mathbb{E}(Y_{ut} - Y_{us})^2 \quad \text{for all } u, t, s;$$
$$\mathbb{E}(X_{ut} - X_{vs})^2 \geq \mathbb{E}(Y_{ut} - Y_{vs})^2 \quad \text{for all } u \neq v \text{ and all } t, s.$$

Then for every $\tau \geq 0$ we have

$$\mathbb{P}\left\{ \inf_{u \in U} \sup_{t \in T} X_{ut} \geq \tau \right\} \leq \mathbb{P}\left\{ \inf_{u \in U} \sup_{t \in T} Y_{ut} \geq \tau \right\}.$$

Consequently,

$$\mathbb{E} \inf_{u \in U} \sup_{t \in T} X_{ut} \leq \mathbb{E} \inf_{u \in U} \sup_{t \in T} Y_{ut}. \tag{7.13}$$

Hint: Use Gaussian Interpolation Lemma 7.3.6 for $f(x) = \prod_i \left[1 - \prod_j h(x_{ij})\right]$ where $h(x)$ is an approximation to the indicator function $\mathbf{1}_{\{x \leq \tau\}}$, as in the proof of Slepian's inequality.

Similarly to Sudakov-Fernique's inequality, it is possible to remove the assumption of equal variances from Gordon's theorem, and still be able to derive (7.13). We will not prove this result.

7.4 Sharp bounds on Gaussian matrices

We will illustrate Gaussian comparison inequalities that we just proved with an application to random matrices. In Section 4.3.2, we studied $m \times n$ random matrices A with independent, sub-gaussian rows. We used the ε-net argument to control the norm of A as follows:

$$\mathbb{E}\, \|A\| \leq \sqrt{m} + C\sqrt{n}$$

where C is a constant. We will now improve this bound for *Gaussian* random matrices, showing that it holds with the sharp constant $C = 1$. Our argument will be based on Sudakov-Fernique's inequality.

Theorem 7.4.1 (Norms of Gaussian random matrices)**.** *Let A be an $m \times n$ matrix with independent $N(0,1)$ entries. Then*

$$\mathbb{E}\,\|A\| \le \sqrt{m} + \sqrt{n}.$$

Proof. We can realize the norm of A as a supremum of a Gaussian process. Indeed,

$$\|A\| = \max_{u \in S^{n-1},\, v \in S^{m-1}} \langle Au, v \rangle = \max_{(u,v) \in T} X_{uv}$$

where T denotes the product set $S^{n-1} \times S^{m-1}$ and

$$X_{uv} := \langle Au, v \rangle \sim N(0,1).$$

(Check!)

To apply Sudakov-Fernique's comparison inequality (Theorem 7.3.10), let us compute the increments of the process (X_{uv}). For any $(u,v), (w,z) \in T$, we have

$$\begin{aligned}
\mathbb{E}(X_{uv} - X_{wz})^2 &= \mathbb{E}\left(\langle Au, v\rangle - \langle Aw, z\rangle\right)^2 = \mathbb{E}\Big(\sum_{i,j} A_{ij}(u_i v_j - w_i z_j)\Big)^2 \\
&= \sum_{i,j} (u_i v_j - w_i z_j)^2 \quad \text{(by independence, mean 0, variance 1)} \\
&= \|uv^{\mathsf{T}} - wz^{\mathsf{T}}\|_F^2 \\
&\le \|u - w\|_2^2 + \|v - z\|_2^2 \quad \text{(see Exercise 7.4.2 below).}
\end{aligned}$$

Let us define a simpler Gaussian process (Y_{uv}) with similar increments as follows:

$$Y_{uv} := \langle g, u \rangle + \langle h, v \rangle, \quad (u,v) \in T,$$

where

$$g \sim N(0, I_n), \quad h \sim N(0, I_m)$$

are independent Gaussian vectors. The increments of this process are

$$\begin{aligned}
\mathbb{E}(Y_{uv} - Y_{wz})^2 &= \mathbb{E}\left(\langle g, u - w\rangle + \langle h, v - z\rangle\right)^2 \\
&= \mathbb{E}\,\langle g, u - w\rangle^2 + \mathbb{E}\,\langle h, v - z\rangle^2 \quad \text{(by independence, mean 0)} \\
&= \|u - w\|_2^2 + \|v - z\|_2^2 \quad \text{(since } g, h \text{ are standard normal).}
\end{aligned}$$

Comparing the increments of the two processes, we see that

$$\mathbb{E}(X_{uv} - X_{wz})^2 \le \mathbb{E}(Y_{uv} - Y_{wz})^2 \quad \text{for all } (u,v),(w,z) \in T,$$

as required in Sudakov-Fernique's inequality. Applying Theorem 7.3.10, we obtain

$$\begin{aligned}
\mathbb{E}\,\|A\| &= \mathbb{E} \sup_{(u,v)\in T} X_{uv} \le \mathbb{E} \sup_{(u,v)\in T} Y_{uv} \\
&= \mathbb{E} \sup_{u\in S^{n-1}} \langle g, u\rangle + \mathbb{E} \sup_{v\in S^{m-1}} \langle h, v\rangle \\
&= \mathbb{E}\,\|g\|_2 + \mathbb{E}\,\|h\|_2 \\
&\le (\mathbb{E}\,\|g\|_2^2)^{1/2} + (\mathbb{E}\,\|h\|_2^2)^{1/2} \quad \text{(by inequality (1.1) for } L_p \text{ norms)} \\
&= \sqrt{n} + \sqrt{m} \quad \text{(recall Lemma 3.2.4)}.
\end{aligned}$$

This completes the proof. □

Exercise 7.4.2. *Prove the following bound used in the proof of Theorem 7.4.1. For any vectors $u, w \in S^{n-1}$ and $v, z \in S^{m-1}$, we have*

$$\|uv^{\mathsf{T}} - wz^{\mathsf{T}}\|_F^2 \le \|u - w\|_2^2 + \|v - z\|_2^2.$$

While Theorem 7.4.1 does not give any tail bound for $\|A\|$, we can automatically deduce a tail bound using concentration inequalities we studied in Section 5.2.

Corollary 7.4.3 (Norms of Gaussian random matrices: tails)**.** *Let A be an $m \times n$ matrix with independent $N(0,1)$ entries. Then for every $t \ge 0$, we have*

$$\mathbb{P}\left\{\|A\| \ge \sqrt{m} + \sqrt{n} + t\right\} \le 2\exp(-ct^2).$$

Proof. This result follows by combining Theorem 7.4.1 with the concentration inequality in the Gauss space, Theorem 5.2.1.

To use concentration, we will view A as a long random vector in $\mathbb{R}^{m\times n}$. In fact, A is the standard normal random vector, $A \sim N(0, I_{nm})$. The Euclidean norm $\|A\|_2$ of such vector is the same as the Frobenius norm $\|A\|_F$, and the operator norm $\|A\|$ is smaller:

$$\|A\| \le \|A\|_F = \|A\|_2.$$

This shows that $A \mapsto \|A\|$ is a Lipschitz function on $\mathbb{R}^{m\times n}$, and its Lipschitz norm is bounded by 1. Then Theorem 5.2.1 yields

$$\mathbb{P}\left\{\|A\| \ge \mathbb{E}\,\|A\| + t\right\} \le 2\exp(-ct^2).$$

The bound on $\mathbb{E}\,\|A\|$ from Theorem 7.4.1 completes the proof. □

Exercise 7.4.4 (Smallest singular values)**.** [Difficulty=5] *Use Gordon's inequality stated in Exercise 7.3.12 to obtain a sharp bound on the smallest singular value of an $m \times n$ random matrix A with independent $N(0,1)$ entries:*

$$\mathbb{E}\, s_n(A) \geq \sqrt{m} - \sqrt{n}.$$

Combine this result with concentration to show the tail bound

$$\mathbb{P}\left\{\|A\| \leq \sqrt{m} - \sqrt{n} - t\right\} \leq 2\exp(-ct^2).$$

Hint: Relate the smallest singular value to the min-max of a Gaussian process:

$$s_n(A) = \min_{u \in S^{n-1}} \max_{v \in S^{m-1}} \langle Au, v \rangle .$$

Apply Gordon's inequality (without the requirement of equal variances, which is noted below Exercise 7.3.12) to show that

$$\mathbb{E}\, s_n(A) \geq \mathbb{E}\, \|h\|_2 - \mathbb{E}\, \|g\|_2 \quad \text{where} \quad g \sim N(0, I_n),\ h \sim N(0, I_m).$$

Combine this with the fact that $f(n) := \mathbb{E}\, \|g\|_2 - \sqrt{n}$ is increasing in dimension n. (Take this fact for granted; it can be proved by a tedious calculation.)

Exercise 7.4.5 (Symmetric random matrices)**.** [Difficulty=6] *Modify the arguments above to bound the norm of a* symmetric *$n \times n$ Gaussian random matrix A whose entries above the diagonal are independent $N(0,1)$ random variables, and the diagonal entries are independent $N(0,2)$ random variables. This distribution of random matrices is called the* Gaussian orthogonal ensemble (GOE)*. Show that*

$$\mathbb{E}\, \|A\| \leq 2\sqrt{n}.$$

Next, deduce the tail bound

$$\mathbb{P}\left\{\|A\| \geq 2\sqrt{n} + t\right\} \leq 2\exp(-ct^2).$$

Bai-Yin's law – state it somewhere, see a section that is commented out here

7.5 Sudakov's minoration inequality

Let us return to studying general Gaussian processes $(X_t)_{t \in T}$. As we observed in Section 7.1.1, the sizes of the increments

$$d(t,s) := \|X_t - X_s\|_2 = \left(\mathbb{E}(X_t - X_s)^2\right)^{1/2} \tag{7.14}$$

defines a distance, or metric on the otherwise abstract index set T. We will call this metric the *canonical metric* associated with the random process.

We also observed in Exercise 7.1.7 that the canonical metric $d(t, s)$ determines the covariance function $\Sigma(t, s)$, which in turn determines the distribution of the process $(X_t)_{t\in T}$. So in principle, we should be able to answer any question about the distribution of a Gaussian process $(X_t)_{t\in T}$ by looking at the geometry of the metric space (T, d). Put plainly, we should be able to study probability via geometry.

Let us then ask an important specific question. How can we evaluate the overall magnitude of the process, namely

$$\mathbb{E} \sup_{t\in T} X_t, \tag{7.15}$$

in terms of the geometry of (T, d)? This turns out to be a difficult a difficult problem, which we will start to study here and continue in Sections **??** ...

Refer.

In this section, we will prove a useful lower bound on (7.15) in terms of the *metric entropy* of the metric space (T, d). Recall from Section 4.2 that for $\varepsilon > 0$, the *covering number*

Refer to a subsection here

$$\mathcal{N}(T, d, \varepsilon)$$

is defined to be the smallest cardinality of an ε-net of T in the metric d. Equivalently, $\mathcal{N}(T, d, \varepsilon)$ is the smallest number[5] of closed balls of radius ε whose union covers T. Recall also that the logarithm of the covering number,

$$\log_2 \mathcal{N}(T, d, \varepsilon)$$

is called the *metric entropy* of T.

Theorem 7.5.1 (Sudakov's minoration inequality)**.** *Let $(X_t)_{t\in T}$ be a Gaussian process. Then, for any $\varepsilon \geq 0$, we have*

$$\mathbb{E} \sup_{t\in T} X_t \geq c\varepsilon \sqrt{\log \mathcal{N}(T, d, \varepsilon)}.$$

where d is the canonical metric (7.14).

Proof. Let us deduce this result from Sudakov-Fernique's comparison inequality (Theorem 7.3.10). Assume that

$$\mathcal{N}(T, d, \varepsilon) =: N$$

is finite; the infinite case will be considered in Exercise 7.5.2. Let $\mathcal{N}$ be a maximal ε-separated subset of T. Then $\mathcal{N}$ is an ε-net of T, and thus

$$|\mathcal{N}| \geq N.$$

[5] If T does not admit a finite ε-net, we set $N(t, d, \varepsilon) = \infty$.

(Check this! Recall the proof of the equivalence between covering and packing numbers in Lemma 4.2.5.) Restricting the process to $\mathcal{N}$, we see that it suffices to show that

$$\mathbb{E} \sup_{t \in \mathcal{N}} X_t \geq \varepsilon \sqrt{\log N}.$$

We can do it by comparing $(X_t)_{t \in \mathcal{N}}$ to a simpler Gaussian process $(Y_t)_{t \in \mathcal{N}}$, which we define as follows:

$$Y_t := \frac{\varepsilon}{\sqrt{2}} g_t, \quad \text{where } g_t \text{ are independent } N(0,1) \text{ random variables.}$$

To use Sudakov-Fernique's comparison inequality (Theorem 7.3.10), we need to compare the increments of the two processes. Fix two different points $t, s \in \mathcal{N}$. By definition, we have

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

while

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

(In the last line, we use that $g_t - g_s \sim N(0,2)$.) So, we have checked that

$$\mathbb{E}(X_t - X_s)^2 \geq \mathbb{E}(Y_t - Y_s)^2 \quad \text{for all } t, s \in \mathcal{N}.$$

Applying Theorem 7.3.10, we obtain

$$\mathbb{E} \sup_{t \in \mathcal{N}} X_t \geq \mathbb{E} \sup_{t \in \mathcal{N}} Y_t = \frac{\varepsilon}{\sqrt{2}} \mathbb{E} \max_{t \in \mathcal{N}} g_t \geq c \sqrt{\log N}.$$

In the last inequality we used that the expected maximum of N standard normal random variables is at least $c\sqrt{\log N}$, see Exercise 2.5.9. The proof is complete. □

Exercise 7.5.2 (Sudakov's minoration for non-compact sets)**.** [Difficulty=2] *Show that if (T,d) is not compact, that is if $N(T,d,\varepsilon) = \infty$ for some ε, then*

$$\mathbb{E} \sup_{t \in T} X_t = \infty.$$

7.5.1 Application for covering numbers of sets in $\mathbb{R}^n$

Sudakov's minoration inequality can be used to estimate the covering numbers of geometric sets

$$T \subset \mathbb{R}^n.$$

To see how to do this, consider a canonical Gaussian process on T, namely

$$X_t := \langle g, t \rangle, \quad t \in T, \quad \text{where } g \sim N(0, I_n).$$

As we observed in Section 7.2.1, the canonical distance for this process is just the Euclidean distance in $\mathbb{R}^n$, i.e.

$$d(t, s) = \|X_t - X_s\|_2 = \|t - s\|_2.$$

Thus Sudakov's inequality can be stated as follows.

Corollary 7.5.3 (Sudakov's minoration inequality in $\mathbb{R}^n$)**.** *Let $T \subset \mathbb{R}^n$. Then, for any $\varepsilon > 0$, we have*

$$\mathbb{E} \sup_{t \in T} \langle g, t \rangle \geq c\varepsilon \sqrt{\log \mathcal{N}(T, \varepsilon)}.$$

Here $\mathcal{N}(T, \varepsilon)$ is the covering number of T by Euclidean balls – the smallest number of Euclidean balls with radii ε and centers in T that cover T, see Section 4.2. Let us illustrate the usefulness of this bound with an example.

Check reference

Consider the unit ball of the ℓ_1 norm in $\mathbb{R}^n$, denoted

$$B_1^n = \{x \in \mathbb{R}^n : \ \|x\|_1 \leq 1\}.$$

The vertices of B_1^n are $\pm e_1, \ldots, \pm e_n$, where (e_i) is the canonical basis. So, in dimension $n = 2$, this set is a diamond, and in dimension $n = 3$, a regular octahedron, see Figure 7.3.

To apply Corollary 7.5.3 for $T = B_1^n$, we compute

$$\mathbb{E} \sup_{t \in B_1^n} \langle g, t \rangle = \mathbb{E} \max_{i \leq n} |g_i| \leq C\sqrt{\log n}. \tag{7.16}$$

Here the first equality follows since the maximum of a linear function $f(t) = \langle g, t \rangle$ on the polyhedron B_1^n is attained on the extremal points (vertices) of B_1^n, which are $\pm e_i$. The final bound in (7.16) then follows from Exercise 2.5.8. Thus, Sudakov's inequality (Corollary 7.5.3) yields

$$\varepsilon \sqrt{\log \mathcal{N}(B_1^n, \varepsilon)} \leq C\varepsilon \sqrt{\log n}.$$

Simplifying this bound, we obtain:

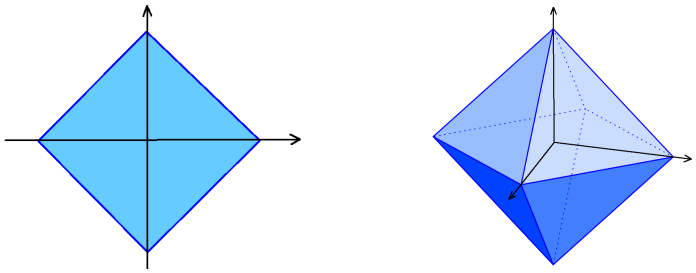


Figure 7.3: The unit ball of the ℓ_1 norm in $\mathbb{R}^n$, denoted B_1^n, is a diamond in dimension $n = 2$ and a regular octahedron in dimension $n = 3$.

Proposition 7.5.4 (Covering numbers of the ℓ_1 ball)**.** *For every $\varepsilon > 0$, we have*

$$\mathcal{N}(B_1^n, \varepsilon) \leq N^{C/\varepsilon^2}.$$

In particular, the covering numbers of B_1^n are *polynomial* in dimension n, for each scale ε.

We may recall that we studied a different method for bounding covering numbers, which is based on volume comparison; see Proposition 4.2.10. Which of the two methods is better? Sometimes Sudakov's inequality gives better results. In the example above, in order to apply the volumetric method for B_1^n one would need to bound the volume of the Minkowski sum $B_1^n + \varepsilon B_2^n$, which is difficult to compute. In other cases – especially for very small $\varepsilon > 0$ – the volumetric bound can be better.

Exercise 7.5.5. [Difficulty=5] *Compare the bounds on the covering numbers for the unit Eucledian ball B_2^n given by Sudakov's minoration and by volume comparison (Proposition 4.2.10). Which method gives better results?*

Exercise 7.5.6 (Covering numbers of polyhedra)**.** *Generalize the proof of Proposition 7.5.4 to arbitrary polyhedra. Let P be a polyhedron in $\mathbb{R}^n$ with N vertices and whose diameter is bounded by 1. Show that for every $\varepsilon > 0$,*

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

Hint: Use the result of Exercise 2.5.8.

7.6 The empirical method for constructing ε-nets

Sudakov's minoration inequality gives only a bound for the covering numbers, but not a *constructive method* to find ε-nets. In this section, we will

give an explicit recipe for finding ε-nets of polyhedra. This construction is known as the *empirical method* of B. Maurey, and it is again based on probability.

The empirical method itself allows one to approximate convex hulls of point sets in $\mathbb{R}^n$. Recall that a *convex combination* of points $z_1, \ldots, z_m \in \mathbb{R}^m$ is a linear combination with coefficients that are non-negative and sum to 1, i.e. a sum of the form

$$\sum_{i=1}^{m} \lambda_i z_i, \quad \text{where} \quad \lambda_i \geq 0 \quad \text{and} \quad \sum_{i=1}^{m} \lambda_i = 1. \tag{7.17}$$

The *convex hull* of a set $T \subset \mathbb{R}^n$ is the set of all convex combinations of all finite collections of points in T:

$$\operatorname{conv}(T) := \{\text{convex combinations of } z_1, \ldots, z_m \in T, \text{ for } m \in \mathbb{N}\},$$

see Figure 7.4 for illustration.

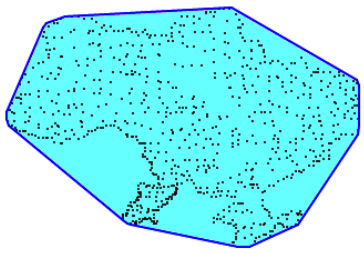


Figure 7.4: The convex hull of a collection of points on the plane.

The number m of elements defining a convex combination in $\mathbb{R}^n$ is not restricted. However, the classical Caratheodory's theorem states that one can always take $m \leq n + 1$.

Theorem 7.6.1 (Caratheodory's theorem)**.** *Every point in the convex hull of a set $T \subset \mathbb{R}^n$ can be expressed as a convex combination of at most $n + 1$ points from T.*

The bound $n + 1$ can not be improved, as it is attained for a simplex T (a set of $n + 1$ points in general position). However, if we only want to approximate x rather than exactly represent it as a convex combination, this is possible with much fewer points.

Theorem 7.6.2 (Approximate Caratheodory's theorem)**.** *Consider a set $T \subset \mathbb{R}^n$ whose diameter is bounded by 1. Then, for every point $x \in \operatorname{conv}(T)$ and every integer k, one can find points $x_1, \dots, x_k \in T$ such that*

$$\Big\| x - \frac{1}{k} \sum_{j=1}^{k} x_j \Big\|_2 \le \frac{1}{\sqrt{k}}.$$

There are two reasons why this result is surprising. First, one can achieve good approximation with convex combinations whose length m does not depend on the dimension n. Second, the coefficients of the convex combinations can be made all equal. (Note however that repetitions among the points x_i are allowed.)

Proof. This argument is known as the *empirical method* of B. Maurey.

Translating T if necessary, we may assume that not only the diameter but also the *radius* of T is bounded by 1, i.e.

$$\|t\|_2 \le 1 \quad \text{for all } t \in T. \tag{7.18}$$

Let us fix a point $x \in \operatorname{conv}(T)$ and express it as a convex combination of some vectors $z_1, \dots, z_m \in T$ as in (7.17). Now we interpret the definition of convex combination (7.17) probabilistically, with λ_i taking the roles of probabilities. Specifically, we can define a random vector Z that takes values z_i with probabilities λ_i:

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

Then

$$\mathbb{E}\, Z = \sum_{i=1}^{m} \lambda_i z_i = x.$$

Consider independent copies $Z_1, Z_2, \dots$ of Z. By the the strong law of large numbers,

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

To get a quantitative form of this result, let us compute the variance of $\frac{1}{k} \sum_{j=1}^{k} Z_j$. (Incidentally, this computation is at the heart of the proof of

the weak law of large numbers). We obtain

$$\mathbb{E}\left\|x - \frac{1}{k}\sum_{j=1}^{k} Z_j\right\|_2^2 = \frac{1}{k^2}\,\mathbb{E}\left\|\sum_{j=1}^{k}(Z_j - x)\right\|_2^2 \quad (\text{since } \mathbb{E}(Z_i - x) = 0)$$

$$= \frac{1}{k^2}\sum_{j=1}^{k}\mathbb{E}\,\|Z_j - x\|_2^2.$$

The last identity is just a higher dimensional version of the basic fact that the variance of a sum of independent random variables equals the sum of variances; see Exercise 7.6.4 below.

It remains to bound the variances of the terms. We have

$$\begin{aligned}\mathbb{E}\|Z_j - x\|_2^2 &= \mathbb{E}\,\|Z - \mathbb{E}\,Z\|_2^2 \\ &= \mathbb{E}\,\|Z\|_2^2 - \|\,\mathbb{E}\,Z\|_2^2 \quad (\text{another variance identity; see Exercise 7.6.4}) \\ &\le \mathbb{E}\,\|Z\|_2^2 \le 1 \quad (\text{since } Z \in T \text{ and using (7.18)}).\end{aligned}$$

We showed that

$$\mathbb{E}\left\|x - \frac{1}{k}\sum_{j=1}^{k} Z_j\right\|_2^2 \le \frac{1}{k}.$$

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

$$\left\|x - \frac{1}{k}\sum_{j=1}^{k} Z_j\right\|_2^2 \le \frac{1}{k}.$$

Since by construction each Z_j takes values in T, the proof is complete. □

We can use the Approximate Caratheodory Theorem 7.6.2 to construct good coverings of geometric sets. Let us give a constructive version of our previous result on coverings for polyhedra in Exercise 7.5.6.

Corollary 7.6.3 (Constructing ε-nets of polyhedra)**.** *Let P be a polyhedron in $\mathbb{R}^n$ with N vertices and whose diameter is bounded by 1. Then, for every $\varepsilon > 0$, we have*

$$\mathcal{N}(P, \varepsilon) \le N^{\lceil 1/\varepsilon^2 \rceil}.$$

Moreover, an ε-net with this cardinality can be constructed as follows. Let $k := \lceil 1/\varepsilon^2 \rceil$ and consider the set

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

Then $\mathcal{N}$ is an ε-net of P, and $|\mathcal{N}| \le N^k$.

Proof. The polyhedron P is the convex hull of its vertices, which we denote by T. Thus, we can apply Theorem 7.6.2 to any point $x \in P = \operatorname{conv}(T)$ and deduce that x is within distance $1/\sqrt{k} \le \varepsilon$ from some point in $\mathcal{N}$. This shows that $\mathcal{N}$ is an ε-net of the polyhedron P.

To bound the cardinality of $\mathcal{N}$, note that there are N^k ways to choose k out of N vertices with repetition. The proof is complete. □

Exercise 7.6.4. [Difficulty=3] *Check the following variance identities that we used in the proof of Approximate Caratheodory's Theorem 7.6.2.*

1. Let $Z_1, \dots, Z_k$ be independent mean zero random vectors in $\mathbb{R}^n$. Show that

$$\mathbb{E} \Big\| \sum_{j=1}^k Z_j \Big\|_2^2 = \sum_{j=1}^k \mathbb{E} \|Z_j\|_2^2.$$

2. Let Z be a random vector in $\mathbb{R}^n$. Show that

$$\mathbb{E} \|Z - \mathbb{E} Z\|_2^2 = \mathbb{E} \|Z\|_2^2 - \|\mathbb{E} Z\|_2^2.$$

7.7 Gaussian width

In our previous study of the Gaussian processes, we encountered an important quantity associated with a general set $T \subset \mathbb{R}^n$. It is the magnitude of the canonical Gaussian process on T, i.e.

$$\mathbb{E} \sup_{t \in T} \langle g, t \rangle \tag{7.19}$$

where the expectation is taken with respect to the Gaussian random vector $g \sim N(0, I_n)$.

We have already seen how the quantity (7.19) can be used to bound covering numbers of T through Sudakov's minoration inequality (Corollary **??**). We will later see that (7.19) plays a central role in high dimensional probability and its applications. In this section, we will give the quantity (7.19) a name and will study its basic properties.

Definition 7.7.1. *The* Gaussian width *of a subset $T \subset \mathbb{R}^n$ is defined as*

$$w(T) := \mathbb{E} \sup_{x \in T} \langle g, x \rangle \quad \text{where} \quad g \sim N(0, I_n).$$

It is useful to think about Gaussian width $w(T)$ as one of the basic geometric quantities associated with subsets of $T \subset \mathbb{R}^n$, such as volume and surface area.

Before we proceed to study the basic properties of Gaussian width, we should mention that several other variants of this concept can be found in the literature, such as

$$\mathbb{E} \sup_{x \in T} |\langle g, x \rangle|, \quad \left(\mathbb{E} \sup_{x \in T} \langle g, x \rangle^2 \right)^{1/2}, \quad \mathbb{E} \sup_{x,y \in T} \langle g, x - y \rangle, \quad \text{etc.}$$

These versions are equivalent, or almost equivalent, to $w(T)$. We will mention some of them in this section.

7.7.1 Basic properties

Proposition 7.7.2 (Gaussian width)**.**

1. *$w(T)$ is finite if and only if T is bounded.*

2. *Gaussian width is invariant under affine transformations. Thus, for every orthogonal matrix U and any vector y, we have*

$$w(UT + y) = w(T).$$

3. *Gaussian width is invariant under taking convex hulls. Thus,*

$$w(\text{conv}(T)) = w(T).$$

4. *Gaussian width respects Minkowski addition of sets and scaling. Thus, for $T, S \in \mathbb{R}^n$ and $a \in \mathbb{R}$ we have*

$$w(T + S) = w(T) + w(S); \quad w(aT) = |a| \, w(T).$$

5. *We have*

$$w(T) = \frac{1}{2} w(T - T) = \frac{1}{2} \mathbb{E} \sup_{x,y \in T} \langle g, x - y \rangle.$$

6. *(Gaussian width and diameter). We have*

$$\frac{1}{\sqrt{2\pi}} \cdot \text{diam}(T) \le w(T) \le \frac{\sqrt{n}}{2} \cdot \text{diam}(T).$$

Proof. Properties 1–4 follow immediately from definition and the rotation invariance of Gaussian distribution. (Check them!)

To prove property 5, we use property 4 twice and get

$$w(T) = \frac{1}{2}\left[w(T) + w(T)\right] = \frac{1}{2}\left[w(T) + w(-T)\right] = \frac{1}{2}w(T-T),$$

as claimed.

To prove the lower bound in property 6, fix a pair of points $x, y \in T$. Then both $x - y$ and $y - x$ lie in $T - T$, so by property 5 we have

$$\begin{aligned} w(T) &\geq \frac{1}{2}\,\mathbb{E}\max\left(\langle x-y, g\rangle\,,\, \langle y-x, g\rangle\right) \\ &= \frac{1}{2}\,\mathbb{E}\,|\,\langle x-y, g\rangle\,| = \frac{1}{2}\sqrt{\frac{2}{\pi}}\|x-y\|_2. \end{aligned}$$

The last identity follows since $\langle x-y, g\rangle \sim N(0, \|x-y\|_2)$ and since $\mathbb{E}\,|X| = \sqrt{2/\pi}$ for $X \sim N(0,1)$. (Check!) It remains to take supremum over all $x, y \in T$, and the lower bound in property 6 follows.

To prove the upper bound in property 6, we again use property 5 to get

$$\begin{aligned} w(T) &= \frac{1}{2}\,\mathbb{E}\sup_{x,y\in T}\langle g, x-y\rangle \\ &\leq \frac{1}{2}\,\mathbb{E}\sup_{x,y\in T}\|g\|_2\|x-y\|_2 \leq \frac{1}{2}\,\mathbb{E}\,\|g\|_2 \cdot \operatorname{diam}(T). \end{aligned}$$

It remains to recall that $\mathbb{E}\,\|g\|_2 \leq (\mathbb{E}\,\|g\|_2^2)^{1/2} = \sqrt{n}$. □

Define diameter formally somewhere.

Exercise 7.7.3 (Gaussian width under linear transformations)**.** [Difficulty=3] *Show that for any $m \times n$ matrix A, we have*

$$w(AT) \leq \|A\|\, w(T).$$

Hint: Use Sudakov-Fernique's comparison inequality.

7.7.2 Geometric meaning of width

The notion of the Gaussian width of a set $T \subset \mathbb{R}^n$ has a nice geometric meaning. The width of T in the direction of a vector $\theta \in S^{n-1}$ is the smallest width of the slab between the parallel hyperplanes that are orthogonal θ and contain T, see Figure 7.5. Analytically, we can express the width in the direction of θ as

$$\sup_{x,y\in T}\langle \theta, x-y\rangle\,.$$

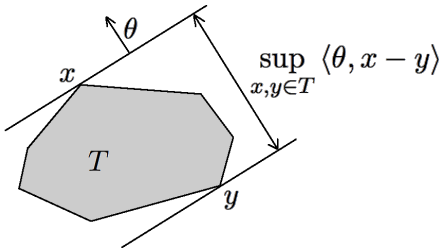


Figure 7.5: The width of a set $T \subset \mathbb{R}^n$ in the direction of a unit vector θ.

(Check!) If we average the width over all unit directions θ in space, we obtain the quantity

$$\mathbb{E} \sup_{x,y \in T} \langle \theta, x - y \rangle = w_s(T - T)$$

Maybe make this a formal definition. We are using spherical width later.

where

$$w_s(T) := \mathbb{E} \sup_{x \in T} \langle \theta, x \rangle \quad \text{where} \quad \theta \sim \mathrm{Unif}(S^{n-1}). \tag{7.20}$$

Analogously to the Gaussian width, $w_s(T)$ is called the *spherical width* or the *mean width* of T.

How different are the Gaussian and spherical widths of T? The difference is what random vector we use to do the averaging; they are $g \sim N(0, I_n)$ for Gaussian width and $\theta \sim \mathrm{Unif}(S^{n-1})$ for spherical width. Both g and θ are rotation invariant, and, as we know, g is approximately $\sqrt{n}$ longer than θ. This makes Gaussian width just a scaling of the spherical width by approximately $\sqrt{n}$. Let us make this relation more precise.

Lemma 7.7.4 (Gaussian vs. spherical widths)**.** *We have*

$$(\sqrt{n} - C)\, w_s(T) \le w(T) \le (\sqrt{n} + C)\, w_s(T).$$

Proof. Let us express the Gaussian vector g through its length and direction:

$$g = \|g\|_2 \cdot \frac{g}{\|g\|_2} =: r\theta.$$

As we observed in Section 3.3.3, r and θ are independent and $\theta \sim \mathrm{Unif}(S^{n-1})$. Thus

$$w(T) = \mathbb{E} \sup_{x \in T} \langle r\theta, x \rangle = (\mathbb{E}\, r) \cdot \mathbb{E} \sup_{x \in T} \langle \theta, x \rangle = \mathbb{E}\, \|g\|_2 \cdot w_s(T).$$

It remains to recall that concentration of the norm implies that

$$\left| \mathbb{E}\, \|g\|_2 - \sqrt{n} \right| \le C,$$

see Exercise 3.1.3. □

Do Urysohn's inequality as an exercise? Commented out.

7.7.3 Examples

Euclidean ball

The Gaussian width of the Euclidean unit sphere and ball is

$$w(S^{n-1}) = w(B_2^n) = \mathbb{E}\,\|g\|_2 = \sqrt{n} \pm C, \tag{7.21}$$

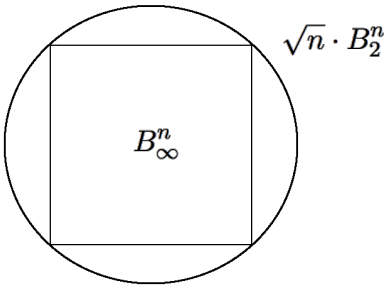
where we used the result of Exercise 3.1.3. Similarly, the spherical width of these sets equals 1.

Cube

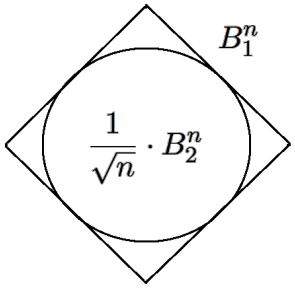
The unit ball of the ℓ_∞ norm on $\mathbb{R}^n$ is $B_\infty^n = [-1,1]^n$. We have

$$\begin{aligned} w(B_\infty^n) &= \mathbb{E}\,\|g\|_1 \quad \text{(check!)} \\ &= \mathbb{E}\,|g_1| \cdot n = \sqrt{\frac{2}{\pi}} \cdot n. \end{aligned} \tag{7.22}$$

Compare this with (7.21). The Gaussian width of the cube B_∞^n and its circumscribed ball $\sqrt{n}B_2^n$ have the same order $\sqrt{n}$, see Figure 7.6a.



(a) The Gaussian widths of the cube and its circumscribed ball are of the same order n.



(b) The Gaussian widths of B_1^n and its inscribed ball are almost of the same order.

Figure 7.6: Gaussian widths of some classical sets in $\mathbb{R}^n$.

The ℓ_1 ball

The Gaussian width of the unit ball of the ℓ_1 norm in $\mathbb{R}^n$ is

$$c\sqrt{\log n} \le w(B_1^n) \le C\sqrt{\log n}. \tag{7.23}$$

The upper bound was already noted in (7.16) where we discussed applications of Sudakov's inequality; the lower bound follows similarly.

Equivalently, the spherical width of B_1^n is

$$w_s(n) \sim \sqrt{\frac{\log n}{n}},$$

Surprisingly, this is much smaller than the diameter of B_1^n, which equals 2! Further, if we compare with (7.21), we see that the Gaussian width of B_1^n is roughly the same (up to a logarithmic factor) as the Gaussian width of its inscribed Euclidean ball $\frac{1}{\sqrt{n}} B_2^n$, see Figure 7.6b. This again might look strange. Indeed, the "octahedron" B_1^n looks much larger than its inscribed ball whose diameter is $\frac{2}{\sqrt{n}}$! Why does Gaussian width behave this way?

Let us try to give an intuitive explanation. In high dimensions, the cube has so many vertices (2^n) that most of the volume is concentrated near them. In fact, the volumes of the cube and its circumscribed ball are both of the order C^n, so these sets are not far from each other from the volumetric point of view. So it should not be very surprising to see that the Gaussian widths of the cube and its circumscribed ball are also of the same order.

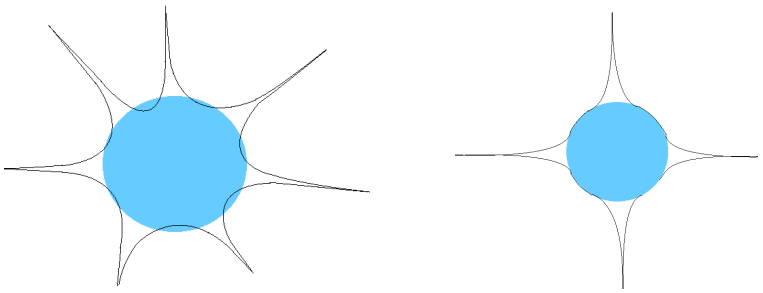
The octahedron B_1^n has much fewer vertices ($2n$) than the cube. A random direction θ in $\mathbb{R}^n$ is likely to be almost orthogonal to all of them. So the width of B_1^n in the direction of θ is not significantly influenced by the presence of vertices. What really determines the width of B_1^n is its "bulk", which is the inscribed Euclidean ball.

A similar picture can be seen from the volumetric viewpoint. There are so few vertices in B_1^n that the regions near them contain very little volume. The bulk of the volume of B_1^n lies much closer to the origin, not far from the inscribed Euclidean ball. Indeed, one can check that the volumes of B_1^n and its inscribed ball are both of the order of $(C/n)^n$. So from the volumetric point of view, the octahedron B_1^n is similar to its inscribed ball; Gaussian width gives the same result.

We can illustrate this phenomenon on Figure 7.7b that shows a "hyperbolic" picture of the B_1^n that is due to V. Milman. Such pictures capture the bulk and outliers very well, but unfortunately they may not accurately show convexity.

Exercise 7.7.5 (Gaussian width of ℓ_p balls)**.** *Let $1 \le p \le \infty$. Consider the unit ball of the ℓ_p norm in $\mathbb{R}^n$:*

$$B_p^n := \{x \in \mathbb{R}^n : \|x\|_p \le 1\}.$$



(a) General convex set (b) The octahedron B_1^n

Figure 7.7: An intuitive, hyperbolic picture of a convex body in $\mathbb{R}^n$. The bulk is a round ball that contains most of the volume.

Check that

$$w(B_p^n) \le C \min\left(\sqrt{p'}\, n^{1/p'},\, \sqrt{\log n}\right).$$

Here p' denotes the conjugate exponent *for p, which is defined by the equation $\frac{1}{p} + \frac{1}{p'} = 1$, and with the convention that $\frac{1}{\infty} = 0$.*

Finite point sets

Exercise 7.7.6. [Difficulty=3] *Let T be a finite set of points in $\mathbb{R}^n$. Show that*

$$w(T) \le \sqrt{\log |T|} \cdot \operatorname{diam}(T).$$

Hint: Use the result of Exercise 2.5.8.

In particular, the spherical mean width of a finite point set T is usually much smaller than the diameter of T, unless there are exponentially many points. This is just a more general observation than we have seen for the ℓ_1 ball B_1^n, which is a convex hull of $2n$ points.

7.7.4 Statistical dimension

The notion of Gaussian width will help us to introduce a more robust version of the classical notion of dimension. In this application, it will be more convenient to work with a closely related *squared* version of the Gaussian width:

$$h(T)^2 := \mathbb{E} \sup_{t \in T} \langle g, t \rangle^2, \quad \text{where} \quad g \sim N(0, I_n). \tag{7.24}$$

The squared and usual versions of the Gaussian width are equivalent, up to constant factor:

Exercise 7.7.7 (Equivalence)**.** [Difficulty=6] *Check that*

$$w(T-T) \le h(T-T) \le w(T-T) + C_1 \operatorname{diam}(T) \le C w(T-T).$$

Hint: Use Gaussian concentration to prove the upper bound.

In particular,

$$\frac{1}{2} w(T) \le h(T-T) \le \frac{C}{2} w(T). \tag{7.25}$$

The usual, linear algebraic, dimension $\dim T$ of a subset $T \subset \mathbb{R}^n$, is the smallest dimension of a linear subspace $E \subset \mathbb{R}^n$ that contains T. The algebraic dimension is highly unstable: it can significantly change (usually upwards) by small perturbations of T. We will now introduce a stable version of dimension, or complexity of T, which is based on Gaussian width.

Definition 7.7.8 (Statistical dimension)**.** *For a bounded set $T \subset \mathbb{R}^n$, the* statistical dimension *of T is defined as*

$$d(T) := \frac{h(T-T)^2}{\operatorname{diam}(T)^2} \sim \frac{w(T)^2}{\operatorname{diam}(T)^2}.$$

We should first note that the statistical dimension is always bounded by the algebraic dimension.

Lemma 7.7.9. *For any set $T \subset \mathbb{R}^n$, we have*

$$d(T) \le \dim(T).$$

Proof. Let $\dim T = k$; this means that T lies in some subspace $E \subset \mathbb{R}^n$ of dimension k. By rotation invariance, we can assume that E is the coordinate subspace, i.e. $E = \mathbb{R}^k = \mathbb{R}^n$. (Why?) By definition, we have

$$h(T-T)^2 = \mathbb{E} \sup_{x,y \in T} \langle g, x-y \rangle^2$$

Since $x - y \in \mathbb{R}^k$ and $\|x-y\|_2 \le \operatorname{diam}(T)$, we have $x - y = \operatorname{diam}(T) \cdot z$ for some $z \in B_2^k$. Thus the quantity above is bounded by

$$\operatorname{diam}(T)^2 \cdot \mathbb{E} \sup_{z \in B_2^k} \langle g, z \rangle^2 = \operatorname{diam}(T)^2 \cdot \mathbb{E} \, \|P_k g\|_2^2$$

where P_k is the orthogonal projection onto $\mathbb{R}^k$. Finally, $P_k g \sim N(0, I_k)$ and

$$\mathbb{E} \, \|P_k g\|_2^2 = k.$$

(Check!) This shows that $h(T-T)^2 \le \operatorname{diam}(T)^2 \cdot k$, as required. □

Exercise 7.7.10. *Show that if T is a Euclidean ball in any subspace of $\mathbb{R}^n$ then*

$$d(T) = \dim(T).$$

In general, the statistical dimension can be much smaller than the algebraic dimension:

Example 7.7.11. Let T be a finite set of points in $\mathbb{R}^n$. Then

$$d(T) \le \log |T|.$$

This follows from the bound on the Gaussian width of T in Exercise 7.7.6, the definition of statistical dimension and the equivalence (7.25).

7.7.5 Stable rank

The statistical dimension is more robust than the algebraic dimension. Indeed, small perturbation of a set T leads to small perturbation of Gaussian width and the diameter of T, and thus the statistical dimension $d(T)$.

To see this on an example, consider the Euclidean ball B_2^n whose both algebraic and statistical dimensions equal n. Let us decrease one of the axes of B_2^n gradually from 1 to 0. The algebraic dimension will stay at n through this process and then jump to $n-1$. The statistical dimension instead gradually slides from n to $n-1$. This is a consequence of the following simple fact.

Exercise 7.7.12 (Ellipsoids)**.** [Difficulty=4] *Let A be an $m \times n$ matrix, and B_2^n denote the unit Euclidean ball. Check that the squared mean width of the ellipsoid AB_2^n is the Frobenius norm of A, i.e.*

$$h(AB_2^n) = \|A\|_F.$$

Deduce that the statistical dimension of the ellipsoid AB_2^n equals

$$d(AB_2^n) = \frac{\|A\|_F^2}{\|A\|^2}. \tag{7.26}$$

This example relates the statistical dimension to the notion of *stable rank* of matrices, which is a robust version of the classical, linear algebraic, rank.

Definition 7.7.13 (Stable rank)**.** *The* stable rank *of an $m \times n$ matrix A is defined as*

$$r(A) := \frac{\|A\|_F^2}{\|A\|^2}.$$

The robustness of stable rank makes it a useful quantity in numerical linear algebra. The usual, algebraic, rank is the algebraic dimension of the image of A; in particular

$$\operatorname{rank}(A) = \dim(AB_2^n).$$

Similarly, (7.26) shows that the *stable* rank is the *statistical* dimension of the image:

$$r(A) = d(AB_2^n).$$

Finally, note that the stable rank is always bounded by the usual rank:

$$r(A) \le \operatorname{rank}(A).$$

(Check this.)

Gaussian complexity

To finish this discussion, let us mention one more cousin of Gaussian width, where instead of squaring $\langle g, x\rangle$ as in (7.24) we take absolute value:

$$\gamma(T) := \mathbb{E} \sup_{x \in T} |\langle g, x\rangle| \quad \text{where} \quad g \sim N(0, I_n). \tag{7.27}$$

We call $\gamma(T)$ *Gaussian complexity* of T. Obviously, we have

$$w(T) \le \gamma(T),$$

with equality if T is origin-symmetric, i.e. $T = -T$. Since $T - T$ is origin-symmetric, property 5 of Proposition 7.7.2 implies that

$$w(T) = \frac{1}{2} w(T - T) = \frac{1}{2} \gamma(T - T). \tag{7.28}$$

In general, Gaussian width and complexity may be different. For example, if T consists of a single point, $w(T) = 0$ but $\gamma(T) > 0$. (What is $\gamma(T)$ here exactly?) Still, these two quantities are very closely related:

Exercise 7.7.14 (Gaussian width vs. Gaussian complexity)**.** [Difficulty=6] *Consider a set $T \subset \mathbb{R}^n$. Show that*

$$\frac{1}{3} [w(T) + \|y\|_2] \le \gamma(T) \le 2 [w(T) + \|y\|_2]$$

for any point $y \in T$. (It is fine if you prove this inequality with other absolute constants instead of 2 and 3.)

In particular, Gaussian width and Gaussian complexity are equivalent for any set T that contains the origin:

$$w(T) \le \gamma(T) \le 2w(T).$$

7.7.6 Computing Gaussian width

How can we compute the Gaussian width $w(T)$ of a given set T? Algorithmically, this is not difficult, as long as we have a good description of T. To see this, recall that Gaussian width is invariant under taking convex hulls by Proposition 7.7.2, so we can assume that T is convex. (Otherwise, one needs to compute the convex hull of T, for which algorithms are available.)

Next, for a fixed $g \in \mathbb{R}^n$, we need to be able to compute.

$$\sup_{x \in T} g(x) \tag{7.29}$$

We can frame this as a convex optimization problem, since we are maximizing a linear function on a convex set.

Finally, we need to average over $g \sim N(0, I_n)$. This can be avoided using Gaussian concentration. Indeed, computing (7.29) with just one realization of g will produce a result close to $w(T)$:

Exercise 7.7.15 (Approximating Gaussian width)**.** [Difficulty=3] *Let $T \subset \mathbb{R}^n$ and $g \sim N(0, I_n)$. Show that for any $u \geq 0$,*

$$\Big| \sup_{x \in T} \langle g, x \rangle - w(T) \Big| \leq u \cdot r(T)$$

with probability at least $1 - 2\exp(-cu^2)$, where

$$r(T) = \sup_{x \in T} \|x\|_2$$

denotes the radius of T. Hint: Deduce this from Gaussian concentration, Theorem 5.2.1.

Is it possible to compute width with *relative* $1 \pm \varepsilon$ error?

Computing the Gaussian width $w(T)$ theoretically for a general given set T, even up to a constant factor, may not be easy. As we know, Sudakov's minoration inequality gives a *lower bound* on the Gaussian width in terms of the covering numbers of T. Indeed, Corollary 7.5.3 states that

$$w(T) \geq c\varepsilon \sqrt{\log \mathcal{N}(T, \varepsilon)}, \quad \varepsilon > 0.$$

In the next section we will develop an *upper bound* on the Gaussian width in terms of the covering numbers of T, known as Dudley's inequality.

7.8 Random projections of sets

Let us consider a set $T \subset \mathbb{R}^n$ and project it onto a random m-dimensional subspace in $\mathbb{R}^n$ chosen uniformly from the Grassmanian $G_{n,m}$ (see Figure 5.2 for illustration. In applications, we might think of T as a collection of data points and P as a means of dimension reduction. What can we say about the geometry of the projected set PT?

If T is a finite set, Johnson-Lindenstrauss Lemma (Theorem 5.3.1) can be helpful. It states that as long as

$$m \gtrsim \log |T|, \tag{7.30}$$

the random projection P acts essentially as scaling of T. Indeed, P shrinks all distances between points in T by a factor $\approx \sqrt{m/n}$, and in particular

$$\operatorname{diam}(PT) \approx \sqrt{\frac{m}{n}} \operatorname{diam}(T). \tag{7.31}$$

If there are too many points in T, or perhaps T is an infinite set, then (7.31) may fail. For example, if $T = B_2^n$ is a Euclidean ball than no projection can shrink the size of T at all, and we have

$$\operatorname{diam}(PT) = \operatorname{diam}(T). \tag{7.32}$$

What happens for a general set T? The following result states that a random projection shrinks T as in (7.31), but it can not shrink beyond the spherical width of T.

Theorem 7.8.1 (Sizes of random projections of sets)**.** *Consider a bounded set $T \subset \mathbb{R}^n$. Let P be a projection in $\mathbb{R}^n$ onto a random m-dimensional subspace $E \sim \operatorname{Unif}(G_{n,m})$. Then, with probability at least $1 - 2e^{-m}$, we have*

$$\operatorname{diam}(PT) \le C\Big[w_s(T) + \sqrt{\frac{m}{n}}\, \operatorname{diam}(T)\Big].$$

For the proof of this result, it will be convenient to change the model for the random projection P.

Exercise 7.8.2 (Equivalent models for random projections)**.** *Let P be a projection in $\mathbb{R}^n$ onto a random m-dimensional subspace $E \sim \operatorname{Unif}(G_{n,m})$. Let Q be an $m \times n$ matrix obtained by choosing the first m rows of a random $n \times n$ orthogonal matrix, which is drawn uniformly from the orthogonal group $U \sim \operatorname{Unif}(O(n))$.*

1. *Show that for any fixed point $x \in \mathbb{R}^n$,*

$$\|Px\|_2 \text{ and } \|Qx\|_2 \text{ have the same distribution.}$$

Hint: Use the singular value decomposition of P.

2. *Show that for any fixed point $z \in S^{m-1}$,*

$$Q^{\mathsf{T}} z \sim \mathrm{Unif}(S^{n-1}).$$

In other words, the map Q^{T} acts a random isometric embedding of $\mathbb{R}^m$ into $\mathbb{R}^n$. Hint: It is enough to check the rotation invariance of the distribution of $Q^{\mathsf{T}} z$.

Proof of Theorem 7.8.1. This proof is one more example of an ε-net argument. Without loss of generality, we may assume that $\mathrm{diam}(T) \le 1$. (Why?)

Step 1: Approximation. By Exercise 7.8.2, it suffices to prove the theorem for Q instead of P. So we are going to bound

$$\mathrm{diam}(QT) = \sup_{x \in T-T} \|Qx\|_2 = \sup_{x \in T-T} \max_{z \in S^{m-1}} \langle Qx, z \rangle .$$

Similarly to our older arguments (for example, in the proof of Theorem 4.3.5 on random matrices), we will discretize the sphere S^{n-1}. Choose an $(1/2)$-net $\mathcal{N}$ of S^{n-1} so that

$$|\mathcal{N}| \le 5^m;$$

this is possible to do by Corollary 4.2.11. We can replace the supremum over the sphere S^{n-1} by the supremum over the net $\mathcal{N}$ paying a factor 2:

$$\mathrm{diam}(QT) \le 2 \sup_{x \in T-T} \max_{z \in \mathcal{N}} \langle Qx, z \rangle = 2 \max_{z \in \mathcal{N}} \sup_{x \in T-T} \left\langle Q^{\mathsf{T}} z, x \right\rangle . \tag{7.33}$$

(Recall Exercise 4.3.2.) In order to control this double maximum, we will first control the quantity

$$\sup_{x \in T-T} \left\langle Q^{\mathsf{T}} z, x \right\rangle . \tag{7.34}$$

for a fixed $z \in \mathcal{N}$ and with high probability, and then take union bound over all z.

Step 2: Concentration. So, let us fix $z \in \mathcal{N}$. By Exercise 7.8.2, $Q^{\mathsf{T}} z \sim \mathrm{Unif}(S^{n-1})$. Then we might recognize the expectation of (7.34) as the spherical width defined in (7.20):

$$\mathbb{E} \sup_{x \in T-T} \left\langle Q^{\mathsf{T}} z, x \right\rangle = w_s(T-T) = 2w_s(T).$$

(The last identity is the spherical version of a similar property of the Gaussian width, see part 5 of Proposition 7.7.2.)

Next, let us check that (7.34) concentrates nicely around its mean $2w_s(T)$. For this, we can use the concentration inequality (5.6) for Lipschitz functions on the sphere. Since we assumed that $\operatorname{diam}(T) \le 1$ in the beginning, once can quickly check that the function

$$\theta \mapsto \sup_{x \in T-T} \langle \theta, x \rangle$$

is a Lipschitz function on the sphere S^{n-1}, and its Lipschitz norm is at most 1. (Do this!) Therefore, applying the concentration inequality (5.6), we obtain

$$\mathbb{P}\left\{ \sup_{x \in T-T} \left\langle Q^{\mathsf{T}} z, x \right\rangle \ge 2w_s(T) + t \right\} \le 2\exp(-cnt^2).$$

Step 3: Union bound. Now we unfix $z \in \mathcal{N}$ by taking the union bound over $\mathcal{N}$. We get

$$\mathbb{P}\left\{ \max_{z \in \mathcal{N}} \sup_{x \in T-T} \left\langle Q^{\mathsf{T}} z, x \right\rangle \ge 2w_s(T) + t \right\} \le |\mathcal{N}| \cdot 2\exp(-cnt^2) \tag{7.35}$$

Recall that $|\mathcal{N}| \le 5^m$. Then, if we choose

$$t = C\sqrt{\frac{m}{n}}$$

with C large enough, the probability in (7.35) can be bounded by $2e^{-m}$. Then (7.35) and (7.33) yield

$$\mathbb{P}\left\{ \frac{1}{2}\operatorname{diam}(QT) \ge 2w(T) + C\sqrt{\frac{m}{n}} \right\} \le e^{-m}.$$

This proves Theorem 7.8.1. □

Exercise 7.8.3 (Gaussian projection)**.** [Difficulty=6] *Prove a version of Theorem 7.8.1 for $m \times n$ Gaussian random matrix G with independent $N(0,1)$ entries. Specifically, show that for any bounded set $T \subset \mathbb{R}^n$, we have*

$$\operatorname{diam}(GT) \le C\left[w(T) + \sqrt{m}\operatorname{diam}(T)\right]$$

with probability at least $1 - 2e^{-m}$. Here $w(T)$ is the Gaussian width of T.

7.8.1 The phase transition

Let us pause to take a closer look at the bound Theorem 7.8.1 gives. We can equivalently write it as

$$\operatorname{diam}(PT) \le C \max\left[w_s(T),\ \sqrt{\frac{m}{n}}\operatorname{diam}(T)\right].$$

Let us compute the dimension m for which the phase transition occurs between the two terms $w_s(T)$ and $\sqrt{\frac{m}{n}}\operatorname{diam}(T)$. Setting them equal to each other and solving for m, we find that the phase transition happens when

$$\begin{aligned} m &= \frac{(\sqrt{n}\, w_s(T))^2}{\operatorname{diam}(T)^2} \\ &\sim \frac{w(T)^2}{\operatorname{diam}(T)^2} \quad \text{(pass to Gaussian width using Lemma 7.7.4)} \\ &\sim d(T) \quad \text{(by Definition 7.7.8 of statistical dimension).} \end{aligned}$$

So we can express the result Theorem 7.8.1 as follows:

$$\operatorname{diam}(PT) \le \begin{cases} C\sqrt{\frac{m}{n}}\operatorname{diam}(T), & m \ge d(T) \\ Cw_s(T), & m \le d(T). \end{cases}$$

Figure 7.8 shows a graph of $\operatorname{diam}(PT)$ as a function of the dimension m.

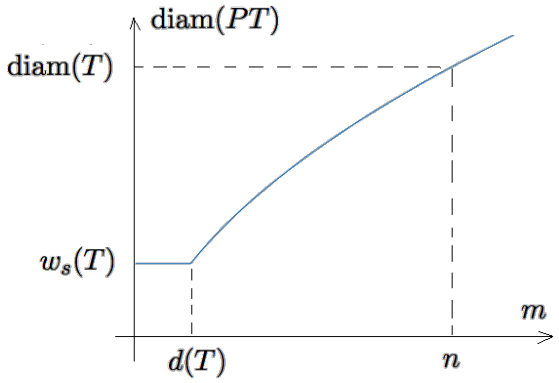


Figure 7.8: The diameter of a random m-dimensional projection of a set T as a function of m.

For large m, the random m-dimensional projection shrinks T by the factor $\sim \sqrt{m/n}$, just like we have seen in (7.31) in the context of Johnson-Lindenstrauss lemma. However, when the dimension m drops below the

statistical dimension $d(T)$, the shrinking stops – it levels off at the spherical width $w_s(T)$. We saw an example of this in (7.32), where a Euclidean ball can not be shrunk by a projection.

Exercise 7.8.4 (The reverse bound)**.** [Difficulty=4] *Show that the bound in Theorem 7.8.1 is optimal: prove the reverse bound*

$$\mathbb{E}\,\mathrm{diam}(PT) \geq c\Big[w_s(T) + \sqrt{\frac{m}{n}}\,\mathrm{diam}(T)\Big]$$

for all bounded sets $T \subset \mathbb{R}^n$. Hint: To obtain the bound $\mathbb{E}\,\mathrm{diam}(PT) \gtrsim w_s(T)$, reduce P to a one-dimensional projection by dropping terms from the singular value decomposition of P. To obtain the bound $\mathbb{E}\,\mathrm{diam}(PT) \geq \sqrt{\frac{m}{n}}\,\mathrm{diam}(T)$, argue about a pair of points in T.

As an application of Theorem 7.8.1, we can obtain a bound on the norms of random projections of matrices.

Corollary 7.8.5 (Random projections of matrices)**.** *Let A be an $n \times k$ matrix.*

1. Let P is a projection in $\mathbb{R}^n$ onto a random m-dimensional subspace chosen uniformly in $G_{n,m}$. Then, with probability at least $1 - 2e^{-m}$, we have

$$\|PA\| \leq C\Big[\frac{1}{\sqrt{n}}\,\|A\|_F + \sqrt{\frac{m}{n}}\,\|A\|\Big].$$

2. Let G be an $m \times n$ Gaussian random matrix with independent $N(0,1)$ entries. Then, with probability at least $1 - 2e^{-m}$, we have

$$\|GA\| \leq C\left(\|A\|_F + \sqrt{m}\,\|A\|\right).$$

Proof. The operator norm is clearly one half the diameter of the image of the unit ball, i.e.

$$\|PA\| = \frac{1}{2}\,\mathrm{diam}(P(AB_2^k)).$$

The Gaussian width of the ellipsoid AB_2^k is equivalent to $\|A\|_F$, see Exercise 7.7.12 and (7.25). Applying Theorem 7.8.1, we obtain the first part of the conclusion. The second part, for Gaussian projections, follows in a similar way from Exercise 7.8.3. □

Chapter 8

Chaining

8.1 Dudley's inequality

Sudakov's minoration inequality that we studied in Section 7.5 gives a *lower bound* on the magnitude

$$\mathbb{E} \sup_{t \in T} X_t$$

of a Gaussian random process $(X_t)_{t \in T}$ in terms of the metric entropy of T. In this section, we will obtain a similar *upper bound*.

This time, we will be able to work not just with Gaussian processes but with more general processes with sub-gaussian increments.

Definition 8.1.1 (Sub-gaussian increments)**.** *Consider a random process $(X_t)_{t \in T}$ on a metric space (T, d). We say that the process has* sub-gaussian increments *if there exists $K \geq 0$ such that*

$$\|X_t - X_s\|_{\psi_2} \leq K d(t, s) \quad \text{for all } t, s \in T. \tag{8.1}$$

An obvious example is a Gaussian process $(X_t)_{t \in T}$ on an abstract set T, and with the metric defined by

$$d(t, s) = \|X_t - X_s\|_2.$$

But in general, the metric d on T may not be induced by the increments of the process as in the example above.

Dudley's inequality gives a bound on general sub-gaussian random processes $(X_t)_{t \in T}$ in terms of the metric entropy of T.

Theorem 8.1.2 (Dudley's integral inequality)**.** *Let $(X_t)_{t\in T}$ be a random process on a metric space (T,d) with sub-gaussian increments as in (8.1). Then*

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

Before we prove Dudley's inequality, it is helpful to compare it with Sudakov's inequality, which for Gaussian processes states that

$$\mathbb{E}\sup_{t\in T} X_t \ge c\sup_{\varepsilon\ge 0}\varepsilon\sqrt{\log \mathcal{N}(T,d,\varepsilon)}.$$

Figure 8.1 illustrates Dudley's and Sudakov's bounds. There is an obvious gap between these two bounds. It can not be closed in terms of the entropy numbers alone; we will explore this point later.

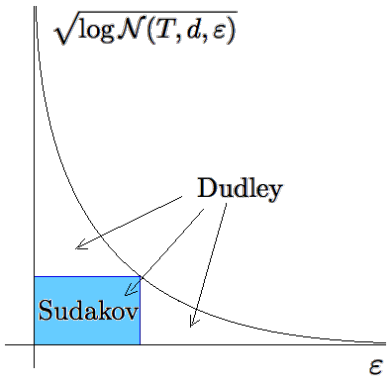


Figure 8.1: Dudley's inequality bounds $\mathbb{E}\sup_{t\in T} X_t$ by the area under the curve. Sudakov's inequality bounds it below by the largest area of a rectangle under the curve, up to constants.

The form of Dudley's inequality might suggest us that $\mathbb{E}\sup_{t\in T} X_t$ is a *multi-scale* quantity, in that we have to examine T at all possible scales ε in order to bound the process. This is indeed so, and the proof will indeed be multi-scale. We will now state and prove a discrete version of Dudley's inequality, where the integral over all positive ε is replaced by a sum over dyadic values $\varepsilon = 2^{-k}$, which somewhat resembles a Riemann sum. Later we will quickly pass to the original form of Dudley's inequality.

Theorem 8.1.3 (Dudley's inequality as a sum)**.** *Let $(X_t)_{t\in T}$ be a random process on a metric space (T,d) with sub-gaussian increments as in (8.1). Then*

Call (X_t) a "sub-gaussian process throughout

$$\mathbb{E} \sup_{t \in T} X_t \le CK \sum_{k \in \mathbb{Z}} 2^{-k} \sqrt{\log \mathcal{N}(T, d, 2^{-k})}. \tag{8.2}$$

Our proof of this theorem will be based on the important technique of *chaining*, which can be useful in many other problems. Chaining is a *multi-scale* version of the ε-net argument that we used successfully in the past, for example in the proofs of Theorems 4.3.5 and 7.8.1.

In the familiar, single-scale ε-net argument, we discretize T by choosing an ε-net $\mathcal{N}$ of t. Then every point $t \in T$ can be approximated by a closest point from the net $\pi(t) \in \mathcal{N}$ with accuracy ε, so that $d(t, \pi(t)) \le \varepsilon$. The increment condition (8.1) yields

$$\|X_t - X_{\pi(t)}\|_{\psi_2} \le CK\varepsilon. \tag{8.3}$$

This gives

$$\mathbb{E} \sup_{t \in T} X_t \le \mathbb{E} \sup_{t \in T} X_{\pi(t)} + \mathbb{E} \sup_{t \in T} (X_t - X_{\pi(t)}).$$

The first term can be controlled by a union bound over $|\mathcal{N}| = \mathcal{N}(T, d, \varepsilon)$ points $\pi(t)$.

To bound the second term, we would like to use (8.3). But it only holds for fixed $t \in T$, and it is not clear how to control the supremum over $t \in T$. To overcome this difficulty, we will not stop here but continue to run the ε-net argument further, building progressively finer approximations $\pi_1(t), \pi_2(t), \dots$ to t with finer nets. Let us now develop formally this technique of chaining.

Proof of Theorem 8.1.3. **Step 1: Chaining set-up.** Without loss of generality, we may assume that $K = 1$ and that T is finite. (Why?) Let us set the dyadic scale

$$\varepsilon_k = 2^{-k}, \quad k \in \mathbb{Z} \tag{8.4}$$

and choose ε_k-nets $\mathcal{N}_k$ of T so that

$$|\mathcal{N}_k| = \mathcal{N}(T, d, \varepsilon_k). \tag{8.5}$$

Change $\mathcal{N}_k$ to T_k; this is the notation adopted in the generic chaining later.

Only a part of the dyadic scale will be needed. Indeed, since T is finite, we can choose κ small enough (for the coarsest net) and K large enough (for the finest net) so that

$$\mathcal{N}_\kappa = \{t_0\} \text{ for some } t_0 \in T, \quad \mathcal{N}_K = T. \tag{8.6}$$

For a point $t \in T$, let $\pi_k(t)$ denote a closest point in $\mathcal{N}_k$, so we have

$$d(t, \pi_k(t)) \le \varepsilon_k. \tag{8.7}$$

Since $\mathbb{E}\, X_{t_0} = 0$, we have

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

We can express $X_t - X_{t_0}$ as a telescoping sum; think about walking from t_0 to t along a chain of points $\pi_k(t)$ that mark progressively finer approximations to t:

$$X_t - X_{t_0} = (X_{\pi_\kappa(t)} - X_{t_0}) + (X_{\pi_{\kappa+1}(t)} - X_{\pi_\kappa(t)}) + \cdots + (X_t - X_{\pi_K(t)}), \quad (8.8)$$

see Figure 8.2 for illustration. The first and last terms of this sum are zero

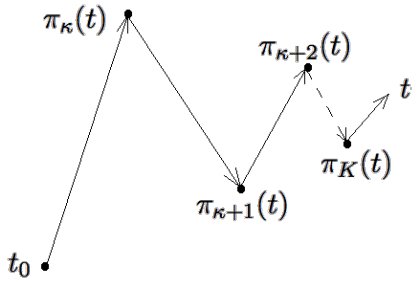


Figure 8.2: Chaining: a walk from a fixed point t_0 to an arbitrary point t in T along elements $\pi_k(T)$ of progressively finer nets of T

by (8.6), so we have

$$X_t - X_{t_0} = \sum_{k=\kappa+1}^{K} (X_{\pi_k(t)} - X_{\pi_{k-1}(t)}). \quad (8.9)$$

Since the supremum of the sum is bounded by the sum of suprema, this yields

$$\mathbb{E} \sup_{t \in T} (X_t - X_{t_0}) \le \sum_{k=\kappa+1}^{K} \mathbb{E} \sup_{t \in T} (X_{\pi_k(t)} - X_{\pi_{k-1}(t)}). \quad (8.10)$$

Step 2: Controlling the increments. Although each term in the bound (8.10) still has a supremum over the entire set T, a closer look reveals that it is actually a maximum over a much smaller set, namely the set all possible pairs $(\pi_k(t), \pi_{k-1}(t))$. The number of such pairs is

$$|\mathcal{N}_k| \cdot |\mathcal{N}_{k-1}| \le |\mathcal{N}_k|^2,$$

a number that we can control through (8.5).

Next, for a fixed t, the increments in (8.10) can be bounded as follows:

$$\begin{aligned}\|X_{\pi_k(t)} - X_{\pi_{k-1}(t)}\|_{\psi_2} &\le d(\pi_k(t),\, \pi_{k-1}(t)) \quad \text{(by (8.1) and since } K = 1) \\ &\le d(\pi_k(t), t) + d(t,\, \pi_{k-1}(t)) \quad \text{(by triangle inequality)} \\ &\le \varepsilon_k + \varepsilon_{k-1} \quad \text{(by (8.7))} \\ &\le 2\varepsilon_{k-1}.\end{aligned}$$

Recall from Exercise 2.5.8 that the expected maximum of N sub-gaussian random variables is at most $CL\sqrt{\log N}$ where L is the maximal ψ_2 norm. Thus we can bound each term in (8.10) as follows:

$$\mathbb{E} \sup_{t \in T} (X_{\pi_k(t)} - X_{\pi_{k-1}(t)}) \le C\varepsilon_{k-1} \sqrt{\log |\mathcal{N}_k|}. \tag{8.11}$$

Step 3: Summing up the increments. We have shown that

$$\mathbb{E} \sup_{t \in T} (X_t - X_{t_0}) \le C \sum_{k=\kappa+1}^{K} \varepsilon_{k-1} \sqrt{\log |\mathcal{N}_k|}. \tag{8.12}$$

It remains substitute the values $\varepsilon_k = 2^{-k}$ from (8.4) and the bounds (8.5) on $|\mathcal{N}_k|$, and conclude that

$$\mathbb{E} \sup_{t \in T} (X_t - X_{t_0}) \le C_1 \sum_{k=\kappa+1}^{K} 2^{-k} \sqrt{\log \mathcal{N}(T, d, 2^{-k})}.$$

Theorem 8.1.3 is proved. □

Let us now deduce the integral form of Dudley's inequality.

Proof of Dudley's integral inequality, Theorem 8.1.2. To convert the sum (8.2) into an integral, we express 2^{-k} as $2\int_{2^{-k-1}}^{2^{-k}} d\varepsilon$. Then

$$\sum_{k \in \mathbb{Z}} 2^{-k} \sqrt{\log \mathcal{N}(T, d, 2^{-k})} = 2 \sum_{k \in \mathbb{Z}} \int_{2^{-k-1}}^{2^{-k}} \sqrt{\log \mathcal{N}(T, d, 2^{-k})}\, d\varepsilon.$$

Within the limits of integral, $2^{-k} \ge \varepsilon$, so $\log \mathcal{N}(T, d, 2^{-k}) \le \log \mathcal{N}(T, d, \varepsilon)$ and the sum is bounded by

$$2 \sum_{k \in \mathbb{Z}} \int_{2^{-k-1}}^{2^{-k}} \sqrt{\log \mathcal{N}(T, d, \varepsilon)}\, d\varepsilon = 2 \int_0^\infty \sqrt{\log \mathcal{N}(T, d, \varepsilon)}\, d\varepsilon.$$

The proof is complete. □

Exercise 8.1.4 (Equivalence of Dudley's integral and sum)**.** [Difficulty=4] *In the proof of Theorem 8.1.2 we bounded Dudley's integral by a sum. Show the reverse bound:*

$$\int_0^\infty \sqrt{\log \mathcal{N}(T,d,\varepsilon)}\, d\varepsilon \le C \sum_{k\in\mathbb{Z}} 2^{-k} \sqrt{\log \mathcal{N}(T,d,2^{-k})}.$$

Dudley's inequality gives a bound on the expectation only, but adapting the argument yields a nice tail bound as well.

Theorem 8.1.5 (Dudley's integral inequality: tail bound)**.** *Let $(X_t)_{t\in T}$ be a random process on a metric space (T,d) with sub-gaussian increments as in (8.1). Then, for every $u \ge 0$, the event*

$$\mathbb{E}\sup_{t\in T} X_t \le CK\Big[\int_0^\infty \sqrt{\log \mathcal{N}(T,d,\varepsilon)}\, d\varepsilon + u\cdot \operatorname{diam}(T)\Big]$$

holds with probability at least $1 - 2\exp(-u^2)$.

Exercise 8.1.6. [Difficulty=6] *Prove Theorem 8.1.5. To this end, first obtain a high-probability version of (8.11):*

$$\sup_{t\in T}(X_{\pi_k(t)} - X_{\pi_{k-1}(t)}) \le C\varepsilon_{k-1}\left[\sqrt{\log|\mathcal{N}_k|} + z\right]$$

with probability at least $1 - 2\exp(-z^2)$.

Use this inequality with $z = z_k$ to control all such terms simultaneously. Summing them up, deduce a bound on $\mathbb{E}\sup_{t\in T} X_t$ with probability at least $1 - 2\sum_k \exp(-z_k^2)$. Finally, choose the values for z_k that give you a good bound; one can set $z_k = u + \sqrt{k-\kappa}$ for example.

8.1.1 Remarks and Examples

Remark 8.1.7 (Limits of Dudley's integral). Although Dudley's integral is formally over $[0,\infty]$, we can clearly make the upper bound equal the diameter of T in the metric d, thus

$$\mathbb{E}\sup_{t\in T} X_t \le CK \int_0^{\operatorname{diam}(T)} \sqrt{\log \mathcal{N}(T,d,\varepsilon)}\, d\varepsilon. \tag{8.13}$$

Indeed, if $\varepsilon > \operatorname{diam}(T)$ then a single point (any point in T) is an ε-net of T, which shows that $\log \mathcal{N}(T,d,\varepsilon) = 0$ for such ε.

Let us apply Dudley's inequality for the canonical Gaussian process, just like we did with Sudakov's inequality in Section 7.5.1. We immediately obtain the following bound.

Theorem 8.1.8 (Dudley's inequality for sets in $\mathbb{R}^n$)**.** *For any set $T \subset R^n$, we have*

$$w(T) \le C \int_0^\infty \sqrt{\log N(T, \varepsilon)}\, d\varepsilon.$$

Example 8.1.9. Let us test Dudley's inequality for the unit Euclidean ball $T = B_2^n$. Recall from (4.6) that

$$N(B_2^n, \varepsilon) \le \left(\frac{3}{\varepsilon}\right)^n \quad \text{for } \varepsilon \in (0, 1]$$

and $N(B_2^n, \varepsilon) = 1$ for $\varepsilon > 1$. Then Dudley's inequality yields a converging integral

$$w(B_2^n) \le C \int_0^1 \sqrt{n \log \frac{3}{\varepsilon}}\, d\varepsilon \le C_1 \sqrt{n}.$$

This is optimal: indeed, as we know from (7.21), the Gaussian width of B_2^n is of the order of $\sqrt{n}$.

Example 8.1.10. Let us do one more test on Dudley's inequality. If $T = P$ is a polyhedron with N vertices, then we know from Corollary 7.6.3 that

$$\mathcal{N}(P, \varepsilon) \le N^{\lceil 1/\varepsilon^2 \rceil}.$$

Substituting this into Dudley's inequality, we get

$$w(P) \le C \int_0^{\operatorname{diam}(P)} \frac{1}{\varepsilon} \sqrt{\log N}\, d\varepsilon.$$

This is, unfortunately, a diverging integral. Note, however, that it is barely divergent; a little better bound on the covering numbers should lead to convergence.

8.1.2 Two-sided Sudakov's inequality

The previous examples suggest (but not prove) that Dudley's inequality should be almost optimal. Although in general it is not true and there is a gap between Dudley's and Sudakov's inequalities, this gap is only logarithmically large in some sense. Let us make this more precise and show that Sudakov's inequality in $\mathbb{R}^n$ is optimal up to $\log n$ factor.

Theorem 8.1.11 (Two-sided Sudakov's inequality)**.** *Let $T \subset \mathbb{R}^n$ and set*

$$s(T) := \sup_{\varepsilon \geq 0} \varepsilon \sqrt{\log \mathcal{N}(T, \varepsilon)}.$$

Then

$$c \cdot s(T) \leq w(T) \leq C \log(n) \cdot s(T).$$

Proof. The main idea here is that because the chaining is expected to converge exponentially fast, $O(\log n)$ steps should suffice to walk from t_0 to somewhere very near t.

As we already noted in (8.13), the coarsest scale in the chaining sum (8.9) can be chosen as the diameter of T. In other words, we can start the chaining at κ which is the smallest integer such that

$$2^{-\kappa} < \operatorname{diam}(T).$$

This is not different from what we did before. What will be different is the finest scale. Instead of going all the way down, let us stop chaining at K which is the largest integer for which

$$2^{-K} \geq \frac{w(T)}{4\sqrt{n}}.$$

(It will be clear why we made this choice in a second.)

Then the last term in (8.8) may not be zero as before, and instead of (8.9) we will need to bound

$$w(T) \leq \sum_{k=\kappa+1}^{K} \mathbb{E} \sup_{t \in T} (X_{\pi_k(t)} - X_{\pi_{k-1}(t)}) + \mathbb{E} \sup_{t \in T} (X_t - X_{\pi_K(t)}). \tag{8.14}$$

To control the last term, recall that $X_t = \langle g, t \rangle$ is the canonical process, so

$$\begin{aligned}
\mathbb{E} \sup_{t \in T} (X_t - X_{\pi_K(t)}) &\leq \mathbb{E} \sup_{t \in T} \langle g, t - \pi_K(t) \rangle \\
&\leq 2^{-K} \cdot \mathbb{E} \, \|g\|_2 \quad (\text{since } \|t - \pi_K(t)\|_2 \leq 2^{-K}) \\
&\leq 2^{-K} \sqrt{n} \\
&\leq \frac{w(T)}{2\sqrt{n}} \cdot \sqrt{n} \quad (\text{by definition of } K) \\
&\leq \frac{1}{2} w(T).
\end{aligned}$$

Putting this into (8.14) and subtracting $\frac{1}{2}\, w(T)$ from both sides, we conclude that

$$w(T) \le 2 \sum_{k=\kappa+1}^{K} \mathbb{E} \sup_{t \in T} (X_{\pi_k(t)} - X_{\pi_{k-1}(t)}). \tag{8.15}$$

Thus, we have removed the last term from (8.14). Each of the remaining terms can be bounded as before. The number of terms in this sum is

$$\begin{aligned} K - \kappa &\le \log_2 \frac{\operatorname{diam}(T)}{w(T)/4\sqrt{n}} \quad \text{(by definition of } K \text{ and } \kappa) \\ &\le \log_2 \left(4\sqrt{n} \cdot \sqrt{2\pi}\right) \quad \text{(by property 6 of Proposition 7.7.2)} \\ &\le C \log n. \end{aligned}$$

Thus we can replace the sum by the maximum in (8.15) by paying a factor $C \log n$. This completes the argument like before, in the proof of Theorem 8.1.3. □

An example of non-optimality of Dudley?

Exercise 8.1.12 (Limits in Dudley's integral)**.** [Difficulty=7] *Prove the following improvement of Dudley's integral inequality (Theorem 8.1.8). For any set $T \subset R^n$, we have*

$$w(T) \le C \int_a^b \sqrt{\log N(T, \varepsilon)}\, d\varepsilon \quad \textit{where} \quad a = \frac{cw(T)}{\sqrt{n}}, \quad b = \operatorname{diam}(T).$$

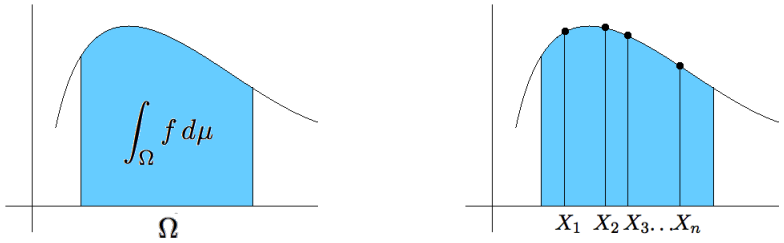
8.2 Application: empirical processes and uniform laws of large numbers

We will give an application of Dudley's inequality to *empirical processes*, which are certain random processes indexed by functions. The theory of empirical processes is a large branch of probability theory, and we will only be able to scratch its surface here. Let us consider a motivating example.

8.2.1 Monte-Carlo method

Suppose we want to evaluate the integral of a function $f : \Omega \to \mathbb{R}$ with respect to some probability measure μ on some domain $\Omega \subset \mathbb{R}^d$:

$$\int_\Omega f\, d\mu,$$



(a) The problem is to compute the integral of f on a domain Ω.

(b) The integral is approximated by the sum $\frac{1}{n}\sum_1^n f(X_i)$ with randomly sampled points X_i.

Figure 8.3: Monte-Carlo method for randomized, numerical integration.

see Figure 8.3a. For example, we could be interested in computing $\int_0^1 f(x)\,dx$ for a function $f : [0,1] \to \mathbb{R}$.

Here is how we can use probability to evaluate integrals. Consider a random vector X that takes values in Ω according to the law μ, i.e.

$$\mathbb{P}\{X \in A\} = \mu(A) \quad \text{for any measurable set } A \subset \Omega.$$

(For example, to evaluate $\int_0^1 f(x)\,dx$, we take $X \sim \text{Unif}[0,1]$.) Then we may interpret the integral as expectation:

$$\int_\Omega f\,d\mu = \mathbb{E}\,f(X).$$

Let $X_1, X_2, \ldots$ be i.i.d. copies of X. The Law of Large Numbers (Theorem 1.3.1) yields that

$$\frac{1}{n}\sum_1^n f(X_i) \to \mathbb{E}\,f(X) \quad \text{almost surely} \tag{8.16}$$

as $n \to \infty$. This means that we can approximate the integral by the sum

$$\int_\Omega f\,d\mu \approx \frac{1}{n}\sum_1^n f(X_i) \tag{8.17}$$

where the points X_i are drawn at random from the domain Ω; see Figure 8.3b. for illustration. This way of numerically computing integrals is called the *Monte-Carlo method*.

Remarkably, we do not need to know the measure μ to evaluate the integral; it suffices to be able to draw random samples X_i according to μ.

Similarly, we do not even need to know f at all points in the domain; a few random points suffice.

The average error of integration is $O(1/\sqrt{n})$, which follows from the rate of convergence in the Law of Large Numbers:

$$\mathbb{E}\left|\frac{1}{n}\sum_{1}^{n} f(X_i) - \mathbb{E}\, f(X)\right| = O\Big(\frac{1}{\sqrt{n}}\Big). \tag{8.18}$$

(Check this by computing the variance of the sum.)

Maybe include this rate in LLN?

8.2.2 A uniform law of large numbers

Can we use one, fixed sample $X_1, \dots, X_n$ to evaluate the integral of *any* function $f : \Omega \to \mathbb{R}$? Of course, not. For a given sample, one can choose a function that oscillates wildly between the sample points, and the approximation (8.17) will fail. However, if we only look at functions f that do not oscillate wildly – for example, Lipschitz functions, then Monte-Carlo method (8.17) will work simultaneously over all such f.

This follows from a *uniform Law of Large Numbers*, which states that (8.16) holds uniformly over all functions in a given function class. In our case, it will be the class of Lipschitz functions

$$\mathcal{F} := \{f : [0,1] \to \mathbb{R},\ \|f\|_{\text{Lip}} \le L\} \tag{8.19}$$

where L is some number.

Theorem 8.2.1 (Uniform Law of Large Numbers)**.** *Let X be a random variable taking values in $[0,1]$, and let $X_1, X_2, \dots, X_n$ be independent copies of X. Then*

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}\, f(X)\right| \le \frac{CL}{\sqrt{n}}. \tag{8.20}$$

Note that this result, which is uniform over *all* Lipschitz f, holds with the same rate of convergence (8.18) as the classical Law of Large numbers that holds for a *single* f.

To prepare for the proof of Theorem 8.2.1, let us view the left side of (8.20) as the magnitude of a random process indexed by functions $f \in \mathcal{F}$. Such random processes are called *empirical processes*.

Definition 8.2.2. *Let $\mathcal{F}$ be a class of real-valued functions $f : \Omega \to \mathbb{R}$ where (Ω, Σ, μ) is a probability space. Let X be a random point in Ω distributed*

according to the law μ, and let $X_1, X_2, \dots, X_n$ be independent copies of X. The random process $(X_f)_{f \in \mathcal{F}}$ defined by

$$X_f := \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E} f(X) \tag{8.21}$$

is called an empirical process *indexed by $\mathcal{F}$.*

Proof of Theorem 8.2.1. Without loss of generality we may assume that $L = 1$ and that

$$f : [0,1] \to [0,1] \quad \text{for all } f \in \mathcal{F}. \tag{8.22}$$

(Why?) We would like to bound the magnitude

$$\mathbb{E} \sup_{f \in \mathcal{F}} |X_f|$$

of the empirical process $(X_f)_{f \in \mathcal{F}}$ defined in (8.21).

Step 1: checking sub-gaussian increments. We will do this using Dudley's inequality, Theorem 8.1.2. To apply this result, we just need to check that the empirical process has sub-gaussian increments. So, fix a pair of functions $f, g \in \mathcal{F}$ and consider

sub-gaussian increments $\mapsto$ sub-gaussian process

$$\|X_f - X_g\|_{\psi_2} = \frac{1}{n} \Big\| \sum_{i=1}^{n} Z_i \Big\|_{\psi_2} \quad \text{where} \quad Z_i := (f-g)(X_i) - \mathbb{E}(f-g)(X).$$

Random variables Z_i are independent and have mean zero. So, by Proposition 2.6.1 we have

$$\|X_f - X_g\|_{\psi_2} \lesssim \frac{1}{n} \Big(\sum_{i=1}^{n} \|Z_i\|_{\psi_2}^2 \Big)^{1/2}.$$

Now, using centering (Lemma 2.6.6) we have

$$\|Z_i\|_{\psi_2} \lesssim \|(f-g)(X_i)\|_{\psi_2} \lesssim \|f-g\|_\infty.$$

It follows that

$$\|X_f - X_g\|_{\psi_2} \lesssim \frac{1}{n} \cdot n^{1/2} \|f-g\|_\infty = \frac{1}{\sqrt{n}} \|f-g\|_\infty.$$

Step 2: applying Dudley's inequality. We found that the empirical process $(X_f)_{f \in \mathcal{F}}$ has sub-gaussian increments with respect to the L_∞ norm.

This allows us to apply Dudley's inequality, Theorem 8.1.2. Note that (8.22) implies that the diameter of $\mathcal{F}$ in L_∞ metric is bounded by 1. Thus

$$\mathbb{E} \sup_{f \in \mathcal{F}} X_f \lesssim \frac{1}{\sqrt{n}} \int_0^1 \sqrt{\log \mathcal{N}(\mathcal{F}, \|\cdot\|_\infty, \varepsilon)}\, d\varepsilon$$

(recall (8.13)).

Using that all functions in $f \in \mathcal{F}$ are Lipschitz with $\|f\|_{\text{Lip}} \le 1$, it is not difficult to bound the covering numbers of $\mathcal{F}$ as follows:

$$\mathcal{N}(\mathcal{F}, \|\cdot\|_\infty, \varepsilon) \le \left(\frac{C}{\varepsilon}\right)^{C/\varepsilon};$$

we will show this in Exercise 8.2.3 below. This bound makes Dudley's integral converge, and we conclude that

$$\mathbb{E} \sup_{f \in \mathcal{F}} X_f \lesssim \frac{1}{\sqrt{n}} \int_0^1 \sqrt{\frac{C}{\varepsilon} \log \frac{C}{\varepsilon}}\, d\varepsilon \lesssim \frac{1}{\sqrt{n}}.$$

Finally, we note that

$$\mathbb{E} \sup_{f \in \mathcal{F}} |X_f| = \mathbb{E} \sup_{f \in \mathcal{F}} X_f$$

(see Exercise 8.2.4); so this quantity is $O(1/\sqrt{n})$ as well. This proves (8.20). □

Exercise 8.2.3 (Metric entropy of the class of Lipschitz functions)**.** *Consider the class of functions*

$$\mathcal{F} := \{f : [0,1] \to [0,1],\ \|f\|_{\text{Lip}} \le 1\}.$$

Show that

$$\mathcal{N}(\mathcal{F}, \|\cdot\|_\infty, \varepsilon) \le \left(\frac{1}{\varepsilon}\right)^{1/\varepsilon} \quad \text{for any } \varepsilon > 0.$$

Hint: put a mesh on the square $[0,1]^2$ with step ε. Given $f \in \mathcal{F}$, show that $\|f - f_0\|_\infty \le \varepsilon$ for some function f_0 whose graph follows the mesh; see Figure 8.4. The number all mesh-following functions f_0 is bounded by $(1/\varepsilon)^{1/\varepsilon}$.

Exercise 8.2.4. *Show that for an empirical process in (8.21) indexed by the function class*

$$\mathcal{F} := \{f : [0,1] \to [0,1],\ \|f\|_{\text{Lip}} \le 1\},$$

we have

$$\mathbb{E} \sup_{f \in \mathcal{F}} |X_f| = \mathbb{E} \sup_{f \in \mathcal{F}} X_f.$$

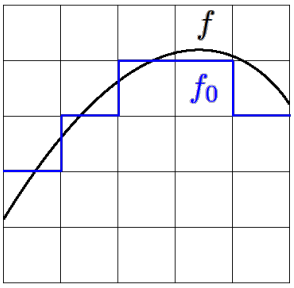


Figure 8.4: Bounding the metric entropy of the class of Lipschitz functions in Exercise 8.2.3. A Lipschitz function f is approximated by a function f_0 on a mesh.

Exercise 8.2.5 (An improved bound on the metric entropy)**.** *Improve the bound in Exercise 8.2.3 to*

$$\mathcal{N}(\mathcal{F}, \|\cdot\|_\infty, \varepsilon) \le e^{C/\varepsilon} \quad \text{for any } \varepsilon > 0.$$

Hint: Use that f is Lipschitz to find a better bound on the number of possible functions f_0.

Exercise 8.2.6 (Higher dimensions)**.** *Consider the class of functions*

$$\mathcal{F} := \left\{ f : [0,1]^d \to \mathbb{R},\ f(0) = 0, \|f\|_{\text{Lip}} \le 1 \right\}.$$

for some dimension $d \ge 1$. Show that

$$\mathcal{N}(\mathcal{F}, \|\cdot\|_\infty, \varepsilon) \le e^{C/\varepsilon^d} \quad \text{for any } \varepsilon > 0.$$

This bound would not allow to simply use Dudley to generalize uniform LLN to higher dimensions: the integral would diverge. See R. van Handel p. 5.31 how to use something like Exercise 8.1.12 to do this.

Empirical measure

Let us take one more look at the Definition 8.2.2 of empirical processes. Consider a probability measure μ_n that is uniformly distributed on the sample $X_1, \dots, X_N$, that is

$$\mu(\{X_i\}) = \frac{1}{n} \quad \text{for every } i = 1, \dots, n.$$

Note that μ_n is a *random* measure. It is called the *empirical measure*.

While the integral of f with respect to the original measure μ is the $\mathbb{E} f(X)$ (the "population" average of f) the integral of f with respect to the empirical measure is $\frac{1}{n}\sum_{i=1}^n f(X_i)$ (the "sample", or empirical, average of

f). In the literature on empirical processes, the population expectation of f is denoted by μf, and the empirical expectation, by $\mu_n f$:

$$\mu f = \int f \, d\mu = \mathbb{E}\, f(X); \quad \mu_n f = \int f \, d\mu_n = \frac{1}{n} \sum_{i=1}^{n} f(X_i).$$

The empirical process X_f in (8.21) thus measures the deviation of sample expectation from the empirical expectation:

$$X_f = \mu f - \mu_n f.$$

The Uniform Law of Large Numbers (8.20) gives a uniform bound on the deviation

$$\mathbb{E} \sup_{f \in \mathcal{F}} |\mu_n f - \mu f| \tag{8.23}$$

over the class of Lipschitz functions $\mathcal{F}$ defined in (8.19).

The quantity (8.23) can be thought as a distance between the measures μ_n and μ. It is called the *Wasserstein's distance* $W_1(\mu, \mu_n)$. The Wasserstein distance has an equivalent formulation as the *transportation cost* of measure μ into measure μ_n, where the cost of moving a mass (probability) $p > 0$ is proportional to p and to the distance moved. The equivalence between the transportation cost and (8.23) is provided by Kantorovich-Rubinstein's duality theorem.

8.3 Application: statistical learning theory

Statistical learning theory, or machine learning, allows one to make predictions based on data. A typical problem of statistical learning can be stated mathematically as follows. Consider a pair of random variables (X, Y) whose distribution is unknown. More generally, X is a random vector or even a point in an abstract probability space Ω like we saw in Section 8.2. It may be helpful to think of Y as a *label* of X.

Suppose we have a sample

$$(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n) \tag{8.24}$$

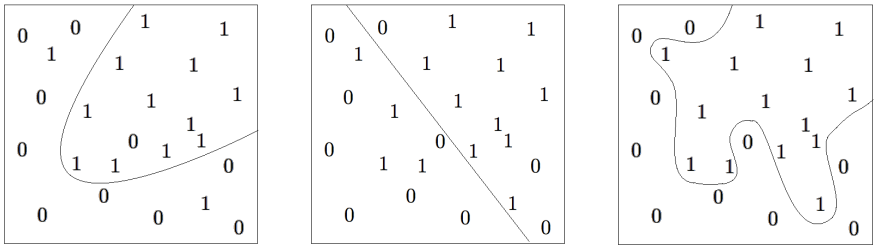
of n independent copies of (X, Y). This sample is called *training data*. Our goal is to learn from the training data how Y depends on X, so we can make a good *prediction* of the label Y for a point X.

Classification problems

Example 8.3.1. Suppose $X \in \mathbb{R}^d$ is a vector of parameters of a patient. For example, X_1 could be body temperature, X_2 blood pressure, etc. Another example is where X encodes d gene expressions. Let the label $Y \in \{0, 1\}$ be the diagnosis of the patient (0=healthy, 1=sick).

We conduct a study on n patients whose parameters and diagnoses are known. This gives us training data (8.24). Our goal is to learn how to predict the diagnosis Y based on the patient's parameters X.

This example belongs to the important family of *classification problems* where the label Y can take finitely many values that encode the class X belongs to. Figure 8.5a illustrates a classification problem where X is a random vector on the plane and the label Y can take values 0 and 1 like in Example 8.3.1. A solution of this classification problem can be described as a partition of the plane into two regions, one where $f(X) = 0$ (healthy) and another where $f(X) = 1$ (sick). Based on this solution, one can diagnose new patients by looking at which region their parameter vectors X fall in.



(a) Right fit (b) Underfitting (c) Overfitting

Figure 8.5: Trade-off between fit and complexity in statistical learning.

Risk, fit and complexity

Generally, a solution to the learning problem can be described as a function $f : \Omega \to \mathbb{R}$. We would naturally want f to minimize the *risk*

$$R(f) := \mathbb{E}(f(X) - Y)^2.$$

Example 8.3.2. For classification problems, we have $Y \in \{0, 1\}$, $f : \Omega \to \{0, 1\}$, and thus

$$R(f) = \mathbb{P}\{f(X) \neq Y\}.$$

(Check!) So the risk is just the probability of misclassification (such as misdiagnosis for a patient).

How much data do we need to learn, i.e how large sample size n needs to be? This of course depends on the *complexity* of the problem. If the dependence of Y on X is very intricate then we need more data, otherwise, less. Usually we do not know the complexity a priori. So we may restrict the complexity of the candidate functions f, insisting that f belong to some given class of functions $\mathcal{F}$.

Definition 8.3.3. *The function f^* that minimizes the risk a given function class $\mathcal{F}$, i.e.*

$$f^* := \arg\min_{f \in \mathcal{F}} R(f),$$

is called the target function*.*

How do we choose the function class $\mathcal{F}$ for a given learning problem? Although there is no general rule, the choice of $\mathcal{F}$ should be based on the *trade-off between fit and complexity*. Suppose we choose $\mathcal{F}$ to be too small; for example, we want the interface between healthy ($f(x) = 0$) and sick diagnoses ($f(x) = 1$) to be a line, like in Figure 8.5b. Although we will can learn such a simple function f with less data, we probably oversimplified the problem and. The linear function do not capture the essential trends in this data, which will reflect in a big risk $R(f)$.

Suppose, on the opposite, that we choose $\mathcal{F}$ to be too large. This will result in *overfitting* where we are trying to fit f to noise like in Figure 8.5c. Plus we will need a lot of data to learn such complicated functions.

A good choice of $\mathcal{F}$ is one that avoids either underfitting or overfitting, just capturing the essential trends in the data just like in Figure 8.5a.

Empirical risk

The target function f^* would give the best solution to the learning problem in the class $\mathcal{F}$, since it minimizes the risk

$$R(f) = \mathbb{E}(f(X) - Y)^2.$$

over all functions in $\mathcal{F}$. Unfortunately, we can not compute $R(f)$ from data (we are not able to take population expectation $\mathbb{E}$), so f^* is not computable either. Instead, we can try to estimate $R(f)$ and thus f from the data.

Definition 8.3.4. *The* empirical risk *for a function $f : \Omega \to \mathbb{R}$ is defined as*

$$R_n(f) := \frac{1}{n} \sum_{i=1}^{n} \left(f(X_i) - Y_i\right)^2.$$

The function f_n^* that minimizes the risk a given function class $\mathcal{F}$, i.e.

$$f_n^* := \arg\min_{f\in\mathcal{F}} R_n(f),$$

is called the empirical target function.

The empirical risk, and thus the empirical target function f_n^*, can be computed from the data. The outcome of learning from the data is thus f_n^*.

The main question: how much data do we need for learning? More precisely, how large is the *excess risk*

$$R(f_n^*) - R(f^*)$$

produced by our having to learn from a finite sample of size n?

Suppose we can approximate the true risk by the empirical risk uniformly in the class $\mathcal{F}$, so we have a bound like

$$\mathbb{E}\sup_{f\in\mathcal{F}} |R_n(f) - R(f)| \le \varepsilon_n \tag{8.25}$$

with ε_n small. Then

$$\begin{aligned} R(f_n^*) &\le R_n(f_n^*) + \varepsilon_n \quad \text{(by (8.25))} \\ &\le R_n(f^*) + \varepsilon_n \quad \text{(since } f_n^* \text{ minimizes } R_n\text{)} \\ &\le R(f^*) + 2\varepsilon_n \quad \text{(again by (8.25))}. \end{aligned}$$

Thus the excess risk is bounded by $2\varepsilon_n$:

$$R(f_n^*) - R(f^*) \le 2\varepsilon_n.$$

We reduced our problem to establishing the uniform deviation inequality (8.25). Let us show how to prove it for a specific example where $\mathcal{F}$ consists of Lipschitz functions:

$$\mathcal{F} := \{f : [0,1] \to \mathbb{R},\ \|f\|_{\text{Lip}} \le L\}. \tag{8.26}$$

For higher dimensions, $[0,1]^d$, see a marginal note for Exercise 8.2.6. I guess $d = 2$ could be done easily there (Dudley's integral barely diverges so can use Exercise 8.1.12), and borrowed here.

Theorem 8.3.5. *Assume that the label Y is in $[0,1]$ almost surely. Then*

$$\mathbb{E}\sup_{f\in\mathcal{F}} |R_n(f) - R(f)| \le \frac{C(L+1)}{\sqrt{n}}.$$

Consequently, the excess risk is

$$R(f_n^*) - R(f^*) \le \frac{C(L+1)}{\sqrt{n}}.$$

In words, the excess risk of learning from data of size n decays like $O(1/\sqrt{n})$.

The proof of Theorem 8.3.5 is similar to the proof of the Uniform Law of Large Numbers, Theorem 8.2.1. It is follows from Dudley's integral inequality for the random process

$$X_f := R_n(f) - R(f). \tag{8.27}$$

One just need to check that the process has sub-gaussian increments, and we will do this in the next exercise.

Exercise 8.3.6. *1. Show that the random process $(X_f)_{f\in\mathcal{F}}$ defined in (8.27) for the Lipschitz functions class $\mathcal{F}$ in (8.26) has sub-gaussian increments:*

$$\|X_f - X_g\|_{\psi_2} \le \frac{CL}{\sqrt{n}}\|f-g\|_\infty \quad \text{for all } f, g \in \mathcal{F}.$$

2. Deduce Theorem 8.3.5.

8.4 Generic chaining

Dudley's and Sudakov's inequalities are simple and useful tools for bounding random processes. These bounds are within $O(\log n)$ from being optimal, as we may recall from Theorem 8.1.11. Unfortunately, it is not possible to close the gap between Dudley's and Sudakov's inequality. There are examples where one inequality gives a sharp bound and the other does not. References? The issue is even more general. The entropy numbers $\mathcal{N}(T, d, \varepsilon)$ do not contain enough information to control the magnitude of $\mathbb{E}\sup_{t\in T} X_t$ up to an absolute constant factor.

8.4.1 A makeover of Dudley's inequality

Fortunately, there is a way to obtain accurate, two-sided bounds on $\mathbb{E}\sup_{t\in T} X_t$ for sub-gaussian processes $(X_t)_{t\in T}$ in terms of the geometry of T. This method is called *generic chaining*, and it is essentially a sharpening of the chaining method we developed in the proof of Dudley's inequality (Theorem 8.1.3). Recall that the outcome of chaining was the bound (8.12):

$$\mathbb{E}\sup_{t\in T} X_t \lesssim \sum_{k=\kappa+1}^{\infty} \varepsilon_{k-1}\sqrt{\log|T_k|}. \tag{8.28}$$

Here ε_k are decreasing positive numbers and T_k are ε_k-nets of T such that $|T_\kappa| = 1$. To be specific, in the proof of Theorem 8.1.3 we chose

$$\varepsilon_k = 2^{-k} \quad \text{and} \quad |T_k| = \mathcal{N}(T, d, \varepsilon_k),$$

so $T_k \subset T$ were the smallest ε_k-nets of T.

In preparation for generic chaining, let us now turn around our choice of ε_k and T_k. Instead of fixing ε_k and operating the smallest possible cardinality of T_k, let us fix the cardinality of T_k and operate with the largest possible ε_k. Namely, fix some subsets $T_k \subset T$ such that

$$|T_0| = 1, \quad |T_k| \le 2^{2^k}, \quad k = 1, 2, \dots \tag{8.29}$$

Do we need to define the distance from a point to a set?

Such sequence of sets $(T_k)_{k=0}^\infty$ is called an *admissible sequence*. Put

$$\varepsilon_k = \sup_{t \in T} d(t, T_k).$$

Then each T_k is an ε_k-net of T. With this choice of ε_k and T_k, the chaining bound (8.28) becomes

$$\mathbb{E} \sup_{t \in T} X_t \lesssim \sum_{k=1}^\infty 2^{k/2} \sup_{t \in T} d(t, T_{k-1}).$$

After re-indexing, we conclude

$$\mathbb{E} \sup_{t \in T} X_t \lesssim \sum_{k=0}^\infty 2^{k/2} \sup_{t \in T} d(t, T_k). \tag{8.30}$$

8.4.2 Talagrand's γ_2 functional and generic chaining

So far, nothing really happened. The bound (8.30) is just an equivalent way to state Dudley's inequality. The important step will come now. The generic chaining will allow us to pull the supremum *outside* the sum in (8.30). The resulting important quantity has a name:

Definition 8.4.1 (Talagrand's γ_2 functional)**.** *Let (T, d) be a metric space. A sequence of subsets $(T_k)_{k=0}^\infty$ is called an* admissible sequence *if the cardinalities of T_k satisfy (8.29). The* γ_2 functional *of T is defined as*

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

where the infimum is with respect to all admissible sequences.

Since the supremum in the γ_2 functional is outside the sum, it is smaller than the Dudley's sum in (8.30). Still, $\gamma_2(T, d)$ controls the magnitude of the random process. Let us state and prove this sharpening of Dudley's inequality.

Theorem 8.4.2 (Generic chaining bound)**.** *Let $(X_t)_{t\in T}$ be a sub-gaussian process on a metric space (T, d) in the sense of (8.1). Then*

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

Proof. We will proceed with the same chaining method that we introduced in the proof of Dudley's inequality Theorem 8.1.3, but we will do chaining more accurately.

Step 1: Chaining set-up. As before, we may assume that $K = 1$ and that T is finite. Let (T_k) be an admissible sequence of subsets of T, and denote $T_0 = \{t_0\}$. We will walk from t_0 to a general point $t \in T$ along the chain

$$t_0 = \pi_0(t) \to \pi_1(t) \to \cdots \to \pi_K(t) = t$$

of points $\pi_k(t) \in T_k$ that are chosen as best approximations to t in T_k, i.e.

$$d(t, \pi_k(t)) = d(t, T_k).$$

The displacement $X_t - X_{t_0}$ can be expressed as a telescoping sum similar to (8.9):

$$X_t - X_{t_0} = \sum_{k=1}^{K} (X_{\pi_k(t)} - X_{\pi_{k-1}(t)}). \tag{8.31}$$

Step 2: Controlling the increments. This is where we need to be more accurate than in Dudley's inequality. We would like to have a uniform bound on the increments, a bound that would state with high probability that

$$\left|X_{\pi_k(t)} - X_{\pi_{k-1}(t)}\right| \le 2^{k/2} d(t, T_k) \quad \forall k \in \mathcal{N}, \quad \forall t \in T. \tag{8.32}$$

Summing these inequalities over all k would lead to a desired bound in terms of $\gamma_2(T, d)$.

To prove (8.32), let us fix k and t first. The sub-gaussian assumption tells us that

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

This means that for every $u \ge 0$, the event

$$\left|X_{\pi_k(t)} - X_{\pi_{k-1}(t)}\right| \le Cu2^{k/2} d(\pi_k(t), \pi_{k-1}(t)) \tag{8.33}$$

holds with probability at least

$$1 - 2\exp(-8u^2 2^k).$$

(To get the constant 8, choose the absolute constant C large enough.) □

We can now unfix $t \in T$ by taking a union bound over

$$|T_k| \cdot |T_{k-1}| \le |T_k|^2 = 2^{2^{k+1}}$$

possible pairs $(\pi_k(t), \pi_{k-1}(t))$. Similarly, we can unfix k by a union bound over all $k \in \mathbb{N}$. Then the probability that the bound (8.33) holds simultaneously for all $t \in T$ and $k \in \mathbb{N}$ is at least

$$1 - \sum_{k=1}^{\infty} 2^{2^{k+1}} \cdot 2\exp(-8u^2 2^k) \ge 1 - 2\exp(-u^2).$$

(Check the last inequality!)

Step 3: Summing up the increments. In the event that the bound (8.33) does holds for all $t \in T$ and $k \in \mathbb{N}$, we can sum up the inequalities over $k \in \mathcal{N}$ and plug the result into the chaining sum (8.31). This yields

$$|X_t - X_{t_0}| \le Cu \sum_{k=1}^{\infty} 2^{k/2} d(\pi_k(t), \pi_{k-1}(t)). \tag{8.34}$$

By triangle inequality, we have

$$d(\pi_k(t), \pi_{k-1}(t)) \le d(t, \pi_k(t)) + d(t, \pi_{k-1}(t)).$$

Using this bound and doing re-indexing, we find that the right hand side of (8.34) can be bounded by $\gamma_2(T, d)$, that is

$$|X_t - X_{t_0}| \le C_1 u \gamma_2(T, d).$$

(Check!) Taking the supremum over T yields

$$\sup_{t \in T} |X_t - X_{t_0}| \le C_2 u \gamma_2(T, d).$$

Recall that this inequality holds with probability at least $1 - 2\exp(-u^2)$. This means that the magnitude in question is a sub-gaussian random variable:

$$\Big\| \sup_{t \in T} |X_t - X_{t_0}| \Big\|_{\psi_2} \le C_3 \gamma_2(T, d).$$

This quickly implies the conclusion of Theorem 8.4.2. (Check!) □

Remark 8.4.3 (Supremum of increments). A quick glance at the end of the proof of Theorem 8.4.2 reveals that the generic chaining method actually yields the bound

$$\mathbb{E} \sup_{t \in T} |X_t - X_{t_0}| \le CK\gamma_2(T, d)$$

for any fixed $t_0 \in T$. Combining it with a similar bound for $X_s - X_{t_0}$ and using triangle inequality, we deduce that

$$\mathbb{E} \sup_{t,s \in T} |X_t - X_s| \le CK\gamma_2(T, d).$$

Note that in either of these two bounds, we need not require the mean zero assumption $\mathbb{E} X_t = 0$. It is required, however, in Theorem 8.4.2. (Why?)

The argument above gives not only a bound on expectation but also a tail bound for $\sup_{t \in T} X_t$. Let us now give a better tail bound, similar to the one we had in Theorem 8.1.5 for Dudley's inequality.

Theorem 8.4.4 (Generic chaining: tail bound). *Let $(X_t)_{t \in T}$ be a subgaussian process on a metric space (T, d) in the sense of (8.1). Then, for every $u \ge 0$, the event*

$$\mathbb{E} \sup_{t \in T} X_t \le CK \Big[\gamma_2(T, d) + u \cdot \operatorname{diam}(T) \Big]$$

holds with probability at least $1 - 2\exp(-u^2)$.

Exercise 8.4.5. *Prove Theorem 8.4.4. Hint: State and use a variant of the increment bound (8.33) with $u + 2^k$ instead of $u2^{k/2}$. In the end of the argument, you will need a bound on the sum of steps $\sum_{k=1}^{\infty} d(\pi_k(t), \pi_{k-1}(t))$. For this, modify the chain $\{\pi_k(t)\}$ by doing a "lazy walk" on it. Stay at the current point $\pi_k(t)$ for a few steps (say, $q - 1$) until the distance to t improves by a factor of 2, that is until*

$$d(t, \pi_{k+q}(t)) \le \frac{1}{2} d(t, \pi_k(t)),$$

then jump to $\pi_{k+q}(t)$. This will make the sum of the steps geometrically convergent.

Exercise 8.4.6 (Dudley's integral vs. γ_2 functional). *Show that γ_2 functional is bounded by Dudley's integral. Namely, show that for any metric space (T, d), one has*

$$\gamma_2(T, d) \le C \int_0^\infty \sqrt{\log \mathcal{N}(T, d, \varepsilon)}\, d\varepsilon.$$

8.5 Talagrand's majorizing measure and comparison theorems

Talagrand's γ_2 functional introduced in Definition 8.4.1 has some advantages and disadvantages over Dudley's integral. A disadvantage is that $\gamma_2(T, d)$ is usually harder to compute than the metric entropy that defines Dudley's integral. Indeed, it could take a real effort to construct a good admissible sequence of sets. However, unlike Dudley's integral, the γ_2 functional gives a bound on Gaussian processes that is *optimal* up to an absolute constant. This is the content of the following theorem.

Theorem 8.5.1 (Talagrand's majorizing measure theorem)**.** *Let $(X_t)_{t\in T}$ be a Gaussian process on a set T. Consider the canonical metric defined on T by (7.14), i.e. $d(t, s) = \|X_t - X_s\|_2$. Then*

$$c \cdot \gamma_2(T, d) \le \mathbb{E} \sup_{t\in T} X_t \le C \cdot \gamma_2(T, d).$$

The upper bound in Theorem 8.5.1 follows directly from generic chaining (Theorem 8.4.2). The lower bound is notably harder to obtain. Its proof, which we will not present in this book, can be thought of as a far reaching, multi-scale strengthening of Sudakov's inequality (Theorem 7.5.1).

Note that the upper bound, as we know from Theorem 8.4.2, holds for any *sub-gaussian* process. Therefore, by combining the upper and lower bounds together, we can deduce that any sub-gaussian process is bounded (via γ_2 functional) by a Gaussian process. Let us state this important comparison result.

Corollary 8.5.2 (Talagrand's comparison inequality)**.** *Let $(X_t)_{t\in T}$ be a random process on a set T and let $(Y_t)_{t\in T}$ be a Gaussian process. Assume that for all $t, s \in T$, we have*

$$\|X_t - X_s\|_{\psi_2} \le K\|Y_t - Y_s\|_2.$$

Then

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

Proof. Consider the canonical metric on T given by $d(t, s) = \|Y_t - Y_s\|_2$. Apply the generic chaining bound (Theorem 8.4.2) followed by the lower bound in the majorizing measure Theorem 8.5.1. Thus we get

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

The proof is complete. □

Corollary 8.5.2 extends Sudakov-Fernique's inequality (Theorem 7.3.10) for sub-gaussian processes. All we pay for such extension is an absolute constant factor.

Let us apply Corollary 8.5.2 for a canonical Gaussian process

$$Y_x = \langle g, x \rangle, \quad x \in T$$

defined on a subset $T \subset \mathbb{R}^n$. Recall from Section 7.7 that the magnitude of this process,

$$w(T) = \mathbb{E} \sup_{x \in T} \langle g, x \rangle$$

is the *Gaussian width* of T. We immediately obtain the following corollary.

Corollary 8.5.3 (Talagrand's comparison inequality: geometric form)**.** *Let $(X_x)_{x \in T}$ be a random process on a subset $T \subset \mathbb{R}^n$. Assume that for all $x, y \in T$, we have*

$$\|X_x - X_y\|_{\psi_2} \le K \|x - y\|_2.$$

Then

$$\mathbb{E} \sup_{x \in T} X_x \le CKw(T).$$

Exercise 8.5.4 (Two-sided bounds)**.** *Show that, in the context of Corollary 8.5.3 (and even without the mean zero assumption) we have*

$$\mathbb{E} \sup_{x \in T} |X_x| \le CK\gamma(T).$$

State explicitly the mean zero assumption in all results where it is required for the process

Recall that $\gamma(T)$ is the Gaussian complexity of T; we introduced this cousin of Gaussian width in (7.27). Hint: Fix $x_0 \in T$ and break the process into two parts: $|X_x| \le |X_x - X_{x_0}| + |X_{x_0}|$. Use Remark 8.4.3 to control the first part and the sub-gaussian condition with $y = 0$ for the second part. Use Exercise 7.7.14 to pass from Gaussian width to Gaussian complexity.

Exercise 8.5.5 (Tail bound)**.** *Show that, in the setting of Corollary 8.5.3, for every $u \ge 0$ we have*

$$\sup_{x \in T} X_x \le CK(w(T) + u \cdot \operatorname{diam}(T))$$

with probability at least $1 - 2\exp(-u^2)$. Hint: Use Theorem 8.4.4.

8.6 Chevet's inequality

Talagrand's comparison inequality (Corollary 8.5.2) has several important consequences. We will cover one application now, others will appear in Chapter **??**.

The theorem we are about to state gives a uniform bound for random quadratic forms, that is a bound on

$$\sup_{x \in T,\, y \in S} \langle Ax, y \rangle \tag{8.35}$$

where A is a random matrix and T and S are general sets.

We already encountered problems of this type when we analyzed the norms of random matrices, namely in the proofs of Theorems 4.3.5 and 7.4.1. In those situations, T and S were Euclidean balls. This time, we will let T and S be arbitrary geometric sets. Our bound on (6.2) will depend on just two geometric parameters of T and S: the *Gaussian width* and the *radius*, defined as

$$\operatorname{rad}(T) := \sup_{x \in T} \|x\|_2. \tag{8.36}$$

Theorem 8.6.1 (Sub-gaussian Chevet's inequality)**.** *Let A be an $m \times n$ random matrix whose entries A_{ij} are independent, mean zero, sub-gaussian random variables. Let $T \subset \mathbb{R}^n$ and $S \subset \mathbb{R}^m$ be arbitrary bounded sets. Then*

$$\mathbb{E} \sup_{x \in T,\, y \in S} \langle Ax, y \rangle \le CK \left[w(T) \operatorname{rad}(S) + w(S) \operatorname{rad}(T) \right]$$

where $K = \max_{ij} \|A_{ij}\|_{\psi_2}$.

Before we prove this theorem, let us make one simple illustration of its use. Setting $T = S^{n-1}$ and $S = S^{m-1}$, we recover a bound on the operator norm of A,

$$\mathbb{E} \|A\| \le CK(\sqrt{n} + \sqrt{m}),$$

which we obtained in Section 4.3.2 using a different method.

Proof of Theorem 8.6.1. We will use the same method as in our proof of the sharp bound on Gaussian random matrices (Theorem 7.4.1). That proof was based on Sudakov-Fernique comparison inequality; this time, we will use the more general Talagrand's comparison inequality.

Without loss of generality, $K = 1$. We would like to bound the random process

$$X_{uv} := \langle Au, v \rangle, \quad u \in T,\, v \in S.$$

Let us first show that this process has sub-gaussian increments. For any $(u, v), (w, z) \in T \times S$, we have

$$\begin{aligned}
\|X_{uv} - X_{wz}\|_{\psi_2} &= \Big\| \sum_{i,j} A_{ij}(u_i v_j - w_i z_j) \Big\|_{\psi_2} \\
&\le \Big(\sum_{i,j} \|A_{ij}(u_i v_j - w_i z_j)\|_{\psi_2}^2 \Big)^{1/2} \quad \text{(by Proposition 2.6.1)} \\
&\le \Big(\sum_{i,j} \|u_i v_j - w_i z_j\|_2^2 \Big)^{1/2} \quad \text{(since } \|A_{ij}\|_{\psi_2} \le K = 1\text{)} \\
&= \|uv^{\mathsf{T}} - wz^{\mathsf{T}}\|_F \\
&= \|(uv^{\mathsf{T}} - wv^{\mathsf{T}}) + (wv^{\mathsf{T}} - wz^{\mathsf{T}})\|_F \quad \text{(adding, subtracting)} \\
&\le \|(u - w)v^{\mathsf{T}}\|_F + \|w(v - z)^{\mathsf{T}}\|_F \quad \text{(by triangle inequality)} \\
&= \|u - w\|\|v\|_2 + \|v - z\|_2\|w\|_2 \\
&\le \|u - w\|_2 \operatorname{rad}(S) + \|v - z\|_2 \operatorname{rad}(T).
\end{aligned}$$

To apply Talagrand's comparison inequality, we need to choose a Gaussian process (Y_{uv}) to compare the process (X_{uv}) to. The outcome of our calculation of the increments of (X_{uv}) suggests the following definition for (Y_{uv}):

$$Y_{uv} := \langle g, u \rangle \operatorname{rad}(S) + \langle h, v \rangle \operatorname{rad}(T),$$

where

$$g \sim N(0, I_n), \quad h \sim N(0, I_m)$$

are independent Gaussian vectors. The increments of this process are

$$\|Y_{uv} - Y_{wz}\|_2^2 = \|u - w\|_2^2 \operatorname{rad}(T)^2 + \|v - z\|_2^2 \operatorname{rad}(S)^2.$$

(Check this as in the proof of Theorem 7.4.1.)

Comparing the increments of the two processes, we find that

$$\|X_{uv} - X_{wz}\|_{\psi_2} \lesssim \|Y_{uv} - Y_{wz}\|_2.$$

(Check!) Applying Talagrand's comparison inequality (Corollary 8.5.3), we conclude that

$$\begin{aligned}
\mathbb{E} \sup_{u \in T,\, v \in S} X_{uv} &\lesssim \mathbb{E} \sup_{u \in T,\, v \in S} Y_{uv} \\
&= \mathbb{E} \sup_{u \in T} \langle g, u \rangle \operatorname{rad}(S) + \mathbb{E} \sup_{v \in S} \langle h, v \rangle \operatorname{rad}(T) \\
&= w(T) \operatorname{rad}(S) + w(S) \operatorname{rad}(T),
\end{aligned}$$

as claimed. □

Chevet's inequality is optimal, up to an absolute constant factor:

Exercise 8.6.2 (Sharpness of Chevet's inequality)**.** *Let A be an $m \times n$ random matrix whose entries A_{ij} are independent $N(0,1)$ random variables. Let $T \subset \mathbb{R}^n$ and $S \subset \mathbb{R}^m$ be arbitrary bounded sets. Show that the reverse of Chevet's inequality holds:*

$$\mathbb{E} \sup_{x \in T,\, y \in S} \langle Ax, y \rangle \geq c \left[w(T) \operatorname{rad}(S) + w(S) \operatorname{rad}(T) \right].$$

Hint: Note that $\mathbb{E} \sup_{x \in T,\, y \in S} \langle Ax, y \rangle \geq \sup_{x \in T} \mathbb{E} \sup_{y \in S} \langle Ax, y \rangle$.

Chevet's inequality can be useful for computing *general* operator norms of random matrices A. The case we have considered so far is where a random matrix $A : \mathbb{R}^n \to \mathbb{R}^m$ acts as a linear operator between the spaces equipped with the Euclidean norm $\|\cdot\|_2$. The norm of A in this case is the classical operator norm; recall Section 4.1.2. Chevet's inequality allows to compute the norm of A even if the spaces $\mathbb{R}^n$ and $\mathbb{R}^m$ are equipped with arbitrary, possibly non-Eucdliean, norms. Let us illustrate this on the example of ℓ_p norms.

definition of general operator norm?

Exercise 8.6.3 ($\ell_p \to \ell_q$ norm of a random matrix)**.** *Let A be an $m \times n$ random matrix whose entries A_{ij} are independent, mean zero, sub-gaussian random variables.*

Check: if p or q is ≥ 2 then the radius is not bounded by 2; does this affect the inequlity?

(a) Show that for every $1 < p \leq \infty$ and $1 \leq q \leq \infty$, one has

$$\mathbb{E} \|A : \ell_p^n \to \ell_q^n\| \leq C_{p,q}(n^{1/p'} + m^{1/q}). \tag{8.37}$$

Here p' denotes the conjugate exponent *for p, which is defined by the equation $\frac{1}{p} + \frac{1}{p'} = 1$, and with the convention that $\frac{1}{\infty} = 0$.*

(b) How does the bound modify if $p = 1$?

(c) Show that the bound (8.37) can be reversed.

Hint: Express

$$\|A : \ell_p^n \to \ell_q^n\| = \sup_{x \in B_p^n,\, y \in B_{q'}^m} \langle Ax, y \rangle$$

and use Exercise 7.7.5.

Exercise 8.6.4 (High probability version of Chevet)**.** *Under the assumptions of Theorem 8.6.1, prove a tail bound for $\sup_{x \in T,\, y \in S} \langle Ax, y \rangle$.*

Hint: Use the result of Exercise 8.5.5.

Exercise 8.6.5 (Gaussian Chevet's inequality)**.** *Suppose the entries of A are $N(0,1)$. Show that Theorem 8.6.1 holds with sharp constant 1, that is*

$$\mathbb{E} \sup_{x \in T,\, y \in S} \langle Ax, y \rangle \leq w(T) \operatorname{rad}(S) + w(S) \operatorname{rad}(T).$$

Hint: Use Sudakov-Fernique's inequality (Theorem 7.3.10) instead of Talagrand's comparison inequality.

Check that this is doable

8.7 Vapnik-Chervonenkis dimension

Write!

Chapter 9

Deviations of random matrices and geometric consequences

This chapter is devoted to a remarkably useful consequence of Talagrand's comparison inequality: a general uniform deviation inequality for random matrices. Let A be an $m \times n$ matrix whose rows A_i are independent, isotropic and sub-gaussian random vectors in $\mathbb{R}^n$. We will essentially show that for any subset $T \subset \mathbb{R}^n$, we have

$$\mathbb{E} \sup_{x \in T} \Big| \|Ax\|_2 - \mathbb{E}\, \|Ax\|_2 \Big| \lesssim w(T). \tag{9.1}$$

We will prove such inequality in Section 9.2 and garner its many applications in Sections **??**.

Let us pause to check the magnitude of $\mathbb{E}\, \|Ax\|_2$ that appears in the inequality (9.1). We should expect that

Extend that exercise to include an approximation like this one.

$$\begin{aligned} \mathbb{E}\, \|Ax\|_2 &\approx (\mathbb{E}\, \|Ax\|_2^2)^{1/2} \quad \text{(by concentration, see Exercise 5.2.5)} \\ &= \Big(\mathbb{E} \sum_{i=1}^m \langle A_i, x \rangle^2 \Big)^{1/2} \\ &= \sqrt{m} \|x\|_2 \quad \text{(by linearity of expectation and isotropy).} \end{aligned} \tag{9.2}$$

In this light, we may expect that the following more informative version of (9.1) should also be true:

$$\mathbb{E} \sup_{x \in T} \Big| \|Ax\|_2 - \sqrt{m} \|x\|_2 \Big| \lesssim w(T). \tag{9.3}$$

We will indeed deduce (9.3) from Talagrand's comparison inequality (Corollary 8.5.3). To apply the comparison inequality, all we need to check that the random process

$$X_x := \|Ax\|_2 - \sqrt{m}\|x\|_2$$

has sub-gaussian increments. We will do this next.

9.1 Sub-gaussian increments of the random matrix process

Theorem 9.1.1 (Sub-gaussian increments)**.** *Let A be an $m \times n$ matrix whose rows A_i are independent, isotropic and sub-gaussian random vectors in $\mathbb{R}^n$. Then the random process*

$$X_x := \|Ax\|_2 - \sqrt{m}\|x\|_2$$

has sub-gaussian increments, namely

$$\|X_x - X_y\|_{\psi_2} \le CK^2\|x - y\|_2 \quad \text{for all } x, y \in \mathbb{R}^n. \tag{9.4}$$

Here $K = \max_i \|A_i\|_{\psi_2}$.

Remark 9.1.2 (Centered process). Applying Centering Lemma 2.6.6 we see that the random process in (9.1), i.e.

$$X'_x := \|Ax\|_2 - \mathbb{E}\,\|Ax\|_2$$

has sub-gaussian increments, too. The conclusion of Theorem 9.1.1 holds for this process, possibly with a different absolute constant.

We are now going to prove Theorem 9.1.1. Although this proof is a bit longer than most of the arguments in this book, we will make it easier by working out simpler, partial cases first and gradually moving toward full generality. We will develop this argument in the next few subsections.

9.1.1 Proof for unit vector x and zero vector y

Assume that

$$\|x\|_2 = 1 \quad \text{and} \quad y = 0.$$

In this case, the inequality in (9.4) we want to prove becomes

$$\Big\| \|Ax\|_2 - \sqrt{m} \Big\|_{\psi_2} \le CK^2. \tag{9.5}$$

Note that Ax is a random vector in $\mathbb{R}^m$ with independent, sub-gaussian coordinates $\langle A_i, x\rangle$, which satisfy $\mathbb{E}\langle A_i, x\rangle^2 = 1$ by isotropy. Then the Concentration of Norm Theorem 3.1.1 yields (9.5).

9.1.2 Proof for unit vectors x, y and for the squared process

Assume now that

$$\|x\|_2 = \|y\|_2 = 1.$$

In this case, the inequality in (9.4) we want to prove becomes

$$\Big\| \|Ax\|_2 - \|Ay\|_2 \Big\|_{\psi_2} \le CK^2 \|x - y\|_2. \tag{9.6}$$

We will first prove a version of this inequality for the *squared* Euclidean norms, which are more convenient to handle. Let us guess what form such inequality should take. We have

$$\begin{aligned} \|Ax\|_2^2 - \|Ay\|_2^2 &= \Big(\|Ax\|_2 + \|Ay\|_2\Big) \cdot \Big(\|Ax\|_2 - \|Ay\|_2\Big) \\ &\lesssim \sqrt{m} \cdot \|x - y\|_2. \end{aligned} \tag{9.7}$$

The last bound should hold with high probability because the typical magnitude of $\|Ax\|_2$ and $\|Ay\|_2$ is $\sqrt{m}$ by (9.2), and since we expect (9.6) to hold.

Now that we guessed the inequality (9.7) for the squared process, let us prove it. We are looking to bound the random variable

$$Z := \frac{\|Ax\|_2^2 - \|Ay\|_2^2}{\|x - y\|_2} = \frac{\langle A(x-y), A(x+y)\rangle}{\|x - y\|_2} = \langle Au, Av\rangle \tag{9.8}$$

where

$$u := \frac{x - y}{\|x - y\|_2} \quad \text{and} \quad v := x + y.$$

The desired bound is

$$|Z| \lesssim \sqrt{m} \quad \text{with high probability.}$$

The coordinates of the vectors Au and Av are $\langle A_i, u\rangle$ and $\langle A_i, v\rangle$. So we can represent Z as a sum of independent random variables

$$Z = \sum_{i=1}^m \langle A_i, u\rangle \langle A_i, v\rangle,$$

Lemma 9.1.3. *The random variables $\langle A_i, u\rangle \langle A_i, v\rangle$ are independent, mean zero, and sub-exponential; more precisely,*

$$\big\| \langle A_i, u\rangle \langle A_i, v\rangle \big\|_{\psi_1} \le 2K^2.$$

Proof. Independence follows from the construction, but the mean zero property is less obvious. Although both $\langle A_i, u\rangle$ and $\langle A_i, v\rangle$ do have zero means, these variables are not necessarily independent from each other. Still, we can check that they are uncorrelated. Indeed,

$$\mathbb{E}\, \langle A_i, x-y\rangle \langle A_i, x+y\rangle = \mathbb{E}\left[\langle A_i, x\rangle^2 - \langle A_i, y\rangle^2\right] = 1 - 1 = 0$$

by isotropy. By definition of u and v, this implies that $\mathbb{E}\, \langle A_i, u\rangle \langle A_i, v\rangle = 0$.

To finish the proof, recall from Lemma 2.7.5 that the product of two sub-gaussian random variables is sub-exponential. So we get

$$\begin{aligned} \big\| \langle A_i, u\rangle \langle A_i, v\rangle \big\|_{\psi_1} &\le \big\| \langle A_i, u\rangle \big\|_{\psi_2} \cdot \big\| \langle A_i, v\rangle \big\|_{\psi_2} \\ &\le K\|u\|_2 \cdot K\|v\|_2 \quad \text{(by sub-gaussian assumption)} \\ &\le 2K^2 \end{aligned}$$

where in the last step we used that $\|u\|_2 = 1$ and $\|v\|_2 \le \|x\|_2 + \|y\|_2 \le 2$. □

To bound Z, we will use Bernstein's inequality (Corollary 2.8.4); recall that it applies for a sum of independent, mean zero, sub-exponential random variables.

Exercise 9.1.4. *Apply Bernstein's inequality (Corollary 2.8.4) and simplify the bound. You should get*

$$\mathbb{P}\left\{|Z| \ge s\sqrt{m}\right\} \le 2\exp\Big(-\frac{cs^2}{K^4}\Big)$$

for any $0 \le s \le \sqrt{m}$. Hint: In this range of s, the sub-gaussian tail will dominate in Bernstein's inequality. Do not forget to apply the inequality for $2K^2$ instead of K because of Lemma 9.1.3.

Recalling the definition of Z, we can see that we obtained the desired bound (9.7).

We wish this bound was true for all s, but for large s it may not hold. The squared process, being sub-exponential, can not have a strictly lighter, sub-gaussian tail everywhere.

9.1.3 Proof for unit vectors x, y and for the original process

Now we would like to remove the squares from $\|Ax\|_2^2$ and $\|Ay\|_2^2$ and deduce inequality (9.6) for unit vectors x and y. Let us state this goal again.

Lemma 9.1.5 (Unit y, original process)**.** *Let $x, y \in S^{n-1}$. Then*

$$\Big\| \|Ax\|_2 - \|Ay\|_2 \Big\|_{\psi_2} \le CK^2 \|x - y\|_2.$$

Proof. Fix $s \ge 0$. The conclusion we want to prove is that

$$p(s) := \mathbb{P}\left\{ \frac{\big| \|Ax\|_2 - \|Ay\|_2 \big|}{\|x - y\|_2} \ge s \right\} \le 4 \exp\Big(- \frac{cs^2}{K^4} \Big). \tag{9.9}$$

We will proceed differently for small and large s.

Case 1: $s \le 2\sqrt{m}$**.** In this range, we will use our results from the previous subsection. They are stated for the squared process though. So, to be able to apply those results, we multiply both sides of the inequality defining $p(s)$ by $\|Ax\|_2 + \|Ay\|_2$. With the same Z as we defined in (9.8), this gives

$$p(s) = \mathbb{P}\left\{ |Z| \ge s\big(\|Ax\|_2 + \|Ay\|_2 \big) \right\} \le \mathbb{P}\left\{ |Z| \ge s \|Ax\|_2 \right\}.$$

As we noted in (9.2), typically we have $\|Ax\|_2 \approx \sqrt{m}$. So it makes sense to break the probability that $|Z| \ge s\|Ax\|_2$ into two cases: one where $\|Ax\| \ge \sqrt{m}/2$ and thus $|Z| \ge s\sqrt{m}/2$, and the other where $\|Ax\| < \sqrt{m}/2$ (and there we will not care about Z). This leads to

$$p(s) \le \mathbb{P}\left\{ |Z| \ge \frac{s\sqrt{m}}{2} \right\} + \mathbb{P}\left\{ \|Ax\|_2 < \frac{\sqrt{m}}{2} \right\} =: p_1(s) + p_2(s).$$

The result of Exercise 9.1.4 gives

$$p_1(s) \le 2 \exp\Big(- \frac{cs^2}{K^4} \Big).$$

Further, the bound (9.5) and triangle inequality gives

$$p_2(s) \le \mathbb{P}\left\{ \big| \|Ax\|_2 - \sqrt{m} \big| > \frac{\sqrt{m}}{2} \right\} \le 2 \exp\Big(- \frac{cs^2}{K^4} \Big).$$

Summing the two probabilities, we conclude a desired bound

$$p(s) \le 4 \exp\Big(- \frac{cs^2}{K^4} \Big).$$

□

Case 2: $s > 2\sqrt{m}$**.** Let us look again at the inequality (9.9) that defines $p(s)$, and slightly simplify it. By triangle inequality, we have

$$\big|\|Ax\|_2 - \|Ay\|_2\big| \le \|A(x-y)\|_2.$$

Thus

$$\begin{aligned} p(s) &\le \mathbb{P}\left\{\|Au\|_2 \ge s\right\} \quad (\text{where } u := \frac{x-y}{\|x-y\|_2} \text{ as before}) \\ &\le \mathbb{P}\left\{\|Au\|_2 - \sqrt{m} \ge s/2\right\} \quad (\text{since } s > 2\sqrt{m}) \\ &\le 2\exp\Big(-\frac{cs^2}{K^4}\Big) \quad (\text{using (9.5) again}). \end{aligned}$$

Therefore, in both cases we obtained the desired estimate (9.9). This completes the proof of the lemma.

9.1.4 Proof of Theorem 9.1.1 in full generality

Finally, we are ready to prove (9.4) for arbitrary $x, y \in \mathbb{R}^n$. By scaling, we can assume without loss of generality that

$$\|x\|_2 = 1 \quad \text{and} \quad \|y\|_2 \ge 1. \tag{9.10}$$

(Why?) Consider the contraction of y onto the unit sphere, see Figure **??**:

$$\bar{y} := \frac{y}{\|y\|_2} \tag{9.11}$$

Use triangle inequality to break the increment in two parts:

$$\|X_x - X_y\|_{\psi_2} \le \|X_x - X_{\bar{y}}\|_{\psi_2} + \|X_{\bar{y}} - X_y\|_{\psi_2}.$$

Since x and $\bar{y}$ are unit vectors, Lemma 9.1.5 may be used to bound the first part. It gives

$$\|X_x - X_{\bar{y}}\|_{\psi_2} \le CK^2\|x - \bar{y}\|_2.$$

To bound the second part, note that $\bar{y}$ and y are collinear vectors, so

$$\|X_{\bar{y}} - X_y\|_{\psi_2} = \|\bar{y} - y\|_2 \cdot \|X_{\bar{y}}\|_{\psi_2}.$$

(Check this identity!) Now, since $\bar{y}$ is a unit vector, (9.5) gives

$$\|X_{\bar{y}}\|_{\psi_2} \le CK^2.$$

Combining the two parts, we conclude that

$$\|X_x - X_y\|_{\psi_2} \le CK^2\big(\|x - \bar{y}\|_2 + \|\bar{y} - y\|_2\big). \tag{9.12}$$

At this point we might get nervous: we need to bound the right hand side by $\|x - y\|_2$, but triangle inequality would give the reverse bound! Nevertheless, looking at Figure 9.1 we may suspect that in our case triangle inequality can be approximately reversed. The next exercise confirms this rigorously.

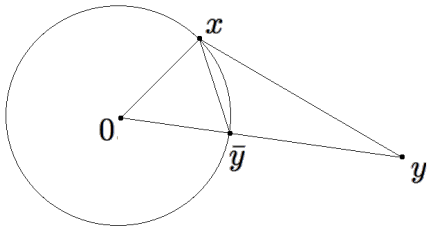


Figure 9.1: Exercise 9.1.6 shows that triangle inequality can be approximately reversed from these three vectors, and we have $\|x - \bar{y}\|_2 + \|\bar{y} - y\|_2 \le \sqrt{2}\|x - y\|_2$.

Exercise 9.1.6 (Reverse triangle inequality)**.** *Consider vectors $x, y, \bar{y} \in \mathbb{R}^n$ be satisfying* (9.10) *and* (9.11)*. Show that*

$$\|x - \bar{y}\|_2 + \|\bar{y} - y\|_2 \le \sqrt{2}\|x - y\|_2.$$

Using the result of this exercise, we deduce from (9.12) the desired bound

$$\|X_x - X_y\|_{\psi_2} \le CK^2\|x - y\|_2.$$

Theorem 9.1.1 is completely proved. □

9.2 Matrix deviation inequality

We will now state and quickly deduce the matrix deviation inequality that we announced in (9.1).

Theorem 9.2.1 (Matrix deviation inequality)**.** *Let A be an $m \times n$ matrix whose rows A_i are independent, isotropic and sub-gaussian random vectors in $\mathbb{R}^n$. Then for any subset $T \subset \mathbb{R}^n$, we have*

$$\mathbb{E} \sup_{x \in T} \Big| \|Ax\|_2 - \sqrt{m}\|x\|_2 \Big| \le CK^2\gamma(T).$$

Here $\gamma(T)$ is the Gaussian complexity introduced in (7.27)*, and $K = \max_i \|A_i\|_{\psi_2}$.*

Proof. By Theorem 9.1.1, the random process

$$X_x := \|Ax\|_2 - \sqrt{m}\|x\|_2.$$

has sub-gaussian increments. This allows us to apply Talagrand's comparison inequality in the form of Exercise 8.5.4. It gives

$$\mathbb{E} \sup_{x \in T} |X_x| \le CK^2 \gamma(T)$$

as announced. □

Exercise 9.2.2 (The centered process)**.** *Show that the conclusion of Theorem 9.2.1 holds with* $\mathbb{E}\,\|Ax\|_2$ *instead* $\sqrt{m}\|x\|_2$*, i.e. we have*

$$\mathbb{E} \sup_{x \in T} \Big| \|Ax\|_2 - \mathbb{E}\,\|Ax\|_2 \Big| \le CK^2 \gamma(T).$$

This form of matrix deviation inequality is dimension-free. Hint: Bound the difference between $\mathbb{E}\,\|Ax\|_2$ *and* $\sqrt{m}\|x\|_2$ *using* (9.5)*.*

Exercise 9.2.3 (Matrix deviation inequality: tail bounds)**.** *1. Prove the following high-probability version of the conclusion of Theorem 9.2.1. For any* $u \ge 0$*, the event*

$$\sup_{x \in T} \Big| \|Ax\|_2 - \sqrt{m}\|x\|_2 \Big| \le CK^2 \left[w(T) + u \cdot \mathrm{rad}(T) \right] \tag{9.13}$$

holds with probability at least $1 - 2\exp(-u^2)$*. Here* $\mathrm{rad}(T)$ *is the radius of* T *introduced in* (8.36)*.*

Hint: Use the high-probability version of Talagrand's comparison inequality from Exercise 8.5.5.

2. Argue that (9.13) *can be further bounded by* $CK^2 u \gamma(T)$ *for* $u \ge 1$*. Conclude that the result of part 1 of this exercise implies the conclusion of Theorem 9.2.1.*

9.3 Bounds on random matrices and sizes of random projections

9.3.1 Two-sided bounds on random matrices

Matrix deviation inequality has a number of important consequences, which we will cover in the next few sections.

To get started, let us apply the matrix deviation inequality for the unit Euclidean sphere $T = S^{n-1}$. In this case, we recover the bounds on random matrices that we proved in Section 4.5.

Indeed, the radius and Gaussian width of $T = S^{n-1}$ satisfy

$$\operatorname{rad}(T) = 1, \quad w(T) \le \sqrt{n}.$$

(Recall (7.21).) Matrix deviation inequality in the form of Exercise 9.2.3 together with triangle inequality imply that the event

$$\sqrt{m} - CK^2(\sqrt{n} + u) \le \|Ax\|_2 \le \sqrt{m} + CK^2(\sqrt{n} + u) \quad \forall x \in S^{n-1}$$

holds with probability at least $1 - 2\exp(-u^2)$.

We can interpret this event as a two-sided bound on the extreme singular values of A (recall (4.2)):

$$\sqrt{m} - CK^2(\sqrt{n} + u) \le s_n(A) \le s_1(A) \le \sqrt{m} + CK^2(\sqrt{n} + u).$$

Thus we recover the result we proved in Theorem 4.5.1.

9.3.2 Sizes of random projections of geometric sets

Let us note an immediate application of matrix deviation inequality. for sizes of random projections of geometric sets we studied in Section 7.8. We will see how matrix deviation inequality can yield a sharper bound.

Proposition 9.3.1 (Sizes of random projections of sets)**.** *Consider a bounded set $T \subset \mathbb{R}^n$. Let A be an $m \times n$ matrix whose rows A_i are independent, isotropic and sub-gaussian random vectors in $\mathbb{R}^n$. Then the scaled matrix*

$$P := \frac{1}{\sqrt{n}} A$$

(a "sub-gaussian projection") satisfies

$$\mathbb{E} \operatorname{diam}(PT) \le \sqrt{\frac{m}{n}} \operatorname{diam}(T) + CK^2 w_s(T).$$

Here $w_s(T)$ is the spherical width of T (recall Section 7.7.2) and $K = \max_i \|A_i\|_{\psi_2}$.

Proof. Theorem 9.2.1 implies via triangle inequality that

$$\mathbb{E} \sup_{x \in T} \|Ax\|_2 \le \sqrt{m} \sup_{x \in T} \|x\|_2 + CK^2 \gamma(T).$$

We can state this inequality in terms of radii of the sets AT and T as

$$\mathbb{E}\operatorname{rad}(AT) \le \sqrt{m}\operatorname{rad}(T) + CK^2\gamma(T).$$

Applying this bound for the difference set $T - T$ instead of T, we can write it as

$$\mathbb{E}\operatorname{diam}(AT) \le \sqrt{m}\operatorname{diam}(T) + CK^2 w(T).$$

(Here we used (7.28) to pass from Gaussian complexity to Gaussian width.) Dividing both sides by $\sqrt{n}$ completes the proof. □

Let us compare Proposition 9.3.1 with our older bounds on sizes of projections, Theorem 7.8.1. We can see that the new result is more general and also sharper. It states that the diameter scales by the exact factor $\sqrt{m/n}$ without an absolute constant in front of it.

Exercise 9.3.2 (Sizes of projections: high-probability bounds)**.** *Use the high-probability version of matrix deviation inequality (Exercise 9.2.3) to obtain a high-probability version of Proposition 9.3.1 with better probability Namely, show that for $\varepsilon > 0$, the bound*

$$\operatorname{diam}(AT) \le (1+\varepsilon)\sqrt{\frac{m}{n}}\operatorname{diam}(T) + CK^2 w_s(T)$$

holds with probability at least $1 - \exp(-c\varepsilon^2 m/K^4)$.

Exercise 9.3.3. *Deduce a version of Proposition 9.3.1 for the original model of P considered in Section 7.8, i.e. for a random projection P onto a random m-dimensional subspace $E \sim \operatorname{Unif}(G_{n,m})$.*

9.4 Johnson-Lindenstrauss Lemma for infinite sets

Let us now apply the matrix deviation inequality for a finite set T. In this case, we recover Johnshon-Lindenstrauss Lemma from Section 5.3 and more.

9.4.1 Recovering the classical Johnson-Lindenstrauss

Let us check that matrix deviation inequality contains the classical Johnson-Lindenstrauss Lemma (Theorem 5.3.1). Let $\mathcal{X}$ be a set of N points in $\mathbb{R}^n$ and define T to be the set of normalized differences of $\mathcal{X}$, i.e.

$$T := \Big\{ \frac{x-y}{\|x-y\|_2} : \ x, y \in \mathcal{X} \text{ are distinct points} \Big\}.$$

Then the radius and Gaussian complexity of T satisfy

$$\operatorname{rad}(T) \le 1, \quad \gamma(T) \le C\sqrt{\log N} \tag{9.14}$$

(Recall Exercise 7.7.6). Then matrix deviation inequality (Theorem 9.2.1) implies that the bound

$$\sup_{x,y\in\mathcal{X}} \left| \frac{\|Ax - Ay\|_2}{\|x - y\|_2} - \sqrt{m} \right| \lesssim \sqrt{\log N} \tag{9.15}$$

holds with high probability. To keep the calculation simple, we will be satisfied here with probability 0.99, which can be obtained using Markov's inequality; Exercise 9.2.3 gives better probability. Also, for simplicity we suppressed the dependence on the sub-gaussian norm K.

Multiply both sides of (9.15) by $\frac{1}{\sqrt{m}}\|x - y\|_2$ and rearrange the terms. We obtain that, with high probability, the scaled random matrix

$$Q := \frac{1}{\sqrt{m}} A$$

is an approximate isometry on $\mathcal{X}$, i.e.

$$(1 - \varepsilon)\|x - y\|_2 \le \|Qx - Qy\|_2 \le (1 + \varepsilon)\|x - y\|_2 \quad \text{for all } x, y \in \mathcal{X}.$$

where

$$\varepsilon \lesssim \sqrt{\frac{\log N}{m}}.$$

Equivalently, if we fix $\varepsilon > 0$ and choose the dimension m such that

$$m \gtrsim \varepsilon^{-2} \log N,$$

then with high probability Q is an ε-isometry on $\mathcal{X}$. Thus we recover the classical Johnson-Lindenstrauss Lemma (Theorem 5.3.1).

Exercise 9.4.1. *In the argument above, quantify the probability of success and dependence on K. Thus, use matrix deviation inequality to give an alternative solution of Exercise 5.3.3.*

9.4.2 Johnson-Lindenstrauss lemma for infinite sets

The argument above does not really depend on $\mathcal{X}$ being a finite set. We only used that $\mathcal{X}$ is finite to bound the Gaussian complexity in (9.14). This means that we can give a version of Johnson-Lindenstrauss lemma for general, not necessarily finite sets. Let us state such version.

Proposition 9.4.2 (Additive Johnson-Lindenstrauss Lemma)**.** *Consider a set $\mathcal{X} \subset \mathbb{R}^n$. Let A be an $m \times n$ matrix whose rows A_i are independent, isotropic and sub-gaussian random vectors in $\mathbb{R}^n$. Then, with high probability (say, 0.99), the scaled matrix*

$$Q := \frac{1}{\sqrt{m}} A$$

satisfies

$$\|x - y\|_2 - \delta \le \|Qx - Qy\|_2 \le \|x - y\|_2 + \delta \quad \text{for all } x, y \in \mathcal{X}$$

where

$$\delta = \frac{CK^2 w(\mathcal{X})}{\sqrt{m}}$$

and $K = \max_i \|A_i\|_{\psi_2}$.

Proof. Choose T to be the difference set, i.e. $T = \mathcal{X} - \mathcal{X}$, and apply matrix deviation inequality (Theorem 9.2.1). It follows that, with high probability,

$$\sup_{x,y \in \mathcal{X}} \big| \|Ax - Ay\|_2 - \sqrt{m}\|x - y\|_2 \big| \le CK^2 \gamma(\mathcal{X} - \mathcal{X}) = 2CK^2 w(\mathcal{X}).$$

(In the last step, we used (7.28).) Dividing both sides by $\sqrt{m}$, we complete the proof. □

Note that the error δ in Proposition 9.4.2 is additive, while the classical Johnson-Lindenstrauss Lemma for finite sets (Theorem 5.3.1) has a multiplicative form of error. This may be a small difference, but in general it is necessary:

Exercise 9.4.3 (Additive error)**.** *Suppose a set $\mathcal{X}$ has a non-empty interior. Check that, in order for the conclusion (5.10) of the classical Johnson-Lindenstrauss lemma to hold, one must have*

$$\operatorname{rank}(Q) = m \ge n.$$

Remark 9.4.4 (Statistical dimension). The additive version of Johnson-Lindenstrauss Lemma can be naturally stated in terms of the statistical dimension of $\mathcal{X}$,

$$d(\mathcal{X}) \sim \frac{w(\mathcal{X})^2}{\operatorname{diam}(\mathcal{X})^2},$$

which we introduced in Section 7.7.4. To see this, let us fix $\varepsilon > 0$ and choose the dimension m so that it *exceeds an appropriate multiple of the statistical dimension*, namely

$$m \geq (CK^4/\varepsilon^2)d(T).$$

Then in Proposition 9.4.2 we have $\delta \leq \varepsilon \operatorname{diam}(\mathcal{X})$. This means that Q *preserves the distances in* $\mathcal{X}$ *to within a small fraction of the maximal distance*, which is the diameter of $\mathcal{X}$.

9.5 Random sections: M^* bound and Escape Theorem

Consider a set $T \subset \mathbb{R}^n$ and a random subspace E with given dimension. How large is the typical intersection of T and E? See Figure 9.2 for illustration. There are two types of answers to this question. In Section 9.5.1 we will give a general bound the expected diameter of $T \cap E$; it is called the M^* bound. The intersection $T \cap E$ can even be empty; this is the content of the *Escape Theorem* which we will prove in Section 9.5.2. Both results are consequences of matrix deviation inequality.

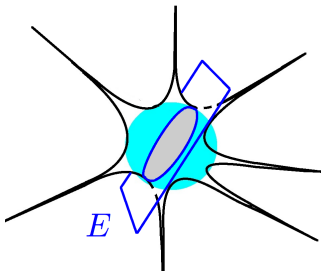


Figure 9.2: Illustration for M^* bound: the intersection of a set T with a random subspace E.

9.5.1 M^* bound

First, it is convenient to realize the random subspace E as a kernel of a random matrix, i.e. set

$$E := \ker A$$

where A is a random $m \times n$ random matrix. We always have

$$\dim(E) \geq n - m,$$

and for continuous distributions we have $\dim(E) = n - m$ almost surely.

Example 9.5.1. Let A be a Gaussian matrix with independent $N(0,1)$ entries. Rotation invariance implies that $E = \ker(A)$ is uniformly distributed in the Grassmanian:

$$E \sim \text{Unif}(G_{n,n-m}).$$

Our main result is the following general bound on the diameters of random sections of geometric sets. For historic reasons, this results is called the M^* bound.

Theorem 9.5.2 (M^*bound)**.** *Consider a set $T \subset \mathbb{R}^n$. Let A be an $m \times n$ matrix whose rows A_i are independent, isotropic and sub-gaussian random vectors in $\mathbb{R}^n$. Then the random subspace $E = \ker A$ satisfies*

$$\mathbb{E}\,\text{diam}(T \cap E) \le \frac{CK^2 w(T)}{\sqrt{m}},$$

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

Proof. Apply Theorem 9.2.1 for $T - T$ and obtain

$$\mathbb{E} \sup_{x,y \in T} \big| \|Ax - Ay\|_2 - \sqrt{m}\|x - y\|_2 \big| \le CK^2 \gamma(\mathcal{X} - \mathcal{X}) = 2CK^2 w(T).$$

If we restrict the supremum to points x, y in the kernel of A, then $\|Ax - Ay\|_2$ disappears since $A(x - y) = 0$, and we have

$$\mathbb{E} \sup_{x,y \in T \cap \ker A} \sqrt{m}\|x - y\|_2 \le 2CK^2 w(T).$$

Dividing by $\sqrt{m}$ yields

$$\mathbb{E}\,\text{diam}(T \cap \ker A) \le \frac{CK^2 w(T)}{\sqrt{m}},$$

which is the bound we claimed. □

Exercise 9.5.3 (Affine sections)**.** *Check that M^* bound holds not only for sections through the origin but for all affine sections as well:*

$$\mathbb{E} \max_{z \in \mathbb{R}^n} \text{diam}(T \cap E_z) \le \frac{CK^2 w(T)}{\sqrt{m}}$$

where $E_z = z + \ker A$.

Remark 9.5.4 (Statistical dimension). Surprisingly, the random subspace E in the M^* bound is not low-dimensional. On the contrary, $\dim(E) \geq n - m$ and we would typically choose $m \ll n$, so E has almost full dimension. This makes the M^* bound a strong and perhaps surprising statement.

It can be enlightening to look at the M^* bound through the lens of the notion of statistical dimension $d(T) \sim w(T)^2/\operatorname{diam}(T)^2$, which we introduced in Section 7.7.4. Fix $\varepsilon > 0$. Then the M^* bound can be stated as

$$\mathbb{E} \operatorname{diam}(T \cap E) \leq \varepsilon \cdot \operatorname{diam}(T)$$

as long as

$$m \geq C(K^4/\varepsilon^2)d(T). \tag{9.16}$$

In words, *the M^* bound becomes non-trivial – the diameter shrinks – as long as the codimension of E exceeds a multiple of the statistical dimension of T.*

Equivalently, the dimension condition states that the sum of dimension of E and a multiple of statistical dimension of T should be bounded by n. This condition should now make sense from the linear algebraic point of view. For example, if T is a centered Euclidean ball in some subspace $F \subset \mathbb{R}^n$ then a non-trivial bond $\operatorname{diam}(T \cap E) < \operatorname{diam}(T)$ is possible only if

$$\dim E + \dim F \leq n.$$

(Why?)

Let us look at one remarkable example of application of the M^* bound.

Example 9.5.5 (The ℓ_1 ball). Let $T = B_1^n$, the unit ball of the ℓ_1 norm in $\mathbb{R}^n$. Since we proved in (7.23) that $w(T) \sim \sqrt{\log n}$, the M^* bound (Theorem 9.5.2) gives

$$\mathbb{E} \operatorname{diam}(T \cap E) \lesssim \sqrt{\frac{\log n}{m}}.$$

For example, if $m = 0.1n$ then

$$\mathbb{E} \operatorname{diam}(T \cap E) \lesssim \sqrt{\frac{\log n}{n}}. \tag{9.17}$$

Comparing this with $\operatorname{diam}(T) = 2$, we see that the diameter shrinks by almost $\sqrt{n}$ as a result of intersecting T with the random subspace E that has almost full dimension (namely, $0.9n$).

For an intuitive explanation of this surprising fact, recall from Section 7.7.3 that the "bulk" the octahedron $T = B_1^n$ is formed by the inscribed ball $\frac{1}{\sqrt{n}} B_2^n$. Then it should not be surprising if a random subspace E tends

to pass through the bulk and miss the "outliers" that lie closer to the vertices of T. This makes the diameter of $T \cap E$ essentially the same as the size of the bulk, which is $1/\sqrt{n}$.

This example indicates what makes a surprisingly strong and general result like M^* bound possible. Intuitively, the random subspace E tends to pass entirely through the bulk of T, which is usually a Euclidean ball with much smaller diameter than T, see Figure 9.2.

Sharper bound – later in Exercise **??**

Exercise 9.5.6 (M^* bound with high probability)**.** *Use the high-probability version of matrix deviation inequality (Exercise 9.2.3) to obtain a high-probability version of the M^* bound.*

9.5.2 Escape theorem

In some circumstances, a random subspace E may completely miss a given set T in $\mathbb{R}^n$. This might happen, for example, if T is a subset of the sphere, see Figure 9.3. In this case, the intersection $T \cap E$ is typically empty under essentially the same conditions as in M^* bound.

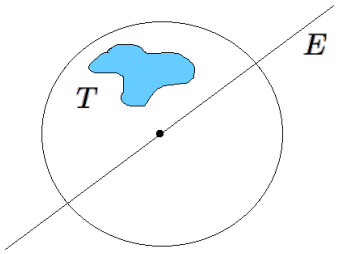


Figure 9.3: Illustration for the Escape theorem: the set T has empty intersection with a random subspace E.

Theorem 9.5.7 (Escape theorem)**.** *Consider a set $T \subset S^{n-1}$. Let A be an $m \times n$ matrix whose rows A_i are independent, isotropic and sub-gaussian random vectors in $\mathbb{R}^n$. If*

$$m \geq CK^4 w(T)^2, \tag{9.18}$$

then the random subspace $E = \ker A$ satisfies

$$T \cap E = \emptyset$$

with probability at least $1 - 2\exp(-cm/K^4)$. Here $K = \max_i \|A_i\|_{\psi_2}$.

Proof. Let us use the high-probability version of matrix deviation inequality from Exercise 9.2.3. It states that the bound

$$\sup_{x \in T} \left| \|Ax\|_2 - \sqrt{m} \right| \le C_1 K^2 (w(T) + u) \tag{9.19}$$

holds with probability at least $1 - 2\exp(-u^2)$. Suppose this event indeed holds and $T \cap E \neq \emptyset$. Then for any $x \in T \cap E$ we have $\|Ax\|_2 = 0$, so our bound becomes

$$\sqrt{m} \le C_1 K^2 (w(T) + u).$$

Choosing $u := \sqrt{m} 2C_1 K^2$, we simplify the bound to

$$\sqrt{m} \le C_1 K^2 w(T) + \frac{\sqrt{m}}{2},$$

which yields

$$\sqrt{m} \le 2C_1 K^2 w(T).$$

But this contradicts the assumption of the Escape theorem, as long as we choose the absolute constant C large enough. This means that the event (9.19) with u chosen as above implies that $T \cap E = \emptyset$. The proof is complete. □

Exercise 9.5.8 (Sharpness of Escape theorem)**.** *Discuss the sharpness of Escape Theorem for the example where T is the unit sphere of some subspace of $\mathbb{R}^n$.*

Exercise 9.5.9 (Escape from a point set)**.** *Prove the following version of Escape theorem with a rotation of a point set instead of a random subspace.*

Consider a set $T \subset S^{n-1}$ and let $\mathcal{X}$ be a set of N points in $\mathbb{R}^n$. Show that, if

$$\sigma_{n-1}(T) < \frac{1}{N}$$

then there exists a rotation $U \in O(n)$ such that

$$T \cap U\mathcal{X} = \emptyset.$$

Here σ_{n-1} denotes the normalized Lebesgue measure (area) on S^{n-1}.

Hint: Consider a random rotation $U \in \mathrm{Unif}(SO(n))$ as in Section 5.2.5. Applying a union bound, show that the probability that there exists $x \in \mathcal{X}$ such that $Ux \in T$ is smaller than 1.

Chapter 10

Sparse Recovery

In this chapter, we study applications of high-dimensional probability to signal recovery problems that are typical for modern data sciences.

10.1 High dimensional signal recovery problems

Mathematically, a *signal* is a vector $x \in \mathbb{R}^n$. Suppose we do not know x, but we have m random, linear, possibly noisy *measurements* of x. Such measurements can be represented a vector $y \in \mathbb{R}^m$ with following form:

$$y = Ax + w. \tag{10.1}$$

Here A is an $m \times n$ known random measurement matrix, and $w \in \mathbb{R}^m$ is an unknown *noise* vector; see Figure 10.1. Our goal is to recover x from A and y as accurately as possible.

Note that the measurements $y = (y_1, \dots, y_m)$ can be equivalently represented as

$$y_i = \langle A_i, x \rangle + w_i, \quad i = 1, \dots, m \tag{10.2}$$

where $A_i^\mathsf{T} \in \mathbb{R}^n$ denote the rows of the matrix A. It is natural to assume that A_i are independent, which makes the observations y_i independent, too.

Example 10.1.1 (Audio sampling). In signal processing applications, x can be a digitized audio signal. The measurement vector y can be obtained by sampling x at m randomly chosen time points, see Figure 10.2.

Example 10.1.2 (Linear regression). The linear regression is one of the major inference problems in Statistics. Here we would like to model the relationship

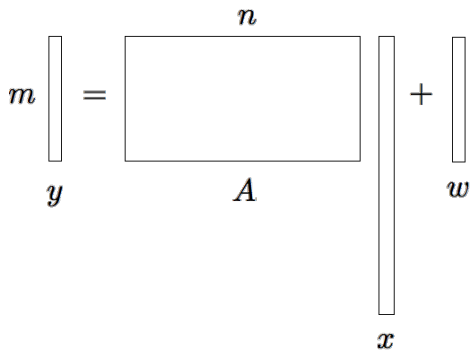


Figure 10.1: Signal recovery problem: recover a signal x from random, linear measurements y.

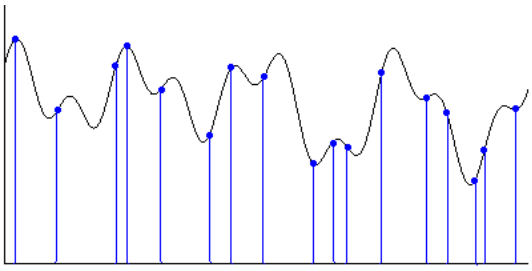


Figure 10.2: Signal recovery problem in audio sampling: recover an audio signal x from a sample of x taken at m random time points.

between n *predictor variables* and a *response variable* using a sample of m observations. The regression problem is usually written as

$$Y = X\beta + w.$$

Here X is an $m \times n$ matrix that contains a sample of predictor variables, $Y \in \mathbb{R}^m$ is a vector that contains a sample of response variables, $\beta \in \mathbb{R}^n$ is a coefficient vector that specifies the relationship that we try to recover, and w is a noise vector.

For example, in genetics one could be interested in predicting a certain disease based on genetic information. One then performs a study on m patients collecting the expressions of their n genes. The matrix X is defined by letting X_{ij} be the expression of gene j in patient i, and the coefficients Y_i of the vector Y can be set to quantify whether or not patient i has the disease (and to what extent). The goal is to recover the coefficients of β, which quantify how each gene affects the disease.

10.1.1 Incorporating prior information about the signal

Many modern signal recovery problems operate in the regime where

$$m \ll n,$$

i.e. we have far fewer measurements than unknowns. For instance, in a typical genetic study like the one described in Example 10.1.2, the number of patients is ~ 100 while the number of genes is $\sim 10,000$.

In this regime, the recovery problem (10.1) is *ill-posed* even in the noiseless case where $w = 0$. It can not be even approximately solved: the solutions form a linear subspace of dimension at least $n - m$. To overcome this difficulty, we can leverage some *prior information* about the signal x – something that we know, believe, or want to enforce about x. Such information can be mathematically be expressed by assuming that

$$x \in T \tag{10.3}$$

where $T \subset \mathbb{R}^n$ is a known set.

The smaller the set T, the fewer measurements m could be needed to recover x. For small t, we can hope that signal recovery can be solved even in the ill-posed regime where $m \ll n$. We will see how this idea works in the next sections.

10.2 Signal recovery based on M^* bound

Let us return to the the recovery problem (10.1). For simplicity, let us first consider the noiseless version of the problem, that it

$$y = Ax, \quad x \in T.$$

To recap, here $x \in \mathbb{R}^n$ is the unknown signal, $T \subset \mathbb{R}^n$ is a known set that encodes our prior information about x, and A is a known $m \times n$ random measurement matrix. Our goal is to recover x from y.

Perhaps the simplest candidate for the solution would be *any* vector x' that is consistent both with the measurements and the prior, so we

$$\text{find } x' : \ y = Ax', \quad x \in T. \tag{10.4}$$

If the set T is convex, this is a convex program (in the feasibility form), and many effective algorithms exists to numerically solve it.

This naïve approach actually works well. We will now quickly deduce this from the M^* bound from Section 9.5.1.

Theorem 10.2.1. *Suppose the rows A_i of A are independent, isotropic and sub-gaussian random vectors. Then a solution $\widehat{x}$ of the program (10.4) satisfies*

$$\mathbb{E} \|\widehat{x} - x\|_2 \le \frac{CK^2 w(T)}{\sqrt{m}},$$

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

Proof. Since $x, \widehat{x} \in T$ and $Ax = A\widehat{x} = y$, we have

$$x, \widehat{x} \in T \cap E_x$$

where $E_x := x + \ker A$. (Figure 10.3 illustrates this situation visually.) Then

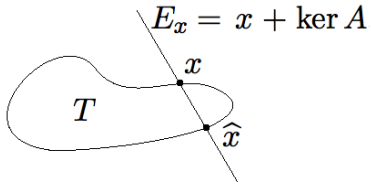


Figure 10.3: Signal recovery: the signal x and the solution $\widehat{x}$ lie in the prior set T and in the affine subspace E_x.

the affine version of the M^* bound (Exercise 9.5.3) yields

$$\mathbb{E}\,\|\widehat{x} - x\|_2 \le \mathbb{E}\,\mathrm{diam}(T \cap E_x) \le \frac{CK^2 w(T)}{\sqrt{m}}.$$

This completes the proof. □

Remark 10.2.2 (Statistical dimension). Arguing as in Remark 9.5.4, we obtain a non-trivial error bound

$$\mathbb{E}\,\|\widehat{x} - x\|_2 \le \varepsilon \cdot \mathrm{diam}(T)$$

provided that the number of measurements m is so that

$$m \ge C(K^4/\varepsilon^2)d(T).$$

In words, *the signal can be approximately recovered as long as the number of measurements m exceeds a multiple of the statistical dimension $d(T)$ of the prior set T.*

Since the statistical dimension can be much smaller than the ambient dimension n, the recovery problem may often be solved even if the high-dimensional, ill-posed regime where

$$m \ll n.$$

We will see some concrete examples of this situation shortly.

Remark 10.2.3 (Convexity). If the prior set T is not convex, we can convexify it by replacing T with its convex hull $\mathrm{conv}(T)$. This makes (10.4) a convex program, and thus computationally tractable. At the same time, the recovery guarantees of Theorem 10.2.1 do not change since

$$w(\mathrm{conv}(T)) = w(T)$$

by Proposition 7.7.2.

Exercise 10.2.4 (Noisy measurements)**.** *Extend the recovery result (Theorem 10.2.1) for the noisy model $y = Ax + w$ we considered in (10.1). Namely, show that*

$$\mathbb{E}\,\|\widehat{x} - x\|_2 \le \frac{CK^2 w(T) + \|w\|_2}{\sqrt{m}}.$$

Hint: Modify the argument that leads to the M^* bound.

Exercise 10.2.5 (Mean squared error)**.** *Prove that the error bound Theorem 10.2.1 can be extended for the mean squared error*

$$\mathbb{E}\,\|\widehat{x} - x\|_2^2.$$

Hint: Modify the M^* bound accordingly.

Exercise 10.2.6 (Recovery by optimization)**.** *Suppose T is the unit ball of some norm $\|\cdot\|_T$ in $\mathbb{R}^n$. Show that the conclusion of Theorem 10.2.1 holds also for the solution of the following optimization program:*

$$\textit{minimize } \|x'\|_T \textit{ s.t. } y = Ax'.$$

10.3 Recovery of sparse signals

10.3.1 Sparsity

Let us give a concrete example of a prior set T. Very often, we believe that x should be *sparse*, i.e. that most coefficients of x are zero, exactly or approximately. For instance, in genetic studies like the one we described in Example 10.1.2, it is natural to expect that very few genes (~ 10) have significant impact on a given disease, and we would like to find out which ones.

In some applications, one needs to change basis so that the signals of interest are sparse. For instance, in the audio recovery problem considered in Example 10.1.1, we typically deal with *band-limited* signals x. Those are the signals whose frequencies (the values of the Fourier transform) are constrained to some small set, such as a bounded interval. While the audio signal x itself is not sparse as is apparent from Figure 10.2, the Fourier transform of x may be sparse. In other words, x may be sparse in the frequency and not time domain.

To quantify the (exact) sparsity of a vector $x \in \mathbb{R}^n$, we consider the size of the support of x which we denote

$$\|x\|_0 := |\operatorname{supp}(x)| = |\{i : \ x_i \neq 0\}|\,.$$

Assume that

$$\|x\|_0 = s \ll n. \tag{10.5}$$

This can be viewed as a special case of a general assumption (10.3) by putting

$$T = \{x \in \mathbb{R}^n : \ \|x\|_0 \leq s\}\,.$$

Then a simple dimension count shows the recovery problem (10.1) could become well posed:

Exercise 10.3.1 (Sparse recovery problem is well posed)**.** *Argue that if $m \geq \|x\|_0$, the solution to the sparse recovery problem (10.1) is unique if it exists.*

Even when the problem (10.1) is well posed, it could be computationally hard. It is easy if one knows the support of x (why?) but usually the support is unknown. An exhaustive search over all possible supports (subsets of a given size s) is impossible since the number of possibilities is exponentially large: $\binom{n}{s} \geq 2^s$.

Fortunately, there exist computationally effective approaches to high-dimensional recovery problems with general constraints (10.3), and the sparse recovery problems in particular. We will cover these approaches next.

Exercise 10.3.2 (The "ℓ_p norms" for $0 \leq p < 1$)**.** *1. Check that $\|\cdot\|_0$ is not a norm on $\mathbb{R}^n$.*

2. Check that $\|\cdot\|_p$ is not a norm on $\mathbb{R}^n$ if $0 < p < 1$. Figure 10.4 illustrates the unit balls for various ℓ_p "norms".

3. Show that, for every $x \in \mathbb{R}^n$,

$$\|x\|_0 = \lim_{p \to 0_+} \|x\|_p.$$

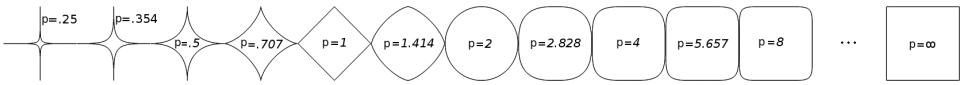


Figure 10.4: The unit balls of ℓ_p for various p in $\mathbb{R}^2$.

10.3.2 Convexifying the sparsity by ℓ_1 norm, and recovery guarantees

Let us specialize the general recovery guarantees developed in Section 10.2 to the sparse recovery problem. To do this, we should choose the prior set T so that it promotes sparsity. In the previous section, we saw that the choice

$$T := \{x \in \mathbb{R}^n : \ \|x\|_0 \leq s\}$$

does not allow for computationally tractable algorithms.

To make T convex, we may replace the "ℓ_0 norm" by the ℓ_p norm with the smallest exponent $p > 0$ that makes this a true norm. This exponent

is obviously $p = 1$ as we can see from Figure 10.4. So let us repeat this important heuristic: *we propose to replace the ℓ_0 "norm" by the ℓ_1 norm.*

Thus it makes sense to choose T to be a scaled ℓ_1 ball:

$$T := \sqrt{s} B_1^n.$$

The scaling factor $\sqrt{s}$ was chosen so that T can accommodate all s-sparse unit vectors:

Exercise 10.3.3. *Check that*

$$\{x \in \mathbb{R}^n : \ \|x\|_0 \le s, \ \|x\|_2 \le 1\} \subset \sqrt{s} B_1^n.$$

For this T, the general recovery program (10.4) becomes

$$\text{Find } x' : \ y = Ax', \quad \|x'\|_1 \le \sqrt{s}. \tag{10.6}$$

Note that this is a convex program, and therefore is computationally tractable. And the general recovery guarantee, Theorem 10.2.1, specialized to our case, implies the following.

Corollary 10.3.4 (Sparse recovery: guarantees)**.** *Assume the unknown s-sparse signal $x \in \mathbb{R}^n$ satisfies $\|x\|_2 \le 1$. Then x can be approximately recovered from the random measurement vector $y = Ax$ by a solution $\widehat{x}$ of the program (10.6). The recovery error satisfies*

$$\mathbb{E} \|\widehat{x} - x\|_2 \le CK^2 \sqrt{\frac{s \log n}{m}}.$$

Proof. Set $T = \sqrt{s} B_1^n$. The result follows from Theorem 10.2.1 and the bound (7.23) on the Gaussian width of the ℓ_1 ball:

$$w(T) = \sqrt{s} w(B_1^n) \le C\sqrt{s \log n}. \quad \square$$

The recovery error becomes small if

$$m \sim s \log n,$$

if the hidden constant here is appropriately large. In words, recovery is possible if *the number of measurements m is almost linear in the sparsity s*, while its dependence on the ambient dimension n is mild (logarithmic). This is good news. It means that for sparse signals, one can solve recovery problems in the high dimensional regime where

$$m \ll n,$$

i.e. with much fewer measurements than the dimension.

Exercise 10.3.5 (Sparse recovery by convex optimization)**.** *1. Show that an unknown s-sparse signal x (without restriction on the norm) can be approximately recovered by solving the convex optimization problem*

$$\text{minimize } \|x'\|_1 \text{ s.t. } y = Ax'. \tag{10.7}$$

The recovery error satisfies

$$\mathbb{E}\,\|\widehat{x} - x\|_2 \le C\sqrt{\frac{s\log n}{m}}\|x\|_2.$$

2. Argue that a similar result holds for approximately sparse signals. State and prove such a guarantee.

10.3.3 The convex hull of sparse vectors, and the logarithmic improvement

The replacement of s-sparse vectors by the octahedron $\sqrt{s}B_1^n$ that we made in Exercise 10.6 is almost sharp. In the following exercise, we show that the convex hull of the set of sparse vectors

$$S_{n,s} := \{x \in \mathbb{R}^n : \ \|x\|_0 \le s,\ \|x\|_2 \le 1\}$$

is approximately the truncated ℓ_1 ball

$$T_{n,s} := \sqrt{s}B_1^n \cap B_2^n = \left\{x \in \mathbb{R}^n : \ \|x\|_1 \le \sqrt{s},\ \|x\|_2 \le 1\right\}.$$

Exercise 10.3.6 (The convex hull of sparse vectors)**.** *1. Check that*

$$\operatorname{conv}(S_{n,s}) \subset T_{n,s}.$$

2. To help us prove a reverse inclusion, fix $x \in T_{n,s}$ and partition the support of x into disjoint subsets $T_1, T_2, \dots$ so that T_1 indexes the largest s elements of x in magnitude, T_2 indexes the next s largest elements, and so on. Show that

$$\sum_{i\ge 1} \|x_{T_i}\|_2 \le 2,$$

where $x_T \in \mathbb{R}^T$ denotes the restriction of x onto a set T.

Hint: Note that $\|x_{T_1}\|_2 \le 1$. Next, for $i \ge 2$, note that each coordinate of x_{T_i} is smaller in magnitude than the average coordinate of $x_{T_{i-1}}$; conclude that $\|x_{T_i}\|_2 \le (1/\sqrt{s})\|x_{T_{i-1}}\|_1$. Then sum up the bounds.

3. Deduce from part 2 that

$$K_{n,s} \subset 2\operatorname{conv}(S_{n,s}).$$

Exercise 10.3.7 (Gaussian width of the set of sparse vectors)**.** *Use Exercise 10.3.6 to show that*

$$w(T_{n,s}) \le 2w(S_{n,s}) \le C\sqrt{s\log(2n/s)}.$$

Improve the logarithmic factor in the error bound for sparse recovery (Corollary 10.3.4) to

$$\mathbb{E}\,\|\widehat{x} - x\|_2 \le C\sqrt{\frac{s\log(2n/s)}{m}}.$$

This shows that

$$m \sim s\log(2n/s)$$

measurements suffice for sparse recovery.

Exercise 10.3.8 (Sharpness)**.** *Use Exercise 10.3.6 to show that*

$$w(T_{n,s}) \ge w(S_{n,s}) \ge c\sqrt{s\log(2n/s)}.$$

Write the hint.

Hint: Covering + Sudakov?

Exercise 10.3.9 (Garnaev-Gluskin's theorem)**.** *Improve the logarithmic factor in the bound (9.5.5) on the sections of the ℓ_1 ball. Namely, show that*

$$\mathbb{E}\,\mathrm{diam}(B_1^n \cap E) \lesssim \sqrt{\frac{\log(2n/m)}{m}}.$$

In particular, this shows that the logarithmic factor in (9.17) is not needed.

Hint: Fix $\rho > 0$ and apply the M^* bound for the truncated octahedron $T_\rho := B_1^n \cap \rho B_2^n$. Use Exercise 10.3.7 to bound the Gaussian width of T_ρ. Furthermore, note that if $\mathrm{rad}(T_\rho \cap E) \le \delta$ for some $\delta \le \rho$ then $\mathrm{rad}(T \cap E) \le \delta$. Finally, optimize in ρ.

10.4 Low-rank matrix recovery

In the following series of exercises, we will establish a *matrix* version of the sparse recovery problem studied in Section 10.3. The unknown signal will now be a $d \times d$ matrix X instead of a signal $x \in \mathbb{R}^n$ considered previously.

There are two natural notions of sparsity for matrices. One is where most of the entries of X are zero, at it is quantifies by the ℓ_0 "norm" $\|X\|_0$, which counts non-zero entries. For this notion, we can directly apply the analysis of sparse recovery from Section 10.3. Indeed, it is enough to vectorize the matrix X and think of it as a long vector in $\mathbb{R}^{d^2}$.

But in this section, we will consider an alternative and equally useful notion of sparsity for matrices: *low rank*. It is quantified by the rank of X,

which we may think of as the ℓ_0 norm of the vector of the singular values of X, i.e.

$$s(X) := (s_i(X))_{i=1}^d. \tag{10.8}$$

Our analysis of the low-rank matrix recovery problem will roughly go along the same lines as the analysis of sparse recovery, but will not be identical to it.

Let us set up a low-rank matrix recovery problem. We would like to recover an unknown $d \times d$ matrix from m random measurements of the form

$$y_i = \langle A_i, X \rangle, \quad i = 1, \dots, m. \tag{10.9}$$

Here A_i are independent $d \times d$ matrices, and $\langle A_i, X \rangle = \operatorname{tr}(A_i^\mathsf{T} X)$ is the canonical inner product of matrices (recall Section 4.1.3). In dimension $d = 1$, the matrix recovery problem (10.9) reduces to the vector recovery problem (10.2).

Since we have m linear equations in $d \times d$ variables, the matrix recovery problem is *ill-posed* if

$$m < d^2.$$

To be able to solve it in this range, we make an additional assumption that X has low rank, i.e.

$$\operatorname{rank}(X) \le r \ll d.$$

10.4.1 The nuclear norm

Like sparsity, the rank is not a convex function. To fix this, in Section 10.3 we replaced the sparsity (i.e. the ℓ_0 "norm") by the ℓ_1 norm. Let us try to do the same for the notion of rank. The rank is the ℓ_0 "norm" of the vector $s(X)$ of the singular values in (10.8). Replacing the ℓ_0 norm by the ℓ_1 norm, we obtain the quantity

$$\|X\|_* := \sum_{i=1}^d s_i(X)$$

which is called the *nuclear norm* of X. (We omit the absolute values since the singular values are non-negative.)

Exercise 10.4.1. *Prove that $\|\cdot\|_*$ is a norm on the space of $d \times d$ matrices.*

Make sure that the solution is not too hard.

Exercise 10.4.2 (Nuclear, Frobenius and operator norms)**.** *Check that*

$$\langle X, Y \rangle \le \|X\|_* \cdot \|Y\| \tag{10.10}$$

Conclude that

$$\|X\|_F^2 \le \|X\|_* \cdot \|X\|.$$

Hint: Think of the nuclear norm $\|\cdot\|_*$, Frobenius norm $\|\cdot\|_F$ and the operator norm $\|\cdot\|$ as matrix analogs of the ℓ_1 norm, ℓ_2 norm and ℓ_∞ norms for vectors, respectively.

Denote the unit ball of the nuclear norm by

$$B_* := \left\{ X \in \mathbb{R}^{d \times d} : \ \|X\|_* \le 1 \right\}.$$

Exercise 10.4.3 (Gaussian width of the unit ball of the nuclear norm)**.** *Show that*

$$w(B_*) \le 2\sqrt{d}.$$

Hint: Use (10.10) followed by Theorem 7.4.1.

The following is a matrix version of Exercise ex: sparse into L1.

Exercise 10.4.4. *Check that*

$$\left\{ X \in \mathbb{R}^{d \times d} : \ \operatorname{rank}(X) \le r, \ \|X\|_F \le 1 \right\} \subset \sqrt{r} B_*.$$

10.4.2 Guarantees for low-rank matrix recovery

It makes sense to try to solve the low-rank matrix recovery problem (10.9) using the matrix version of the convex program (10.6), i.e.

$$\text{Find } X' : \ y_i = \langle A_i, X' \rangle \ \forall i = 1, \dots, m; \quad \|X'\|_* \le \sqrt{r}. \tag{10.11}$$

Exercise 10.4.5 (Low-rank matrix recovery: guarantees)**.** *Suppose the random matrices A_i are independent and have all independent, sub-gaussian entries.*[1] *Assume the unknown $d \times d$ matrix X with rank r satisfies $\|X\|_F \le 1$. Then X can be approximately recovered from the random measurements $y_i r$ by a solution $\widehat{X}$ of the program (10.11). The recovery error satisfies*

$$\mathbb{E} \|\widehat{X} - X\|_2 \le CK^2 \sqrt{\frac{rd}{m}}.$$

The recovery error becomes small if

$$m \sim rd,$$

[1] The independence of entries can be relaxed. How?

if the hidden constant here is appropriately large. This allows us to recover low-rank matrices even when the number of measurements m is too small, i.e. when

$$m \ll d^2$$

and the matrix recovery problem (without rank assumption) is ill-posed.

Exercise 10.4.6. *Extend the matrix recovery result for* approximately *low-rank matrices. (Quantify approximately small rank in a convenient way.)*

The following is a matrix version of Exercise 10.7.

Exercise 10.4.7 (Low-rank matrix recovery by convex optimization)**.** *Show that an unknown matrix X or rank r can be approximately recovered by solving the convex optimization problem*

$$\text{minimize } \|X'\|_* \text{ s.t. } y_i = \langle A_i, X' \rangle \ \forall i = 1, \dots, m.$$

Exercise 10.4.8 (Rectangular matrices)**.** *Extend the matrix recovery result from quadratic to rectangular, $d_1 \times d_2$ matrices.*

10.5 Exact recovery

The guarantees for sparse recovery that we just developed can be dramatically improved. The recovery error of sparse signals x can actually be *zero*. In this section, we will deduce an exact recovery result from Escape Theorem 9.5.7.

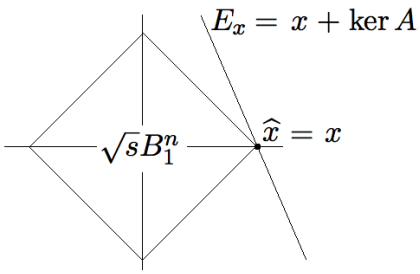
To see why such a strong, counter-intuitive result can make sense, let us look at the recovery problem from a geometric viewpoint illustrated by Figure 10.3. A solution $\widehat{x}$ of the program (10.6) must lie in the intersection of the prior set T, which in our case is the ℓ_1 ball $\sqrt{s}B_1^n$, and the affine subspace $E_x = x + \ker A$.

The ℓ_1 ball is a polyhedron, and the s-sparse unit vector x lies on the $s-1$-dimensional edge of that polyhedron, see Figure 10.5a.

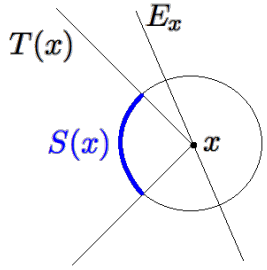
It could happen with non-zero probability that the random subspace E_x is *tangent* to the polyhedron at the point x. If this does happen, x is the only point of intersection between the ℓ_1 ball and E_x. In this case, it follows that the solution $\widehat{x}$ to the program (10.6) is exact:

$$\widehat{x} = x.$$

To justify this argument, all we need to check is that a random subspace E_x is tangent to the ℓ_1 ball with high probability. We can do this using



(a) Exact sparse recovery happens when the random subspace E_x is tangent to the ℓ_1 ball at the point x.



(b) The tangency occurs iff E_x is disjoint from the spherical part $S(x)$ of the tangent cone $T(x)$ of the ℓ_1 ball at point x.

Figure 10.5: Exact sparse recovery

Escape Theorem 9.5.7. To see a connection, look at what happens in a small neighborhood around the tangent point, see Figure 10.5b. The subspace E_x is tangent if and only if the *tangent cone* $T(x)$ (formed by all rays emanating from x toward the points in the ℓ_1 ball) intersects E_x at a single point x. Equivalently, this happens if and only if the *spherical part* $S(x)$ of the cone (the intersection of $T(x)$ with a small sphere centered at x) is disjoint from E_x. But this is exactly the conclusion of Escape Theorem 9.5.7!

Let us now formally state the exact recovery result. We shall consider the noiseless sparse recovery problem

$$y = Ax.$$

and try to solve it using the optimization program (10.7), i.e.

$$\text{minimize } \|x'\|_1 \text{ s.t. } y = Ax'. \tag{10.12}$$

Theorem 10.5.1 (Exact sparse recovery)**.** *Suppose the rows A_i of A are independent, isotropic and sub-gaussian random vectors, and let $K := \max_i \|A_i\|_{\psi_2}$. Then the following happens with probability at least $1 - 2\exp(-cm/K^4)$.*

Assume an unknown signal $x \in \mathbb{R}^n$ is s-sparse and the number of measurements m satisfies

$$m \geq CK^4 s \log n.$$

Then a solution $\widehat{x}$ of the program (10.12) is exact, i.e.

$$\widehat{x} = x.$$

To prove the theorem, we would like to show that the recovery error

$$h := \widehat{x} - x$$

is zero. Let us examine the vector h more closely. First we show that h has more "energy" on the support of x than outside it.

Lemma 10.5.2. *Let $S := \operatorname{supp}(x)$. Then*

$$\|h_{S^c}\|_1 \le \|h_S\|_1.$$

Here $h_S \in \mathbb{R}^S$ denotes the restriction of the vector $h \in \mathbb{R}^S$ onto a subset of coordinates $S \subset \{1, \dots, n\}$.

Proof. Since $\widehat{x}$ is the minimizer in the program (10.12), we have

$$\|\widehat{x}\|_1 \le \|x\|_1. \tag{10.13}$$

But there is also a lower bound

$$\begin{aligned}
\|\widehat{x}\|_1 &= \|x + h\|_1 \\
&= \|x_S + h_S\|_1 + \|x_{S^c} + h_{S^c}\|_1 \\
&= \|x + h_S\|_1 + \|h_{S^c}\|_1 \quad (\text{since } x_S = 0 \text{ and } x_{S^c} = 0) \\
&\ge \|x\|_1 - \|h_S\|_1 + \|h_{S^c}\|_1 \quad (\text{by triangle inequality}).
\end{aligned}$$

Substitute this bound into (10.13) and cancel $\|x\|_1$ on both sides to complete the proof. □

Lemma 10.5.3. *The error vector satisfies*

$$\|h\|_1 \le 2\sqrt{s}\|h\|_2.$$

Proof. Using Lemma 10.5.2 and then Hölder's inequality, we obtain

$$\|h\|_1 = \|h_S\|_1 + \|h_{S^c}\|_1 \le 2\|h_S\|_1 \le 2\sqrt{s}\|h_S\|_2.$$

Since trivially $\|h_S\|_2 \le \|h\|_2$, the proof is complete. □

Proof of Theorem 10.5.1. Assume that the recovery is not exact, i.e.

$$h = \widehat{x} - x \neq 0.$$

By Lemma 10.5.3, the normalized error $h/\|h\|_2$ lies in the set

$$T_s := \left\{ z \in S^{n-1} : \ \|z\|_1 \le 2\sqrt{s} \right\}.$$

Since also

$$Ah = A\widehat{x} - Ax = y - y = 0,$$

we have

$$\frac{h}{\|h\|_2} \in T_s \cap \ker A. \tag{10.14}$$

Escape Theorem 9.5.7 states that this intersection is empty with high probability, as long as

$$m \geq CK^4 w(T_s)^2.$$

Now,

$$w(T_s) \leq 2\sqrt{s} w(B_1^n) \leq C\sqrt{s \log n}, \tag{10.15}$$

where we used the bound (7.23) on the Gaussian width of the ℓ_1 ball. Thus, if $m \geq CK^4 s \log n$, the intersection in (10.14) is empty with high probability, which means that the inclusion in (10.14) can not hold. This contradiction implies that our assumption that $h \neq 0$ is false with high probability. The proof is complete. □

Exercise 10.5.4 (Improving the logarithmic factor)**.** *Show that the conclusion of Theorem 10.5.1 holds under a weaker assumption on the number of measurements, which is*

$$m \geq CK^4 s \log(2n/s).$$

Hint: Use the result of Exercise 10.3.7.

Exercise 10.5.5. *Give a geometric interpretation of the proof of Theorem 10.5.1, using Figure 10.5b. What does the proof say about the tangent cone $T(x)$? Its spherical part $S(x)$?*

Do we need to relax the constraint $y = Ax'$ in the program?

Exercise 10.5.6 (Noisy measuremments)**.** *Extend the result on sparse recovery (Theorem 10.5.1) for noisy measurements, where*

$$y = Ax + w.$$

Remark 10.5.7. Theorem 10.5.1 shows that one can effectively solve *underdetermined systems of linear equations* $y = Ax$ with $m \ll n$ equations in n variables, if the solution is sparse.

10.6 Lasso algorithm for sparse regression

10.6.1 Statistical formulation

In this section we will analyze an alternative method for sparse recovery. This method was originally developed in statistics for the equivalent problem of *sparse linear regression*, and it is called Lasso ("least absolute shrinkage and selection operator").

Let us recall the classical linear regression problem, which we described in Example 10.1.2. It is

$$Y = X\beta + w \tag{10.16}$$

where X is a known $m \times n$ matrix that contains a sample of predictor variables, $Y \in \mathbb{R}^m$ is a known vector that contains a sample of the values of the response variable, $\beta \in \mathbb{R}^n$ is an unknown coefficient vector that specifies the relationship between predictor and response variables, and w is a noise vector. We would like to recover β.

If we do not assume anything else, the regression problem can be solved by the method of *ordinary least squares*, which minimizes the ℓ_2-norm of the error over all candidates for β:

$$\text{minimize } \|Y - X\beta'\|_2 \text{ s.t. } \beta' \in \mathbb{R}^n. \tag{10.17}$$

Now let us make an extra assumption that β' is *sparse*, so that the response variable depends only on a few of the n predictor variables (e.g. the cancer depends on few genes). So, like in (10.5), we assume that

$$\|\beta\|_0 \leq s$$

for some $s \ll n$. As we argued in Section 10.3, the ℓ_0 is not convex, and its convex proxy is the ℓ_1 norm. This prompts us to modify the ordinary least squares program (10.17) by including a restriction on the ℓ_1 norm, which promotes sparsity in the solution:

$$\text{minimize } \|Y - X\beta'\|_2 \text{ s.t. } \|\beta'\|_1 \leq R, \tag{10.18}$$

where R is a parameter which specifies a desired sparsity level of the solution. The program (10.18) is one of the formulations of Lasso, the most popular statistical method for sparse linear regression. It is a convex program, and therefore is computationally tractable.

10.6.2 Mathematical formulation and guarantees

It would be convenient to return to the notation we used for sparse recovery instead of using the statistical notation in the previous section. So let us restate the linear regression problem (10.16) as

$$y = Ax + w$$

where A is a known $m \times n$ matrix, $y \in \mathbb{R}^m$ is a known vector, $x \in \mathbb{R}^n$ is an unknown vector that we are trying to recover, and $w \in \mathbb{R}^m$ is noise which is either fixed or random and independent of A. Then Lasso program (10.18) becomes

$$\text{minimize } \|y - Ax'\|_2 \text{ s.t. } \|x'\|_1 \le R. \tag{10.19}$$

We will prove the following guarantee of the performance of Lasso.

Theorem 10.6.1 (Performance of Lasso)**.** *Suppose the rows A_i of A are independent, isotropic and sub-gaussian random vectors, and let $K := \max_i \|A_i\|_{\psi_2}$. Then the following happens with probability at least $1 - 2\exp(-s \log n)$.*

Assume an unknown signal $x \in \mathbb{R}^n$ is s-sparse and the number of measurements m satisfies

$$m \ge CK^4 s \log n. \tag{10.20}$$

Then a solution $\widehat{x}$ of the program (10.19) with $R := \|x\|_1$ is accurate:

$$\|\widehat{x} - x\|_2 \le C\sigma \sqrt{\frac{s \log n}{m}},$$

where σ is such that the noise satisfies $\|w\|_2 \le \sigma\sqrt{m}$.

Remark 10.6.2 (Noise). Let us clarify the dependence of the recovery error on the noise. The condition $\|w\|_2 \le \sigma\sqrt{m}$ simply bounds the *average noise per measurement* by σ, since we can rewrite this condition as

$$\frac{\|w\|_2^2}{m} = \frac{1}{m} \sum_{i=1}^{m} w_i^2 \le \sigma^2.$$

Then, if the number of measurements is

$$m \sim s \log n,$$

Theorem 10.6.1 bounds the recovery error by the average noise per measurement σ. And if m is larger, the recovery error gets smaller.

Remark 10.6.3 (Exact recovery). In the noiseless model $y = Ax$ we have $w = 0$ and thus Lasso recovers x exactly, i.e.

$$\widehat{x} = x.$$

The proof of Theorem 10.6.1 will be similar to our proof of Theorem 10.5.1 on exact recovery, although instead of the Escape theorem we will use Matrix Deviation Inequality (Theorem 9.2.1) directly this time.

We would like to bound the norm of the error vector

$$h := \widehat{x} - x.$$

Exercise 10.6.4. *Check that h satisfies the conclusions of Lemmas 10.5.2 and 10.5.3, so we have*

$$\|h\|_1 \le 2\sqrt{s}\|h\|_2. \tag{10.21}$$

Hint: The proofs of these lemmas are based on the fact that $\|\widehat{x}\|_1 \le \|x\|_1$, which holds in our situation as well.

In case where the noise w is nonzero, we can not expect to have $Ah = 0$ like in Theorem 10.5.1. (Why?) Instead, we can give an upper and a lower bounds for $\|Ah\|_2$.

Lemma 10.6.5 (Upper bound on $\|Ah\|_2$)**.** *We have*

$$\|Ah\|_2^2 \le 2\left\langle h, A^{\mathsf{T}} w \right\rangle. \tag{10.22}$$

Proof. Since $\widehat{x}$ is the minimizer of Lasso program (10.19), we have

$$\|y - A\widehat{x}\|_2 \le \|y - Ax\|_2.$$

Let us express both of this inequality in terms of h and w, using that $y = Ax + w$ and $h = \widehat{x} - x$:

$$\begin{aligned} y - A\widehat{x} &= Ax + w - A\widehat{x} = w - Ah; \\ y - Ax &= w. \end{aligned}$$

So we have

$$\|w - Ah\|_2 \le \|w\|_2.$$

Square both sides:

$$\|w\|_2^2 - 2\langle w, Ah\rangle + \|Ah\|_2^2 \le \|w\|_2^2.$$

Simplifying this bound completes the proof. □

Lemma 10.6.6 (Lower bound on $\|Ah\|_2$)**.** *With probability at least* $1 - 2\exp(-4s\log n)$*, we have*

$$\|Ah\|_2^2 \geq \frac{m}{4}\|h\|_2^2.$$

Proof. By (10.21), the normalized error $h/\|h\|_2$ lies in the set

$$T_s := \left\{ z \in S^{n-1} : \ \|z\|_1 \leq 2\sqrt{s} \right\}.$$

Use Matrix Deviation Inequality in its high-probability form (Exercise 9.2.3) with $u = 2\sqrt{s\log n}$. It yields that, with probability at least $1-2\exp(-4s\log n)$,

$$\begin{aligned} \sup_{z\in T_s} \left| \|Az\|_2 - \sqrt{m} \right| &\leq C_1 K^2 \left(w(T_s) + 2\sqrt{s\log n} \right) \\ &\leq C_2 K^2 \sqrt{s\log n} \quad \text{(recalling (10.15))} \\ &\leq \frac{\sqrt{m}}{2} \quad \text{(by assumption on } m\text{)}. \end{aligned}$$

To make the last line work, choose the absolute constant C in (10.20) large enough. By triangle inequality, this implies that

$$\|Az\|_2 \geq \frac{\sqrt{m}}{2} \quad \text{for all } z \in T_s.$$

Substituting $z := h/\|h\|_2$, we complete the proof. □

The last piece we need to prove Theorem 10.6.1 is an upper bound on the right hand side of (10.22).

Lemma 10.6.7. *With probability at least* $1 - 2\exp(-4s\log n)$*, we have*

$$\left\langle h, A^{\mathsf{T}} w \right\rangle \leq CK\|h\|_2\|w\|_2\sqrt{s\log n}. \tag{10.23}$$

Proof. As in the proof of Lemma 10.6.6, the normalized error satisfies

$$z = \frac{h}{\|h\|_2} \in T_s.$$

So, dividing both sides of (10.23) by $\|h\|_2$, we see that it is enough to bound the supremum random process

$$\sup_{z\in T_s} \left\langle z, A^{\mathsf{T}} w \right\rangle$$

with high probability. We are going to use Talagrand's comparison inequality (Corollary 8.5.3). This result applies for random processes with subgaussian increments, so let us check this condition first.

Exercise 10.6.8. *Show that the random process*

$$X_t := \left\langle t, A^{\mathsf{T}} w \right\rangle, \quad t \in \mathbb{R}^n,$$

has sub-gaussian increments, and

$$\|X_t - X_s\|_{\psi_2} \le CK\|w\|_2 \cdot \|t - s\|_2.$$

Hint: Recall the proof of sub-gaussian Chevet's inequality (Theorem 8.6.1).

Now we can use Talagrand's comparison inequality in the high-probability form (Exercise 8.5.5) for $u = 2\sqrt{s \log n}$. We obtain that, with probability at least $1 - 2\exp(-4s \log n$,

$$\begin{aligned} \sup_{z \in T_s} \left\langle z, A^{\mathsf{T}} w \right\rangle &\le C_1 K \|w\|_2 \left(w(T_s) + 2\sqrt{s \log n} \right) \\ &\le C_2 K \|w\|_2 \sqrt{s \log n} \quad \text{(recalling (10.15)).} \end{aligned}$$

This completes the proof of Lemma 10.23. □

Proof of Theorem 10.6.1. Put together the bounds in Lemmas 10.6.5, 10.6.6 and 10.23. By union bound, we have that with probability at least $1 - 2\exp(-4s \log n)$,

$$\frac{m}{4} \|h\|_2^2 \le CK \|h\|_2 \|w\|_2 \sqrt{s \log n}.$$

Solving for $\|h\|_2$, we obtain

$$\|h\|_2 \le CK \frac{\|w\|_2}{\sqrt{m}} \cdot \sqrt{\frac{s \log n}{m}}.$$

This completes the proof of Theorem 10.6.1. □

Exercise 10.6.9 (Improving the logarithmic factor)**.** *Show that Theorem 10.6.1 holds if* $\log n$ *is replaced by* $\log(n/s)$*, thus giving a stronger guarantee.*

Hint: Use the result of Exercise 10.3.7.

Exercise 10.6.10. *Deduce the exact recovery guarantee (Theorem 10.5.1) directly from the Lasso guarantee (Theorem 10.6.1). The probability that you get could be a bit weaker.*

Another popular form of Lasso program (10.19) is the following *unconstrained version*:

$$\text{minimize } \|y - Ax'\|_2 + \lambda \|x'\|_1, \tag{10.24}$$

This is a convex optimization problem, too. Here λ is a parameter which can be adjusted depending on the desired level of sparsity. The method of Lagrange multipliers shows that the constrained and unconstrained versions of Lasso are equivalent for appropriate R and λ. This however does not immediately tell us how to choose λ. The following exercise settles this question.

Consider providing hints from folder Lasso-unconstrained

Exercise 10.6.11 (Unconstrained Lasso)**.** *Assume number of measurements satisfy*

$$m \gtrsim s \log n.$$

Choose the parameter λ so that $\lambda \gtrsim \sqrt{\log n}\|w\|_2$. Then, with high probability, the solution $\widehat{x}$ of unconstrained Lasso (10.24) satisfies

$$\|\widehat{x} - x\|_2 \lesssim \frac{\lambda\sqrt{s}}{m}.$$

Chapter 11

Supplement: Dvoretzky-Milman's Theorem

Write intro

11.1 Deviations of random matrices with respect to general norms

In this section we generalize the matrix deviation inequality from Section 9.2. We will replace the Euclidean norm by any positive homogeneous, subadditive function.

Definition 11.1.1. *Let V be a vector space. A function $f : V \to \mathbb{R}$ is called* positive homogeneous *if*

$$f(\alpha x) = \alpha f(x) \quad \text{for all } \alpha \geq 0 \text{ and } x \in V.$$

The function f is called subadditive *if*

$$f(x + y) \leq f(x) + f(y) \quad \text{for all } x, y \in V.$$

Note that despite being called "positive homogeneous", f is allowed to take negative values. ("Positive" here applies to the multiplier α in the definition.)

Example 11.1.2. 1. Any *norm* on a vector space is positive homogeneous and subadditive. The subadditivity is nothing else than triangle inequality in this case.

2. Clearly, any *linear functional* on a vector space is positive homogeneous and subadditive. In particular, for any fixed vector $y \in \mathbb{R}^m$, the function $f(x) = \langle x, y \rangle$ is a positive homogeneous and subadditive on $\mathbb{R}^m$.

3. Consider a bounded set $S \subset \mathbb{R}^m$ and define the function

$$f(x) := \sup_{y \in S} \langle x, y \rangle, \quad x \in \mathbb{R}^m. \tag{11.1}$$

   Then f is a positive homogeneous and subadditive on $\mathbb{R}^m$. This function is sometimes called the *support function* of S.

Exercise 11.1.3. *Check that the function $f(x)$ in part 11.1 of Example 11.1.2 is positive homogeneous and subadditive.*

Exercise 11.1.4. *Let $f : V \to \mathbb{R}$ be a subadditive function on a vector space V. Show that*

$$f(x) - f(y) \le f(x - y) \quad \text{for all } x, y \in V. \tag{11.2}$$

We are ready to state the main result of this section.

Theorem 11.1.5 (General matrix deviation inequality)**.** *Let A be an $m \times n$ Gaussian random matrix with i.i.d. $N(0,1)$ entries. Let $f : \mathbb{R}^m \to \mathbb{R}$ be a positive homogenous and subadditive function, and let $b \in \mathbb{R}$ be such that*

$$f(x) \le b\|x\|_2 \quad \text{for all } x \in \mathbb{R}^n. \tag{11.3}$$

Then for any subset $T \subset \mathbb{R}^n$, we have

$$\mathbb{E} \sup_{x \in T} \Big| f(Ax) - \mathbb{E}\, f(Ax) \Big| \le Cb\gamma(T).$$

Here $\gamma(T)$ is the Gaussian complexity introduced in (7.27).

This theorem generalizes the matrix deviation inequality (in the form we gave in Exercise 9.2.2) to arbitrary seminorms.

Exactly as in Section 9.2, Theorem 11.1.5 would follow from Talagrand's comparison inequality once we show that the random process $X_x := f(Ax) - \mathbb{E}\, f(Ax)$ has sub-gaussian increments. Let us do this now.

Theorem 11.1.6 (Sub-gaussian increments)**.** *Let A be an $m \times n$ Gaussian random matrix with i.i.d. $N(0,1)$ entries, and let $f : \mathbb{R}^m \to \mathbb{R}$ be a positive homogenous and subadditive function satisfying (11.3). Then the random process*

$$X_x := f(Ax) - \mathbb{E}\, f(Ax)$$

has sub-gaussian increments with respect to the Euclidean norm, namely

$$\|X_x - X_y\|_{\psi_2} \le Cb\|x - y\|_2 \quad \text{for all } x, y \in \mathbb{R}^n. \tag{11.4}$$

Exercise 11.1.7. *Deduce the general matrix deviation inequality (Theorem 11.1.5) from Talagrand's comparison inequality (in the form of Exercise 8.5.4) and Theorem 11.1.6.*

Proof of Theorem 11.1.6. Without loss of generality we may assume that $b = 1$. (Why?) Just like in the proof of Theorem 9.1.1, let us first assume that

$$\|x\|_2 = \|y\|_2 = 1.$$

In this case, the inequality in (11.4) we want to prove becomes

$$\|f(Ax) - f(Ay)\|_{\psi_2} \le C\|x - y\|_2. \tag{11.5}$$

Step 1. Creating independence. Consider the vectors

$$u := \frac{x+y}{2}, \quad v := \frac{x-y}{2} \tag{11.6}$$

(see Figure 11.1). Then

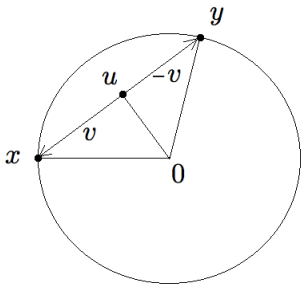


Figure 11.1: Signal recovery: the signal x and the solution $\widehat{x}$ lie in the prior set T and in the affine subspace E_x.

$$x = u + v, \quad y = u - v$$

and thus

$$Ax = Au + Av, \quad Ay = Au - Av.$$

Since the vectors u and v are orthogonal (check!), the Gaussian random vectors Au and Av are independent.

Include such exercise to the gaussian section, and refer to it.

Step 2. Using Gaussian concentration. Let us condition on $a := Au$ and study the conditional distribution of

$$f(Ax) = f(a + Av).$$

refer

By rotation invariance, $a + Av$ is a Gaussian random vector that we can expressed as

$$a + Av = a + \|v\|_2 \, g, \quad \text{where} \quad g \sim N(0, I_m).$$

We claim that $f(a + \|v\|_2 \, g)$ as a function of g is Lipschitz with respect to the Euclidean norm on $\mathbb{R}^m$, and

$$\|f\|_{\text{Lip}} \le \|v\|_2. \tag{11.7}$$

To check this, fix $t, s \in \mathbb{R}^m$ and note that

$$\begin{aligned} f(t) - f(s) &= f(a + \|v\|_2 \, t) - f(a + \|v\|_2 \, s) \\ &\le f(\|v\|_2 \, t - \|v\|_2 \, s) \quad \text{(by (11.2))} \\ &= \|v\|_2 \, f(t - s) \quad \text{(by positive homogeneity)} \\ &\le \|v\|_2 \, \|t - s\|_2 \quad \text{(using (11.3) with } b = 1), \end{aligned}$$

and (11.7) follows.

Concentration in the Gauss space (Theorem 5.2.1) then yields

$$\|f(g) - \mathbb{E} \, f(g)\|_{\psi_2(a)} \le C\|v\|_2,$$

or

$$\big\|f(a + Av) - \mathbb{E}_a \, f(a + Av)\big\|_{\psi_2(a)} \le C\|v\|_2, \tag{11.8}$$

where the index "a" reminds us that these bounds are valid for the conditional distribution, with $a = Au$ fixed.

Step 2. Removing the conditioning. Since random vector $a - Av$ has the same distribution as $a + Av$ (why?), it satisfies the same bound.

$$\big\|f(a - Av) - \mathbb{E}_a \, f(a - Av)\big\|_{\psi_2(a)} \le C\|v\|_2. \tag{11.9}$$

Subtract (11.9) from (11.8), use triangle inequality and the fact that the expectations are the same; this gives

$$\big\|f(a + Av) - f(a - Av)\big\|_{\psi_2(a)} \le 2C\|v\|_2.$$

This bound is for the conditional distribution, and it holds for any fixed realization of a random variable $a = Au$. Therefore, it holds for the original distribution, too:

$$\big\|f(Au + Av) - f(Au - Av)\big\|_{\psi_2} \le 2C\|v\|_2.$$

(Why?) Passing back to the x, y notation by (11.6), we obtain the desired inequality (11.5).

The proof is complete for the unit vectors x, y; Exercise 11.1.8 below extends it for the general case. □

Exercise 11.1.8 (Non-unit x, y)**.** *Extend the proof above to general (not necessarily unit) vectors x, y. Hint: Follow the argument in Section 9.1.4.*

Remark 11.1.9. It is an open question if Theorems 11.1.5 and 11.1.6 hold for general sub-gaussian matrices A.

Exercise 11.1.10 (Anisotropic distributions)**.** *Extend Theorems 11.1.5 to $m \times n$ matrices A whose columns are independent $N(0, \Sigma)$ random vectors, where Σ is a general covariance matrix. Show that*

$$\mathbb{E} \sup_{x \in T} \Big| f(Ax) - \mathbb{E}\, f(Ax) \Big| \le Cb\gamma(\Sigma^{1/2} T).$$

Exercise 11.1.11 (Tail bounds)**.** *Prove a high-probability version of Theorem 11.1.5. Hint: Follow Exercise 9.2.3.*

11.2 Johnson-Lindenstrauss embeddings and sharper Chevet inequality

Just like the original matrix deviation inequality from Chapter 9, the general Theorem 9.2.1 has many consequences. In this section we consider an application related to Johnson-Lindenstrauss lemma and a sharpened form of Chevet's inequality.

11.2.1 Johnson-Lindenstrauss Lemma for general norms

Applying the general matrix deviation inequality similarly to Section 9.4, it is quite straightforward to deduce the following.

Exercise 11.2.1. *State and prove a version of Johnson-Lindenstrauss Lemma for a general norm (as opposed to the Euclidean norm) on $\mathbb{R}^m$.*

Exercise 11.2.2 (Johnson-Lindenstrauss Lemma for ℓ_1 norm)**.** *Specialize the previous exerise to the ℓ_1 and ℓ_∞ norms. Thus, $\mathcal{X}$ be a set of N points in $\mathbb{R}^n$, let A be an $m \times n$ Gaussian matrix with i.i.d. $N(0,1)$ entries, and let $\varepsilon \in (0,1)$.*

Suppose that

$$m \geq C(\varepsilon) \log N.$$

Show that with high probability the matrix $Q := \sqrt{\pi/2} \cdot m^{-1} A$ satisfies

$$(1-\varepsilon)\|x-y\|_2 \leq \|Qx - Qy\|_1 \leq (1+\varepsilon)\|x-y\|_2 \quad \text{for all } x, y \in \mathcal{X}.$$

This conclusion is very similar to the original Johnson-Lindenstrauss Lemma (Theorem 5.3.1), except the distance between the projected points is measured in the ℓ_1 norm.

Exercise 11.2.3 (Johnson-Lindenstrauss embedding into ℓ_∞)**.** *Use the same notation as in the previous exercise, but assume this time that*

$$m \geq N^{C(\varepsilon)}.$$

Show that with high probability the matrix $Q := (\log m)^{-1/2} A$ satisfies

$$(1-\varepsilon)\|x-y\|_2 \leq \|Qx - Qy\|_\infty \leq (1+\varepsilon)\|x-y\|_2 \quad \text{for all } x, y \in \mathcal{X}.$$

Note that in this case $m \geq N$, so Q gives an almost isometric embedding *(rather than a projection) of the set $\mathcal{X}$ into ℓ_∞.*

11.2.2 Two-sided Chevet's inequality

The general matrix deviation inequality will help us sharpen Chevet's inequality, which we originally proved in Section 8.6.

Theorem 11.2.4 (General Chevet's inequality)**.** *Let A be an $m \times n$ Gaussian random matrix with i.i.d. $N(0,1)$ entries. Let $T \subset \mathbb{R}^n$ and $S \subset \mathbb{R}^m$ be arbitrary bounded sets. Then*

$$\mathbb{E} \sup_{x \in T} \Big| \sup_{y \in S} \langle Ax, y \rangle - w(S)\|x\|_2 \Big| \leq C\gamma(T) \operatorname{rad}(S).$$

This is a sharper, two-sided form of Chevet's inequality (Theorem 8.6.1).

Proof. Let us apply general matrix deviation inequality (Theorem 11.1.5) for the function f defined in (11.1), i.e. for

$$f(x) := \sup_{y \in S} \langle x, y \rangle .$$

To do this, we need to compute b for which (11.3) holds Fix $x \in \mathbb{R}^m$ and use Cauchy-Schwarz inequality to get

$$f(x) \le \sup_{y \in S} \|x\|_2 \|y\|_2 = \mathrm{rad}(S)\|x\|_2.$$

Thus (11.3) holds with $b = \mathrm{rad}(S)$.

It remains to compute $\mathbb{E} f(Ax)$ appearing in the conclusion of Theorem 11.1.5. By rotation invariance of Gaussian distribution, the random vector Ax has the same distribution as $g\|x\|_2$ where $g \in N(0, I_m)$. Then refer

$$\begin{aligned} \mathbb{E} f(Ax) &= \mathbb{E} f(g)\, \|x\|_2 \quad \text{(by positive homogeneity)} \\ &= \mathbb{E} \sup_{y \in S} \langle g, y \rangle\ \|x\|_2 \quad \text{(by definition of } f\text{)} \\ &= w(S)\|x\|_2 \quad \text{(by definition of the Gaussian width).} \end{aligned}$$

Substituting this into the conclusion of Theorem 11.1.5, we complete the proof. □

11.3 Dvoretzky-Milman's Theorem

Dvoretzky-Milman's Theorem is a remarkable result, which says if you project a bounded set in $\mathbb{R}^n$ onto a random subspace of suitably low dimension, then the convex hull of the projection will be *approximately a round ball* with high probability. Figure... shows a random projection of a unit cube.... Write, add figure

We will deduce Dvoretzky-Milman's Theorem from the two-sided Chevet's inequality, which we proved in Section 11.2.2.

11.3.1 Gaussian images of sets

It will be more convenient for us to work with Gaussian random projections than with ordinary projections. Here is a very general result that compares the Gaussian projection of a general set to a Euclidean ball. refer

Theorem 11.3.1 (Random projections of sets)**.** *Let A be an $m \times n$ Gaussian random matrix with i.i.d. $N(0,1)$ entries, and $T \subset \mathbb{R}^n$ be a bounded set. Then the following holds with probability at least* 0.99*:*

$$r_- B_2^m \subset \mathrm{conv}(AT) \subset r_+ B_2^m$$

where

$$r_\pm := w(T) \pm C\sqrt{m}\ \mathrm{rad}(T).$$

The lest inclusion holds only if $r_- \ge 0$; the right inclusion, always.

We will shortly deduce this theorem from two-sided Chevet's inequality. The following exercise will provide the link between the two results. It asks you to show that the support function (11.1) of general set S is the ℓ_2 norm if and only if S is the Euclidean ball; there is also a stability version of this equivalence.

Exercise 11.3.2 (Almost Euclidean balls and support functions)**.** *1. Let $V \subset \mathbb{R}^m$ be a bounded set. Show that $V = B_2^m$ if and only if*

$$\sup_{x \in V} \langle x, y \rangle = \|y\|_2 \quad \text{for all } y \in \mathbb{R}^m.$$

2. Let $V \subset \mathbb{R}^m$ be a bounded set and $r_-, r_+ \geq 0$. Show that the inclusion

$$r_- B_2^m \subset \operatorname{conv}(V) \subset r_+ B_2^m$$

holds if and only if

$$r_- \|y\|_2 \leq \sup_{x \in V} \langle x, y \rangle \leq r_+ \|y\|_2 \quad \text{for all } y \in \mathbb{R}^m.$$

Proof of Theorem 11.3.1. Let us write the two-sided Chevet's inequality in the following form:

$$\mathbb{E} \sup_{y \in S} \left| \sup_{x \in T} \langle Ax, y \rangle - w(T) \|y\|_2 \right| \leq C\gamma(S) \operatorname{rad}(T).$$

where $T \subset \mathbb{R}^n$ and $S \subset \mathbb{R}^m$. (To get this form, use Theorem 11.2.4 for T and S swapped with each other and for A^T instead of A – do this!)

Exercise?

Choose S to be the sphere S^{m-1} and recall that it Gaussian complexity $\gamma(S) \leq \sqrt{m}$. Then, by Markov's inequality, the following holds with probability at least 0.99:

refer

$$\left| \sup_{x \in T} \langle Ax, y \rangle - w(T) \|y\|_2 \right| \leq C\sqrt{m} \operatorname{rad}(T) \quad \text{for every } y \in S^{m-1}.$$

Use triangle inequality and recall the definition of $r_\pm$ to get

$$r_- \leq \sup_{x \in T} \langle Ax, y \rangle \leq r_+ \quad \text{for every } y \in S^{m-1}.$$

By homogeneity, this is equivalent to

$$r_- \|y\|_2 \leq \sup_{x \in T} \langle Ax, y \rangle \leq r_+ \|y\|_2 \quad \text{for every } y \in \mathbb{R}^m.$$

(Why?) Finally, note that

$$\sup_{x \in T} \langle Ax, y \rangle = \sup_{x \in AT} \langle x, y \rangle$$

and apply Exercise 11.3.2 for $V = AT$ to complete the proof. □

11.3.2 Dvoretzky-Milman's Theorem

Theorem 11.3.3 (Dvoretzky-Milman's theorem: Gaussian form)**.** *Let A be an $m \times n$ Gaussian random matrix with i.i.d. $N(0,1)$ entries, $T \subset \mathbb{R}^n$ be a bounded set, and let $\varepsilon \in (0,1)$. Suppose*

$$m \le c\varepsilon^2 d(T)$$

where $d(T)$ is the statistical dimension of T introduced in Section 7.7.4. Then with probability at least 0.99, we have

$$(1-\varepsilon)B \subset \operatorname{conv}(AT) \subset (1+\varepsilon)B$$

where B is a Euclidean ball with radius $w(T)$.

Proof. Translating T is necessary, we can assume that T contains the origin. Apply Theorem 11.3.1. All that remains to check is that $r_- \ge (1-\varepsilon)w(T)$ and $r_+ \le (1+\varepsilon)w(T)$, which by definition would follow if

$$C\sqrt{m}\,\operatorname{rad}(T) \le \varepsilon w(T). \tag{11.10}$$

To check this inequality, recall that by assumption and Definition 7.7.8 we have

$$m \le c\varepsilon^2 d(T) \le \frac{\varepsilon^2 w(T)^2}{\operatorname{diam}(T)^2}$$

provided the absolute constant $c > 0$ is chosen sufficiently small. Next, since T contains the origin, $\operatorname{rad}(T) \le \operatorname{diam}(T)$. (Why?) This implies (11.10) and completes the proof. □

Remark 11.3.4. As is obvious from the proof, if T contains the origin then the Euclidean ball B can be centered at the origin, too. Otherwise, the center of B can be chosen as Tx_0, where $x_0 \in T$ is any fixed point.

With high probability?

Example 11.3.5 (Projections of the cube). Consider the cube

$$T = [-1,1]^n = B_\infty^n.$$

Recall that

$$w(T) = \sqrt{\frac{2}{\pi}} \cdot n;$$

recall (7.22). Since $\operatorname{diam}(T) = 2\sqrt{n}$, that the statistical dimension of the cube is

$$d(T) \sim \frac{d(T)^2}{\operatorname{diam}(T)^2} \sim n.$$

Apply Theorem 11.3.3. If $m \le c\varepsilon^2 n$ then with high probability we have

$$(1-\varepsilon)B \subset \operatorname{conv}(AT) \subset (1+\varepsilon)B$$

where B is a Euclidean ball with radius $\sqrt{2/\pi} \cdot n$.

In words, *a random Gaussian projection of the cube onto a subspace of dimension $m \sim n$ is close to a round ball.* Figure 11.2 illustrates this remarkable fact.

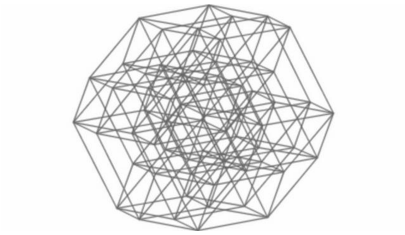


Figure 11.2: A random projection of a 6-dimensional cube onto the plane

Exercise 11.3.6 (Gaussian cloud)**.** *Consider a Gaussian cloud of n points in $\mathbb{R}^m$, which is formed by i.i.d. random vectors $g_1, \dots, g_n \sim N(0, I_m)$. Supose that $n \ge \exp(Cn)$ with large enough absolute constant C. Show that with high probability, the convex hull the Gaussian cloud is approximately a Euclidean ball with radius $\sim \log n$. See Figure 11.3 for illustration.*

Hint: Set T to be the canonical basis $\{e_1, \dots, e_n\}$ in $\mathbb{R}^n$, represent the points as $g_i = Te_i$, and apply Theorem 11.3.3.

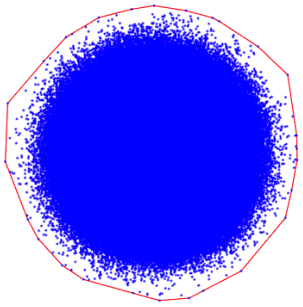


Figure 11.3: A gaussian cloud of 10^7 points on the plane, and its convex hull.

Exercise 11.3.7 (Projections of ellipsoids)**.** *Consider the ellipsoid $\mathcal{E}$ in $\mathbb{R}^n$ given as a linear image of the unit Euclidean ball, i.e.*

$$\mathcal{E} = S(B_2^n)$$

where S is an $n \times n$ matrix. Let A be the $m \times n$ Gaussian matrix with i.i.d. $N(0,1)$ entries. Suppose that

$$m \gtrsim r(S)$$

where $r(S)$ is the stable rank of S (recall Definition 7.7.13). Show that with high probability, the Gaussian projection $A(\mathcal{E})$ of the ellipsoid is almost a round ball with radius $\|S\|_F$:

$$A(\mathcal{E}) \approx \|S\|_F B_2^n.$$

Hint: First replace in Theorem 11.3.3 the Gaussian width $w(T)$ with the quantity $h(T) = (\mathbb{E} \sup_{t \in T} \langle g, t \rangle^2)^{1/2}$, which we discussed in (7.24) and which is easier to compute for ellipsoids.

Exercise 11.3.8 (Random projection in the Grassmanian)**.** *Prove a version of Dvoretzky-Milman's theorem for the projection P onto a random m-dimensional subspace in $\mathbb{R}^n$. Under the same assumptions, the conclusion should be that*

$$(1-\varepsilon)B \subset \operatorname{conv}(AT) \subset (1+\varepsilon)B$$

where B is a Euclidean ball with radius $w_s(T)$. (Recall that $w_s(T)$ is the spherical width of T, which we introduced in (7.20).)

Summary of random projections of geometric sets

It is useful to compare Dvoretzky-Milman's theorem to our earlier estimates on the diameter of random projections of geometric sets, which we developed in Section 7.8. We found that a random projection P of a set T onto an m-dimensional subspace in $\mathbb{R}^n$ satisfies a phase transition. For $m \gtrsim d(T)$, the projection shrinks the size of T by the factor of order $\sqrt{m/n}$, i.e.

$$\operatorname{diam}(PT) \lesssim \sqrt{\frac{m}{n}} \quad \text{if } m \geq d(T).$$

For smaller dimensions, $m \lesssim d(T)$, the size surprisingly stops shrinking. All we can say is that

$$\operatorname{diam}(PT) \lesssim w_s(T) \sim \frac{w(T)}{\sqrt{n}} \quad \text{if } m \leq d(T),$$

see Section 7.8.1.

Dvoretzky-Milman's theorem explains why the size of T stops shrinking for $m \lesssim d(T)$. Indeed, in this regime the projection PT is *approximately*

the round ball of radius of order $w_s(T)$ (see Exercise 11.3.8), regardless how small m is.

Let us summarize our findings. *A random projection of a set T in $\mathbb{R}^n$ onto an m-dimensional subspace shrinks the size of T by the factor $\sqrt{m/n}$ if $m \gtrsim d(T)$. For smaller m, the projection becomes approximately a round ball of diameter $\sim w_s(T)$, and does not shrink with m.*

Bibliography

[1] S. Boucheron, G. Lugosi, P. Massart, *Concentration inequalities. A nonasymptotic theory of independence.* With a foreword by Michel Ledoux. Oxford University Press, Oxford, 2013.

[2] M. Ledoux, *The concentration of measure phenomenon.* Mathematical Surveys and Monographs, 89. American Mathematical Society, Providence, RI, 2001.

[3] M. Ledoux, M. Talagrand, *Probability in Banach spaces. Isoperimetry and processes.* Ergebnisse der Mathematik und ihrer Grenzgebiete (3), 23. Springer-Verlag, Berlin, 1991.

[4] V. D. Milman, G. Schechtman, *Asymptotic theory of finite-dimensional normed spaces.* With an appendix by M. Gromov. Lecture Notes in Mathematics, 1200. Springer-Verlag, Berlin, 1986.

Index