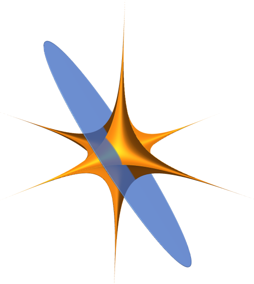
High-Dimensional Probability

An Introduction with Applications in Data Science

Second Edition

Roman Vershynin

November 28, 2025



https://www.math.uci.edu/~rvershyn/

This material will be published by Cambridge University Press as *High-Dimensional Probability, 2nd Edition* by Roman Vershynin. This pre-publication version is free to view and download for personal use only. Not for re-distribution, re-sale or use in derivative works. ©Roman Vershynin 202X.

If you want to be notified when the printed version is available, let me know: rvershyn@uci.edu

Contents

Preface to the Second Edition

Who is this book for?

This textbook is aimed at doctoral students, advanced master's students, and beginning researchers in mathematics, statistics, computer science, electrical engineering, and related fields, who seek to deepen their understanding of probabilistic methods commonly used in modern data science research. It can be used for self-study or as a textbook for a second probability course with data science applications.

Why this book?

Data science is evolving rapidly, and probabilistic methods are key to these advances. A typical graduate probability course no longer provides the mathematical sophistication needed for early-career data science researchers. This book aims to fill that gap, presenting essential probabilistic methods and results for mathematical data scientists.

What is this book about?

High-dimensional probability studies random objects in $\mathbb{R}^n$ – like random vectors and matrices – where the dimension n can be very large. This book builds some foundational tools for analyzing such objects, including concentration inequalities, covering and packing, decoupling and symmetrization, chaining and comparison for stochastic processes, empirical processes, VC theory, and more.

The theory is integrated with applications to covariance estimation, semidefinite programming, networks, elements of statistical learning, error-correcting codes, clustering, dimension reduction, and more.

This book covers only a fraction of high-dimensional probability, with just a few data science examples. Each chapter ends with a *Notes* section that points to other resources on the topic.

Are you ready?

To read this book, you will need a solid knowledge of probability theory (at the masters or doctoral level), strong undergraduate linear algebra, and some familiarity with metric, normed, and Hilbert spaces. Measure theory is not required.

Exercises and hints

Each chapter ends with exercises, most of which have *hints* in the back of the book. The difficulty of exercises is indicated by coffee cups you need to solve them, ranging from trivial (☕) to challenging (☕☕☕☕). If you are studying this book on your own, make sure to work through the exercises – they will help build your skills and reveal many extra storylines you won't find in the main text!

What's new in the second edition?

I thoroughly revised the book with the student — you — in mind. The text is now better suited for beginners: more streamlined, self-contained, and focused. Here are the major updates:

1. This edition adds 200 new exercises, many covering tools and applications beyond the main text – such as expanders, Le Cam's method, entropy, the cut norm and its semidefinite relaxation, Gaussian mixture models, matrix sketching, the nuclear norm, one-bit quantization, the small ball method, Lasso regression, and much more.
2. All exercises have been moved to the end of each chapter. The book is now more self-contained and the material flows more smoothly.
3. Hints are now more extensive and are provided for most nontrivial problems.
4. To help beginners, Chapter 1 (on analysis and probability) has been significantly expanded and Section 4.1 (on linear algebra) fully rewritten.
5. The final three chapters are now merged into a streamlined Chapter 9, with minimal content loss.
6. The entire text has been revised for a friendlier tone and clearer focus.

But it's still the same book. I worked hard to keep it thin – short enough for one semester or so.

Acknowledgements

The feedback from many colleagues and students was instrumental in shaping this book and its second edition. I am thankful to Pedro Abdalla, Karthik Abinav, Diego Armentano, Florent Benaych-Georges, Aaron Berk, Jennifer Bryson, Kwanghee Choi, Soham De, Hu Fu, Lukas Grätz, Remi Gribonval, Nick Harvey, Yiyun He, Harrie Hendriks, Ping Hsu, Mike Izbicki, T.S. Jayram, Adam Jozefiak, Valeriya Kotelnikova, Jana Krägel, Joris Kühl, Yi Li, Chris Liaw, George Linderman, Galyna Livshyts, Hengrui Luo, Cong Ma, Pete Morcos, Jelani Nelson, Adrien Peltzer, Yaniv Plan, Sikander Randhawa, Karthik Sankararaman, Ekkehard Schnoor, Yifei Shen, Nils Siebken, Martin Spindler, Dominik Stöger, Ho Boon Suan, Tim Sullivan, Terence Tao, Ulrich Terstiege, Morten Thonack, Joel Tropp, Katarzyna Wyczesany, Xinyuan Xie, Haoshu Xu, Robert Zimmerman, who generously pointed out typos, omissions and inaccuracies, and made helpful suggestions.

Special thanks to the colleagues and students who took the time to carefully proofread the book or shared extensive feedback that led to major improvements: Sjoerd Dirksen, Larry Goldstein, Wu Han, Franz Haniel, Ichiro Hashimoto, Jake Knigge, Aryeh Kontorovich, Wai-Kit Lam, Yi Li, Mark Meckes, Abbas Mehrabian, Holger Rauhut, Michael Scheutzow, Mahdi Soltanolkotabi, and Han Wu.

The second edition grew out of my experience teaching high-dimensional probability remotely at Kyiv National University in 2022 and 2023, during the Russian invasion of Ukraine. The extraordinary commitment of Ukrainian students to studying mathematics in the midst of war deeply moved me.

If you are willing to share your experiences studying or teaching from this book, or if you have suggestions for improving it, I'd love to hear from you:

rvershyn@uci.edu

With warm wishes and high hopes,

Roman Vershynin
Irvine, California

AI usage disclosure: ChatGPT was used to rephrase and polish text (including this disclosure), and occasionally to assist in generating MATLAB and Mathematica code for some figures.

Appetizer: Using Probability to Cover a Set

Let's begin with an elegant example of how probabilistic reasoning can help in geometry – feel free to skip it if you'd rather jump to the main content.

A *convex combination* of points $z_1, \ldots, z_m \in \mathbb{R}^n$ is a linear combination with coefficients that are nonnegative and sum to 1, i.e. it is a sum of the form

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

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 0.1 for illustration.

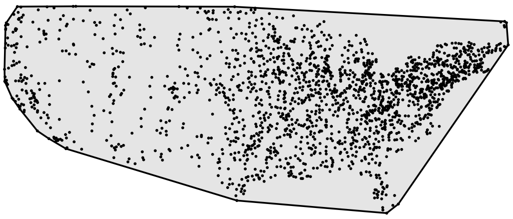


Figure 0.1 The shaded area is the convex hull of the U.S. cities.

The number m of elements defining a convex combination in $\mathbb{R}^n$ is not restricted a priori. However, the classical Caratheodory theorem states that one can always take $m \leq n + 1$:

Theorem 0.0.1 (Caratheodory 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 $n + 1$ is tight – for example, it is exactly what you need for a simplex (a set of $n + 1$ points in general position). But what if we are okay with just approximating the point instead of representing it exactly? Then we can actually get away with using way fewer points – and surprisingly, the number does not even depend on the dimension!

Theorem 0.0.2 (Approximate Caratheodory theorem)**.** *Consider a set $T \subset \mathbb{R}^n$ that is contained in the unit Euclidean ball. Then, for every point $x \in \mathrm{conv}(T)$ and every $k \in \mathbb{N}$, one can find points $x_1, \dots, x_k \in T$ such that*[1]

$$\Big\| x - \frac{1}{k} \sum_{j=1}^{k} x_j \Big\|_2 \le \frac{1}{\sqrt{k}}.$$

This result is surprising for two reasons: first, the number of points k does not depend on the dimension n; and second, all the convex weights are equal. (Though repetitions among the points x_i are allowed.)

Proof This argument is known as the *empirical method* of B. Maurey.

Fix $x \in \mathrm{conv}(T)$ and write it as a convex combination of $z_1, \dots, z_m \in T$ as in (0.1). Let us interpret (0.1) probabilistically. Define a random vector Z taking values z_i with probabilities λ_i:

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

(This is possible because the coefficients λ_i are nonnegative and sum to one, so they can be interpreted as probabilities.) The expected value of Z is

$$\mathbb{E}\, Z = \sum_{i=1}^{m} \lambda_i z_i = x.$$

Consider independent copies $Z_1, Z_2, \dots$ of Z, i.e. independent random vectors with the same distribution as Z. The strong law of large numbers tells us that

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

For a more quantitative result, let's compute the mean squared error:

$$\mathbb{E} \Big\| x - \frac{1}{k} \sum_{j=1}^{k} Z_j \Big\|_2^2 = \frac{1}{k^2} \mathbb{E} \Big\| \sum_{j=1}^{k} (Z_j - x) \Big\|_2^2 = \frac{1}{k^2} \sum_{j=1}^{k} \mathbb{E} \|Z_j - x\|_2^2$$

The last identity is just a higher-dimensional version of the fact that the variance of a sum of independent random variables equals the sum of their variances (check this in Exercise 0.3!).

It remains to bound the variances of the terms. We have

$$\begin{aligned} \mathbb{E}\|Z_j - x\|_2^2 &= \mathbb{E}\|Z - \mathbb{E}\, Z\|_2^2 \\ &= \mathbb{E}\|Z\|_2^2 - \|\mathbb{E}\, Z\|_2^2 \quad \text{(by another variance identity, see Exercise 0.1(a))} \\ &\le \mathbb{E}\|Z\|_2^2 \le 1 \quad \text{(since } Z \in T \text{ and by the assumption on } T.) \end{aligned}$$

[1] Here and elsewhere in this book, $\|u\|_2$ denotes the Euclidean norm of a vector $u \in \mathbb{R}^n$, thus $\|u\|_2^2 = u_1^2 + \dots + u_n^2$.

We showed that

$$\mathbb{E}\Big\|x - \frac{1}{k}\sum_{j=1}^{k} Z_j\Big\|_2^2 \le \frac{1}{k}.$$

Therefore, there exists a realization of the random variables $Z_1, \ldots, Z_k$ such that

$$\Big\|x - \frac{1}{k}\sum_{j=1}^{k} Z_j\Big\|_2^2 \le \frac{1}{k}.$$

Since by construction each random variable Z_j takes values in T, the proof is complete. □

0.0.1 Covering geometric sets

Let's give a geometric application of Theorem 0.0.2. To cover a given set $P \subset \mathbb{R}^n$ with balls of a given radius, finding the minimal number is often tricky. Six balls cover the polygon in Figure 0.2, but is it clear that five balls won't suffice?

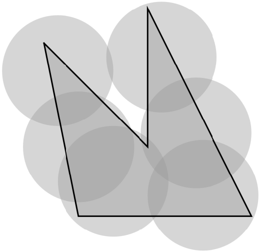


Figure 0.2 The covering problem

Theorem 0.0.2 can help us find economical coverings for polyhedra in high dimensions, called polytopes:

Corollary 0.0.3 (Covering polytopes by balls)**.** *Let P be a polytope in $\mathbb{R}^n$ with N vertices, contained in the unit Euclidean ball. Then, for every $k \in \mathbb{N}$, the polytope P can be covered by at most N^k Euclidean balls of radii $1/\sqrt{k}$.*

Proof Consider the set

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

We claim that the family of balls with radii $1/\sqrt{k}$ centered at points in $\mathcal{N}$ satisfy the conclusion of the corollary. To check this, note that $P \subset \operatorname{conv}(P) = \operatorname{conv}(T)$ where $T = \{\text{vertices of } P\}$. Thus we can apply Theorem 0.0.2 to any point $x \in P \subset \operatorname{conv}(T)$ and deduce that x is within distance $1/\sqrt{k}$ from some point in $\mathcal{N}$. This shows that the balls with radii $1/\sqrt{k}$ centered at $\mathcal{N}$ indeed cover P.

To bound the cardinality of $\mathcal{N}$, note that there are N^k ways to choose k out of N vertices with repetition. Thus $|\mathcal{N}| \le N^k$. The proof is complete. □

Covering techniques are helpful in several settings. We relate covering to packing in Section 4.2, to entropy and coding in Section 4.3, and to random processes in Chapters 7–8. But for now, let us show how to use covering to study volume.

You may remember the formula for the volume of some special shapes like a parallelopiped or a prism. But computing the volume of a general polyhedron in not easy, especially in high dimensions. Nevertheless, we have a simple bound:

Theorem 0.0.4 (Volume of a polytope)**.** *Let P be a polytope with N vertices, which is contained in the unit Euclidean ball of $\mathbb{R}^n$, denoted by B. Then*

$$\frac{\operatorname{Vol}(P)}{\operatorname{Vol}(B)} \le \left(3\sqrt{\frac{\log N}{n}}\right)^n. \tag{0.2}$$

Proof Corollary 0.0.3 says that the polytope P can be covered by at most N^k balls of radius $1/\sqrt{k}$. The volume of each ball is $(1/\sqrt{k})^n \operatorname{Vol}(B)$. (This is because of the particular way the volume scales in dimension n: increasing the radius of the ball by a factor of r increases the volume by a factor of r^n.) The volume of P is bounded by the total volume of the balls that cover P, thus

$$\operatorname{Vol}(P) \le N^k (1/\sqrt{k})^n \operatorname{Vol}(B),$$

and rearranging the terms gives

$$\frac{\operatorname{Vol}(P)}{\operatorname{Vol}(B)} \le \frac{N^k}{k^{n/2}}. \tag{0.3}$$

This bound is true for every $k \in \mathbb{N}$. So let us find the optimal k, the one that minimizes the right hand side. To do this, we take the logarithm, differentiate, and set the derivative to zero. This produces the optimal value

$$k_0 = \frac{n}{2\log N}. \tag{0.4}$$

(Check!) Substitute $k = k_0$ into (0.3), simplify the expression, and get

$$\frac{\operatorname{Vol}(P)}{\operatorname{Vol}(B)} \le \left(\sqrt{\frac{2e\log N}{n}}\right)^n.$$

This bound is even better than we claimed. But there is a slight inaccuracy in our proof. Can you spot it without reading further?

The value of k needs to be integer, and there is no guarantee that the optimal value (0.4) is integer. To fix this inaccuracy, we can take $k = \lceil k_0 \rceil$. Substitute this k into (0.3), use that $k_0 \le k \le k_0 + 1$ and get the corrected bound

$$\frac{\operatorname{Vol}(P)}{\operatorname{Vol}(B)} \le \frac{N^{k_0+1}}{k_0^{n/2}} \le N\left(\sqrt{\frac{2e\log N}{n}}\right)^n.$$

If $N \le e^{n/9}$ then the right had side is bounded by $\left(3\sqrt{\log(N)/n}\right)^n$, and the proof is complete. (Check!) On the other hand, if $N > e^{n/9}$, then the right hand side of the bound (0.2) is greater than 1, and such a bound holds trivially since $P \subset B$. So either way, the proof is complete. □

Remark 0.0.5 (A high-dimensional surprise)**.** Theorem 0.0.4 leads to the counterintuitive conclusion: *polytopes with a modest number vertices have extremely small volume.* Indeed, if you remember from the beginning of the proof how the volume scales, you can read the bound (0.2) as follows: the polytope P has volume as small as the Euclidean ball of radius $3\sqrt{\log(N)/n}$, and maybe even smaller. So if N grows slower than exponentially in n, that radius shrinks to zero, meaning P takes up a tiny fraction of the unit ball.

You will encounter other surprising phenomena in this book. As you develop more insight into high dimensions, such facts will begin to make intuitive sense.

Now, your turn – verify the variance formulas we used in the proof of Theorem 0.0.2 (Exercises 0.1, 0.3), make your own probabilistic proof of a deterministic result (Exercise 0.4), show that approximate Caratheodory is tight (Exercise 0.5), check a very handy bounds on binomial sums (Exercise 0.6 – don't miss that one!), discover another counterintuitive fact about high dimensions (Exercises 0.7–0.8) and improve the bound on the volume of polytopes (Exercise 0.9).

Notes

In this section we gave an illustration of the *probabilistic method*, where one employs randomness to construct a useful object. The book [17] presents many illustrations of the probabilistic method, mainly in combinatorics.

The empirical method of B. Maurey presented in this section was originally published in [271] and has found many applications since then. B. Carl used it to get bounds on covering numbers [76], as we did in Corollary 0.0.3.

A weaker version of the approximate Caratheodory theorem (Theorem 0.0.2), in which we do not insist that all weights of the convex combination be equal, is still non-trivial. It can be proved without using probability, instead employing a version of the Frank-Wolfe algorithm–a deterministic, iterative, greedy algorithm, see [43, Lemma 2.6].

Like Caratheodory theorem, several other results in combinatorial geometry can be made dimension-free by allowing them to be approximate rather than exact [10].

Theorem 0.0.4 and its strengthening in Exercise 0.9 was first proved by B. Carl and A. Pajor [77]. By considering random polytopes, N. Dafnis, A. Giannopoulos and A. Tsolomitis [90] showed that the bound in Exercise 0.9 is optimal in the entire interesting range $n \le N \le e^n$.

Exercises

0.1 ☕☕ (Two variance formulas)

(a) Recall that the variance of a random variable X satisfies $\operatorname{Var}(X) = \mathbb{E}(X - \mathbb{E}\,X)^2 = \mathbb{E}\,X^2 - (\mathbb{E}\,X)^2$. Let us prove a higher-dimensional version of this identity. Check that any random vector Z in $\mathbb{R}^n$ satisfies

$$\mathbb{E}\|Z - \mathbb{E}\,Z\|_2^2 = \mathbb{E}\|Z\|_2^2 - \|\mathbb{E}\,Z\|_2^2.$$

(b) Let Z be a random vector in $\mathbb{R}^n$, and Z' be an independent copy of Z, i.e. a random vector independent of Z and with the same distribution as Z. Check that

$$\mathbb{E}\|Z - \mathbb{E}\,Z\|_2^2 = \frac{1}{2}\,\mathbb{E}\|Z - Z'\|_2^2.$$

0.2 ☕☕☕ (Expectation minimizes the mean squared error) The variance of a random variable X has the following extremal property:

$$\mathrm{Var}(X) = \mathbb{E}\min_{a\in\mathbb{R}}(X-a)^2.$$

Let us prove a more general, high-dimensional version of this fact. Check that a random vector Z in $\mathbb{R}^n$ with finite $\mathbb{E}\|Z\|_2^2$ satisfies

$$\mathbb{E}\|Z - \mathbb{E}\,Z\|_2^2 = \min_{a\in\mathbb{R}^n}\mathbb{E}\|Z-a\|_2^2.$$

0.3 ☕☕ (Variance of a sum) Recall that the variance of a sum of independent random variables equals the sum of variances. Let us prove a higher-dimensional version of this identity. Check that any independent mean-zero random vectors $Z_1,\ldots,Z_k$ in $\mathbb{R}^n$ satisfy

$$\mathbb{E}\Big\|\sum_{j=1}^k Z_j\Big\|_2^2 = \sum_{j=1}^k \mathbb{E}\|Z_j\|_2^2.$$

0.4 ☕☕ (Balancing vectors) Let $x_1,\ldots,x_n$ be vectors in $\mathbb{R}^n$ that lie within the unit Euclidean ball centered at the origin.

(a) Prove that it is possible to assign a sign ± to each vector such that the sum $\pm x_1 \pm x_2 \pm \cdots \pm x_n$ lies within a Euclidean ball of radius $\sqrt{n}$ centered at the origin.

(b) Explain why the value $\sqrt{n}$ cannot be reduced in general.

0.5 ☕☕☕ (Approximate Caratheodory is asymptotically tight) Demonstrate by example that the bound in Theorem 0.0.2 is almost tight. Specifically, for every n find a set $T \subset \mathbb{R}^n$ such that for any convex combination $\sum_{j=1}^k \lambda_j x_j$ of any k points $x_1,\ldots,x_k \in T$, one has

$$\Big\|x - \sum_{j=1}^k \lambda_j x_j\Big\|_2 \ge \sqrt{\frac{1}{k} - \frac{1}{n}}.$$

Let $n\to\infty$ while keeping k fixed to see that Theorem 0.0.2 is asymptotically tight in high dimensions.

0.6 ☕☕ (Bounds on binomial coefficients) Prove the inequalities

$$\Big(\frac{n}{k}\Big)^k \le \binom{n}{k} \le \sum_{j=0}^k \binom{n}{j} \le \Big(\frac{en}{k}\Big)^k$$

for any integers $1 \le k \le n$.

0.7 ☕☕ (Thin shell phenomenon) Let us prove a counterintuitive fact that most of the volume of the high-dimensional ball lies near the surface. Consider the points inside the unit Euclidean ball of $\mathbb{R}^n$ that lie within distance $5/n$ from the surface of the ball, see Figure 0.3. Prove that such points make up over 99% of the volume of the ball.

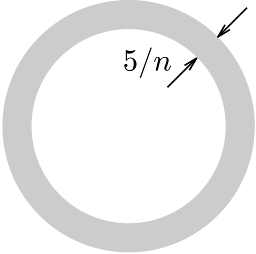


Figure 0.3 Over 99% of the volume of the unit ball in $\mathbb{R}^n$ lies within distance $5/n$ from the surface (Exercise 0.7)

0.8 ☕☕ (Thin shell phenomenon, continued) Let X be a random vector that is uniformly distributed[2] in the Euclidean unit ball of $\mathbb{R}^n$. Prove that

$$\mathbb{E}\|X\|_2 = \frac{n}{n+1}.$$

0.9 ☕☕☕ (Carl-Pajor theorem) Let's improve Theorem 0.0.4 by replacing N with N/n. Let P be a polytope with $N \geq n$ vertices, which is contained in the unit Euclidean ball of $\mathbb{R}^n$, denoted by B. Prove that

$$\frac{\mathrm{Vol}(P)}{\mathrm{Vol}(B)} \leq \left(C \sqrt{\frac{\log(eN/n)}{n}} \right)^n$$

where $C > 0$ is an absolute constant.[3]

[2] This means that X takes values in the unit ball B, and $\mathbb{P}\{X \in A\} = \mathrm{Vol}(A \cap B)/\,\mathrm{Vol}(B)$ for any measurable subset $A \subset B$.

[3] Feel free to choose any absolute constant C – even something huge like 100 or 10^{10} –whatever makes things easier. You can even leave C unspecified. Just make sure that C does not depend on anything like N, n, or P.

1

A Quick Refresher on Analysis and Probability

Most of the material we recall in this chapter is taught in basic analysis and probability courses. You may wish to skip this chapter altogether, or skim through it, if you are well prepared. Either way, try the exercises at the end of the chapter, especially the more difficult ones.

1.1 Convex sets and functions

A subset $K \subset \mathbb{R}^n$ is a *convex set* if, for any pair of points in K, the line segment connecting these two points is also contained in K:

$$\lambda x + (1-\lambda)y \in K \quad \forall x, y \in K,\ \lambda \in [0,1].$$

Let $K \subset \mathbb{R}^n$ be a convex subset. A function $f : K \to \mathbb{R}$ is a *convex function* if

$$f(\lambda x + (1-\lambda)y) \le \lambda f(x) + (1-\lambda) f(y) \quad \forall x, y \in K,\ \lambda \in [0,1]. \tag{1.1}$$

In other words, f is convex if, when restricted to a line segment connecting any pair of points in K, the graph of f lies below the line segment connecting these two points. Figure 1.1 illustrates the definitions of convex sets and functions.

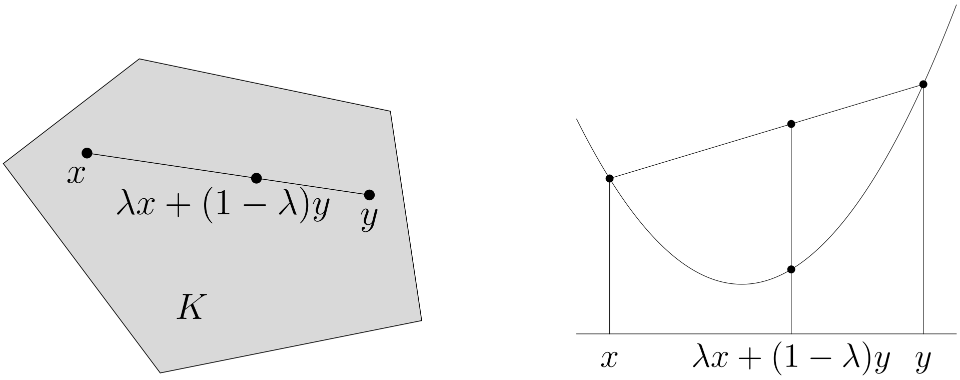


Figure 1.1 Definitions of a convex set and a convex function

[0] This material will be published by Cambridge University Press as *High-Dimensional Probability, 2nd Edition* by Roman Vershynin. This pre-publication version is free to view and download for personal use only. Not for re-distribution, re-sale or use in derivative works. ©Roman Vershynin 202X.
If you want to be notified when the printed version is available, let me know: rvershyn@uci.edu

Concave functions are defined similarly, except that the inequality in (1.1) is reversed. Equivalently, f is concave if and only if $-f$ is convex.

The *maximum principle* states that a convex function defined on the convex set $K = \mathrm{conv}(x_1, \ldots, x_n)$ attains its maximum at some point x_i (try Exercise 1.4).

1.2 Norms and inner products

You should already be familiar with the definitions of *metrics*, *norms*, and *inner products*. To see if you remember them well, take a minute to prove that (a) a norm is a convex function and (b) the unit ball of a normed space is a convex set.

The most popular example of a norm and an inner product on $\mathbb{R}^n$ is the Euclidean norm and the dot product, defined as

$$\|x\|_2 = \Big(\sum_{i=1}^n x_i^2\Big)^{1/2} \quad \text{and} \quad \langle x, y\rangle = x^\mathsf{T} y = \sum_{i=1}^n x_i y_i. \tag{1.2}$$

The Euclidean norm agrees with the dot product in the sense that $\|x\|_2^2 = \langle x, x\rangle$.

More generally, for any exponent $p \in [1, \infty]$, we can define the ℓ^p *norm* on $\mathbb{R}^n$ by

$$\|x\|_p = \Big(\sum_{i=1}^n |x_i|^p\Big)^{1/p} \text{ for } p \in [1, \infty), \quad \text{and } \|x\|_\infty = \max_{i \le n} |x_i|.$$

The *Minkowski inequality* states that the triangle inequality holds for the ℓ^p norm: for any vectors $x, y \in \mathbb{R}^n$ we have

$$\|x + y\|_p \le \|x\|_p + \|y\|_p.$$

It follows that the ℓ^p norm indeed defines a norm on $\mathbb{R}^n$ for every $p \in [1, \infty]$. (Check!)

The best way to visualize the geometry of an ℓ^p norm is to look at the unit ball of the space $(\mathbb{R}^n, \|\cdot\|_p)$, that is

$$B_p^n = \{x \in \mathbb{R}^n : \ \|x\|_p \le 1\}.$$

For example, B_2^n is the unit Euclidean ball – a *round ball* with unit radius, B_∞^n is a *cube*, and B_1^n a *cross-polytope* – a high-dimensional version of an octahedron:

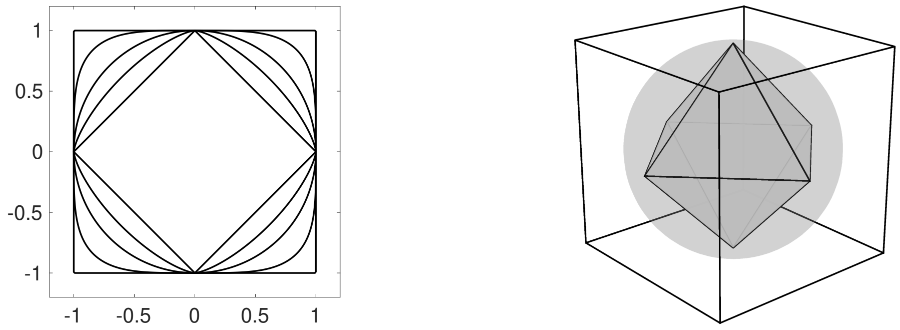
$$B_\infty^n = [-1, 1]^n, \quad B_1^n = \mathrm{conv}\left(\{\pm e_1, \ldots, \pm e_n\}\right), \tag{1.3}$$

where $e_1, \ldots, e_n$ denotes the standard basis of $\mathbb{R}^n$ (Exercise 1.6). Figure 1.2 illustrates the unit ℓ^p balls for different exponents p.

The ℓ^p norms of a given vector x are decreasing in p:

$$\|x\|_q \le \|x\|_p \quad \text{whenever } p \le q, \tag{1.4}$$

see Exercise 1.17. Equivalently, the unit ℓ^p balls are increasing in p, that is $B_p^n \subset B_q^n$ whenever $p \le q$. (Why?)



(a) ℓ^p balls for $p = 1, 1.4, 2, 4, \infty$

(b) ℓ^p balls for $p = 1, 2, \infty$

Figure 1.2 Some unit ℓ^p balls in dimensions $n = 2$ (left) and $n = 3$ (right)

The *Cauchy-Schwarz inequality* states that for all vectors $x, y \in \mathbb{R}^n$ we have

$$|\langle x, y\rangle| \le \|x\|_2 \|y\|_2.$$

The *Hölder inequality* generalizes this result to the ℓ^p norms:

$$|\langle x, y\rangle| \le \|x\|_p \|y\|_{p'} \quad \text{if } \frac{1}{p} + \frac{1}{p'} = 1. \tag{1.5}$$

A pair of numbers $p, p' \in [1, \infty]$ satisfying the condition in (1.5) is called *conjugate exponents.*[1]

The Hölder inequality is tight. In Exercise 1.19 you will check that for any vector x one can find a vector $y \ne 0$ for which the Hölder inequality becomes an equality. In other words, the following *duality formula* holds for the ℓ^p norms:

$$\max\left\{\langle x, y\rangle :\ y \in B^n_{p'}\right\} = \|x\|_p. \tag{1.6}$$

1.3 Random variables and random vectors

In a basic course in probability theory, we learned about the two most important quantities associated with a random variable X, namely the *expectation*[2] (also called *mean*), and *variance*. They are be denoted by[3]

$$\mathbb{E}\, X \qquad \text{and} \qquad \mathrm{Var}(X) = \mathbb{E}(X - \mathbb{E}\, X)^2. \tag{1.7}$$

The *linearity of expectation* guarantees that for any random variables $X_1, \dots, X_n$, whether independent or not, and for any fixed numbers $a_1, \dots, a_n$, we have

$$\mathbb{E}\left[a_1 X_1 + \cdots + a_n X_n\right] = a_1\, \mathbb{E}\, X_1 + \cdots + a_n\, \mathbb{E}\, X_n.$$

[1] We agree that $\frac{1}{0} = \infty$ and $\frac{1}{\infty} = 0$ to make $(1, \infty)$ and $(\infty, 1)$ pairs of conjugate exponents.

[2] If you have studied measure theory, you know that the expectation $\mathbb{E}\, X$ of a random variable X on a probability space $(\Omega, \Sigma, \mathbb{P})$ is defined as Lebesgue integral of the function $X : \Omega \to \mathbb{R}$. This allows all Lebesgue integration theorems to apply for expectations of random variables.

[3] Throughout this book, we drop the brackets in the notation $\mathbb{E}[f(X)]$ and simply write $\mathbb{E}\, f(X)$ instead. Thus, nonlinear functions bind before expectation.

Variance generally does not satisfy a similar property. However, if the random variables X_i are independent (or even uncorrelated), then

$$\operatorname{Var}(a_1X_1 + \cdots + a_nX_n) = a_1^2 \operatorname{Var}(X_1) + \cdots + a_n^2 \operatorname{Var}(X_n). \tag{1.8}$$

The simplest example of a random variable is the *indicator* of a given event E, which is denoted by $\mathbf{1}_E$. The random variable $\mathbf{1}_E$ takes value 1 if event E occurs and 0 if E does not occur. The expectation of an indicator obviously satisfies

$$\mathbb{E}\,\mathbf{1}_E = \mathbb{P}(E). \tag{1.9}$$

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 as $\mathbb{E}\, X^p$, and the *p-th absolute moment* is $\mathbb{E}|X|^p$. If we take p-th root of the absolute moment, we arrive at the notion of the L^p *norm* of a random variable:

$$\|X\|_{L^p} = (\mathbb{E}|X|^p)^{1/p} \text{ for } p \in (1,\infty), \quad \text{and } \|X\|_{L^\infty} = \operatorname{ess\,sup}|X|, \tag{1.10}$$

where ess sup denotes the essential supremum.

The normed space consisting of all random variables on a given probability space that have finite L^p norms is called the L^p space:

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

Minkowski inequality, which we recall in (1.21), implies that the L^p norm indeed defines a norm on the space L^p for each $p \in [1,\infty]$.

The exponent $p = 2$ is special: L^2 is not only a normed but also an inner product space. The inner product is given by

$$\langle X, Y\rangle_{L^2} = \mathbb{E}\, XY, \tag{1.11}$$

and it agrees with the L^2 norm in the sense that $\|X\|_{L^2}^2 = \langle X, X\rangle_{L^2}$.

The *standard deviation* of a random variable X is

$$\sigma(X) = \sqrt{\operatorname{Var}(X)} = \|X - \mathbb{E}\, X\|_{L^2}. \tag{1.12}$$

The *covariance* of random variables of X and Y is

$$\operatorname{cov}(X,Y) = \mathbb{E}(X - \mathbb{E}\, X)(Y - \mathbb{E}\, Y) = \langle X - \mathbb{E}\, X, Y - \mathbb{E}\, Y\rangle_{L^2}. \tag{1.13}$$

The concept of a random variable can be generalized to higher dimensions. We can define a *random vector* $X = (X_1, \ldots, X_n)$ taking values in $\mathbb{R}^n$ as a vector whose all n coordinates X_i are random variables. The expected value of X is defined coordinate-wise, that is

$$\mathbb{E}\, X = (\mathbb{E}\, X_1, \ldots, \mathbb{E}\, X_n).$$

What about *variance in high dimensions*? If we want the traditional definition of variance (1.7) to make sense for random vectors, we have to decide how to square vectors $x \in \mathbb{R}^n$, or how to multiply vectors x by themselves. This can be done in two ways: as a dot product, or “inner product” $x^\mathsf{T} x = \|x\|_2^2$, and as an "outer product" xx^T. The first interpretation leads to defining the variance of X as $\mathbb{E}\|X - \mathbb{E}\,X\|_2^2$, the quantity that played a prominent role in the Appetizer. The second interpretation leads to a more informative result: it gives rise to the notion of the *covariance matrix*

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

which is an $n \times n$ matrix whose (i,j)-th entry equals $\operatorname{cov}(X_i, X_j)$. We will work with covariance matrices in Section 3.2.

1.4 Union bound

If E_i are disjoint events, then the additivity axiom of probability tells us that

$$\mathbb{P}\Big(\bigcup_{i=1}^n E_i\Big) = \sum_{i=1}^n \mathbb{P}(E_i).$$

If the events E_i are not disjoint, this identity may fail: each point belonging to several events E_i is counted as one on the left hand side, but several times on the right hand side. Thus, instead of an equality, we get an inequality:

Lemma 1.4.1 (Union bound)**.** *For any events $E_1, \ldots, E_n$, we have*

$$\mathbb{P}\Big(\bigcup_{i=1}^n E_i\Big) \le \sum_{i=1}^n \mathbb{P}(E_i).$$

Proof If the event $\bigcup_{i=1}^n E_i$ occurs, then at least one of the events E_i occurs. Thus the indicators satisfy

$$\mathbf{1}_{\bigcup_{i=1}^n E_i} \le \sum_{i=1}^n \mathbf{1}_{E_i}. \tag{1.14}$$

(The right hand side counts how many E_i occur.) Now take expectations and use (1.9) and linearity of expectation to complete the proof. □

The union bound is typically used to bound the probability of a bad event.

Example 1.4.2 (Dense random graphs have no isolated vertices)**.** Imagine that $n \ge 2$ freshmen arrive on campus. Friendships form randomly, with each pair of students becoming friends with probability p independent of all other pairs. Let us show that if $p \ge 4\ln(n)/n$ then there are no friendless students with probability at least $1 - 1/n$.

Proof Call the students $1, \ldots, n$, and let E_i denote the event that student i is friendless. This means that none of the other $n-1$ students are friends with i, and these $n-1$ sad events are independent and have probability $1-p$ each. Thus $\mathbb{P}(E_i) = (1-p)^{n-1}$.

We would like to bound the probability of the bad event

$$B = \{\text{there exists a friendless student}\} = \bigcup_{i=1}^n E_i.$$

The union bound gives

$$\mathbb{P}(B) \le \sum_{i=1}^{n} \mathbb{P}(E_i) = n(1-p)^{n-1}.$$

To simplify this bound, use the inequality $1-p \le e^{-p}$ and recall the assumptions on n and p. A little computation yields $n(1-p)^{n-1} \le 1/n$. Thus, the complement of the bad event B happens with probability at least $1 - 1/n$. □

1.5 Conditioning

The conditioning trick often helps us to calculate probabilities and expectations. This method is based on the concept of *conditional probability* of an event E given event F, denoted by

$$\mathbb{P}(E \mid F) = \frac{\mathbb{P}(E \cap F)}{\mathbb{P}(F)},$$

and the concept of *conditional expectation* of a random variable X given another random variable Y, denoted by $\mathbb{E}[X|Y]$.

The *law of total expectation* says that

$$\mathbb{E}\, X = \mathbb{E}\left[\mathbb{E}[X|Y]\right]. \tag{1.15}$$

Informally, this identity says that in order to compute the expectation of X, we can first average with respect to X while keeping the value of Y fixed, and then average the result with respect to Y. This is the essence of the conditioning trick.

The conditioning trick can also help us compute probabilities. If E is any event and Y is a random variable, we can define the conditional probability of E given Y as

$$\mathbb{P}(E \mid Y) := \mathbb{E}\left[\mathbf{1}_E | Y\right],$$

where $\mathbf{1}_E$ denotes the indicator of the event E. Since $\mathbb{E}\,\mathbf{1}_E = \mathbb{P}(E)$, the law of total expectation in this case reads as

$$\mathbb{P}(E) = \mathbb{E}\left[\mathbb{P}(E \mid Y)\right]. \tag{1.16}$$

Thus, in order to compute probabilities using the conditioning trick, we first need to find the probability while keeping the value of Y fixed, and then average the result with respect to Y.

Suppose we have a decomposition of the sample space Ω into mutually disjoint events $F_1, F_2, \ldots$ In other words, suppose that each outcome belongs to one, and only one, event F_i. Then the *law of total probability* allows us to compute the probability of any event E as follows:

$$\mathbb{P}(E) = \sum_i \mathbb{P}(E \mid F_i)\mathbb{P}(F_i). \tag{1.17}$$

The law of total probability can be quickly deduced from the law of total expectation. To do so, apply (1.16) for the random variable Y that takes value i if event F_i

occurs. Note that the random variable $\mathbb{P}(E|Y)$ takes value $\mathbb{P}(E|Y=i) = \mathbb{P}(E|F_i)$ with probability $\mathbb{P}\{Y=i\} = \mathbb{P}(F_i)$.

Everything we said here about random variables X and Y is also true for random vectors.

Example 1.5.1 (Probability of a perfect cancelation)**.** Let $a_1, \ldots, a_n$ be real numbers, not all of which are zero. What is the probability that

$$\pm a_1 \pm \cdots \pm a_n = 0$$

where the signs are chosen at random? Let us show that this probability is always bounded by 1/2. To state this problem rigorously, we can model the random signs as independent *Rademacher random variables* $X_1, \ldots, X_n$, i.e. random variables that take values -1 and 1 with probability 1/2 each. We claim that

$$\mathbb{P}\{S_n = 0\} \le \frac{1}{2} \quad \text{where} \quad S_n = \sum_{i=1}^n a_i X_i.$$

Proof We can assume without loss of generality that $a_n \ne 0$ by rearranging the terms if necessary. The proof is by exposing the last term $a_n X_n$ of the sum, while keeping all the previous terms fixed by conditioning. So let us condition on the random variables $X_1, \ldots, X_{n-1}$, or to be pedantic, on the random vector $(X_1, \ldots, X_{n-1})$. The conditioning fixes the values of $X_1, \ldots, X_{n-1}$, and therefore it fixes the value of $S_{n-1} = \sum_{i=1}^{n-1} a_i X_i$. All randomness is now left with X_n. Since $S_n = S_{n-1} + a_n X_n$, we have

$$\mathbb{P}\{S_n = 0 \,|\, X_1, \ldots, X_{n-1}\} = \mathbb{P}\Big\{X_n = -\frac{S_{n-1}}{a_n} \,\Big|\, X_1, \ldots, X_{n-1}\Big\} \le \frac{1}{2}.$$

The inequality holds because X_n is independent of $X_1, \ldots, X_{n-1}$, the value $u = -S_{n-1}/a_n$ is fixed by conditioning, and the definition of the Rademacher distribution implies that $\mathbb{P}\{X_n = u\} \le 1/2$ for any fixed value $u \in \mathbb{R}$.

We can finish the proof by applying the law of total expectation (1.16):

$$\mathbb{P}\{S_n = 0\} = \mathbb{E}\Big[\mathbb{P}\{S_n = 0 \,|\, X_1, \ldots, X_{n-1}\}\Big] \le \mathbb{E}\,\frac{1}{2} = \frac{1}{2}. \quad \square$$

By the way, the result in Example 1.5.1 is sharp. If there are exactly two nonzero coefficients a_i and they are equal to each other, then we have $\mathbb{P}\{S_n = 0\} = 1/2$. (Check!)

1.6 Probabilistic inequalities

Jensen inequality states that for any random variable X and a convex function $f : \mathbb{R} \to \mathbb{R}$, we have

$$f(\mathbb{E}\,X) \le \mathbb{E}\,f(X). \tag{1.18}$$

More generally, (1.18) holds for any random vector X taking values in $\mathbb{R}^n$ and any convex function $f : \mathbb{R}^n \to \mathbb{R}$. In Exercise 1.3, you will prove this for random vectors taking finitely many values; the general case can be deduced by approximation. In particular, since any norm on $\mathbb{R}^n$ is a convex function, Jensen inequality yields

$$\|\mathbb{E}\, X\| \le \mathbb{E}\|X\|. \tag{1.19}$$

The L^p norms of a given random variable X are increasing[4] in p:

$$\|X\|_{L^p} \le \|X\|_{L^q} \quad \text{whenever } p \le q, \tag{1.20}$$

see Exercise 1.11.

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

$$\|X + Y\|_{L^p} \le \|X\|_{L^p} + \|Y\|_{L^p}. \tag{1.21}$$

In other words, the L^p norm satisfies the triangle inequality.

Cauchy-Schwarz inequality states that for any random variables $X, Y \in L^2$, we have

$$\|XY\|_{L^1} \le \|X\|_{L^2}\|Y\|_{L^2}.$$

The *Hölder inequality* generalizes this result to the L^p norms. For any pair of conjugate exponents $p, p' \in [1, \infty]$ (introduced in (1.5)) and any pair of random variables $X \in L^p$ and $Y \in L^{p'}$, we have

$$\|XY\|_{L^1} \le \|X\|_{L^p}\|Y\|_{L^{p'}}. \tag{1.22}$$

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

$$F_X(t) = \mathbb{P}\{X \le t\}, \quad t \in \mathbb{R}.$$

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

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

The following result allows us to compute expectation in terms of the tails.

Lemma 1.6.1 (Integrated tail formula). *Any nonnegative random variable X satisfies*

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

The two sides of this identity are either finite or infinite simultaneously.

[4] This might sound a little bit confusing when we recall from (1.4) that the ℓ^p norms are *decreasing* in p. There is no contradiction because the ℓ^p and L^p norms are normalized differently. The ℓ^p norm of a vector $x = (x_1, \ldots, x_n)$ is $\|x\|_{\ell^p} = \big(\sum_{i=1}^n |x_i|^p\big)^{1/p}$, while the L^p norm of a random variable X that takes values $x_1, \ldots, x_n$ with probabilities $1/n$ each is $\|X\|_{L^p} = \big(\frac{1}{n}\sum_{i=1}^n |x_i|^p\big)^{1/p}$.

Proof We can represent any nonnegative real number x via the identity

$$x = \int_0^x 1\,dt = \int_0^\infty \mathbf{1}_{\{t<x\}}\,dt.$$

Replace x with the random variable X and take the expectation on both sides. This gives

$$\mathbb{E}\,X = \mathbb{E}\int_0^\infty \mathbf{1}_{\{t<X\}}\,dt = \int_0^\infty \mathbb{E}\,\mathbf{1}_{\{t<X\}}\,dt = \int_0^\infty \mathbb{P}\{t<X\}\,dt.$$

To change the order of expectation and integration in the second equality, we used Fubini-Tonelli theorem. The proof is complete. □

Conversely, Markov inequality bounds the tails in terms of expectation:

Proposition 1.6.2 (Markov inequality)**.** *For any nonnegative random variable X and $t>0$, we have*

$$\mathbb{P}\{X \ge t\} \le \frac{\mathbb{E}\,X}{t}.$$

Proof Fix $t>0$. We can represent any real number x via the identity

$$x = x\mathbf{1}_{\{x\ge t\}} + x\mathbf{1}_{\{x<t\}}.$$

Replace x with the random variable X and take expectation:

$$\mathbb{E}\,X = \mathbb{E}\,X\mathbf{1}_{\{X\ge t\}} + \mathbb{E}\,X\mathbf{1}_{\{X<t\}} \ge \mathbb{E}\,t\mathbf{1}_{\{X\ge t\}} + 0 = t\cdot\mathbb{P}\{X\ge t\}.$$

Divide both sides by t to complete the proof. □

Markov's inequality gives the best possible tail bound if all we know is $\mathbb{E}\,X$. But if we also know the variance, we get a better bound that decays quadratically in t – and learn how tightly X concentrates around its mean:

Corollary 1.6.3 (Chebyshev inequality)**.** *Let X be a random variable with mean μ and variance σ^2. Then, for any $t>0$, we have*

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

Proof Square both sides of the bound $|X-\mu|\ge t$ and apply Markov inequality for the random variable $(X-\mu)^2$. □

1.7 Limit theorems

The study of *sums of independent random variables* lies at the heart of classical probability theory. Let $X_1,\dots,X_N$ be independent and identically distributed (i.i.d.) random variables with mean μ and variance σ^2. Then the formula for the variance of a sum (1.8) gives

$$\operatorname{Var}\left(\frac{1}{N}\sum_{i=1}^N X_i\right) = \frac{\sigma^2}{N}. \tag{1.23}$$

Thus, the variance of the *sample mean* $\frac{1}{N}\sum_{i=1}^{N} X_i$ of the sample $\{X_1, \ldots, X_N\}$ shrinks to zero as the size of the sample N increases to infinity. This indicates that for large N, we should expect that the sample mean concentrates tightly about its expectation μ. One of the most important results in probability theory – the law of large numbers – states precisely this.

Theorem 1.7.1 (Strong law of large numbers)**.** *Let $X_1, X_2, \ldots$ be a sequence of 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 central limit theorem makes one step further. It identifies the limiting distribution of the (properly scaled) sum S_N as a *normal* distribution, sometimes also called a *Gaussian* distribution.

Definition 1.7.2 (Normal distribution)**.** A random variable X is a standard normal random variable, denoted by

$$X \sim N(0,1),$$

if X has density

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad x \in \mathbb{R}. \tag{1.24}$$

The random variable X has mean zero and variance 1.

More generally, a random variable X has a normal distribution

$$X \sim N(\mu, \sigma^2)$$

if X can be expressed as $X = \mu + \sigma Z$ for some fixed numbers $\mu \in \mathbb{R}$ and $\sigma > 0$, and where $Z \sim N(0,1)$. The density of X equals

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

The random variable X has mean μ and variance σ^2.

Theorem 1.7.3 (Lindeberg-Lévy central limit theorem)**.** *Let $X_1, X_2, \ldots$ be a sequence of 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:

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

Then, as $N \to \infty$,

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

Convergence in distribution means that the CDF of Z_N converges pointwise to the CDF of $g \sim N(0,1)$. In terms of tails, this means that for every $t \in \mathbb{R}$,

$$\mathbb{P}\{Z_N > t\} \to \mathbb{P}\{g > t\} = \frac{1}{\sqrt{2\pi}} \int_t^\infty e^{-x^2/2}\, dx \quad \text{as } N \to \infty.$$

Example 1.7.4 (Bernoulli and binomial distributions)**.** One remarkable special case of the central limit theorem is where X_i are Bernoulli random variables with some fixed parameter $p \in (0,1)$, denoted

$$X_i \sim \mathrm{Ber}(p).$$

This means that X_i take values 1 and 0 with probabilities p and $1-p$ respectively. One can easily check that

$$\mathbb{E}\, X_i = p \text{ and } \mathrm{Var}(X_i) = p(1-p)$$

The sum $S_N := X_1 + \cdots + X_N$ of independent $\mathrm{Ber}(p)$ random variables X_i is said to have the *binomial distribution*, denoted

$$S_N \sim \mathrm{Binom}(N,p).$$

The central limit theorem (Theorem 1.7.3) yields that as $N \to \infty$,

$$\frac{S_N - Np}{\sqrt{Np(1-p)}} \to N(0,1) \quad \text{in distribution.} \tag{1.25}$$

This special case of the central limit theorem is called *de Moivre-Laplace theorem.*

Now consider independent random variables $X_i \sim \mathrm{Ber}(p_i)$ whose parameters p_i *decay to zero* as $N \to \infty$ so fast that the sum S_N has mean $O(1)$ instead of being proportional to N. The central limit theorem fails in this regime. A different result we are about to state says that S_N still converges, but to the *Poisson distribution* instead of the normal distribution.

Definition 1.7.5 (Poisson distribution)**.** A random variable Z has *Poisson distribution* with parameter $\lambda > 0$, denoted

$$Z \sim \mathrm{Pois}(\lambda),$$

if Z takes values in $\{0,1,2,\dots\}$ with probabilities

$$\mathbb{P}\{Z = k\} = e^{-\lambda} \frac{\lambda^k}{k!}, \quad k = 0,1,2,\dots \tag{1.26}$$

Theorem 1.7.6 (Poisson limit theorem)**.** *Consider independent random variables $X_{N,i} \sim \mathrm{Ber}(p_{N,i})$ for $N = 1,2,\dots$ and $1 \le i \le N$. Let*

$$S_N = X_{N,1} + \cdots + X_{N,N}.$$

Assume that, as $N \to \infty$,

$$\max_{i \le N} p_{N,i} \to 0 \quad \textit{and} \quad \mathbb{E}\, S_N = p_{N,1} + \cdots + p_{N,N} \to \lambda < \infty.$$

Then, as $N \to \infty$,

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

The probability mass function (1.26) of the Poisson distribution includes the factorial $k!$, which is not an easy function to work with. For large k, we can simplify it using Stirling approximation:

Lemma 1.7.7 (Stirling approximation)**.** *We have*

$$n! = \sqrt{2\pi n}\Big(\frac{n}{e}\Big)^n (1 + o(1)) \quad \textit{as } n \to \infty.$$

Using Stirling approximation in (1.26), we see that for any fixed parameter $\lambda > 0$, a Possion random variable $Z \sim \mathrm{Pois}(\lambda)$ satisfies

$$\mathbb{P}\{Z = k\} = \frac{e^{-\lambda}}{\sqrt{2\pi k}} \left(\frac{e\lambda}{k}\right)^k (1 + o(1)) \quad \text{as } k \to \infty, \tag{1.27}$$

so this decays roughly as k^{-k} – slightly faster than exponentially.

There are non-asymptotic versions of Stirling approximation which hold for any given value of n as opposed to the limit as $n \to \infty$. Here is one of them:

Lemma 1.7.8 (Bounds on the factorial)**.** *For any $n \in \mathbb{N}$, we have*

$$\Big(\frac{n}{e}\Big)^n \le n! \le en\Big(\frac{n}{e}\Big)^n. \tag{1.28}$$

Proof If we recall the Taylor series for e^x and drop all terms except n-th, we get $e^x \ge x^n/n!$. Substitute $x = n$ and rearrange the terms to obtain the lower bound in (1.28). To prove the upper bound, note that

$$\ln(n!) = \sum_{k=1}^{n} \ln k \le \int_1^n \ln x \, dx + \ln n = n(\ln n - 1) + 1 + \ln n. \tag{1.29}$$

(The inequality in (1.29) follows by comparing the areas as in the usual proof of the integral test – do it!) Exponentiating and rearranging the terms yields the upper bound in (1.28). □

Remark 1.7.9 (Gamma function)**.** The *gamma function* extends the notion of factorial for all real numbers, and even for all complex numbers z whose real part is positive. It is defined as follows:

$$\Gamma(z) := \int_0^\infty t^{z-1} e^{-t} dt. \tag{1.30}$$

Repeated integration by parts (do it!) reveals that

$$\Gamma(n + 1) = n! \quad \text{for all } n = 0, 1, 2, \ldots$$

Stirling approximation (Lemma 1.7.7) remains valid for the gamma function:

$$\Gamma(z + 1) = \sqrt{2\pi z}\Big(\frac{z}{e}\Big)^z (1 + o(1)) \quad \text{as } \mathbb{R} \ni z \to \infty. \tag{1.31}$$

1.8 Notes

The question we raise in Example 1.5.1 is known as the *Littlewood-Offord problem*. Originally considered by Littlewood and Offord [217] and Erdös [122], this problem and its variants was studied extensively since then; see the surveys [318, 291, 258].

Proofs of the strong law of large numbers (Theorem 1.7.1) and Lindeberg-Lévy central limit theorem (Theorem 1.7.3) can be found e.g. in [116, Sections 1.7 and 2.4] and [42, Sections 6 and 27].

Both Proposition 1.6.2 and Corollary 1.6.3 are due to Chebyshev. However, following the established tradition, we call Proposition 1.6.2 Markov inequality.

In modern language, Example 1.4.2, Exercises 1.9, and Exercise 1.10 identify the threshold for the existence of isolated vertices in the Erdős-Rényi model $G(n,p)$. This problem was first studied in seminal paper of Erdös and Rényi [121]. There are now more general and more precise results about the degrees of random graphs, see [47, Chapter 3] and [131, Chapter 3]. The second moment method showcased in Exercise 1.10 is ubiquitous in combinatorics and theoretical computer science, see [17, Chapter 4]. The result of Exercise 1.8 is asymptotically optimal: the largest cardinality of an independent set in an Erdős-Rényi random graph $G \sim G(n,p)$ is approximately $2\log_b n$ where $b = 1/(1-p)$, see [131, Section 7.2].

A short proof of Stirling approximation (Lemma 1.7.7) can be found in [286] and [123, II.9]. This proof yields the following non-asymptotic result, which holds for every any $n \in \mathbb{N}$:

$$\sqrt{2\pi n}\Big(\frac{n}{e}\Big)^n e^{\frac{1}{12n+1}} \le n! \le \sqrt{2\pi n}\Big(\frac{n}{e}\Big)^n e^{\frac{1}{12n}}.$$

This result implies the asymptotic form of Stirling approximation (Lemma 1.7.8) and also improves upon Lemma 1.7.8. Short proofs of the Stirling approximation for the gamma function (1.31) can be found in [44, 101]; for non-asymptotic versions, see [172].

Exercises

1.1 ☕ Consider any subset $T \subset \mathbb{R}^n$. Check that $\operatorname{conv}(T)$ is a convex set.

1.2 ☕ Check that the pointwise maximum of a finite number of convex functions is a convex function.

1.3 ☕☕ (Jensen inequality)

(a) The definition of a convex function (1.1) involves convex combinations of two points x and y. Let us extend it to arbitrarily many points. Let $K \subset \mathbb{R}^n$ be a convex subset. Prove that a function $f : K \to \mathbb{R}$ is convex if and only if the following holds. For any $m \in \mathbb{N}$, any vectors $x_i \in K$ and any numbers $\lambda_i \ge 0$ with $\sum_{i=1}^m \lambda_i = 1$, we have

$$f\Big(\sum_{i=1}^m \lambda_i x_i\Big) \le \sum_{i=1}^m \lambda_i f(x_i).$$

(b) Let X be a random vector in $\mathbb{R}^n$ that takes finitely many values, and let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function. Deduce from part (a) Jensen inequality:

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

1.4 ☕ (A maximum principle) Prove for any convex function f and a subset $T \subset \mathbb{R}^n$:

$$\sup_{x \in \operatorname{conv}(T)} f(x) = \sup_{x \in T} f(x).$$

1.5 ☕☕ (Expressing a cube as a convex hull of its vertices) It seems almost obvious that the cube is the convex hull of its vertices:

$$[-1,1]^n = \operatorname{conv}\left(\{-1,1\}^n\right).$$

Prove this by expressing any point in the cube as a convex combination of the vertices.

1.6 ☕☕ (Expressing a cross-polytope as a convex hull of its vertices) Check that the unit ball corresponding to the ℓ^1 norm in $\mathbb{R}^n$ is the *absolute convex hull* of the standard basis $e_1, \ldots, e_n$ in $\mathbb{R}^n$, that is

$$B_1^n = \operatorname{conv}\left(\{\pm e_1, \ldots, \pm e_n\}\right).$$

Write down a formula that expresses any point $x \in B_1^n$ as a convex combination of the vectors $\pm e_1, \ldots, \pm e_n$.

1.7 ☕☕ (Random graphs with random number of vertices) Suppose that in Example 1.4.2, the number n of freshmen who arrive on campus is a random variable that has Poisson distribution with mean λ. As before, each pair of students becomes friends with probability p independent of all other pairs. Show that if $p \geq 2\ln(\lambda)/\lambda$ then there are no friendless students with probability at least $1 - 1/\lambda$.

1.8 ☕☕ (Independent sets in random graphs) Call a group of people *independent* if no two members are friends. Suppose $n \geq 7$ students enroll in a class on high-dimensional probability, with each pair becoming friends independently with probability $1/2$. Show that, with probability at least $1 - 1/n$, this class has no independent subsets of more than $2\log_2 n$ students.

1.9 ☕ (Dense random graphs have no isolated vertices) Let us refine the result of Example 1.4.2. Suppose n freshmen arrive on campus, with each pair becoming friends independently with probability p_n. Fix any $\varepsilon > 0$ and assume that

$$p_n > \frac{(1+\varepsilon)\ln n}{n} \quad \text{for every } n \in \mathbb{N}.$$

Prove that there are no friendless students with probability that converges to 1 as $n \to \infty$.

1.10 ☕☕☕☕ (Sparse random graphs have isolated vertices) Let us prove a converse to Exercise 1.9. Fix any $\varepsilon > 0$ and assume that

$$p_n < \frac{(1-\varepsilon)\ln n}{n} \quad \text{for every } n \in \mathbb{N}.$$

Then there exists at least one friendless student with probability that converges to 1 as $n \to \infty$. You will prove this result using the so-called *second moment method*:

(a) Denote the number of friendless students by S_n and express it as $S_n = X_1 + \ldots + X_n$ where X_i is the indicator of the event that student i is friendless. Show that

$$\mu_n = \mathbb{E}\, S_n \to \infty.$$

Thus the expected number of friendless students is large. But this does not automatically imply that there exists even one friendless student with high probability! (Why?)

(b) Compute the second moment $\mathbb{E}\, S_n^2$ by expanding the square. Conclude that

$$\frac{\operatorname{Var}(S_n)}{\mu_n^2} \to 0.$$

(c) Use Chebyshev inequality to complete the proof.

1.11 ☕☕ (Monotonicity of the L^p norm)

(a) Let X be a random variable. Show that $\|X\|_{L^p}$ is an increasing function in p:

$$\|X\|_{L^p} \le \|X\|_{L^q} \quad \text{for any } 0 \le p \le q \le \infty.$$

(b) Demonstrate that the inequality in part (a) can not be reversed: for any $0 \le p < q \le \infty$, find an example of a random variable X with $\|X\|_{L^p} < \infty$ and $\|X\|_{L^q} = \infty$.

1.12 ☕ (Interpolation between L^1 and L^∞) We know the L^p norm of any random variable X is bounded by the L^∞ norm. We can get an even better bound if we also know that the L^1 norm of X is small. Show that

$$\|X\|_{L^p} \le \|X\|_{L^1}^{\frac{1}{p}} \, \|X\|_{L^\infty}^{1-\frac{1}{p}} \quad \text{for any } 1 < p < \infty.$$

1.13 ☕☕☕ (Expectation of a maximum) Let $X_1, \ldots, X_n$ be nonnegative random variables.

(a) Prove that

$$\max_{i \le n} \mathbb{E}\, X_i \le \mathbb{E} \max_{i \le n} X_i \le n \cdot \max_{i \le n} \mathbb{E}\, X_i.$$

(b) Demonstrate that both inequalities in part (a) may be optimal. Specifically, find random variables $X_1, \ldots, X_n$ satisfying $\max_i \mathbb{E}\, X_i = \mathbb{E} \max_i X_i > 0$ and random variables Y_i satisfying $\mathbb{E} \max_i Y_i = n \cdot \max_i \mathbb{E}\, Y_i > 0$.

(c) Demonstrate that the upper bound in part (a) may be approximately optimal even for independent random variables. Specifically, find independent random variables $X_1, \ldots, X_n$ satisfying $\mathbb{E} \max_i X_i > cn \cdot \max_i \mathbb{E}\, X_i$, where $c > 0$ is an absolute constant.[5]

1.14 ☕☕ Let $X_1, \ldots, X_n$ be nonnegative random variables. Prove that for any $1 \le p < \infty$, we have

$$\left(\sum_{i=1}^n (\mathbb{E}\, X_i)^p \right)^{1/p} \le \mathbb{E} \left(\sum_{i=1}^n X_i^p \right)^{1/p} \le \left(\sum_{i=1}^n \mathbb{E}\left(X_i^p \right) \right)^{1/p}.$$

1.15 ☕ (Integrated tail formulas) Prove the following more general versions of Lemma 1.6.1.

[5] You are free to choose any value for $c > 0$ as long as it is an *absolute constant*, which means that c must not depend on anything, in particular on the distribution of the random variables Y_i or their number n. From the context it should be clear that choosing a small value such as $c = 0.001$ would make your job easier than choosing a large value such as $c = 1000$.

(a) Let X be any random variable, not necessarily nonnegative. Then

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

(b) Let X be a nonnegative random variable. Let $f : \mathbb{R}^+ \to \mathbb{R}^+$ be an increasing, differentiable function satisfying $f(0) = 0$. Then

$$\mathbb{E}\,f(X) = \int_0^\infty \mathbb{P}\{X > t\} f'(t) dt.$$

(c) Let X be a random variable, not necessarily nonnegative. Deduce that for every $p \in (0, \infty)$ we have

$$\mathbb{E}|X|^p = \int_0^\infty \mathbb{P}\{|X| > t\} p t^{p-1}\,dt$$

1.16 ☕☕ (Paley-Zygmund inequality) Markov inequality says a random variable is unlikely to be much bigger than its expectation. But what about the reverse? Can a nonnegative random variable be much smaller than its expectation with high probability? In general, yes (example?), but not if the second moment isn't too large. Let X be a nonnegative random variable with finite variance. Show that for any $\varepsilon \in [0, 1]$:

$$\mathbb{P}\{X > \varepsilon\,\mathbb{E}\,X\} \geq (1 - \varepsilon)^2 \frac{(\mathbb{E}\,X)^2}{\mathbb{E}[X^2]}.$$

1.17 ☕☕☕ (Comparison of the ℓ^p norms) Let $0 \leq p \leq q \leq \infty$.

(a) Prove that for any vector $x \in \mathbb{R}^n$ we have

$$\|x\|_q \leq \|x\|_p \leq n^{\frac{1}{p} - \frac{1}{q}} \|x\|_q.$$

(b) Demonstrate that both inequalities in part (a) can be optimal. Specifically, find nonzero vectors $x, y \in \mathbb{R}^n$ satisfying $\|x\|_p = \|x\|_q$ and $\|y\|_p = n^{\frac{1}{p} - \frac{1}{q}} \|y\|_q$.

1.18 ☕ (The ℓ^∞ norm is the limit of the ℓ^p norms) Consider any vector $x \in \mathbb{R}^n$.

(a) Prove that

$$\|x\|_p \to \|x\|_\infty \text{ as } p \to \infty.$$

(b) In fact, p does not need to be too large for the ℓ^p norm to get reasonably close to the ℓ^∞ norm. Show that if $p \geq \ln n$, then

$$\|x\|_\infty \leq \|x\|_p \leq e\|x\|_\infty.$$

1.19 ☕☕☕ (Duality of the ℓ_p norms) Let $p, p' \in [1, \infty]$ be conjugate exponents.

(a) Show that Hölder inequality is tight: for any vector x there exists a vector $y \neq 0$ for which $\langle x, y\rangle = \|x\|_p \|y\|_{p'}$.

(b) Conclude that for every vector $x \in \mathbb{R}^n$, we have

$$\max\left\{\langle x, y\rangle :\ y \in B^n_{p'}\right\} = \|x\|_p.$$

2

Concentration of Sums of Independent Random Variables

In this chapter, we step into the vast world of concentration inequalities. We start with some motivation in Section 2.1, then move on to key results – Hoeffding inequality (Sections 2.2 and 2.7), Chernoff inequality (Section 2.3), Bernstein inequality (Section 2.9) and Khintchine inequality (Section 2.7).

We also introduce two important classes of distributions: subgaussian (Section 2.6) and subexponential (Section 2.8). They are "natural habitats" for many results in high-dimensional probability and its applications.

We give two applications of concentration inequalities: for robust mean estimation in Section 2.4, and for random graphs in Section 2.5. More applications appear later in the book.

Be sure to try the exercises at the end of this chapter! They offer a guided tour of many more concepts, methods, results, and applications: the Mills ratio (Exercise 2.3), small ball probabilities (Exercise 2.7), a version of Le Cam's two-point method (Exercise 2.17), the expander mixing lemma for random graphs (Exercise 2.20), stochastic dominance (Exercises 2.27 and 2.28), Orlicz norms (Exercise 2.42), Bennett inequality (Exercise 2.48), and more.

2.1 Why concentration inequalities?

Concentration inequalities measure the deviation of a random variable X from its mean $\mathbb{E}\, X = \mu$. They typically provide two-sided bounds on the tails of $X - \mu$ such as:

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

The simplest concentration inequality is Chebyshev inequality (Corollary 1.6.3). It is very general but often too weak. Let us illustrate this with the example of the binomial distribution.

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

[0] This material will be published by Cambridge University Press as *High-Dimensional Probability, 2nd Edition* by Roman Vershynin. This pre-publication version is free to view and download for personal use only. Not for re-distribution, re-sale or use in derivative works. ©Roman Vershynin 202X.
If you want to be notified when the printed version is available, let me know: rvershyn@uci.edu

Let S_N denote the number of heads. Then S_N has binomial distribution: $S_N \sim \operatorname{Binom}(N, 1/2)$, and thus

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

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

$$\mathbb{P}\Big\{ S_N \ge \frac{3}{4}N \Big\} \le \mathbb{P}\Big\{ \Big| S_N - \frac{N}{2} \Big| \ge \frac{N}{4} \Big\} \le \frac{4}{N}. \tag{2.1}$$

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

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

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

where X_i are independent Bernoulli random variables with parameter $1/2$. (These X_i are the indicators of heads.) De Moivre-Laplace central limit theorem (1.25) states that the distribution of the normalized number of heads

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

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

$$\mathbb{P}\Big\{ S_N \ge \frac{3}{4}N \Big\} = \mathbb{P}\Big\{ Z_N \ge \sqrt{N/4} \Big\} \approx \mathbb{P}\Big\{ g \ge \sqrt{N/4} \Big\} \tag{2.2}$$

where $g \sim N(0, 1)$. Sadly, the Gaussian tail $\mathbb{P}\{g \ge t\}$ cannot be calculated analytically for general values of t. It is one of those "special functions" that cannot be expressed in terms of elementary functions. However, accurate approximations exist for the Gaussian tail:

Proposition 2.1.2 (Gaussian tails)**.** *Let $g \sim N(0, 1)$. Then for all $t > 0$, we have*

$$\frac{t}{t^2+1} \cdot \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \le \mathbb{P}\{g \ge t\} \le \frac{1}{t} \cdot \frac{1}{\sqrt{2\pi}} e^{-t^2/2}.$$

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

$$\mathbb{P}\{g \ge t\} \le \frac{1}{\sqrt{2\pi}} e^{-t^2/2}. \tag{2.3}$$

Proof To obtain an upper bound on the tail

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

make a change of 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. You will prove the lower bound in Exercise 2.2. □

Remark 2.1.3 (Tighter bounds)**.** Proposition 2.1.2 is sufficient for most purposes. If you ever need a more precise approximation, check out Exercise 2.3.

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

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

This quantity decays to zero *exponentially* fast in N, far better than the linear decay in (2.1) that follows from Chebyshev inequality.

However, the bound (2.4) does not rigorously follow from the central limit theorem. While the normal approximation in (2.2) is valid, the approximation error decays too slowly. This can be seen in the following quantitative version of the central limit theorem:

Theorem 2.1.4 (Berry-Esseen central limit theorem)**.** *In the setting of Theorem 1.7.3, for every $N \in \mathbb{N}$ and every $t \in \mathbb{R}$ we have*

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

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

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

Can the approximation error in the central limit theorem approximation error be improved? Generally, no. For even N, the probability of exactly half the flips being heads equals

$$\mathbb{P}\{S_N = N/2\} = 2^{-N} \binom{N}{N/2} \approx \sqrt{\frac{2}{\pi N}}.$$

The last estimate follows from Stirling approximation (Lemma 1.7.7). (Check!) Hence

$$\mathbb{P}\{Z_N = 0\} \approx \sqrt{\frac{2}{\pi N}} \quad \text{while} \quad \mathbb{P}\{g = 0\} = 0;$$

the last equality holds because the normal distribution is continuous. Thus the approximation error has to be of the order of $1/\sqrt{N}$.

In summary, the central limit theorem approximates $S_N = X_1 + \cdots + X_N$ by a normal distribution, which is known for its light, exponentially decaying tails. However, the approximation error decays too slowly, hindering the proof that

S_N has such light tails. To address this problem, we will develop an alternative approach to concentration that bypasses the central limit theorem.

2.2 Hoeffding inequality

We start with a particularly simple concentration inequality for sums of independent Rademacher random variables. Recall that a random variable X has the *Rademacher* distribution if it takes values -1 and 1 with probability $1/2$ each:

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

Theorem 2.2.1 (Hoeffding inequality)**.** *Let $X_1,\ldots,X_N$ be independent Rademacher random variables, and let $a=(a_1,\ldots,a_N)\in\mathbb{R}^N$ be fixed. Then, for any $t\geq 0$, we have*

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

Proof This argument might be called the *exponential moment method.*

Remember how we proved Chebyshev inequality (Corollary 1.6.3): we squared both sides and applied Markov inequality. Let us do something similar here. But instead of squaring both sides, let us multiply by a fixed parameter $\lambda\geq 0$ (whose exact value we will choose later) and exponentiate. This gives us

$$\begin{aligned}\mathbb{P}\Big\{\sum_{i=1}^N a_iX_i\geq t\Big\}&=\mathbb{P}\Big\{\exp\Big(\lambda\sum_{i=1}^N a_iX_i\Big)\geq\exp(\lambda t)\Big\}\\&\leq e^{-\lambda t}\,\mathbb{E}\exp\Big(\lambda\sum_{i=1}^N a_iX_i\Big).\end{aligned}\tag{2.5}$$

In the last step we applied Markov inequality (Proposition 1.6.2).

We have reduced the problem to bounding the *moment generating function* (MGF) of the sum $\sum_{i=1}^N a_iX_i$. As we remember from a basic probability course, the MGF of the sum is the product of the MGFs of the terms; this follows from independence. So

$$\mathbb{E}\exp\Big(\lambda\sum_{i=1}^N a_iX_i\Big)=\prod_{i=1}^N\mathbb{E}\exp(\lambda a_iX_i).\tag{2.6}$$

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

$$\mathbb{E}\exp(\lambda a_iX_i)=\frac{1}{2}\exp(\lambda a_i)+\frac{1}{2}\exp(-\lambda a_i)=\cosh(\lambda a_i).$$

Now we use the numeric inequality

$$\cosh(x)\leq\exp(x^2/2)\quad\text{for all } x\in\mathbb{R},\tag{2.7}$$

which can be verified by comparing the Taylor expansions of both sides (Exercise 2.5). This gives

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

Substituting into (2.6) and then into (2.5), we obtain

$$\begin{aligned}\mathbb{P}\Big\{\sum_{i=1}^N a_i X_i \ge t\Big\} \le e^{-\lambda t} \prod_{i=1}^N \exp(\lambda^2 a_i^2/2) &= \exp\Big(-\lambda t + \frac{\lambda^2}{2}\sum_{i=1}^N a_i^2\Big) \\ &= \exp\Big(-\lambda t + \frac{\lambda^2}{2}\|a\|_2^2\Big). \qquad (2.8)\end{aligned}$$

This bound holds for any $\lambda \ge 0$. Let us choose the best value of λ, the one for which the expression (2.8) is smallest. The minimum of (2.8) is attained for $\lambda = t/\|a\|_2^2$, and it equals equals $\exp(-t^2/2\|a\|_2^2)$. (Check!) This completes the proof of Hoeffding inequality. □

Remark 2.2.2 (Exponentially light tails)**.** Hoeffding inequality can be seen as a concentration version of the central limit theorem. With normalization $\|a\|_2 = 1$, it gives exponentially light tails $e^{-t^2/2}$, nearly matching the standard normal tail in (2.3).

Remark 2.2.3 (Non-asymptotic theory)**.** Unlike the classical limit theorems of probability theory, Hoeffding inequality is *non-asymptotic* in that it holds for all fixed N as opposed to $N \to \infty$. As we will see later, non-asymptotic results are attractive for applications in the data sciences, where N often corresponds to the sample size.

Remark 2.2.4 (The probability of $\frac{3}{4}N$ heads)**.** Using Hoeffding inequality, we can now return to Question 2.1.1 and bound the probability of at least $\frac{3}{4}N$ heads in N tosses of a fair coin. Note that if $Y \sim \text{Ber}(1/2)$, then $2Y - 1$ is Rademacher random variable. Since number of heads S_N is a sum of N independent Ber(1/2) variables, $2S_N - N$ is a sum of N independent Rademacher variables. Hence

$$\mathbb{P}\Big\{\text{at least } \frac{3}{4}N \text{ heads}\Big\} = \mathbb{P}\Big\{S_N \ge \frac{3}{4}N\Big\} = \mathbb{P}\Big\{2S_N - N \ge \frac{N}{2}\Big\} \le \exp(-N/8).$$

In other words, the probability is *exponentially small* in the number of tosses. We have obtained a rigorous bound that is quite close to our heuristic guess (2.4).

We can easily extend Hoeffding inequality for *two-sided tails* $\mathbb{P}\{|S_N| \ge t\}$ where $S_N = \sum_{i=1}^N a_i X_i$. To do this, express the event $|S_N| \ge t$ as the union of two events: $S_N \ge t$ and $-S_N \ge t$. The union bound gives

$$\mathbb{P}\{|S_N| \ge t\} \le \mathbb{P}\{S_N \ge t\} + \mathbb{P}\{-S_N \ge t\}.$$

Applying Hoeffding inequality twice, for random variables X_i and $-X_i$, we obtain a two-sided tail after paying only a factor of 2:

Theorem 2.2.5 (Hoeffding inequality, two-sided)**.** *Let $X_1, \ldots, X_N$ be independent Rademacher random variables, and $a = (a_1, \ldots, a_N) \in \mathbb{R}^N$. Then, for any $t > 0$, we have*

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

The exponential moment method we used to prove Hoeffding inequality is quite flexible. It applies far beyond the canonical example of the Rademacher distribution. For example, in Exercise 2.10 you will prove the following extension of Hoeffding inequality for general bounded random variables:

Theorem 2.2.6 (Hoeffding inequality for bounded random variables)**.** *Let $X_1, \ldots, X_N$ be independent random variables such that $X_i \in [a_i, b_i]$ for every i. Then, for any $t > 0$, we have*

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

2.3 Chernoff inequality

Hoeffding inequality is quite sharp for Rademacher random variables, but it may be too conservative for general bounded random variables. For example, when Bernoulli random variables X_i have small means p_i, their sum approximates a Poisson distribution (Theorem 1.7.6). Hoeffding inequality (Theorem 2.2.6) does not account for small p_i, yielding a Gaussian bound that is far from the true Poisson tail. Let's prove an inequality that is sensitive to the sizes of p_i:

Theorem 2.3.1 (Chernoff 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*

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

Proof We will use the same exponential moment method as in the proof of Hoeffding inequality (Theorem 2.2.1). Just like before, we multiply the inequality $S_N \geq t$ by $\lambda \geq 0$, exponentiate, then apply Markov inequality and use independence to get (2.6):

$$\mathbb{P}\{S_N \geq t\} \leq e^{-\lambda t} \prod_{i=1}^N \mathbb{E} \exp(\lambda X_i). \tag{2.9}$$

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

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

In the last step, we used the inequality $1 + x \le e^x$. So,

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

Substituting this into (2.9), we obtain

$$\mathbb{P}\{S_N \ge t\} \le e^{-\lambda t} \exp\left[(e^\lambda - 1)\mu\right] = \exp\left[-\lambda t + (e^\lambda - 1)\mu\right].$$

This bound holds for any $\lambda > 0$. The minimum of the expression on the right hand side is attained for $\lambda = \ln(t/\mu)$ (check!), which is nonnegative by the assumption that $t \ge \mu$. Substitute this value of λ and simplify to complete the proof. □

Remark 2.3.2 (Chernoff inequality: left tails)**.** By slightly modifying the proof of Chernoff inequality, we can also obtain a bound on the left tail:

$$\mathbb{P}\{S_N \le t\} \le e^{-\mu}\left(\frac{e\mu}{t}\right)^t \quad \text{for every } 0 < t \le \mu.$$

You will prove this bound in Exercise 2.11.

Remark 2.3.3 (Poisson tails)**.** Does the expression $e^{-\mu}(e\mu/t)^t$ in Chernoff inequality seem obscure? Think about what happens for very small p_i. The Poisson limit theorem (Theorem 1.7.6) suggests $S_N \approx \mathrm{Pois}(\mu)$. Using Stirling approximation as in (1.27), we expect that

$$\mathbb{P}\{S_N = t\} \approx \frac{e^{-\mu}}{\sqrt{2\pi t}}\left(\frac{e\mu}{t}\right)^t \quad \text{for every fixed integer } t > 0.$$

Chernoff inequality gives a similar result, but rigorous and non-asymptotic. It essentially bounds the *entire* tail $\mathbb{P}\{S_N \ge t\}$ by *one* value $\mathbb{P}\{S_N = t\}$ in that tail!

The Poisson tails are heavier than Gaussian, decaying as $t^{-t} = e^{-t\log t}$, which exceeds e^{-t^2} for large t. Fortunately, this occurs only for t much larger than the mean μ. For small deviations, the Poisson tail resembles the Gaussian:

Corollary 2.3.4 (Chernoff inequality: small deviations)**.** *In the setting of Theorem 2.3.1, we have*

$$\mathbb{P}\{|S_N - \mu| \ge \delta\mu\} \le 2\exp\left(-\frac{\delta^2\mu}{3}\right) \quad \textit{for every } 0 \le \delta \le 1.$$

Proof Using Theorem 2.3.1 with $t = (1+\delta)\mu$ and simplifying, we get

$$\mathbb{P}\{S_N \ge (1+\delta)\mu\} \le \exp\left[-\mu\left((1+\delta)\ln(1+\delta) - \delta\right)\right]. \tag{2.10}$$

Now expand the expression inside exponent into a Taylor series:

$$(1+\delta)\ln(1+\delta) - \delta = \frac{\delta^2}{2} - \frac{\delta^3}{2\cdot 3} + \frac{\delta^4}{3\cdot 4} - \frac{\delta^5}{4\cdot 5} + \cdots \ge \frac{\delta^2}{3}.$$

(To check the last bound, subtract $\delta^2/3$ from both sides – this makes a series with alternating signs and decreasing terms.) Plug this into (2.10) and get

$$\mathbb{P}\{S_N \ge (1+\delta)\mu\} \le \exp\left(-\frac{\delta^2\mu}{3}\right).$$

The left tail $\mathbb{P}\{S_N \le (1-\delta)\mu\}$ can be bounded similarly, using Remark 2.3.2. (Try it! You should get even a better bound: $\exp(-\delta^2\mu/2)$.) Combine the two tails by a union bound to complete the proof. □

To see a Gaussian tail in Corollary 2.3.4, try normalizing S_N roughly as in the central limit theorem. If we set $Z_N = (S_N - \mu)/\sqrt{\mu}$, we can rewrite the conclusion of Corollary 2.3.4 as follows:

$$\mathbb{P}\{|Z_N| \ge t\} \le 2\exp\left(-t^2/3\right) \quad \text{for every } 0 \le t \le \sqrt{\mu}.$$

Remark 2.3.5 (Small and large deviations)**.** Figure 2.1 shows the probability mass function of $\mathrm{Binom}(N, \mu/N)$ with $N = 200$ and $\mu = 10$. Near the mean μ, the bell-shaped curve reflects Gaussian behavior. Far to the right, the slower decay reflects Poisson behavior. These two regimes align with the central limit theorem and the Poisson limit theorem, respectively.

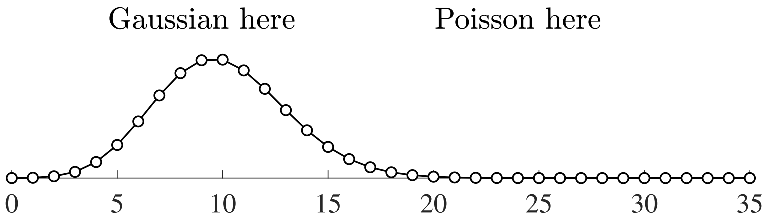


Figure 2.1 The probability mass function of the distribution $\mathrm{Binom}(N, \mu/N)$ with $N = 200$ and $\mu = 10$. It is approximately normal near the mean μ, but it is heavier far from the mean.

2.4 Application: median-of-means estimator

A fundamental data science problem is learning an unknown distribution from a sample – for example inferring a country's income distribution from a survey of N people.

The most basic task is estimating the mean. Let X be a random variable with mean μ and variance σ^2 (representing a population), and let $X_1, \dots, X_N$ be independent copies[1] of X (representing a sample). We would like to find an estimator $\widehat{\mu} = \widehat{\mu}(X_1, \dots, X_N)$ that satisfies $\widehat{\mu} \approx \mu$ with high probability.

The simplest and most popular mean estimator is the *sample mean*

$$\widehat{\mu} := \frac{1}{N}\sum_{i=1}^{N} X_i. \tag{2.11}$$

The expected value and variance of this estimator are

$$\mathbb{E}\,\widehat{\mu} = \mu; \quad \mathrm{Var}(\widehat{\mu}) = \frac{1}{N^2}\sum_{i=1}^{N}\mathrm{Var}(X_i) = \frac{\sigma^2}{N}. \tag{2.12}$$

[1] Independent copies of X are independent random variables that have the same distribution as X.

Then Chebyshev inequality gives

$$\mathbb{P}\Big\{|\widehat{\mu}-\mu| \ge \frac{t\sigma}{\sqrt{N}}\Big\} \le \frac{1}{t^2} \quad \text{for every } t > 0. \tag{2.13}$$

For example, (2.13) ensures the error is at most $10\sigma/\sqrt{N}$ with at least 99% probability – an acceptable solution to the mean estimation problem.

But is this solution optimal? Is it possible for the probability in (2.13) to decay more quickly than $1/t^2$? For the Gaussian distribution, the answer is yes. If $X \sim N(\mu, \sigma^2)$, then $\widehat{\mu} \sim N(\mu, \sigma^2/N)$, and $\frac{\widehat{\mu}-\mu}{\sigma/\sqrt{N}} \sim N(0,1)$. Using the Gaussian tail bound (2.3) twice (for the right and left tails) gives

$$\mathbb{P}\Big\{|\widehat{\mu}-\mu| \ge \frac{t\sigma}{\sqrt{N}}\Big\} \le \sqrt{\frac{2}{\pi}}e^{-t^2/2} \quad \text{for every } t \ge 1.$$

For example, the error is at most $3\sigma/\sqrt{N}$ with at least 99% probability.

One might suspect that such a strong, Gaussian tail decay requires Gaussian-like assumptions. But surprisingly, a mean estimator exists with Gaussian tail decay that works for *any* distribution with finite variance!

Theorem 2.4.1 (Median-of-means estimator). *Let X be a random variable with mean μ and variance σ^2, and let $X_1, \dots, X_N$ be independent copies of X. For any $0 \le t \le \sqrt{N}$, there exists an estimator $\widehat{\mu} = \widehat{\mu}(X_1, \dots, X_N)$ that satisfies*

$$\mathbb{P}\Big\{|\widehat{\mu}-\mu| \ge \frac{t\sigma}{\sqrt{N}}\Big\} \le 2e^{-ct^2},$$

where $c > 0$ is an absolute constant.

The estimator $\widehat{\mu}$ we will construct is known as the *median-of-means estimator*. A median of a finite set of real numbers $\{x_1, \dots, x_n\}$, denoted

$$\mathrm{Med}(x_1, \dots, x_n),$$

is a value M such that at least half the numbers[2] satisfy $x_i \le M$ and at least half the numbers satisfy $x_i \ge M$.

While the median lacks some convenient properties of the mean, such as linearity, it has a major advantage: robustness. If you send one sample point x_i off to infinity, the mean becomes infinite – but the median stays put or just shifts a bit (at most to the next point – why?).

Proof of Theorem 2.4.1 Assume for simplicity that $N = BL$ for some integers B and L. Divide the sample $X_1, \dots, X_N$ into B blocks of length L, compute each block's sample mean, and take their median:

$$\mu_b = \frac{1}{L}\sum_{i=(b-1)L+1}^{bL} X_i, \quad \widehat{\mu} = \mathrm{Med}(\mu_1, \dots, \mu_B).$$

[2] The median may not be unique: for the set $\{1,2,3,4\}$, any value in $[2,3]$ is a median. This ambiguity can be resolved by taking the midpoint of the interval of all possible medians, like 2.5 in this case.

Arguing as in (2.12) we see that each random variable μ_b has expected value μ and variance σ^2/L. Then Chebyshev inequality yields

$$\mathbb{P}\Big\{\mu_b \geq \mu + \frac{t\sigma}{\sqrt{N}}\Big\} \leq \frac{N}{t^2 L} = \frac{B}{t^2} = \frac{1}{4}$$

if we choose the number of blocks to be $B = t^2/4$. By definition of median,

$$\mathbb{P}\Big\{\widehat{\mu} \geq \mu + \frac{t\sigma}{\sqrt{N}}\Big\} \leq \mathbb{P}\Big\{\text{at least half of the numbers } \mu_1, \dots, \mu_B \text{ are} \geq \mu + \frac{t\sigma}{\sqrt{N}}\Big\}.$$

We are looking at B independent events, each occurring with probability at most $1/4$. The probability that at least half of them occur is bounded by $\exp(-c_0 B)$ for some absolute constant $c_0 > 0$, following from Hoeffding inequality (Theorem 2.2.6) – check! Hence

$$\mathbb{P}\Big\{\widehat{\mu} \geq \mu + \frac{t\sigma}{\sqrt{N}}\Big\} \leq \exp(-c_0 B) = \exp(-c_0 t^2/4).$$

Similarly, the probability that $\widehat{\mu} \leq \mu - \frac{t\sigma}{\sqrt{N}}$ has the same bound. Combining these two bounds completes the proof.

There is a slight inaccuracy in our argument though. The number of blocks B must be an integer that divides N, while our choice $B = t^2/4$ ensures only that $0 \leq B \leq N$ by assumption on t. You will fix this issue in Exercise 2.16. □

2.5 Application: degrees of random graphs

We apply Chernoff inequality for a classic combinatorial object: a random graph.

The simplest model of a random graph is the *Erdős-Rényi model* $G(n, p)$, which is constructed on a set of n vertices by connecting every pair of distinct vertices independently with probability p. Figure 2.2 shows two examples. The Erdős-Rényi model model often serves as the simplest stochastic model for large real-world networks.

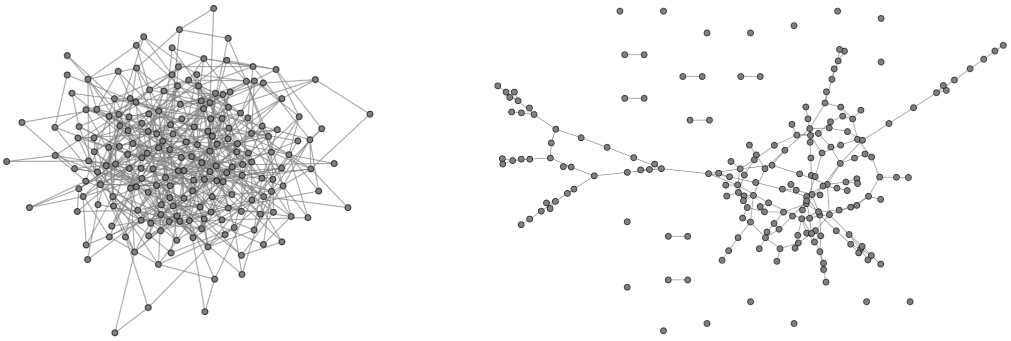


Figure 2.2 Examples of random graphs in the Erdős-Rényi model $G(n, p)$ with $n = 200$ vertices and connection probabilities $p = 0.03$ (left) and $p = 0.01$ (right).

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

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

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

Proposition 2.5.1 (Dense graphs are almost regular)**.** *There is an absolute constant C such that the following holds. Consider a random graph $G \sim G(n, p)$ with expected degree satisfying $d \geq C \log n$. Then, with probability at least* 0.99*, the following event occurs: all vertices of G have degrees between* $0.9d$ *and* $1.1d$*.*

Proof The argument is a combination of concentration with a union bound. Let us fix a vertex i of the graph. The degree of i, which we denote d_i, is a sum of $n - 1$ independent $\text{Ber}(p)$ random variables (the indicators of the edges incident to i). Applying Chernoff inequality (Corollary 2.3.4) we get

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

This bound holds for each fixed vertex i. Next, we can "unfix" i by taking the union bound (Lemma 1.4.1) over all n vertices. We obtain

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

If $d \geq C \log n$ for a sufficiently large absolute constant C, the probability is bounded by 0.01. This means that with probability 0.99, the complementary event occurs: $\forall i \leq n$: $|d_i - d| < 0.1d$. The proof is complete. □

Remark 2.5.2 (Sparse graphs are far from regular)**.** You may wonder if the condition $d \gtrsim \log n$ in Proposition 2.5.1 is optimal. It is: if $d < (1 - \varepsilon) \ln n$, an isolated vertex appears ("a friendless student" in Exercise 1.10), making the minimum degree zero. If you have not done Exercises 1.9 and 1.10 yet, give them a try them now, along with Exercises 2.18–2.20.

2.6 Subgaussian distributions

Let us take a fresh look at Hoeffding inequality for sums of independent random variables X_i:

$$\mathbb{P}\Big\{\Big|\sum_{i=1}^{N} a_i X_i\Big| \geq t\Big\} \leq 2 \exp\left(-\frac{ct^2}{\|a\|_2^2}\right) \quad \text{for all } t \geq 0. \tag{2.14}$$

This result holds for Rademacher random variables X_i (Theorem 2.2.1), more generally for any mean-zero bounded random variables X_i (Theorem 2.2.6), and for standard normal random variables X_i (why?).

This makes us wonder: what is the largest class of distributions for which

Hoeffding inequality holds? If the sum $\sum_{i=1}^N a_i X_i$ consists of a single term X_i, (2.14) becomes

$$\mathbb{P}\{|X_i| > t\} \le 2e^{-ct^2} \quad \text{for all } t \ge 0. \tag{2.15}$$

We will soon show that this condition is also sufficient.

Distributions that satisfy (2.15) are referred to *subgaussian distributions*. They represent a natural and often canonical class for deriving results in high-dimensional probability theory and its applications.

Because of the importance of subgaussian distributions, it would be useful to find other ways to express (2.15). For inspiration, let us turn to the standard normal random variable $X \sim N(0,1)$. Its moment generation function is

$$\mathbb{E}\exp(\lambda X) = e^{\lambda^2/2} \quad \text{for all } \lambda \in \mathbb{R}, \tag{2.16}$$

and in Exercise 2.22 you will check that the absolute moments of X satisfy

$$\|X\|_{L^p} = (\mathbb{E}|X|^p)^{1/p} \le C\sqrt{p} \quad \text{for all } p \ge 1. \tag{2.17}$$

It turns out that these properties – the subgaussian tail decay like in (2.15), the growth of the moment generating function like in (2.16), and the growth of moments like in (2.17) – are all equivalent for general distributions. They all convey the same message: a given distribution is bounded by a normal distribution.

Proposition 2.6.1 (Subgaussian properties). *Let X be a random variable. The following properties are equivalent, with the parameters $K_i > 0$ differing by at most an absolute constant factor.*[3]

(i) (Tails) There exists $K_1 > 0$ such that

$$\mathbb{P}\{|X| \ge t\} \le 2\exp(-t^2/K_1^2) \quad \textit{for all } t \ge 0.$$

(ii) (Moments) There exists $K_2 > 0$ such that

$$\|X\|_{L^p} = (\mathbb{E}|X|^p)^{1/p} \le K_2\sqrt{p} \quad \textit{for all } p \ge 1.$$

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

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

Moreover, if $\mathbb{E}X = 0$ then properties (i)–(iii) are equivalent to the following one:

(iv) (MGF) There exists $K_4 > 0$ such that

$$\mathbb{E}\exp(\lambda X) \le \exp(K_4^2\lambda^2) \quad \textit{for all } \lambda \in \mathbb{R}.$$

Proof The proof is a bit long but very instructive: you will learn how to transform one type of information about random variables into another.

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

(i)⇒(ii) Assume property (i) holds. Rescaling X to X/K_1 we can assume that $K_1 = 1$. (Check!) The integrated tail formula (Lemma 1.6.1) for $|X|^p$ gives

$$\begin{aligned}
\mathbb{E}|X|^p &= \int_0^\infty \mathbb{P}\{|X|^p \geq u\}\,du \\
&= \int_0^\infty \mathbb{P}\{|X| \geq t\}\,pt^{p-1}\,dt \quad \text{(change of variables } u = t^p\text{)} \\
&\leq \int_0^\infty 2e^{-t^2} pt^{p-1}\,dt \quad \text{(by property (i))} \\
&= p\Gamma(p/2) \quad \text{(set } t^2 = s \text{ and use definition (1.30) of Gamma function)} \\
&\leq 3p(p/2)^{p/2} \quad \text{(since } \Gamma(x) \leq 3x^x \text{ for all } x \geq 1/2\text{. Check!)}
\end{aligned}$$

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

(ii)⇒(iii) Assume property (ii) holds. Rescaling X to X/K_2 we can assume that $K_2 = 1$. (Check!) Using the Taylor series expansion of the exponential function, we obtain

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

Property (ii) guarantees that $\mathbb{E}[X^{2p}] \leq (2p)^p$, while $p! \geq (p/e)^p$ by Lemma 1.7.8. Substituting these two bounds, we get

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

if we choose $\lambda = 1/2\sqrt{e}$. This yields property (iii) with $K_3 = 2\sqrt{e}$.

(iii)⇒(i) Assume property (iii) holds. As before, we may assume that $K_3 = 1$. Then, exponentiating and using Markov inequality, we get

$$\mathbb{P}\{|X| \geq t\} = \mathbb{P}\{e^{X^2} \geq e^{t^2}\} \leq e^{-t^2}\,\mathbb{E}\,e^{X^2} \leq 2e^{-t^2}.$$

This proves property (i) with $K_1 = 1$.

To prove the "moreover" part, we show that (iii) ⇒ (iv) and (iv) ⇒ (i).
(iii)⇒(iv) Assume that property (iii) holds. As before, assume that $K_3 = 1$. Let us use the numeric inequality

$$e^x \leq 1 + x + \frac{x^2}{2} e^{|x|}$$

which follows from the Taylor theorem with the Lagrange form of the remainder.

Substitute $x = \lambda X$ to get

$$\begin{aligned}\mathbb{E}\, e^{\lambda X} &\le 1 + \frac{\lambda^2}{2}\, \mathbb{E}\, X^2 e^{|\lambda X|} \quad \text{(since } \mathbb{E}\, X = 0 \text{ by assumption)}\\ &\le 1 + \frac{\lambda^2}{2} e^{\lambda^2/2}\, \mathbb{E}\, e^{X^2} \quad \text{(since } x^2 \le e^{x^2/2} \text{ and } |\lambda x| \le \lambda^2/2 + x^2/2)\\ &\le (1+\lambda^2) e^{\lambda^2/2} \quad \text{(since } \mathbb{E}\, e^{X^2} \le 2 \text{ by assumption (iii))}\\ &\le e^{3\lambda^2/2} \quad \text{(since } 1 + z \le e^z).\end{aligned}$$

This proves property (iv) with $K_4 = \sqrt{3/2}$.

(iv)⇒(i) Assume property (iv) holds; we can assume that $K_4 = 1$. Let us apply the exponential moment method, which we first learned in the proof of Hoeffding inequality (Theorem 2.2.1). Let $\lambda > 0$ be a parameter to be chosen later. Exponentiating and applying Markov inequality, we get

$$\mathbb{P}\{X \ge t\} = \mathbb{P}\{e^{\lambda X} \ge e^{\lambda t}\} \le e^{-\lambda t}\, \mathbb{E}\, e^{\lambda X} \le e^{-\lambda t} e^{\lambda^2} = e^{-\lambda t + \lambda^2}.$$

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 obtain $\mathbb{P}\{X \le -t\} \le e^{-t^2/4}$. Combining these two bounds yields $\mathbb{P}\{|X| \ge t\} \le 2e^{-t^2/4}$. Thus property (i) holds with $K_1 = 2$. The proposition is proved. □

Remark 2.6.2 (Zero mean)**.** You might wonder why we assumed $\mathbb{E}\, X = 0$ in property (iv). In Exercise 2.23 you will show that any random variable X satisfying (iv) *must* have zero mean.

Remark 2.6.3 (On constant factors)**.** The constant 2 in properties (i) and (iii) does not have any special meaning; it can be replaced with any absolute constant greater than 1. (Check!)

2.6.1 The subgaussian norm

Definition 2.6.4 (Subgaussian distributions)**.** A random variable X is called *subgaussian* if it satisfies any of the equivalent properties (i)–(iii) in Proposition 2.6.1. Its *subgaussian norm*, denoted $\|X\|_{\psi_2}$, is the smallest K_3 satisfying property (iii). In other words, we define

$$\|X\|_{\psi_2} = \inf\left\{K > 0 :\ \mathbb{E} \exp(X^2/K^2) \le 2\right\}. \tag{2.18}$$

In Exercise 2.42, you will confirm that $\|\cdot\|_{\psi_2}$ indeed defines a norm on the space of subgaussian random variables. A key part of this statement is the triangle inequality: any random variables X and Y, not necessarily independent, satisfy

$$\|X + Y\|_{\psi_2} \le \|X\|_{\psi_2} + \|Y\|_{\psi_2}.$$

Example 2.6.5. The following random variables are subgaussian; you will compute their subgaussian norms in Exercises 2.24, 2.33:

(a) normal,
(b) Rademacher,
(c) Bernoulli,
(d) Binomial,
(e) any bounded random variable.

Conversely, the exponential, Poisson, geometric, chi-squared, Gamma, Cauchy, and Pareto distributions are *not* subgaussian (Exercise 2.25).

In light of the definition of the subgaussian norm, Proposition 2.6.1 yields:

Proposition 2.6.6 (Subgaussian bounds)**.** *Every subgaussian random variable X satisfies the following bounds.*

(i) (Tails) $\mathbb{P}\{|X| \geq t\} \leq 2\exp\left(-ct^2/\|X\|_{\psi_2}^2\right)$ *for all* $t \geq 0$.

(ii) (Moments) $\|X\|_{L^p} \leq C\|X\|_{\psi_2}\sqrt{p}$ *for all* $p \geq 1$.

(iii) (MGF of X^2) $\mathbb{E}\exp\left(X^2/\|X\|_{\psi_2}^2\right) \leq 2$.

(iv) (MGF) If $\mathbb{E}\,X = 0$ *then* $\mathbb{E}\exp(\lambda X) \leq \exp\left(C\lambda^2\|X\|_{\psi_2}^2\right)$ *for all* $\lambda \in \mathbb{R}$.

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

There is a number of other equivalent ways to describe subgaussian distributions – discover some of them in Exercises 2.26–2.28, 2.39. Also, there is a sharper way to define a subgaussian norm, useful if you don't want to lose any absolute constant factors – try Exercise 2.40 now!

2.7 Subgaussian Hoeffding and Khintchine inequalities

After our hard work characterizing subgaussian distributions in the previous section, let us explore how these results are useful.

The fundamental property of variance (1.8) implies that independent mean-zero random variables $X_1, \ldots, X_N$ obey the Pythagorean theorem:

$$\Big\|\sum_{i=1}^N X_i\Big\|_{L^2}^2 = \sum_{i=1}^N \|X_i\|_{L^2}^2. \tag{2.19}$$

A similar but slightly weaker property holds for the subgaussian norm:

Proposition 2.7.1 (Subgaussian norm of a sum)**.** *Let $X_1, \ldots, X_N$ be independent, mean-zero, subgaussian random variables. Then*

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

where C is an absolute constant.

Proof Let us compute the moment generating function of the sum $S_N = \sum_{i=1}^N X_i$. For any $\lambda \in \mathbb{R}$, we have

$$\begin{aligned}\mathbb{E}\exp(\lambda S_N) &= \prod_{i=1}^N \mathbb{E}\exp(\lambda X_i) \quad \text{(by independence)} \\ &\le \prod_{i=1}^N \exp\left(C\lambda^2 \|X_i\|_{\psi_2}^2\right) \quad \text{(by Proposition 2.6.6(iv))} \\ &= \exp(\lambda^2 K^2) \quad \text{where } K^2 = C\sum_{i=1}^N \|X_i\|_{\psi_2}^2.\end{aligned}$$

Due to (iii)⇔(iv) in Proposition 2.6.1, this gives

$$\mathbb{E}\exp(cS_N^2/K^2) \le 2$$

where $c > 0$ is an absolute constant. By definition of the subgaussian norm (2.18), it follows that $\|S_N\|_{\psi_2} \le K/\sqrt{c}$, finishing the proof. □

Remark 2.7.2. (Is there a reverse bound?) Given the Pythagorean identity (2.19), you might wonder if the inequality in Proposition 2.7.1 can be reversed. You will show in Exercises 2.33 and 2.34 that the answer is "yes" for identical distributions of X_i, but "no" in general.

2.7.1 Subgaussian Hoeffding inequality

We can rephrase Proposition 2.7.1 in terms of subgaussian tails, using Proposition 2.6.6(i):

Theorem 2.7.3 (Subgaussian Hoeffding inequality)**.** *Let $X_1, \ldots, X_N$ be independent, mean-zero, subgaussian random variables. Then, for every $t \ge 0$, we have*[4]

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

Example 2.7.4 (Recovering classical Hoeffding)**.** Let X_i follow the Rademacher distribution, and apply Theorem 2.7.3 to the random variables $a_i X_i$ instead of X_i. Noting that $\|a_i X_i\|_{\psi_2} = |a_i| \cdot \|X_i\|_{\psi_2}$ and that $\|X_i\|_{\psi_2}$ is an absolute constant (why?), we get

$$\mathbb{P}\left\{\left|\sum_{i=1}^N a_i X_i\right| \ge t\right\} \le 2\exp\left(-\frac{ct^2}{\|a\|_2^2}\right). \tag{2.20}$$

This is our old friend – Hoeffding inequality for the Rademacher distribution (Theorem 2.2.1), just with a different absolute constant c instead of $1/2$. The same reasoning can be used to recover Hoeffding inequality for general bounded random variables (Theorem 2.2.6), again up to an absolute constant (Exercise 2.29).

[4] From now on, we will occasionally omit the phrase "where $c > 0$ is an absolute constant," as this is always implied.

2.7.2 Subgaussian Khintchine inequality

Let us establish another classic result: a two-sided bound on the L^p norms of sums of independent random variables.

Theorem 2.7.5 (Khintchine inequality)**.** *Let $X_1, \ldots, X_N$ be independent subgaussian random variables with zero means and unit variances, and let $a_1, \ldots, a_N \in \mathbb{R}$. Then, for every $p \in [2, \infty)$, we have*

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

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

Proof For $p = 2$, we actually have an equality, since the Pythagorean identity (2.19) combined with the unit variance assumption gives

$$\Big\|\sum_{i=1}^N a_i X_i\Big\|_{L^2} = \Big(\sum_{i=1}^N a_i^2 \|X_i\|_{L^2}^2\Big)^{1/2} = \Big(\sum_{i=1}^N a_i^2\Big)^{1/2}.$$

Then the lower bound in the theorem follows from the monotonicity of the L^p norms (1.20). For the upper bound, use Proposition 2.7.1:

$$\Big\|\sum_{i=1}^N a_i X_i\Big\|_{\psi_2} \le C\Big(\sum_{i=1}^N a_i^2 \|X_i\|_{\psi_2}^2\Big)^{1/2} \le CK\Big(\sum_{i=1}^N a_i^2\Big)^{1/2},$$

and finish by applying Proposition 2.6.6(ii). □

You will extend Khintchine's inequality for $p \in [1, 2]$ in Exercise 2.36.

2.7.3 Maximum of subgaussians

So far, we have focused on *sums* of random variables. But what about other, possibly nonlinear, functions? We will explore them in depth in Chapter 5. For now, let us give one example: a version of Proposition 2.7.1 for the *maximum.*

Proposition 2.7.6 (Maximum of subgaussians)**.** *Let $X_1, \ldots, X_N$ be subgaussian random variables for some $N \ge 2$, that are not necessarily independent. Then*

$$\big\|\max_{i\le N} X_i\big\|_{\psi_2} \le C\sqrt{\log N}\, \max_{i\le N} \|X_i\|_{\psi_2}. \tag{2.21}$$

In particular, we have

$$\mathbb{E}\max_{i\le N} X_i \le CK\sqrt{\log N} \tag{2.22}$$

where $K = \max_{i\le N}\|X_i\|_{\psi_2}$. The same bounds obviously hold for $\max_i |X_i|$. (Why?)

Proof We give two different proofs of (2.21); pick the one you prefer.

First proof is based on the union bound. Without loss of generality, we can assume that $\max_i \|X_i\|_{\psi_2} = 1$. (Why?) For any $t \ge 0$, we have

$$\mathbb{P}\Big\{\max_{i \le N} X_i \ge t\Big\} \le \sum_{i=1}^N \mathbb{P}\{X_i \ge t\} \le 2N \exp(-ct^2)$$

where the subgaussian tail comes from Proposition 2.6.6(i). If $N \le \exp(ct^2/2)$, then the probability above is bounded by $2\exp(-ct^2/2)$, which is stronger than we need. If $N > \exp(ct^2/2)$, then the probability of any event is trivially bounded by $2\exp(-ct^2/3\ln N)$, since this quantity exceeds 1. Thus, in either case,

$$\mathbb{P}\Big\{\max_{i \le N} X_i \ge t\Big\} \le 2\exp\left(-\frac{ct^2}{3\ln N}\right) \quad \text{for any } t \ge 0.$$

Due to (i)⇔(iii) in Proposition 2.6.6, we get $\|\max_{i \le N} X_i\|_{\psi_2} \le C\sqrt{\ln N}$ as claimed.

Second proof of (2.21) is based on replacing the maximum by the sum. Assume as before that $\max_i \|X_i\|_{\psi_2} = 1$ and denoting $Z = \max_{i \le N} |X_i|$, we have

$$\mathbb{E}\, e^{Z^2} = \mathbb{E} \max_{i \le N} e^{X_i^2} \le \mathbb{E} \sum_{i=1}^N e^{X_i^2} = \sum_{i=1}^N \mathbb{E}\, e^{X_i^2} \le 2N.$$

Since $M := \sqrt{2\ln(2N)} \ge 1$, Jensen inequality yields

$$\mathbb{E}\, e^{Z^2/M^2} \le \left(\mathbb{E}\, e^{Z^2}\right)^{1/M^2} \le (2N)^{\frac{1}{2\ln(2N)}} = \sqrt{e} < 2.$$

This gives $\|\max_{i \le N} |X_i|\|_{\psi_2} \le M = \sqrt{2\ln(2N)}$, proving (2.21).

The bound (2.22) follows from (2.21) via Proposition 2.6.6(ii) for $p = 1$. □

Remark 2.7.7 (Gaussian samples have no outliers)**.** The factor $\sqrt{\log N}$ appearing in Proposition 2.7.6 is unavoidable: in Exercise 2.38 you will see that i.i.d. random variables $g_i \sim N(0,1)$ satisfy

$$\mathbb{E} \max_{i \le N} |g_i| \approx \sqrt{2\ln N}.$$

The good news is that the logarithmic factor grows slowly and is often negligible. This is great for sampling because it helps prevent extreme outliers. On average, the farthest point in an N-point sample from a normal distribution is only $\sqrt{2\ln N}$ standard deviations away from the mean!

2.7.4 Centering

Many results in probability, such as the subgaussian Hoeffding inequality, require the random variables X_i to have zero mean. When they do not, we can center X_i by subtracting their means. Such centering can only reduce the L^2 norm:

$$\|X - \mathbb{E}\, X\|_{L^2} \le \|X\|_{L^2}. \tag{2.23}$$

This follows from an extremal property of the variance (Exercise 0.2). Let us check that centering also does not destroy the subgaussian norm:

Lemma 2.7.8 (Centering)**.** *Any subgaussian random variable X satisfies*

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

Proof Since $\|\cdot\|_{\psi_2}$ is a norm (Exercise 2.42), triangle inequality gives

$$\|X - \mathbb{E}\, X\|_{\psi_2} \le \|X\|_{\psi_2} + \|\mathbb{E}\, X\|_{\psi_2}. \tag{2.24}$$

We only need to bound the second term. Note that for any constant random variable a, we trivially have[5] $\|a\|_{\psi_2} \lesssim |a|$ (Exercise 2.24(b)). Using this for $a = \mathbb{E}\, X$ and using Jensen inequality for $f(x) = |x|$, we get

$$\|\mathbb{E}\, X\|_{\psi_2} \lesssim |\mathbb{E}\, X| \le \mathbb{E}|X| = \|X\|_{L^1} \lesssim \|X\|_{\psi_2}$$

where in the last step we used Proposition 2.6.6(ii) with $p = 1$. Substitute into (2.24) to complete the proof. □

2.8 Subexponential distributions

The class of subgaussian distributions is natural and quite broad. Nevertheless, it leaves out some important distributions whose tails are heavier than Gaussian. For example, consider a standard normal random vector $g = (g_1, \ldots, g_N) \in \mathbb{R}^N$, whose coordinates g_i are independent $N(0, 1)$ random variables. Let us look at the Euclidean norm of g:

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

Does $\|g\|_2$ concentrate around its expected value? On the one hand, $\|g\|_2$ is a sum of independent random variables g_i^2, so we might expect some concentration. On the other hand, while g_i are subgaussian random variables, g_i^2 are not. Indeed, recalling how the Gaussian tails behave (Proposition 2.1.2), we can see that[6]

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

So the tails of g_i^2 behave like those of the exponential distribution, and are strictly heavier than subgaussian tails. This prevents us from using the tools we have established so far, like Hoeffding inequality (Theorem 2.7.3), when studying the concentration of $\|g\|_2$.

2.8.1 Subexponential properties

With this in mind, let us look at distributions with exponential or lighter tail decay, known as subexponential distributions. Our analysis will be pretty similar to what we did for subgaussian distributions in Section 2.6, so we will move a bit quicker. Here is a version of Proposition 2.6.1:

[5] In this proof and later, the notation $a \lesssim b$ means that $a \le Cb$ where C is some absolute constant.

[6] Here we ignored the prefactor $1/t$, as it has little impact on the exponent for large t.

Proposition 2.8.1 (Subexponential properties)**.** *Let X be a random variable. The following properties are equivalent, with the parameters $K_i > 0$ differing by at most an absolute constant factor.*[7]

(i) (Tails) There exists $K_1 > 0$ such that

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

(ii) (Moments) There exists $K_2 > 0$ such that

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

(iii) (MGF of $|X|$) There exists $K_3 > 0$ such that

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

Moreover, if $\mathbb{E}\,X = 0$ then properties (i)–(iii) are equivalent to the following one:

(iv) (MGF) There exists $K_4 > 0$ such that

$$\mathbb{E}\exp(\lambda X) \leq \exp(K_4^2\lambda^2) \quad \textit{for all } \lambda \textit{ such that } |\lambda| \leq \frac{1}{K_4}.$$

Proof The equivalence of properties (i)–(iii) can be proved just like in Proposition 2.6.1; you will tackle this in Exercise 2.41. The equivalence of (iv) to the other properties is somewhat different, so we will prove it now.

(iii)⇒(iv) Assume that property (iii) holds. Without loss of generality we may assume that $K_3 = 1$. (Why?) Let us use the numeric inequality

$$e^x \leq 1 + x + \frac{x^2}{2}e^{|x|}$$

which follows from the Taylor theorem with the Lagrange form of the remainder. Assume that $|\lambda| \leq 1/2$ and substitute $x = \lambda X$ to get

$$\begin{aligned}
\mathbb{E}\,e^{\lambda X} &\leq 1 + \frac{\lambda^2}{2}\,\mathbb{E}\,X^2 e^{|\lambda X|} \quad (\text{since } \mathbb{E}\,X = 0 \text{ by assumption})\\
&\leq 1 + 2\lambda^2\,\mathbb{E}\,e^{|X|} \quad (\text{since } x^2 \leq 4e^{|x|/2} \text{ and } e^{|\lambda x|} \leq e^{|x|/2})\\
&\leq 1 + 4\lambda^2 \quad (\text{since } \mathbb{E}\,e^{|X|} \leq 2 \text{ by assumption (iii)})\\
&\leq e^{4\lambda^2}.
\end{aligned}$$

This yields property (iv) with $K_4 = 2$.

(iv)⇒(i) Assume that property (iv) holds with $K_4 = 1$ without loss of generality. Exponentiate, apply Markov inequality and use (iv) for $\lambda = 1$ to get

$$\mathbb{P}\{X \geq t\} = \mathbb{P}\{e^X \geq e^t\} \leq e^{-t}\,\mathbb{E}\,e^X \leq e^{1-t}.$$

In a similar way, we get $\mathbb{P}\{-X \geq t\} \leq e^{1-t}$. By the union bound, we conclude that $\mathbb{P}\{|X| \geq t\} \leq 2e^{1-t}$, which is further bounded by $2e^{-t/3}$ if $t \geq 3/2$. And

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

if $t < 3/2$, we have $2e^{-t/3} \geq 1$ and the probability is trivially bounded by this quantity. This establishes property (i) with $K_1 = 3$. □

Remark 2.8.2 (MGF near the origin)**.** You may be surprised to see the same bound on the MGF near the origin for subgaussian and subexponential distributions (property (iv) in Propositions 2.6.1 and 2.8.1.) But this is actually expected for any random variable X with mean zero. To see why, assume for simplicity that X is bounded and has unit variance. Approximate the MGF using the first two terms of the Taylor expansion:

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

as $\lambda \to 0$. For the standard *normal* distribution $N(0,1)$, this approximation becomes an equality, see (2.16). For *subgaussian* distributions, Proposition 2.6.1(iv) says that a bound like this holds for all λ, characterizing subgaussians. And for *subexponential* distributions, Proposition 2.8.1(iv) says that says that a bound like this holds for small λ, characterizing subexponentials.

Remark 2.8.3 (MGF far from the origin)**.** For subexponentials, the MGF bound can only be guaranteed near zero. Indeed, the MGF for an exponential random variable $X \sim \mathrm{Exp}(1)$ is infinite for $\lambda \geq 1$. (Check!)

2.8.2 The subexponential norm

Definition 2.8.4 (Subexponential distributions)**.** A random variable X is called a *subexponential* if it satisfies any of the equivalent properties (i)–(iii) in Proposition 2.8.1. Its *subexponential norm*, denoted $\|X\|_{\psi_1}$, is the smallest K_3 in property (iii). In other words,

$$\|X\|_{\psi_1} = \inf\left\{K > 0 : \ \mathbb{E}\exp(|X|/K) \leq 2\right\}. \tag{2.25}$$

In Exercise 2.42, you will confirm that $\|\cdot\|_{\psi_1}$ indeed defines a norm on the space of subexponential random variables.

Subgaussian and subexponential distributions are closely connected. Their definition directly imply the following (check!):

Lemma 2.8.5 (Subexponential is subgaussian squared)**.** *X is subgaussian if and only if X^2 is subexponential, and*

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

More generally:

Lemma 2.8.6 (Subgaussian × subgaussian = subexponential)**.** *If X and Y are subgaussian then XY is subexponential, and*

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

Proof Without loss of generality, we may assume that $\|X\|_{\psi_2} = \|Y\|_{\psi_2} = 1$. (Why?) By definition, this implies that $\mathbb{E}\exp(X^2) \le 2$ and $\mathbb{E}\exp(Y^2) \le 2$. Then

$$\begin{aligned}
\mathbb{E}\exp(|XY|) &\le \mathbb{E}\exp\left(\frac{X^2}{2} + \frac{Y^2}{2}\right) \quad (\text{since } |ab| \le \frac{a^2}{2} + \frac{b^2}{2}) \\
&= \mathbb{E}\left[\exp\left(\frac{X^2}{2}\right)\exp\left(\frac{Y^2}{2}\right)\right] \\
&\le \frac{1}{2}\,\mathbb{E}\left[\exp(X^2) + \exp(Y^2)\right] \quad (\text{using that } ab \le \frac{a^2}{2} + \frac{b^2}{2} \text{ again}) \\
&\le \frac{1}{2}(2+2) = 2 \quad (\text{by assumption}).
\end{aligned}$$

By definition, it follows that $\|XY\|_{\psi_2} \le 1$. □

Example 2.8.7. The following random variables are subexponential (check it!):

(a) any subgaussian random variable,
(b) the square of any subgaussian random variable (Lemma 2.8.6),
(c) exponential,
(d) Poisson,
(e) geometric,
(f) chi-squared,
(g) Gamma.

Conversely, the Cauchy and Pareto distributions are *not* subexponential. (Check!)

Many properties of subgaussian distributions extend to subexponentials. One of such properties is centering (Lemma 2.7.8), whose subexponential version is

$$\|X - \mathbb{E}X\|_{\psi_1} \le C\|X\|_{\psi_1}. \tag{2.26}$$

You will check this and some other properties in Exercise 2.44.

Remark 2.8.8 (All the norms together)**.** We have introduced many properties of random variables. How do they all fit together? Here is a chain of implications:

$$\begin{aligned}
&X \text{ is bounded a.s.} \Rightarrow X \text{ is subgaussian} \Rightarrow X \text{ is subexponential} \\
&\Rightarrow X \text{ has moments of all orders} \Rightarrow X \text{ has finite variance} \Rightarrow X \text{ has finite mean.}
\end{aligned}$$

Quantitatively, this corresponds to the following chain of inequalities:

$$\|X\|_{L^1} \le \|X\|_{L^2} \le \|X\|_{L^p} \lesssim \|X\|_{\psi_1} \lesssim \|X\|_{\psi_2} \lesssim \|X\|_{L^\infty}.$$

This holds for each $p \in [2, \infty)$, where the $\lesssim$ signs hide an $O(p)$ factor in one of the inequalities and absolute constant factors in the other two inequalities. (Explain why each inequality holds!)

Remark 2.8.9 (Going more general: ψ_α and Orlicz norms)**.** Subgaussian and subexponential distributions are part of the broader family of ψ_α distributions. An even more general framework is provided by Orlicz spaces and norms. You will explore this in Exercises 2.43 and 2.42.

2.9 Bernstein inequality

We are ready to state and prove a version of Hoeffding's inequality (Theorem 2.7.3) that works for subexponential distributions:

Theorem 2.9.1 (Subexponential Bernstein inequality)**.** *Let $X_1, \dots, X_N$ be independent, mean-zero, subexponential random variables. Then, for every $t \geq 0$, we have*

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

where $c > 0$ is an absolute constant.

If this bound seems a bit intimidating, do not worry—we will break it down and simplify it after the proof.

Proof We use the exponential moment method, which we learned in the proof of Hoeffding and Chernoff inequalities (Theorems 2.2.1 and 2.3.1). Denote $S_N = \sum_{i=1}^N X_i$, multiply both sides of the inequality $S_N \geq t$ by a parameter $\lambda > 0$, exponentiate, and use Markov inequality and independence. This leads to (2.9):

$$\mathbb{P}\{S_N \geq t\} \leq e^{-\lambda t} \prod_{i=1}^N \mathbb{E}\exp(\lambda X_i). \tag{2.27}$$

To bound the MGF of each term X_i, we use property (iv) in Proposition 2.8.1. It says that if λ is small enough so that

$$|\lambda| \leq \frac{c}{\max_i \|X_i\|_{\psi_1}}, \tag{2.28}$$

then[8] $\mathbb{E}\exp(\lambda X_i) \leq \exp\left(C\lambda^2 \|X_i\|_{\psi_1}^2\right)$. Substituting this into (2.27), we obtain

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

Now we minimize this expression in λ subject to the constraint (2.28). The optimal choice is $\lambda = \min\left(\frac{t}{2C\sigma^2}, \frac{c}{\max_i \|X_i\|_{\psi_1}}\right)$, for which we obtain

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

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

To make Theorem 2.9.1 more handy, let us apply it for $a_i X_i$ instead of X_i. This gives us a version of (2.20) for subexponential distributions:

[8] Recall that by Proposition 2.8.1 and definition of the subexponential norm, property (iv) holds for a value of K_4 that is within an absolute constant factor of $\|X\|_{\psi_1}$.

Corollary 2.9.2 (Subexponential Bernstein inequality, simplified)**.** *Let $X_1, \ldots, X_N$ be independent, mean-zero, subexponential random variables, and $a = (a_1, \ldots, a_N) \in \mathbb{R}^N$. Then, for every $t \geq 0$, we have*

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

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

Remark 2.9.3 (Why two tails?)**.** Unlike Hoeffding inequality (Theorem 2.7.3), Bernstein inequality (Theorem 2.9.1) has *two tails* – gaussian and exponential. The gaussian tail is not surprising: it is what we would expect from the central limit theorem. The exponential tail is also necessary, since even a single subexponential term X_i can have tail as large as $\exp\big(-ct/\|X_i\|_{\psi_1}\big)$, which is strictly heavier than a gaussian tail. What is surprising (and amazing) is that the exponential tail in Theorem 2.9.1 is no worse than what you would get from a *single* term X_i – the one with the largest subexponential norm.

Remark 2.9.4 (Small and large deviations)**.** Normalizing the sum in Corollary 2.9.2 like in the central limit theorem, we get[9]

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

In the *small deviations* range ($t \leq \sqrt{N}$), we get a gaussian tail bound. This range grows with N, reflecting the increasing strength of the central limit theorem. Meanwhile, in the *large deviations* range ($t \geq \sqrt{N}$), the tail bound remains heavy and exponential, driven by a single dominant term X_i. This is illustrated in Figure 2.3; we observed a similar phenomenon in Remark 2.3.5.

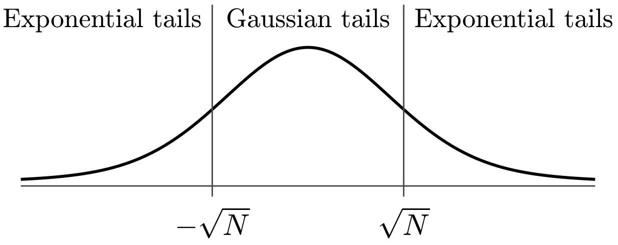


Figure 2.3 Bernstein inequality exhibits a mixture of two tails: gaussian for small deviations and exponential for large deviations.

To conclude this chapter, we mention a version of Bernstein inequality that is sensitive to the variances of the terms X_i. However, this comes at the cost of a stronger assumption that the terms X_i are bounded a.s.

[9] For simplicity, we suppressed the dependence on K by allowing the constant $c > 0$ depend on K.

Theorem 2.9.5 (Bernstein inequality for bounded distributions)**.** *Let $X_1, \dots, X_N$ be independent, mean-zero random variables satisfying $|X_i| \le K$ all i. Then, for every $t \ge 0$, we have*

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

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

You will prove this version of Bernstein inequality in Exercise 2.47.

2.10 Notes

Concentration inequalities cover a broad area, and we will explore them further in Chapter 5. Various forms of Hoeffding, Chernoff, and Bernstein inequalities, as well as related results, can be found in [52], [344, Chapter 2], [17, Appendix A], [249, Chapter 4], [210], [127, Chapter 7], [21, Section 3.5.4], [284, Chapter 1], [24, Chapter 4].

The proof of the upper bound on the Gaussian tails in Proposition 2.1.2 is borrowed from [116, Theorem 1.4]. To learn more about the Mills ratio mentioned in Exercise 2.3, see [133].

Berry-Esseen central limit theorem (Theorem 2.1.4) with an extra factor 3 on the right hand side can be found e.g. in [116, Section 2.4.d]; the best currently known factor is ≈ 0.47 [302].

The exponential moment method, used to derive the concentration inequalities in this chapter, was pioneered by S. Bernstein [38, 39, 40]. Early forms of Chernoff inequality (Theorem 2.3.1) appear in [83]. Hoeffding inequality (Theorem 2.2.6) was first proved in [161].

The survey [219] explores various approaches to the mean estimation problem introduced in Section 2.4. Median-of-means estimators have been employed since the 1980s, they appear in the early work of A. Nemirovsky and D. Yudin [256] and M. Jerrum, L. Valiant, and V. Vazirani [177]. The analysis of the median-of-means estimator in Theorem 2.4.1 largely follows [219]. Various impossibility results of the type presented in Exercise 2.17 can be found in [99, 219].

Section 2.5 scratches the surface of the rich theory of random graphs. The books [47, 174, 131] present comprehensive introductions to random graph theory. Vertex degrees of random graphs, which we discuss in Section 2.5 and in Exercises 2.18 and 2.19, have been extensively explored. For many asymptotically sharp results, see [131, Chapter 3]. The expansion property explored in Exercise 2.20 is most closely related to the expander mixing lemma. To explore a rich theory of expander graphs, refer to the survey [164].

Subgaussian distributions, discussed in Section 2.6, were introduced by J. P. Kahane [179]. Some of their basic properties discussed in Sections 2.6–2.7 were initially established in [179, 68]. In those early works, subgaussian distributions required mean zero, and the definition of subgaussian norm corresponds to the notion of the exact subgaussian norm discussed in Exercise 2.40. For a modern exposition of basic properties of the exact subgaussian norm (Exercise 2.40) and its relation with Orlicz norms (Exercise 2.42), see [285]. The presentation in Sections 2.6–2.9 mostly follows [340].

The original version of Khintchine inequality (Theorem 2.7.5, Exercise 2.36) for Rademacher distributions appears in [183, 216]. Sharp versions of Khintchine inequality for Rademacher distributions and related results can be found in [311, 152, 192, 252, 160], as well as [127, Theorem 8.5]; see [21, Section 3.7] for historical remarks.

Several forms of Bernstein inequality appeared in the original work of S. Bernstein's work [38, 39, 40]; see [21, Section 3.7] for a historical account of S. Bernstein's contributions. The dependence on the subexponential norm in Bernstein inequality (Corollary 2.9.2) can often be improved [176]. Concentration inequalities for general ψ_α distributions can be found e.g. in [79, Theorem 1.2.8] and [210].

Bennett inequality (Exercise 2.48) was likely first published in [35, 36]; see also [161].

Exercises

2.1 ☕☕☕ (Products of i.i.d. random variables do not concentrate) Many random variables concentrate around the mean; in particular, the probability of exceeding the mean is usually constant. Here is an example of poor concentration, where this probability is exponentially small. Let $X_1, \ldots, X_n$ be independent random variables uniformly distributed in $[0, 1]$. Prove that their product $Y_n := X_1 \cdots X_n$ satisfies

$$(0.5)^n \le \mathbb{P}\{Y_n \ge \mathbb{E}\, Y_n\} \le (0.95)^n.$$

2.2 ☕ (Gaussian tails: a lower bound) The lower bound in Proposition 2.1.2 was left unproved; let us prove it now. Let $g \sim N(0, 1)$. Show that for all $t > 0$, we have

$$\mathbb{P}\{g \ge t\} \ge \frac{t}{t^2+1} \cdot \frac{1}{\sqrt{2\pi}} e^{-t^2/2}.$$

To do this, consider the difference between the left and right hand sides as a function of t. Check that this function decreases to 0 as t increases to infinity.

2.3 ☕☕☕ (Mills ratio) Although the Gaussian tail $\mathbb{P}\{g > t\}$ for $g \sim N(0, 1)$ cannot be calculated analytically for all t, it can be expanded into series as follows:

$$\frac{\mathbb{P}\{g > t\}}{f(t)} = \frac{1}{t} - \frac{1}{t^3} + \frac{1 \cdot 3}{t^5} - \frac{1 \cdot 3 \cdot 5}{t^7} + \cdots \quad \text{for } t > 1, \tag{2.29}$$

where $f(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$ is the density of g. Moreover, the ratio is sandwiched between any pair of consecutive partial sums of this series. For example,

$$\frac{1}{t} - \frac{1}{t^3} \le \frac{\mathbb{P}\{g > t\}}{f(t)} \le \frac{1}{t} - \frac{1}{t^3} + \frac{3}{t^5} \quad \text{for } t > 0. \tag{2.30}$$

Prove (2.30) by doing the following steps:

(a) Check that $f'(x) + x f(x) = 0$ for all x.

(b) Use the equation from part (a) to integrate $\mathbb{P}\{g > t\} = \int_t^\infty f(x)\,dx$ by parts. Repeat.

2.4 ☕ (Truncated gaussian moments) Show that $g \sim N(0, 1)$ satisfies:

(a) $\mathbb{E}\, g \mathbf{1}_{\{g>t\}} = \dfrac{1}{\sqrt{2\pi}} e^{-t^2/2}$ for all $t > 0$.

(b) $\mathbb{E}\, g^2 \mathbf{1}_{\{g>t\}} \le \left(t + \dfrac{1}{t}\right) \dfrac{1}{\sqrt{2\pi}} e^{-t^2/2}$ for all $t \in \mathbb{R}$.

2.5 ☕☕ (Completing the proof of Hoeffding inequality) Prove the numerical bound (2.7):

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

2.6 ☕ (Gaussian tail by the exponential moment method) Use the exponential moment method to prove that $g \sim N(0, 1)$ satisfies

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

Although this bound has a larger constant factor than (2.3), it is still sufficient for many purposes.

2.7 ☕☕ (Small ball probability) Let $X_1, \dots, X_N$ be nonnegative independent random variables with continuous distributions. Assume that the probability density functions of X_i are all uniformly bounded by K. Show that for any $\varepsilon > 0$ we have

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

2.8 ☕☕☕ (An MGF comparison inequality) Let X and Y be random variables with the same mean. Assume that X takes values in an interval $[a, b]$, while Y takes values in the two-point set $\{a, b\}$. Prove that

$$\mathbb{E}\, e^{\lambda X} \le \mathbb{E}\, e^{\lambda Y} \quad \text{for all } \lambda \in \mathbb{R}.$$

2.9 ☕☕☕ (Hoeffding lemma) Prove that any random variable X that takes values in an interval $[a, b]$ satisfies

$$\mathbb{E}\, e^{\lambda(X - \mathbb{E}\, X)} \le \exp\left(\frac{\lambda^2 (b-a)^2}{8} \right) \quad \text{for all } \lambda \in \mathbb{R}.$$

To prove this, follow these steps:

(a) Argue that we may assume without loss of generality that X has mean zero, that $b - a = 1$, and that X takes values in the two-point set $\{a, b\}$.

(b) Compute the *cumulant generating function* $K(\lambda) := \log \mathbb{E}\, e^{\lambda X}$ and check that $K(0) = K'(0) = 0$ and $K''(\lambda) \le 1/4$. Conclude that $K(\lambda) \le \lambda^2/8$ for all $\lambda \in \mathbb{R}$.

2.10 ☕ (Hoeffding inequality for bounded random variables) Deduce Theorem 2.2.6 from Hoeffding lemma (Exercise 2.9).

2.11 ☕ (Chernoff inequality: left tails) Prove the result mentioned in Remark 2.3.2: under the assumptions of Theorem 2.3.1, we have

$$\mathbb{P}\{S_N \le t\} \le e^{-\mu} \left(\frac{e\mu}{t} \right)^t \quad \text{for every } 0 < t \le \mu.$$

2.12 ☕☕☕ (Reverse Chernoff) Show that a Binomial random variable $S_N \sim \mathrm{Binom}(N, \mu/N)$ satisfies the "reverse Chernoff" inequality":

$$\mathbb{P}\{S_N = t\} \ge e^{-\mu} \left(\frac{\mu}{t} \right)^t \quad \text{for every integer } t \in [\mu, N].$$

2.13 ☕☕ (Poisson tails) Thanks to the Poisson limit theorem (Theorem 1.7.6)), it is reasonable to expect that Chernoff inequality works for Poisson variables too. And it does! Use the exponential moment method to prove the following tail bounds for $X \sim \mathrm{Pois}(\mu)$:

(a) (Version of Theorem 2.3.1) $\mathbb{P}\{X \ge t\} \le e^{-\mu}(e\mu/t)^t$ for every $t \ge \mu$.

(b) (Version of Remark 2.3.2) $\mathbb{P}\{X \le t\} \le e^{-\mu}(e\mu/t)^t$ for every $0 < t \le \mu$.

(c) (Version of Corollary 2.3.4) $\mathbb{P}\{|X - \mu| \ge \delta\mu\} \le 2\exp(-\delta^2\mu/3)$ for every $0 \le \delta \le 1$.

(d) (Version of Exercise 2.12) $\mathbb{P}\{X = t\} \ge e^{-\mu}(\mu/t)^t$ for every integer $t > 0$.

2.14 ☕☕ (Chernoff inequality: small deviations) Let's extend Corollary 2.3.4. 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 $\mu = \mathbb{E}\, S_N$. Show that

$$\mathbb{P}\{|S_N - \mu| \ge \delta\mu\} \le 2\exp\Big(-\frac{\delta^2\mu}{2+\delta}\Big) \quad \text{for every } \delta \ge 0.$$

2.15 ☕ (Chernoff inequality for bounded random variables) Argue that all versions of Chernoff inequality (Theorem 2.3.1, Remark 2.3.2, and Corollary 2.3.4) are valid for any independent random variables X_i that take values in the interval $[0,1]$ and have means p_i.

2.16 ☕☕ (Median-of-means: fixing the proof) The proof of Theorem 2.4.1 is slightly flawed: $B = t^2/4$ may not be an integer that divides N. Fix this.

2.17 ☕☕☕ (No mean estimator is subgaussian everywhere) You may wonder if Theorem 2.4.1 holds for all $t > 0$. It does not. In fact, no mean estimator can provide subgaussian confidence for all quantiles. *Disprove* the following claim:

"There exists an absolute constant $c > 0$ such that the following holds. For any integer $N > 0$ and every $t \ge 0$, one can find a function $\widehat{\mu} : \mathbb{R}^N \to \mathbb{R}$ satisfying

$$\mathbb{P}\Big\{\Big|\widehat{\mu}(X_1,\ldots,X_N) - \mu\Big| \ge \frac{t\sigma}{\sqrt{N}}\Big\} \le 2e^{-ct^2}$$

for all i.i.d. random variables $X_1,\ldots,X_N$ with mean μ and variance σ^2."

Do this by working out a version of *Le Cam's two-point method*:

(a) Consider two random vectors $X = (X_1,\ldots,X_N)$ and $Y = (Y_1,\ldots,Y_N)$ with independent Laplace[10] coordinates: $X_i \sim \mathrm{Lap}(0,1)$ and $Y_i \sim \mathrm{Lap}(\mu,1)$ for some $\mu > 0$. Check that $\mathbb{P}\{X \in B\} \le e^{\mu}\,\mathbb{P}\{Y \in B\}$ for any measurable subset $B \subset \mathbb{R}^N$.

(b) Assume for contradiction that the claim holds. Apply it to bound the probabilities of the events $|\hat{\mu}(X)| \ge \mu/2$ and $|\hat{\mu}(Y) - \mu| \ge \mu/2$.

(c) Choose μ large enough and replace Y with X using (a) to make both events $|\hat{\mu}(X)| < \mu/2$ and $|\hat{\mu}(X) - \mu| < \mu/2$ likely. Use triangle inequality to arrive at a contradiction.

2.18 ☕ (Most degrees of sparse random graphs are OK) In Exercise 1.10, we saw that a typical sparse random graph $G \sim G(n,p)$ has isolated vertices ("friendless students"), resulting in a minimum degree of zero. However, most vertex degrees are still close to the expected degree $d = (n-1)p$. Prove that there exists an absolute constant $C > 0$ such that if $d \ge C$, then with probability at least 0.99 the following event holds: at least 99% of the vertices have their degrees between $0.9d$ and $1.1d$.

2.19 ☕☕☕☕ (Maximum degrees of sparse random graphs) In Exercise 1.10, we observed that a typical sparse random graph $G \sim G(n,p)$ has the minimum degree zero. But what about the maximum degree ("the number of friends of the most popular student"), denoted

[10] The Laplace distribution $\mathrm{Lap}(\mu,1)$ has probability density function $\frac{1}{2}e^{-|x-\mu|}$ for $x \in \mathbb{R}$. The mean is μ and the variance is 2. (Check!)

$\Delta(G)$? Show that there exist absolute constants $c, c_1, c_2 > 0$ such that the following holds: If $n \geq 3$ and the expected degree $d = (n-1)p$ satisfies $d \leq c(\log n)^{0.99}$, then

$$c_1 \frac{\log n}{\log \log n} \leq \Delta(G) \leq c_2 \frac{\log n}{\log \log n}$$

with probability at least 0.99.

2.20 ☕☕ (Expansion property of random graphs) Random graphs $G(n,p)$ exhibit a remarkable property: the number of edges[11] $e(S,T)$ connecting any two vertex subsets S and T is proportional to the sizes of these sets, provided they are not too small. Prove that there exists an absolute constant $C > 0$ such that, with probability at least $1 - 2^n$, the following event occurs:

$$0.9p \leq \frac{e(S,T)}{|S||T|} \leq 1.1p \quad \text{for all disjoint subsets of vertices } S,\, T \text{ with } |S||T| \geq \frac{Cn}{p}.$$

2.21 ☕☕ (Boosting randomized algorithms) Imagine we have an algorithm for solving some decision problem, like checking if p is prime. Suppose the algorithm makes decisions randomly, giving the correct result with probability $\frac{1}{2} + \varepsilon$ for some $\varepsilon > 0$, only slightly better than pure guessing. To improve the performance, run the algorithm N times and take the majority vote. Show that, for any $\delta \in (0,1)$, the answer is correct with probability at least $1 - \delta$, as long as

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

2.22 ☕ (Absolute moments of the normal distribution) Let $g \sim N(0,1)$.

(a) Express the absolute moments of g in terms of the gamma function as follows:

$$\mathbb{E}|g|^p = \frac{2^{p/2}}{\sqrt{\pi}} \Gamma\left(\frac{p+1}{2}\right) \quad \text{for each } p \geq 1.$$

(b) Deduce that the L^p-norm of g satisfies

$$\|g\|_{L^p} = \left(\mathbb{E}|g|^p\right)^{1/p} = \sqrt{\frac{p}{e}}\,(1 + o(1)) \quad \text{as } p \to \infty.$$

2.23 ☕☕ (Subgaussian MGF requires zero mean) You might wonder why we assumed that $\mathbb{E}\, X = 0$ in property (iv) of Proposition 2.6.1. Show that any random variable X satisfying this property *must* have zero mean.

2.24 ☕ (Examples of subgaussian distributions) Check the following.

(a) (Constant) If $X = c$ a.s. for some constant c, then $\|X\|_{\psi_2} = c/\sqrt{\ln 2}$.

(b) (Bounded) If X bounded a.s. then $\|X\|_{\psi_2} \leq \|X\|_\infty / \sqrt{\ln 2}$.

(c) (Rademacher) A Rademacher random variable X satisfies $\|X\|_{\psi_2} = 1/\sqrt{\ln 2}$.

(d) (Normal) $X \sim N(0, \sigma^2)$ satisfies $\|X\|_{\psi_2} = \sigma\sqrt{8/3}$.

(e) (Bernoulli) $X \sim \mathrm{Ber}(p)$ satisfies $\|X\|_{\psi_2} = 1/\sqrt{\ln(1 + 1/p)}$.

[11] Formally, $e(S,T)$ denotes the number of edges between pairs of vertices $s \in S$ and $t \in T$.

2.25 ☕ (Examples of non-subgaussian distributions) Explain why the exponential, Poisson, geometric, chi-squared, Gamma, Cauchy, and Pareto distributions are not subgaussian.

2.26 ☕☕ (Subgaussian characterization: MGF of X^2) In Proposition 2.6.1 we saw a few equivalent ways to describe subgaussian distributions. Here is one more. Prove that a random variable X is subgaussian if and only if there exists $K > 0$ such that

$$\mathbb{E}\exp(\lambda^2 X^2) \le \exp(\lambda^2 K^2) \quad \text{for all } \lambda \text{ such that } |\lambda| \le \frac{1}{K}. \tag{2.31}$$

More precisely, show that if X is subgaussian, then (2.31) holds with $K = \|X\|_{\psi_2}$. Conversely, prove that if (2.31) holds, then X is subgaussian with $\|X\|_{\psi_2} \le 2K$.

2.27 ☕☕ (Subgaussian characterization: almost stochastic dominance) For two random variables X and Y, let us say that $X \preceq Y$ if

$$\mathbb{P}\{X \ge t\} \le 2\mathbb{P}\{Y \ge t\} \quad \text{for all } t \in \mathbb{R}. \tag{2.32}$$

(a) Prove that X is subgaussian if and only if there exists $K > 0$ such that

$$|X| \preceq K|g| \quad \text{where } g \sim N(0,1). \tag{2.33}$$

More precisely, show that if X is subgaussian, then (2.33) holds with some $K \le C\|X\|_{\psi_2}$. Conversely, prove that if (2.33) holds, then X is subgaussian with $\|X\|_{\psi_2} \le CK$. As always, C stands for an absolute constant of your choice.

(b) Show by example that part (a) may fail if the factor 2 in (2.32) is replaced with 1.

2.28 ☕☕☕ (Subgaussian characterization: convex dominance) For two random variables X and Y, let us say that $X \precsim Y$ if $\mathbb{E}\,\Phi(X) \le \mathbb{E}\,\Phi(Y)$ for any convex, increasing function $\Phi : \mathbb{R} \to \mathbb{R}$. Prove that X is subgaussian if and only if there exists $K > 0$ such that

$$|X| \precsim K|g| \quad \text{where } g \sim N(0,1). \tag{2.34}$$

More precisely, show that if X is subgaussian, then (2.34) holds with some $K \le C\|X\|_{\psi_2}$. Conversely, prove that if (2.34) holds, then X is subgaussian with $\|X\|_{\psi_2} \le CK$.

2.29 ☕ (Hoeffding inequality for bounded random variables) Deduce Hoeffding inequality for bounded random variables (Theorem 2.2.6) from the subgaussian Hoeffding inequality (Theorem 2.7.3), possibly with some other absolute constant instead of 2 in the exponent.

2.30 ☕☕ (A reverse Hoeffding inequality) Let $X_1, \dots, X_N$ be independent subgaussian random variables with zero means and unit variances, and let $a_1, \dots, a_N \in \mathbb{R}$. Prove that

$$\mathbb{P}\left\{\left|\sum_{i=1}^N a_i X_i\right| \ge \frac{1}{2}\Big(\sum_{i=1}^N a_i^2\Big)^{1/2}\right\} \ge \frac{c}{K^4}$$

where $K = \max_{i \le N} \|X_i\|_{\psi_2}$ and $c > 0$ is an absolute constant. This result is a general, quantitative version of Example 1.5.1, though it does not attain the exact probability of 1/2.

2.31 ☕☕ (Hoeffding requires zero mean) Let $X_1, X_2, \dots$ be i.i.d. random variables that satisfy the Hoeffding-like inequality (2.14) for any N, any coefficient vector $a \in \mathbb{R}^N$, and for some $c > 0$ independent of N and a. Prove that $\mathbb{E}\, X_i = 0$.

2.32 ☕☕ (Subgaussian norm of a sum) Prove that any two independent, mean-zero, subgaussian random variables X and Y satisfy

$$\|X + Y\|_{\psi_2} \asymp \|X\|_{\psi_2} + \|Y\|_{\psi_2}$$

where "$\asymp$" indicates equivalence up to absolute constant factors.[12]

2.33 ☕☕☕ (Subgaussian norm of i.i.d. sums) In light of the Pythagorean identity (2.19), it is natural to ask if the inequality in Proposition 2.7.1 can be reversed. In this and the next exercise, you will demonstrate that the answer is "yes" for identical distributions of X_i, but "no" in general.

(a) Let $X, X_1, X_2, \dots, X_N$ be i.i.d. mean-zero, subgaussian random variables. Prove that

$$\Big\| \sum_{i=1}^{N} X_i \Big\|_{\psi_2} \asymp \sqrt{N}\, \|X\|_{\psi_2}$$

where the sign "$\asymp$" indicates equivalence up to absolute constant factors.

(b) Deduce that the binomial random variable $S_N \sim \mathrm{Binom}(N, p)$ satisfies

$$\|S_N - Np\|_{\psi_2} \asymp \sqrt{\frac{N}{\log(2/p)}}.$$

2.34 ☕☕☕☕ (Subgaussian norm of non-i.i.d. sums)

(a) Find independent, mean-zero, subgaussian random variables $X_1, X_2, \dots$ such that for any N and any coefficient vector $a = (a_1, \dots, a_N) \in \mathbb{R}^N$, we have

$$\Big\| \sum_{i=1}^{N} a_i X_i \Big\|_{\psi_2} \asymp \|a\|_\infty$$

where "$\asymp$" indicates equivalence up to absolute constant factors.

(b) Conclude that the inequality in Proposition 2.7.1 cannot be reversed in general.

2.35 ☕☕☕ (Interpolating between L^1 and L^∞) In Exercise 1.12, we showed how to bound the L^p norm of a random variable X using bounds on its L^1 and L^∞ norms. Now, let us prove a similar result for the subgaussian norm.

(a) Prove that if $\|X\|_{L^1} = a$ and $\|X\|_{L^\infty} = b$, then

$$\|X\|_{\psi_2} \le \frac{Cb}{\sqrt{\log(2b/a)}}$$

where C is an absolute constant.

(b) Give an example showing that the bound in part (a) is tight (up to an absolute constant) for any a and b.

[12] We write $a \asymp b$ to indicate that $c_1 b \le a \le c_2 b$ where $c_1, c_2 > 0$ are absolute constants.

2.36 ☕☕ (Khintchine inequality for $p < 2$) In Theorem 2.7.5, we proved Khintchine inequality for $p \in [2, \infty)$. Let us now extend it to $p \in [1, 2]$.

(a) (Extrapolation) Show that any random variable Z with finite L^3 norm satisfies the inequality

$$\|Z\|_{L^2} \le \|Z\|_{L^1}^{1/4} \, \|Z\|_{L^3}^{3/4}.$$

(b) Let $X_1, \dots, X_N$ be independent subgaussian random variables with zero means and unit variances, and let $a_1, \dots, a_N \in \mathbb{R}$. Prove that

$$cK^{-3} \Big(\sum_{i=1}^N a_i^2 \Big)^{1/2} \le \mathbb{E} \Big| \sum_{i=1}^N a_i X_i \Big| \le \Big(\sum_{i=1}^N a_i^2 \Big)^{1/2}.$$

where $K = \max_i \|X_i\|_{\psi_2}$ and $c > 0$ is an absolute constant.

(c) Conclude that the same inequality holds for the L^p norm of the sum for any $p \in [1, 2]$.

2.37 ☕☕☕ (A maximal inequality) Prove the following strengthening of Proposition 2.7.6. Let $X_1, X_2, \dots$ be a sequence of subgaussian random variables, not necessarily independent. Then

$$\Bigg\| \sup_k \frac{X_k}{\sqrt{\log(2k)}} \Bigg\|_{\psi_2} \le C \sup_k \|X_k\|_{\psi_2}$$

where C is an absolute constant.

2.38 ☕☕☕☕ (Maximum of Gaussians) Let us prove an asymptotically sharp version of the maximal inequality (2.22) for standard normal random variables.

(a) Let $g_1, \dots, g_N$ be $N(0, 1)$ random variables for some $N \ge 2$, that are not necessarily independent. Prove that

$$\mathbb{E} \max_{i \le N} g_i \le \sqrt{2 \ln N} \quad \text{and} \quad \mathbb{E} \max_{i \le N} |g_i| \le \sqrt{2 \ln(2N)}.$$

(b) Prove that if $g_1, g_2, \dots$ are independent $N(0, 1)$, then we have as $N \to \infty$:

$$\mathbb{E} \max_{i \le N} g_i = \sqrt{2 \ln N} \, (1 + o(1)) \quad \text{and} \quad \mathbb{E} \max_{i \le N} |g_i| = \sqrt{2 \ln N} \, (1 + o(1)) \, .$$

2.39 ☕☕☕ (Subgaussian characterization: a maximal inequality) Prove that a random variable X is subgaussian if and only if there exists $K > 0$ such that

$$\mathbb{E} \max_{i \le N} |X_i| \le K \sqrt{\log N} \quad \text{for any } N = 2, 3, \dots \tag{2.35}$$

where X_i are independent copies of X. We have already proved half of this statement: Proposition 2.7.6 shows that if X is subgaussian, then (2.35) holds with $K \le C \|X\|_{\psi_2}$. Now prove the converse: if (2.35) holds, then X is subgaussian and $\|X\|_{\psi_2} \le CK$.

2.40 ☕☕☕☕ (A surgeon's view: the exact subgaussian norm) You may have noticed that most results involving the subgaussian norm hold up to an absolute constant factor rather than exactly. To refine these results, we can redefine the subgaussian norm using the equivalent property (iv) from Proposition 2.6.1. For motivation, recall that the MGF of a normal random variable $X \sim N(\mu, \sigma^2)$ satisfies $\mathbb{E}e^{\lambda(X-\mu)} = e^{\sigma^2\lambda^2/2}$ for all $\lambda \in \mathbb{R}$. Inspired by this, we can define the *subgaussian variance* of a random variable X as

$$\mathrm{Var}_G(X) := \inf\Big\{\sigma^2 : \ \mathbb{E}\, e^{\lambda(X-\mathbb{E}\,X)} \le e^{\sigma^2\lambda^2/2} \text{ for all } \lambda \in \mathbb{R}\Big\}.$$

Recall that the L^2 norm of X can be expressed as $\|X\|_{L^2}^2 = \mathbb{E}[X^2] = \mathrm{Var}(X) + (\mathbb{E}\,X)^2$. Inspired by this, define the *exact subgaussian norm* of a random variable X by the identity

$$\|X\|_G^2 := \mathrm{Var}_G(X) + (\mathbb{E}\,X)^2.$$

Prove the following.

(a) The exact subgaussian norm indeed defines a norm on the space of subgaussian random variables.

(b) The exact and standard subgaussian norms are equivalent up to absolute constant factors: $c_1\|X\|_{\psi_2} \le \|X\|_G \le c_2\|X\|_{\psi_2}$.

(c) We have $\mathrm{Var}(X) \le \mathrm{Var}_G(X)$ and $\|X\|_{L^2} \le \|X\|_G$, with equalities holding if X is normally distributed.

(d) If $X_1, \ldots, X_N$ are independent and mean-zero, then

$$\mathrm{Var}_G\Big(\sum_{i=1}^N X_i\Big) \le \sum_{i=1}^N \mathrm{Var}_G(X_i).$$

Rewriting this as $\big\|\sum_{i=1}^N X_i\big\|_G^2 \le \sum_{i=1}^N \|X_i\|_G^2$ gives an exact version of Proposition 2.7.1.

(e) An exact version of centering (Lemma 2.7.8): $\|X - \mathbb{E}\,X\|_G \le \|X\|_G$.

2.41 ☕☕ (Subexponential properties) Prove the equivalence of properties (i)–(iii) in Proposition 2.8.1 by modifying the proof of Proposition 2.6.1.

2.42 ☕☕☕ (A bird's eye view: Orlicz norms) Here is a general framework that covers most norms we have seen so far. Let $\psi : [0, \infty) \to [0, \infty)$ be a nondecreasing, convex function with $\psi(0) = 0$. The Orlicz norm of a random variable X is defined as

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

(a) Verify that this indeed defines a norm on the set of all random variables (on the same probability space) for which $\|X\|_\psi$ is finite.

(b) Explain why the following are examples of Orlicz norms: the L^p norm for any $p \in [1, \infty)$, the subgaussian norm $\|X\|_{\psi_2}$, and the subexponential norm $\|X\|_{\psi_1}$.

2.43 ☕☕ (ψ_α distributions) Consider the distributions whose tails decay at a rate of $\exp(-ct^\alpha)$ or faster, where $\alpha \in (0, \infty)$ is a fixed parameter. When $\alpha = 2$, these are subgaussian distributions, and when $\alpha = 1$, they are subexponential. State and prove a version of Proposition 2.8.1(i)–(iii) for such distributions, and define the ψ_α norm.

2.44 ☕☕ Many properties of subgaussian distributions extend to subexponential distributions, with some modifications. State and prove for subexponentials:

(a) A version of centering (Lemma 2.7.8), already stated in (2.26).

(b) A version of the maximal inequality (Exercise 2.39).

(c) A version of convex dominance (Exercise 2.28).

2.45 ☕ (Restating Bernstein inequality) Here is a popular way to state results like Theorem 2.9.1. Let $X_1, \dots, X_N$ be independent, mean-zero, subexponential random variables. Check that, for every $u \ge 0$, we have

$$\mathbb{P}\Big\{\Big|\sum_{i=1}^N X_i\Big| \ge C\left(\sigma\sqrt{u} + Ku\right)\Big\} \le 2\exp(-u)$$

where $\sigma^2 = \sum_{i=1}^N \|X_i\|_{\psi_1}^2$ and $K = \max_i \|X_i\|_{\psi_1}$.

2.46 ☕☕ (Subexponential Khintchine inequality) Prove the following subexponential version of Theorem 2.7.5. Let $X_1, \dots, X_N$ be independent, mean-zero, subexponential random variables, and $a = (a_1, \dots, a_N) \in \mathbb{R}^N$. Then, for every $p \in [2, \infty)$, we have

$$\Big\|\sum_{i=1}^N a_i X_i\Big\|_{L^p} \le CK\big(\sqrt{p}\|a\|_2 + p\|a\|_\infty\big)$$

where $K = \max_i \|X_i\|_{\psi_1}$ and C is an absolute constant.

2.47 ☕☕ (Bernstein inequality for bounded distributions)

(a) Let X be a mean-zero random variable satisfying $|X| \le K$ a.s. Prove the following bound on the MGF of X:

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

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

(b) Deduce Theorem 2.9.5 by the exponential moment method.

2.48 ☕☕☕ (Bennett inequality) Let us prove the following strengthening of Bernstein inequality (Theorem 2.9.5). Let $X_1, \dots, X_N$ be independent, mean-zero random variables satisfying $|X_i| \le K$ all i. Then, for every $t \ge 0$, we have

$$\mathbb{P}\left\{\Big|\sum_{i=1}^N X_i\Big| \ge t\right\} \le 2\exp\left[-\frac{\sigma^2}{K^2}\,h\left(\frac{Kt}{\sigma^2}\right)\right] \tag{2.36}$$

where $\sigma^2 = \sum_{i=1}^N \mathbb{E}\,X_i^2$ and $h(u) = (1+u)\ln(1+u) - u$. Prove this result as follows:

(a) Let X be a mean-zero random variable satisfying $|X| \le K$. Show that the MGF of X satisfies the following bound for any $\lambda > 0$:

$$\mathbb{E}\exp(\lambda X) \le \exp\left[\frac{\mathbb{E}\,X^2}{K^2}\left(e^{\lambda K} - 1 - \lambda K\right)\right].$$

(b) Use the exponential moment method to deduce (2.36).

(c) Check that in the small deviations range, where $u = Kt/\sigma^2 \approx 0$, we have $h(u) \approx u^2/2$, so Bennett inequality gives approximately the Gaussian tail bound $\exp(-t^2/\sigma^2)$.

(d) Check that we always have $h(u) \ge \frac{1}{2}u\ln u$, so Bennett inequality gives a Poisson tail bound $2(\sigma^2/Kt)^{t/2K}$ in the large deviations range, similar to Exercise 2.13.

3

Random Vectors in High Dimensions

In this chapter, we study random vectors $X = (X_1, \dots, X_n) \in \mathbb{R}^n$, where the dimension n is typically very large. High-dimensional distributions are common in data science. For instance, computational biologists study gene expressions for $n \sim 10^4$ genes in the human genome, modeling these as a random vector, with each coordinate X_i representing the expression of a specific gene in a randomly selected individual from a population.

Life in high dimensions presents new challenges because there is *exponentially more room* in higher dimensions than in low dimensions. For instance, a cube in $\mathbb{R}^n$ with side 2 has 2^n times the volume of a unit cube (see Figure 3.1). This abundance of room makes many algorithms exponentially harder, a phenomenon known as the *curse of dimensionality.*

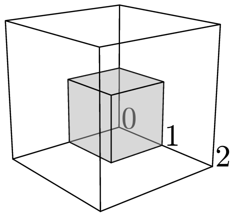


Figure 3.1 The abundance of room in high dimensions: the volume of the larger cube is exponentially bigger than the volume of the smaller cube.

High dimensional probability often helps bypass these difficulties. We start by studying the Euclidean norm of a random vector X with independent coordinates, showing in Section 3.1 that it tightly concentrates around its mean. Sections 3.2, 3.3 and 3.4 introduce basic concepts, results and examples of high-dimensional distributions, with an application to principal component analysis, a powerful data exploration tool. In Section 3.5, we give a probabilistic proof of the classical Grothendieck inequality and apply it to semidefinite optimization. In Section 3.6, we explore a semidefinite relaxation for the classical hard optimization problem – maximum cut – and present the celebrated Goemans-Williamson randomized

[0] This material will be published by Cambridge University Press as *High-Dimensional Probability, 2nd Edition* by Roman Vershynin. This pre-publication version is free to view and download for personal use only. Not for re-distribution, re-sale or use in derivative works. ©Roman Vershynin 202X.
If you want to be notified when the printed version is available, let me know: rvershyn@uci.edu

approximation algorithm. In Section 3.7, we provide an alternative proof of the Grothendieck inequality with nearly the best known constant using the kernel trick, a method widely used in machine learning.

Don't miss the exercises! You will compute the ℓ^p norm of a random vector (Exercises 3.5 and 3.6); explore classical invariant ensembles in random matrix theory (Exercises 3.18 and 3.19), get introduced to entropy (Exercise 3.46), and dive into the fantastic Grothendieck inequality (Exercise 3.57) and its role in semidefinite programming (Exercise 3.58).

Along the way, you will come across some surprising facts: the Gaussian distribution is close to the uniform distribution on the sphere (3.16), a huge number of random points are typically in convex position (Exercise 3.23), a random vector is likely to get close to a cube but unlikely to hit a cube even 100 times larger (Exercise 3.28), there are exponentially many almost orthogonal vectors (Exercise 3.41), and more.

3.1 Concentration of the norm

Where in the space $\mathbb{R}^n$ is a random vector $X = (X_1, \ldots, X_n)$ likely to be found? Assume the coordinates X_i are independent random variables with zero mean and unit variance. What is the typical length of X? We have

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

So we should expect the length of X to be

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

We will now show that X is indeed very likely to be close to $\sqrt{n}$.

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

$$\Big\| \|X\|_2 - \sqrt{n} \Big\|_{\psi_2} \le CK^2 \tag{3.1}$$

where $K = \max_i \|X_i\|_{\psi_2}$ and C is an absolute constant.[1]

Proof Using Proposition 2.6.6, we can rewrite (3.1) as the gaussian tail bound:

$$\mathbb{P}\big\{ \big| \|X\|_2 - \sqrt{n} \big| \ge t \big\} \le 2\exp\left(-\frac{ct^2}{K^4} \right) \quad \text{for all } t \ge 0. \tag{3.2}$$

We are going to prove this tail bound using Bernstein inequality. First, look at

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

[1] From now on, we will always denote various positive absolute constants by C, C_1, c, c_1 without saying this explicitly.

which is a sum of independent, mean-zero random variables. Since X_i are subgaussian, $X_i^2 - 1$ are subexponential. More precisely, applying centering (2.26) and using Lemma 2.8.5, we see that

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

Applying Bernstein inequality (Corollary 2.9.2) for $N = n$ and $a_i = 1/n$, we obtain for any $u \geq 0$ that

$$\begin{aligned} \mathbb{P}\left\{\left|\frac{1}{n}\|X\|_2^2 - 1\right| \geq u\right\} &\leq 2\exp\left[-c_1 \min\left(\frac{u^2 n}{K^4}, \frac{un}{K^2}\right)\right] \\ &\leq 2\exp\left[-\frac{cn}{K^4}\min(u^2, u)\right] \end{aligned} \tag{3.3}$$

(In the last step, we used that K is bounded below by an absolute constant, since $1 = \|X_1\|_{L^2} \leq C\|X_1\|_{\psi_2} \leq CK$ by Proposition 2.6.6(ii).)

We obtained a concentration inequality for $\|X\|_2^2$, and will now use it to deduce one for $\|X\|_2$, using on the following simple observation valid for all $z, \delta \geq 0$:

$$|z - 1| \geq \delta \quad \text{implies} \quad |z^2 - 1| \geq \max(\delta, \delta^2). \tag{3.4}$$

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

$$\begin{aligned} \mathbb{P}\left\{\left|\frac{1}{\sqrt{n}}\|X\|_2 - 1\right| \geq \delta\right\} &\leq \mathbb{P}\left\{\left|\frac{1}{n}\|X\|_2^2 - 1\right| \geq \max(\delta, \delta^2)\right\} \quad \text{(by (3.4))} \\ &\leq 2\exp\left(-\frac{cn}{K^4}\delta^2\right) \quad \text{(by (3.3) for } u = \max(\delta, \delta^2)). \end{aligned}$$

Change variables to $t = \delta\sqrt{n}$ to obtain the desired subgaussian tail (3.2). □

Remark 3.1.2 (Thin shell phenomenon)**.** Theorem 3.1.1 shows that random vectors in $\mathbb{R}^n$ mostly stay in a shell of constant thickness around the sphere of radius $\sqrt{n}$. This might seem surprising, so let's give an intuitive explanation. The square of the norm, $S_n := \|X\|_2^2$, has mean n and standard deviation $O(\sqrt{n})$ (why?) Thus $\|X\|_2 = \sqrt{S_n}$ ought to deviate by $O(1)$ around $\sqrt{n}$, because

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

see Figure 3.2 for an illustration. To get a better feel for the thin shell phenomenon, try Exercises 3.1–3.3 to find out that, typically,

$$\operatorname{Var}(\|X\|_2) = \Theta(1) \quad \text{and} \quad \mathbb{E}\|X\|_2 = \sqrt{n} - \Theta(1/\sqrt{n}).$$

3.2 Covariance matrices and principal component analysis

Assuming a random vector has independent coordinates is not always realistic. So to handle more general cases, let's quickly review basic concepts about high-dimensional probability distributions – points we covered briefly in Section 1.3.

The *covariance matrix* of a random vector X taking values in $\mathbb{R}^n$ is defined as

$$\operatorname{cov}(X) = \mathbb{E}(X - \mu)(X - \mu)^\mathsf{T} = \mathbb{E}\, XX^\mathsf{T} - \mu\mu^\mathsf{T}, \quad \text{where } \mu = \mathbb{E}\, X.$$

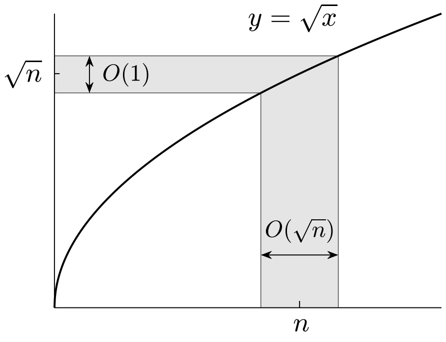


Figure 3.2 Concentration of the norm of a random vector X in $\mathbb{R}^n$. While $\|X\|_2^2$ deviates by $O(\sqrt{n})$ around n, $\|X\|_2$ deviates by $O(1)$ around $\sqrt{n}$.

(Check the identity above!) Thus, $\mathrm{cov}(X)$ is an $n \times n$ symmetric, positive semidefinite matrix. It is a high-dimensional generalization of the notion of variance of a random variables Z, which is

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

The entries of the covariance matrix $\mathrm{cov}(X)$ are the *covariances* of the pairs of coordinates of $X = (X_1, \ldots, X_n)$:

$$\mathrm{cov}(X)_{ij} = \mathbb{E}(X_i - \mathbb{E}\, X_i)(X_j - \mathbb{E}\, X_j). \tag{3.5}$$

It is sometimes helpful to disregard the mean of X and thus to consider the *second moment matrix*

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

a higher dimensional generalization of the second moment $\mathbb{E}\, Z^2$ of a random variable Z. By translation (replacing X with $X - \mu$), many problems can be reduced to the mean-zero case, where the covariance and second moment matrices are the same:

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

Thus, we will mostly focus on the second moment matrix $\Sigma = \Sigma(X)$ in the future.

3.2.1 What can we learn from the covariance matrix?

The covariance matrix can tell us much more about the random vector X than just the covariances of its coordinates. The next result shows how to gain access to: (a) the variance of one-dimensional marginals of X, i.e. the random variables $\langle V, v\rangle$ obtained by projecting X onto a given direction $v \in \mathbb{R}^n$; (b) the Euclidean norm of X, and (c) the angle between two independent copies of X.

Proposition 3.2.1 (Covariance matrix helps compute interesting quantities)**.** *Let X be a random vector in $\mathbb{R}^n$ with second moment matrix $\Sigma = \mathbb{E}\, X X^{\mathsf{T}}$. Then* *(a) (1D marginals) For any fixed vector* $v \in \mathbb{R}^n$*, we have*

$$\mathbb{E}\langle X, v\rangle^2 = v^\mathsf{T}\Sigma v. \tag{3.6}$$

(b) (Norm)[2] $\mathbb{E}\|X\|_2^2 = \operatorname{tr}(\Sigma)$.

(c) If Y *is an independent copy of* X*, then*[3] $\mathbb{E}\langle X, Y\rangle^2 = \|\Sigma\|_F^2$.

Proof (a) Using the linearity of expectation, we have

$$\mathbb{E}\langle X, v\rangle^2 = \mathbb{E}\left(v^\mathsf{T}X\right)\left(X^\mathsf{T}v\right) = v^\mathsf{T}\,\mathbb{E}\left[XX^\mathsf{T}\right]v = v^\mathsf{T}\Sigma v.$$

(b) The diagonal entries of the second moment matrix are $\Sigma_{ii} = \mathbb{E}\,X_{ii}^2$. Thus

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

(c) Do it yourself: write the inner product as a sum and expand the square. □

3.2.2 Principal component analysis

The most interesting insight into a random vector X lies in the eigenvalues and the eigenvectors of its covariance matrix $\Sigma = \operatorname{cov}(X)$. Since Σ is symmetric, the spectral theorem tells us that the eigenvalues λ_i of Σ are real, and there exists an orthonormal basis of $\mathbb{R}^n$ consisting of eigenvectors v_i of Σ. Writing the identity matrix in $\mathbb{R}^n$ as $I_n = \sum_{i=1}^n v_i v_i^\mathsf{T}$ (check!), multiplying on both sides by Σ and using that $\Sigma v_i = \lambda_i v_i$, we obtain the *spectral decomposition* of Σ:

$$\Sigma = \sum_{i=1}^n \lambda_i v_i v_i^\mathsf{T}. \tag{3.7}$$

We usually arrange the eigenvalues λ_i in (weakly) decreasing order.

There is a handy optimization-based approach to eigenvalues. The largest eigenvalue λ_1 can be obtained by maximizing the quadratic form $v^\mathsf{T}\Sigma v$ over all unit vectors $v \in \mathbb{R}^n$, and the maximum is attained at the top eigenvector $v = v_1$. Once we remove v_1 and maximize the quadratic form over the unit vectors that are orthogonal to v_1, we get the second largest eigenvalue λ_2, and the maximum is attained at the eigenvector v_2, and so on. More formally:

Proposition 3.2.2 (Optimization-based characterization of eigenvalues)**.** *Let* Σ *be an* $n \times n$ *symmetric matrix with eigenvalues* $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ *and corresponding unit eigenvectors* $v_1, \ldots, v_n$*. Then, for every* $k = 1, \ldots, n$*, we have*

$$\lambda_k = \max_{v \perp \{v_1, \ldots, v_{k-1}\},\, \|v\|_2 = 1} v^\mathsf{T}\Sigma v, \tag{3.8}$$

and the maximum is attained at v_k*.*

[2] Here $\operatorname{tr}(\Sigma) = \sum_{i=1}^n \Sigma_{ii}$ is the trace of Σ.

[3] Here $\|\Sigma\|_F = (\sum_{i,j=1}^n \Sigma_{ij}^2)^{1/2}$ is the Frobenius norm of Σ.

Proof Consider any unit vector $v \in \mathbb{R}^n$ that is orthogonal to $\{v_1, \ldots, v_{k-1}\}$. Using spectral decomposition (3.7), we get

$$\begin{aligned} v^\mathsf{T} \Sigma v = v^\mathsf{T} \Big(\sum_{i=1}^n \lambda_i v_i v_i^\mathsf{T} \Big) v &= \sum_{i=1}^n \lambda_i (v^\mathsf{T} v_i)(v_i^\mathsf{T} v) \\ &= \sum_{i=k}^n \lambda_i \langle v, v_i \rangle^2 \quad \text{(due to the orthogonality assumption)} \\ &\le \lambda_k \sum_{i=k}^n \langle v, v_i \rangle^2 \quad \text{(since } \lambda_i \text{ are weakly decreasing)} \\ &\le \lambda_k \quad \text{(by Bessel inequality, since } v_i \text{ are orthonormal)} \end{aligned}$$

Moreover, we have $v_k^\mathsf{T} \Sigma v_k = v_k^\mathsf{T} (\lambda_k v_k) = \lambda_k$. The proof is complete. □

Combining (3.8) with (3.6), we get the following interpretation of the eigenvalues and eigenvectors of the covariance matrix:

Corollary 3.2.3 (Principal component analysis)**.** *Let X be a random vector in $\mathbb{R}^n$ whose covariance matrix has eigenvalues $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_n \ge 0$ and eigenvectors $u_1, \ldots, u_n$. Then*

$$\lambda_k = \max_{v \perp \{v_1, \ldots, v_{k-1}\},\, \|v\|_2 = 1} \operatorname{Var}(\langle X, v \rangle),$$

and the maximum is attained at v_k.

Imagine that a random vector $X \in \mathbb{R}^n$ represents data, like the genetic data on p. 58. According to Corollary 3.2.3, the top eigenvector v_1 of the covariance matrix gives the first *principal component*, indicating the direction where the data spreads out the most, with λ_1 as the variance in that direction. The next eigenvector v_2 shows the second-best direction for capturing the remaining variance of the data, which is λ_2, and so on (see Figure 3.3). Try Exercise 3.4 for a more general interpretation of explained variance.

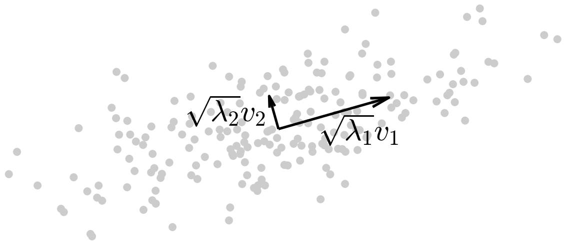


Figure 3.3 This plot shows 200 points sampled from a distribution, with the top two principal components v_i scaled by the standard deviations $\sqrt{\lambda_i}$.

Remark 3.2.4 (Dimension reduction)**.** It often happens with real data that only a few eigenvalues λ_i are large and informative, while the rest are small and treated as noise. In such cases, just a few principal components capture most of the data's variability. Even though the data lives in a high-dimensional space $\mathbb{R}^n$, it is essentially *low-dimensional*, clustering around the subspace E spanned by the few top principal components v_i.

Principal Component Analysis (PCA) is a basic data analysis method that finds the first few principal components v_i and projects the data onto the subspace E they span. This reduces the data's dimension and makes analysis easier. And if E is two- or three-dimensional, PCA can help visualize the data.

3.2.3 Isotropic distributions

You may remember from a basic probability course how it is helpful to assume that random variables have zero mean and unit variance. This idea extends to higher dimensions, where isotropy generalizes the concept of unit variance.

Definition 3.2.5 (Isotropic random vectors)**.** A random vector X in $\mathbb{R}^n$ is called *isotropic* if

$$\mathbb{E}\, X X^{\mathsf{T}} = I_n$$

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

Proposition 3.2.1 implies that X is isotropic if and only if

$$\mathbb{E}\langle X, v\rangle^2 = \|v\|_2^2 \quad \text{for any fixed vector } v \in \mathbb{R}^n. \tag{3.9}$$

(Check the"only if"!) Since the right-hand side does not depend on the direction of v, (3.9) basically says is that *isotropic distributions spread equally in all directions.*

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

$$Z = \frac{X - \mu}{\sqrt{\operatorname{Var}(X)}}, \quad \text{so we can write} \quad X = \mu + \operatorname{Var}(X)^{1/2} Z.$$

This idea extends to higher dimensions. Any random vector X with invertible variance matrix can be reduced by translation and dilation to the *standard score*[4]

$$Z = \operatorname{cov}(X)^{-1/2}(X - \mu), \quad \text{so we can write} \quad X = \mu + \operatorname{cov}(X)^{1/2} Z. \tag{3.10}$$

This often allows us to assume, without loss of generality, that random vectors have zero means and are isotropic. And even if the covariance matrix is not invertible, the idea still holds. Any random vector X can still be written as $X = \mu + \operatorname{cov}(X)^{1/2} Z$ for some mean-zero, isotropic random vector Z (Exercise 3.10 – try it now!)

[4] If you are unfamiliar with the concept of the square root of matrix, define it using the spectral decomposition: if $\Sigma = \sum_i \lambda_i v_i v_i^{\mathsf{T}}$ set $\Sigma^{1/2} = \sum_i \lambda_i^{1/2} v_i v_i^{\mathsf{T}}$. This idea extends beyond the square root: for example, $\Sigma^{-1/2} = \sum_i \lambda_i^{-1/2} v_i v_i^{\mathsf{T}}$. We will explore "matrix calculus" in Section 5.4.1.

3.3 Examples of high-dimensional distributions

Let's give a few basic examples of high-dimensional distributions.

3.3.1 Standard normal

The most iconic high-dimensional distribution is Gaussian, or multivariate normal. A random vector $Z = (Z_1, \dots, Z_n)$ has the *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 $N(0,1)$. The density of Z is the product of the n standard normal densities (1.24):

$$f_Z(z) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} e^{-z_i^2/2} = \frac{1}{(2\pi)^{n/2}}\, e^{-\|z\|_2^2/2}, \quad z \in \mathbb{R}^n. \tag{3.11}$$

The standard normal distribution is isotropic. (Why?)

Note that $f_Z(z)$ depends only on the length of the vector z, not its direction. So, the standard normal density is *rotation invariant* – it stays the same under any rotation. Formally, this means:

Proposition 3.3.1 (Rotation invariance)**.** *Consider a random vector $Z \sim N(0, I_n)$ and a fixed $n \times n$ orthogonal matrix U. Then*

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

In particular, if we focus on the first coordinate of UZ, we obtain $(UZ)_1 = \langle U_1, Z\rangle \sim N(0,1)$, where U_1 denotes the first row of U. Since this row can be an arbitrary unit vector in $\mathbb{R}^n$, we conclude that all 1D marginals of the standard normal distribution are standard normal. More generally, after rescaling we get:

Corollary 3.3.2 (1D marginals of the standard normal distribution)**.** *Consider a random vector $Z \sim N(0, I_n)$ and a fixed vector $v \in \mathbb{R}^n$. Then*

$$\langle Z, v\rangle \sim N(0, \|v\|_2^2).$$

This yields a classic result you might remember from a basic probability course:

Corollary 3.3.3 (The sum of independent normals is normal)**.** *Consider independent random variables $X_i \sim N(\mu_i, \sigma_i^2)$. Then*

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

Proof We can write $X_i = \mu_i + \sigma_i Z_i$, where Z_i are independent standard normal random variables. Then

$$\sum_{i=1}^n X_i = \mu + \sum_{i=1}^n \sigma_i Z_i = \mu + \langle Z, v\rangle \quad \text{where} \quad v = (\sigma_1, \dots, \sigma_n).$$

By Corollary 3.3.2, we have $\langle Z, v\rangle \sim N(0, \sigma^2)$, so $\mu + \langle Z, v\rangle \sim N(\mu, \sigma^2)$. □

3.3.2 General normal

Recall from Section 1.7 that a random variable X is called normally distributed if X can be obtained by translating and dilating some standard normal random variable $Z \sim N(0,1)$, i.e. if X can be represented as

$$X = \mu + \sigma Z$$

for some fixed $\mu \in \mathbb{R}^n$ and $\sigma > 0$. Such X has mean μ and variance σ^2, and we write $X \sim N(\mu, \sigma^2)$. Let us extend this idea to higher dimensions:

Definition 3.3.4 (General normal distribution)**.** A random vector X in $\mathbb{R}^n$ is normally distributed if it can be obtained by an affine transformation of some standard normal random vector $Z \sim N(0, I_k)$, i.e. if X can be expressed as

$$X = \mu + AZ$$

for some fixed vector $\mu \in \mathbb{R}^n$ and $n \times k$ matrix A. Such X has mean μ and covariance matrix $\Sigma = AA^\mathsf{T}$ (check!), and we write $X \sim N(\mu, \Sigma)$.

Proposition 3.3.5 (Uniqueness)**.** *The distribution of X is uniquely determined by μ and Σ. Specifically, X has the same distribution as*

$$Y = \mu + \Sigma^{1/2} Z' \quad \textit{where } \Sigma = AA^\mathsf{T} \textit{ and } Z' \sim N(0, I_n). \tag{3.12}$$

Proof We will use a version of the *Cramér-Wold device*, which says that the distributions of all 1D marginals uniquely determine the distribution in $\mathbb{R}^n$. More formally, if X and Y are random vectors in $\mathbb{R}^n$ and $\langle X, u\rangle$ and $\langle Y, u\rangle$ have the same distribution for every $u \in \mathbb{R}^n$, then X and Y have the same distribution.[5]

We want to check that AZ and $\Sigma^{1/2} Z'$ have the same distribution. By Corollary 3.3.2, for each $v \in \mathbb{R}^n$ we have

$$\langle AZ, v\rangle = \langle Z, A^\mathsf{T} v\rangle \sim N(0, \|A^\mathsf{T} v\|_2^2) \quad \text{and} \quad \langle \Sigma^{1/2} Z', v\rangle \sim N(0, \|\Sigma^{1/2} v\|_2^2).$$

Moreover, $\|A^\mathsf{T} v\|_2^2 = \|\Sigma^{1/2} v\|_2^2$ since $\Sigma = AA^\mathsf{T}$. (Check!) So, all 1D marginals of AZ and $\Sigma^{1/2} Z'$ match, and Cramér-Wold device completes the proof. □

If Σ is invertible, the density of X exists and is can be computed it in terms of μ and Σ. The formula may look complicated, but it simply says that a general normal density is an affine transformation of the standard normal density (3.11):

Proposition 3.3.6 (General normal density)**.** *If Σ is invertible, the probability density function of a random vector $X \sim N(\mu, \Sigma)$ equals*

$$f(x) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} \exp\Big(-\frac{1}{2} (x-\mu)^\mathsf{T} \Sigma^{-1} (x-\mu) \Big), \quad x \in \mathbb{R}^n. \tag{3.13}$$

Here $|\Sigma|$ denotes the determinant of Σ.

You will prove this in Exercise 3.15 by a change of variables. Figure 3.4 shows examples of two densities of multivariate normal distributions.

[5] The Cramér-Wold device works because the characteristic function uniquely determines the distribution in $\mathbb{R}^n$, thanks to the Fourier inversion formula. If you do not want to use Cramér-Wold, try proving (3.12) with linear algebra instead (hint: use the singular value decomposition of A).

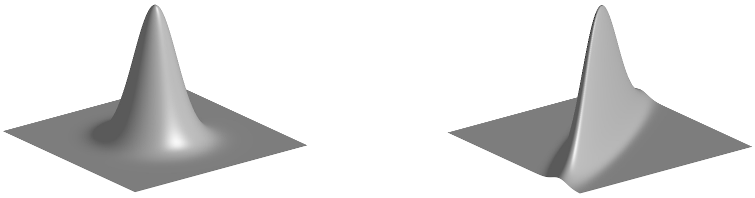


Figure 3.4 The densities of the isotropic distribution $N(0, I_2)$ and an anisotropic distribution $N(0, \Sigma)$. One is obtained by affine transformation of the other. The level sets of the isotropic density are circles, and the level sets of the anisotropic distribution are ellipses.

You might remember that independent random variables are always uncorrelated (why?), but the reverse is not always true (example?) However, the reverse is true for jointly normal random variables:

Corollary 3.3.7 (Jointly normal random variables)**.** *Random variables $X_1, \ldots, X_n$ are called jointly normal if the random vector $X = (X_1, \ldots, X_n)$ is normally distributed. Jointly normal random variables are independent if and only if they are uncorrelated.*

Proof If X_i are uncorrelated, the covariance matrix of X is diagonal. Then the density (3.13) factors, i.e. can be expressed as $f(x) = f_1(x_1) \cdots f_n(x_n)$ for all $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$. (Check!) Now recall a classic criterion of independence: the joint density of random variables X_i factors if and only if X_i are independent. □

Warning: some normal random variables are not *jointly* normal, and can be uncorrelated but not independent (Exercise 3.17).

3.3.3 Uniform on the sphere

The coordinates of an isotropic random vector are always uncorrelated, but not necessarily independent, such as in this example:

Proposition 3.3.8 (A sphere is isotropic)**.** *The uniform distribution*[6] *on the Euclidean sphere in $\mathbb{R}^n$ with radius $\sqrt{n}$ centered at the origin, i.e.* $\mathrm{Unif}(\sqrt{n}S^{n-1})$, *is isotropic.*

Proof Let $X = (X_1, \ldots, X_n) \sim \mathrm{Unif}(S^{n-1})$. By symmetry, each pair of coordinates (X_i, X_j) with distinct i, j has the same distribution as $(-X_i, X_j)$. So, $\mathbb{E}\, X_i X_j = -\mathbb{E}\, X_i X_j$, hence $\mathbb{E}\, X_i X_j = 0$. Next, we always have $\|X\|_2 = 1$, so

$$1 = \mathbb{E}\|X\|_2^2 = \mathbb{E}\, X_1^2 + \cdots + \mathbb{E}\, X_n^2.$$

Since X_i are identically distributed, all n terms in this sum are the same, which

[6] We say that a random vector X is uniformly distributed on a sphere if, for every (Borel) subset E of the sphere, the probability $\mathbb{P}\{X \in E\}$ is the fraction of the $(n-1)$-dimensional area of E relative to the sphere.

yields $\mathbb{E}\,X_i^2 = 1/n$. So, the coordinates of $\sqrt{n}X$ are uncorrelated with second moment equal 1, making $\sqrt{n}X$ isotropic. □

Isotropic random vectors are almost orthogonal

High-dimensional spaces can be counterintuitive. Pick two random points, and they will probably be almost orthogonal!

To be more concrete, let X, Y be independent random vectors uniformly distributed on the unit sphere S^{n-1}. Then $\sqrt{n}X$ and $\sqrt{n}Y$ are independent, identically distributed, and isotropic by Proposition 3.3.8. By Proposition 3.2.1(c), we have $\mathbb{E}\langle\sqrt{n}X, \sqrt{n}Y\rangle^2 = \operatorname{tr}(I_n) = n$. Dividing by n^2, we obtain

$$\mathbb{E}\langle X, Y\rangle^2 = \frac{1}{n}.$$

Then Markov inequality yields

$$|\langle X, Y\rangle| = O(1/\sqrt{n}) \quad \text{with high probability,} \tag{3.14}$$

showing that X and Y typically are almost orthogonal. This contrasts with low dimensions, where a pair of random directions in the plane have the (smaller) angle averaging $\pi/4$. (Check!) In higher dimensions, there is more room, so vectors spread out like lone stars in the sky.

Gaussian and spherical distributions are similar

Both the uniform distribution on S^{n-1} and the standard normal distribution are rotation-invariant. So, if $g \sim N(0, I_n)$, the normalized vector $g/\|g\|_2$ has a rotation-invariant probability distribution on S^{n-1}. But such a distribution is unique (why?), so:

$$g \sim N(0, I_n) \quad \Longrightarrow \quad \frac{g}{\|g\|_2} \sim \operatorname{Unif}\left(S^{n-1}\right). \tag{3.15}$$

The density (3.11) of the standard normal distribution $N(0, I_n)$ is highest at the origin (Figure 3.4), which might suggest that $g \sim N(0, I_n)$ is concentrated near the origin. However, as we saw in Section 3.1, that is not the case! Instead, the normal distribution is concentrated around a thin spherical shell at radius $\sqrt{n}$:

$$\|g\|_2 \approx \sqrt{n} \quad \text{with high probability.} \tag{3.16}$$

Combining this with (3.15), we can informally say that the standard normal distribution is roughly like the uniform distribution on the sphere of radius $\sqrt{n}$:

$$N(0, I_n) \approx \operatorname{Unif}\left(\sqrt{n}S^{n-1}\right). \tag{3.17}$$

This defies our low-dimensional intuition (see Figure 3.5). The catch? There is almost no volume near the origin. A ball of radius $o(\sqrt{n})$ has exponentially small volume (you will compute it in Exercise 4.27), offsetting the density peak and keeping random vectors away from 0. Try Exercise 3.7 to see this in action.

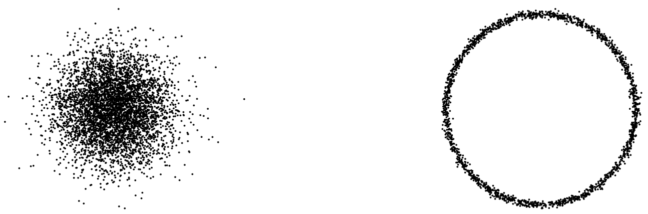


Figure 3.5 Sampling from the standard normal distribution in 2D (left) and its heuristic visualization in high dimensions (right). In high dimensions, the standard normal distribution closely resembles the uniform distribution on a sphere of radius $\sqrt{n}$.

1D projections of the sphere are approximately normal

A rigorous take on the heuristic (3.17) is that 1D marginals of the uniform distribution on the sphere are approximately normal. Even though we won't need this beautiful fact later, let us prove it, thinking of it as a "spherical" version of the Berry-Esseen central limit theorem 2.1.4.

Theorem 3.3.9 (Projective central limit theorem)**.** *Let X be a random vector uniformly distributed on the unit sphere on $\mathbb{R}^n$, i.e $X \sim \mathrm{Unif}\,(S^{n-1})$. Then*

$$\sqrt{n}\langle X, v\rangle \to N(0,1) \quad \textit{in distribution}$$

as $n \to \infty$. In fact, the CDFs converge uniformly:

$$\sup_{v \in S^{n-1}} \sup_{t \in \mathbb{R}} \bigl|\mathbb{P}\bigl\{\sqrt{n}\langle X, v\rangle \le t\bigr\} - \mathbb{P}\{g_1 \le t\}\bigr| \to 0$$

where $g_1 \sim N(0,1)$.

Proof By (3.15), we can assume $X = g/\|g\|_2$ with $g \sim N(0, I_n)$. By rotation invariance, the distribution of $\langle X, v\rangle$ is the same for all v, so we can choose $v = (1, 0, 0, \ldots, 0)$ and get

$$\langle X, v\rangle = \frac{g_1}{\|g\|_2}.$$

The heuristic (3.16) will let us swap $\|g\|_2$ with $\sqrt{n}$, completing the proof. To do this rigorously, we use a *decomposition trick*: split into a "good event" where the heuristic holds (so we use it) and a "bad event" where it does not (which happens with small probability and can be ignored). By (3.2), the "good event"

$$E_n := \bigl\{\bigl|\|g\|_2 - \sqrt{n}\bigr| \le \ln n\bigr\} \quad \text{is likely:} \quad p_n := \mathbb{P}(E_n^c) \to 0.$$

If E_n occurs and $t \ge 0$ (which we can always assume by symmetry – check!), then the event of interest $\sqrt{n}\langle X, v\rangle \le t$ implies

$$g_1 \le \frac{t\|g\|_2}{\sqrt{n}} \le t\Bigl(1 + \frac{\ln n}{\sqrt{n}}\Bigr) =: t_n.$$

Splitting the event based on whether E_n occurs, we get

$$\begin{aligned}\mathbb{P}\{\sqrt{n}\langle X,v\rangle \le t\} &\le \mathbb{P}\{\sqrt{n}\langle X,v\rangle \le t \text{ and } E_n\} + \mathbb{P}(E_n^c)\\ &\le \mathbb{P}\{g_1 \le t_n\} + p_n.\end{aligned}$$

Hence

$$\mathbb{P}\{\sqrt{n}\langle X,v\rangle \le t\} - \mathbb{P}\{g_1 \le t\} \le \mathbb{P}\{g_1 \in [t,t_n]\} + p_n. \tag{3.18}$$

The density of g_1 on $[t,t_n]$ is bounded by $e^{-t^2/2}$, so

$$(3.18) \le e^{-t^2/2}(t_n - t) + p_n = e^{-t^2/2} t \frac{\ln n}{\sqrt{n}} + p_n \le \frac{C \ln n}{\sqrt{n}} + p_n.$$

The right-hand side does not depend on v or t, and it goes to zero as $n \to \infty$. A similar argument gives a bound on $\mathbb{P}\{g_1 \le t\} - \mathbb{P}\{\sqrt{n}\langle X,v\rangle \le t\}$ that also goes to zero. (Do it!) Combining the two bounds completes the proof. □

Remark 3.3.10 (The density of the 1D marginals of the sphere)**.** The density of the 1D marginals of the uniform distribution on the sphere of radius $\sqrt{n}$ can be computed. As you will see in Exercise 3.27, it is proportional to $(1 - x^2/n)^{\frac{n-3}{2}}$. For large n, this approximates $e^{-x^2/2}$, consistent with the Gaussian limit.

3.3.4 Uniform on a convex set

This example comes from convex and computational geometry. A bounded convex set $K \subset \mathbb{R}^n$ with a non-empty interior is called a *convex body*. Let X be a random vector uniformly distributed in K:

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

The density of X equals $1/\operatorname{Vol}(K)$ on K and zero elsewhere. The mean of X,

$$\mu = \mathbb{E}\, X = \frac{1}{\operatorname{Vol}(K)} \int_K x\, dx,$$

is the center of gravity of K. If $\Sigma = \operatorname{cov}(X)$ is the covariance matrix of K, then the standard score $Z := \Sigma^{-1/2}(X - \mu)$ is an an isotropic random vector, as we noticed in (3.10). Also, Z is uniformly distributed in the affinely transformed copy of K:

$$Z \sim \operatorname{Unif}\left(\Sigma^{-1/2}(K - \mu)\right).$$

(Why?) In summary, we found an affine transformation T that makes the uniform distribution on $T(K)$ isotropic. The convex body $T(K)$ is often itself called an isotropic convex body.

In algorithmic convex geometry, we can think of the isotropic convex body $T(K)$ as a well-conditioned version of K, with T acting like a *preconditioner* (see Figure 3.6). Algorithms such as finding the volume of K usually work better when K is well-conditioned.

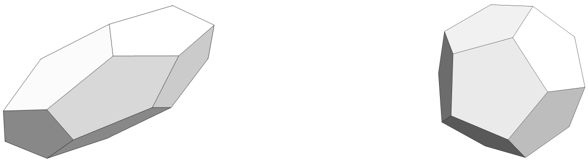


Figure 3.6 A convex body K (left) is transformed into an isotropic convex body $T(K)$ (right). The preconditioner is $T = \Sigma^{-1/2}$, where Σ is the covariance matrix of K.

3.3.5 Frames

The notion of a frame, widely used in signal processing, extends the concept of a basis but drops the requirement of linear independence. Frames are intimately connected to discrete isotropic distributions:

Proposition 3.3.11 (Parseval frames). *For any vectors $u_1, \ldots, u_N \in \mathbb{R}^n$, the following are equivalent:*

(i) (Parseval identity) $\|x\|_2^2 = \sum_{i=1}^N \langle u_i, x\rangle^2$ for each $x \in \mathbb{R}^n$.
(ii) (Frame expansion) $x = \sum_{i=1}^N \langle u_i, x\rangle u_i$ for each $x \in \mathbb{R}^n$.
(iii) (Decomposition of identity) $I_n = \sum_{i=1}^N u_i u_i^\mathsf{T}$.
(iv) (Isotropy) The random vector $X \sim \mathrm{Unif}\{\sqrt{N}u_1, \ldots, \sqrt{N}u_N\}$ is isotropic.

A set of vectors satisfying these equivalent properties is called a Parseval frame*.*

Proof (i)⇒(iv) The identity in (i) can be written as

$$\|x\|_2^2 = \frac{1}{N}\sum_{i=1}^N \left\langle \sqrt{N}u_i, x \right\rangle^2 = \mathbb{E}\langle X, x\rangle^2.$$

Since this holds for all $x \in \mathbb{R}^n$, the random vector X is isotropic by (3.9).

(iv)⇒(iii) because isotropy of X means that

$$I_n = \mathbb{E}\, XX^\mathsf{T} = \frac{1}{N}\sum_{i=1}^N \left(\sqrt{N}u_i\right)\left(\sqrt{N}u_i\right)^\mathsf{T} = \sum_{i=1}^N u_i u_i^\mathsf{T}.$$

(iii)⇒ (ii) follows by multiplying on both sides by x. Finally, (ii)⇒ (i) comes from taking the inner product with x. □

Example 3.3.12 (Coordinate distribution)**.** The standard basis $\{e_1, \ldots, e_n\}$ in $\mathbb{R}^n$ is a Parseval frame. Therefore, a *coordinate random vector*

$$X \sim \mathrm{Unif}\left\{\sqrt{n}e_1, \ldots, \sqrt{n}e_n\right\},$$

is isotropic. Among all high-dimensional distributions, Gaussian is often the easiest to work with – think of it as “the best.” The coordinate distribution, being highly discrete, is often “the worst.”

Example 3.3.13 (The Mercedes-Benz frame)**.** An iconic example of a Parseval frame that is not linearly independent is the set of N equispaced points on the circle of radius $\sqrt{2/N}$, such as the one in Figure 3.7 (try Exercise 3.30).

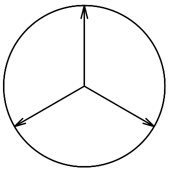


Figure 3.7 A Mercedez-Benz frame: three equispaced points on the circle of radius $\sqrt{2/3}$ form a Parseval frame in $\mathbb{R}^2$.

Finally, here are two more examples of isotropic distributions.

Example 3.3.14 (Uniform on the discrete cube)**.** A *Rademacher random vector* $X = (X_1, \ldots, X_n)$ has independent Rademacher coordinates X_i. Equivalently, X is uniformly distributed on the unit discrete cube in $\mathbb{R}^n$:

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

The Rademacher distribution is isotropic. (Check!)

Example 3.3.15 (Product distributions)**.** More generally, any random vector $X = (X_1, \ldots, X_n)$ whose coordinates X_i are independent random variables with zero mean and unit variance is isotropic. (Why?)

3.4 Subgaussian distributions in higher dimensions

Let's extend the concept of subgaussian distributions, introduced in Section 2.6, to higher dimensions. To get inspired, note that the multivariate normal distribution is fully determined by its *one-dimensional marginals*, or projections onto lines: a random vector X in $\mathbb{R}^n$ is normal if and only if $\langle X, v\rangle$ is normal for any $v \in \mathbb{R}^n$ (see Exercise 3.16). This suggests a natural way to define multivariate subgaussian distributions:

Definition 3.4.1 (Subgaussian random vectors)**.** A random vector X in $\mathbb{R}^n$ is called *subgaussian* if the one-dimensional marginals $\langle X, v\rangle$ are subgaussian random variables for all $v \in \mathbb{R}^n$. The *subgaussian norm* of X is defined by taking the maximal subgaussian norm of $\langle X, v\rangle$ over all unit vectors v:

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

Let's look as some basic examples.

3.4.1 Gaussian, Rademacher, and more

Random vectors with independent subgaussian coordinates give a lot of examples:

Lemma 3.4.2 (Distributions with independent subgaussian coordinates)**.** *Let*

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

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

Proof The lower bound comes from picking v as a standard basis vector in Definition 3.4.1. For the upper bound, fix any $v = (v_1, \dots, v_n) \in S^{n-1}$. Then

$$\begin{aligned} \|\langle X, v\rangle\|_{\psi_2}^2 &= \Big\|\sum_{i=1}^n v_i X_i\Big\|_{\psi_2}^2 \le C \sum_{i=1}^n \|v_i X_i\|_{\psi_2}^2 \quad \text{(by Proposition 2.7.1)} \\ &= C \sum_{i=1}^n v_i^2 \|X_i\|_{\psi_2}^2 \le C \max_{i \le n} \|X_i\|_{\psi_2}^2 \quad \Big(\text{using that } \sum_{i=1}^n v_i^2 = 1\Big). \end{aligned}$$

Since v is arbitrary, the proof is complete. □

Example 3.4.3 (Rademacher)**.** As a consequence, a Rademacher normal random vector (which we introduced in Example 3.3.14) is subgaussian, and

$$c_1 \le \|X\|_{\psi_2} \le c_2$$

where $c_1, c_2 > 0$ are absolute constants.

Example 3.4.4 (Normal)**.** The same goes for the standard normal random vector $X \sim N(0, I_n)$. In Exercise 3.38 you will sharpen and generalize this to $N(0, \Sigma)$.

3.4.2 Uniform on the sphere

The projective central limit Theorem 3.3.9 tells us that the uniform distribution on the sphere of radius $\sqrt{n}$ in $\mathbb{R}^n$ has approximately Gaussian 1D marginals. So the next natural question is: are these marginals subgaussian, with their subgaussian norm bounded by a constant? Let us show that they are. We can think of this fact as a concentration version of the projective central limit theorem, similar to how Hoeffding's inequality in Section 2.2 is a concentration version of the classical central limit theorem.

Theorem 3.4.5 (The uniform distribution on the sphere is subgaussian)**.** *Let X be a random vector uniformly distributed on the unit sphere on $\mathbb{R}^n$, i.e $X \sim$ Unif (S^{n-1}). Then, for any $v \in S^{n-1}$ and $t \ge 0$, we have*

$$\mathbb{P}\{\langle X, v\rangle \ge t\} \le 2\exp\Big(-\frac{t^2 n}{2}\Big). \tag{3.19}$$

In particular, X is subgaussian, and $\|X\|_{\psi_2} \le C/\sqrt{n}$.

Proof By rotation invariance, we can assume

$$X = \frac{g}{\|g\|_2} \quad \text{where } g \sim N(0, I_n),$$

as we mentioned in (3.15). By the same reason, the distribution of $\langle X, v\rangle$ does not depend on v, so we can take $v = (1, 0, 0, \dots, 0)$, which makes $\langle X, v\rangle = X_1$.

Thus, the inequality $\langle X, v\rangle \geq t$ becomes $g_1 \geq t\|g\|_2$. Squaring both sides, moving g_1^2 from the right to the left hand side and simplifying gives

$$g_1 \geq s\|\bar{g}\|_2 \quad \text{where } s = \frac{t}{\sqrt{1-t^2}} \text{ and } \bar{g} = (g_2, \ldots, g_n).$$

To find the probability of this event, we apply the conditioning trick from Section 1.5. First, we fix $\|\bar{g}\|_2$ by conditioning on $\bar{g}$, which does not alter the distribution of g_1 since g_1 and $\bar{g}$ are independent. Then we "uncondition" by taking the expectation over $\bar{g}$. The law of total expectation (1.16) yields

$$\mathbb{P}\{\langle X, v\rangle \geq t\} = \mathbb{P}\{g_1 \geq s\|\bar{g}\|_2\} = \mathbb{E}\left[\mathbb{P}\{g_1 \geq s\|\bar{g}\|_2 \mid \bar{g}\}\right]. \tag{3.20}$$

Since $s\|\bar{g}\|_2$ is fixed by conditioning, the conditional probability in (3.20) reduces to a Gaussian tail. Using the bound $\mathbb{P}\{g_1 \geq u\} \leq \exp(-u^2/2)$ from Exercise 2.6 (if you haven't done it yet, give it a try – it is simple!), we get

$$(3.20) \leq \mathbb{E}\exp\Big(-\frac{s^2\|\bar{g}\|_2^2}{2}\Big) = \Big[\mathbb{E}\exp\Big(-\frac{s^2 g_1^2}{2}\Big)\Big]^{n-1}. \tag{3.21}$$

To get the last identity, write $\|\bar{g}\|_2^2 = g_2^2 + \cdots + g_n^2$ and note that all g_i are independent and distributed identically with $g_1 \sim N(0,1)$. Now, check that

$$\mathbb{E}\exp(-s^2 g_1^2/2) = 1/\sqrt{1+s^2}$$

by expressing the expectation as an integral and changing variables. Thus

$$(3.21) = \Big(\frac{1}{1+s^2}\Big)^{\frac{n-1}{2}} = (1-t^2)^{\frac{n-1}{2}} \leq \exp\Big(-\frac{t^2(n-1)}{2}\Big),$$

since $1 - x \leq e^{-x}$ for all x. To finish the proof, note that the probability in (3.19) is zero for $t \geq 1$ since the Cauchy-Schwarz inequality always gives $\langle X, v\rangle \leq \|X\|_2\|v\|_2 = 1$, while for $t \leq 1$ we have $\exp(-t^2(n-1)/2) \leq e^{1/2}\exp(-t^2 n/2)$. □

3.4.3 Non-examples

Some distributions in $\mathbb{R}^n$ are subgaussian but have a huge subgaussian norm, making it impractical to work with them as subgaussian. Here are a few examples.

Example 3.4.6 (Uniform on a convex body)**.** Let K is a convex body in $\mathbb{R}^n$ and

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

be isotropic as in Section 3.3.4. Qualitatively, X is always subgaussian since K is bounded. But what about quantitatively – is the subgaussian norm of X bounded by an absolute constant?

This is true for *some* isotropic convex bodies like the unit cube $[-1,1]^n$ (thanks to Lemma 3.4.2) and the Euclidean ball of radius $\sqrt{n+2}$ (see Exercises 3.25 and 3.42). But for other convex bodies like the isotropic cross-polytope (a ball in the ℓ^1 norm), the subgaussian norm of X may grow with n (Exercise 3.44).

Even so, a weaker result always holds: X has *subexponential* marginals, and

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

for all unit vectors v. This comes from C. Borell's lemma, which follows from the Brunn-Minkowski inequality, see [134, Section 2.2.b$_3$].

Example 3.4.7 (Coordinate distribution). Recall "the worst" isotropic distribution from Example 3.3.12, given by a random vector

$$X \sim \mathrm{Unif}\left\{\sqrt{n}e_1, \dots, \sqrt{n}e_n\right\}$$

where $\{e_1, \dots, e_n\}$ is the standard basis of $\mathbb{R}^n$. Is X subgaussian? In a qualitative sense, yes: any distribution with finitely many values is. But the subgaussian norm of X grows with n: you will see in Exercise 3.43 that

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

So, quantitatively, it is not really useful to think of X as subgaussian.

Example 3.4.8 (Discrete distributions). Some isotropic discrete distributions have subgaussian norm bounded by a constant, like the Rademacher distribution in Example 3.4.3. However, such distributions must take exponentially many values (Exercise 3.46 – go try it!)

In particular, this rules out *frames* (see Section 3.3.5) as good subgaussian distributions, unless they have exponentially many terms, in which case they are mostly useless in practice.

3.5 Application: Grothendieck inequality and semidefinite programming

In this section and the next, we use high-dimensional Gaussians to tackle problems that do not seem related to probability at all. We start with a probabilistic proof of Grothendieck inequality, a remarkable result which we will use later in the analysis of computationally hard problems.

Theorem 3.5.1 (Grothendieck inequality). *Consider an $m \times n$ matrix (a_{ij}) of real numbers. Assume that*

$$\Big|\sum_{i,j} a_{ij} x_i y_j\Big| \le 1 \quad \textit{for any numbers } x_i, y_j \in \{-1, 1\}.$$

Then, for any Hilbert space H, we have:

$$\Big|\sum_{i,j} a_{ij} \langle u_i, v_j\rangle\Big| \le K \quad \textit{for any unit vectors } u_i, v_j \in H.$$

Here $K \le 1.783$ is an absolute constant.

There is nothing random in the statement of this theorem, but our proof will be probabilistic. In fact, we will give two proofs of Grothendieck inequality. The one here gives a much worse bound, $K \le 14.1$, while an alternative approach in Section 3.7 improves it to $K \le 1.783$ as stated in Theorem 3.5.1.

Before diving into the first argument, let us note a simple observation.

Remark 3.5.2 (A homogeneous form of Grothendieck inequality)**.** The assumption of Grothendieck inequality can be equivalently stated as

$$\Big|\sum_{i,j} a_{ij} x_i y_j\Big| \le \max_i |x_i| \cdot \max_j |y_j| \tag{3.22}$$

for any real numbers x_i and y_j. (You will check this in Exercise 3.47.) The conclusion of Grothendieck inequality can be equivalently stated as

$$\Big|\sum_{i,j} a_{ij} \langle u_i, v_j \rangle\Big| \le K \max_i \|u_i\| \cdot \max_j \|v_j\| \tag{3.23}$$

for any Hilbert space H and any vectors $u_i, v_j \in H$. (Check this by rescaling.)

Proof of Theorem 3.5.1 with $K \le 288$. **Step 1: Reductions.** Note that Grothendieck inequality becomes trivial if we allow the value of K depend on the matrix $A = (a_{ij})$. (For example, $K = \sum_{ij} |a_{ij}|$ would work – check!) Let $K = K(A)$ be the smallest number that makes the conclusion (3.23) hold for a given matrix A and any Hilbert space H and any vectors $u_i, v_j \in H$. Our goal is to show that K is actually *independent* of A and the dimensions m and n.

Without loss of generality,[7] we may show this for a specific Hilbert space H, namely for $\mathbb{R}^N$ equipped with the Euclidean norm $\|\cdot\|_2$. By definition of $K = K(A)$, there exist vectors $u_i, v_j \in \mathbb{R}^N$ satisfying

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

Step 2: Introducing randomness. The key idea of the proof is to express the vectors u_i, v_j using Gaussian random variables

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

Then U_i and V_j are standard normal random variables whose correlations follow exactly the inner products of the vectors u_i and v_j:

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

This comes directly from Corollary 3.3.2 and Exercise 3.9. (If you haven't done this exercise yet, try it now!) Thus

$$K = \sum_{i,j} a_{ij} \langle u_i, v_j \rangle = \mathbb{E} \sum_{i,j} a_{ij} U_i V_j. \tag{3.24}$$

[7] This works because we can first replace H with the subspace spanned by the vectors u_i and v_j, which has dimension at most $N = m + n$ and inherits the norm from H. Then, we use the basic fact that all N-dimensional Hilbert spaces are isometric to $\mathbb{R}^N$ with the usual Euclidean norm $\|\cdot\|_2$. This isometry can be built by matching a given orthonormal basis of H with the canonical basis of $\mathbb{R}^N$.

Suppose for a moment that the random variables $|U_i|$ and $|V_j|$ were almost surely bounded by some constant, say R. Then, from the assumption (3.22), we would get $|\sum_{i,j} a_{ij} U_i V_j| \le R^2$ almost surely. Plugging this into (3.24) would give $K \le R^2$, finishing the proof.

Step 3: Truncation. This reasoning is flawed, of course, because the Gaussian random variables $U_i, V_j \sim N(0,1)$ are unbounded. But their tails are light enough that they are close to being bounded. To act on this heuristic, we use a *truncation* trick. Pick a level $R \ge 1$ and split the random variables like this:

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

We similarly decompose $V_j = V_j^- + V_j^+$. Now U_i^- and V_j^- are bounded by R, as desired. The remainder terms U_i^+ and V_j^+ are small in the L^2 norm: a Gaussian tail bound (Exercise 2.4(b)) gives

$$\|U_i^+\|_{L^2}^2 \le 2\Big(R + \frac{1}{R}\Big) \frac{1}{\sqrt{2\pi}} e^{-R^2/2} < \frac{4}{R^2}, \tag{3.25}$$

A similar bound holds for V_j^+.

Step 4: Breaking up the sum. Replacing $U_i V_j$ with $(U_i^- + U_i^+)(V_i^- + V_j^+)$ in (3.24) and expanding the sum, we get

$$K = \underbrace{\mathbb{E} \sum_{i,j} a_{ij} U_i^- V_j^-}_{S_-} + \underbrace{\mathbb{E} \sum_{i,j} a_{ij} U_i^+ V_j^-}_{S_\pm} + \underbrace{\mathbb{E} \sum_{i,j} a_{ij} U_i^- V_j^+}_{S_\mp} + \underbrace{\mathbb{E} \sum_{i,j} a_{ij} U_i^+ V_j^+}_{S_+}.$$

Now let us bound each term. S_- is the easiest: by construction, $|U_i^-|$ and $|V_j^-|$ are bounded by R, so just as we explained in Step 2, we get

$$S_- \le R^2.$$

We cannot use the same reasoning for $S_\pm$, since the random variable U_i^+ is unbounded. Instead, let us treat the random variables U_i^+ and V_j^- as elements of the Hilbert space L^2 with inner product $\langle X, Y \rangle_{L^2} = \mathbb{E}\, XY$, and thus write

$$S_\pm = \sum_{i,j} a_{ij} \langle U_i^+, V_j^- \rangle_{L^2}.$$

We have $\|U_i^+\|_{L^2} < 2/R$ by (3.25) and $\|V_j^-\|_{L^2} \le \|V_j\|_{L^2} = 1$ by construction. Then, applying the conclusion (3.23) for the Hilbert space $H = L^2$, we find that[8]

$$S_\pm \le K \cdot \frac{2}{R}.$$

The last two terms, $S_\mp$ and S_+, can be bounded just like $S_\pm$. (Check!)

[8] It might seem odd that we are using the inequality we are trying to prove. But remember, we picked $K = K(A)$ at the start as the smallest value to make Grothendieck inequality work. That is the K we are using here.

Step 5: Putting everything together. Plugging the bounds on all four terms into (3.24), we conclude that

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

Setting $R = 12$ and rearranging the terms gives $K \le 288$. A little finer analysis, skipping the rough $4/R^2$ bound in (3.25), yields $K \le 14.1$ (Exercise 3.48). □

Remark 3.5.3 (Quadratic Grothendieck: $x_i = y_i$)**.** The assumption of Grothendieck inequality can often be relaxed by assuming $x_i = y_i$, letting us bound a quadratic instead of a bilinear form. Let $A = (a_{ij})$ be an $n \times n$ matrix, either symmetric positive-semidefinite or diagonal-free. Assume that

$$\Big|\sum_{i,j} a_{ij} x_i x_j\Big| \le 1 \quad \text{for any numbers } x_i \in \{-1, 1\}.$$

Then, for any Hilbert space H, we have:

$$\Big|\sum_{i,j} a_{ij} \langle u_i, v_j \rangle\Big| \le 2K \quad \text{for any unit vectors } u_i, v_j \in H.$$

Here K is the absolute constant from Grothendieck inequality. You will check this in Exercises 3.49 and 3.50 – go ahead and try them now!

3.5.1 Semidefinite programming

Some hard computational problems can be relaxed into easier, more computationally tractable problems. Relaxation is often achieved through semidefinite programming, and Grothendieck inequality can help guarantee the its quality. Let's see how.

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

$$\text{maximize } \langle A, X \rangle : \quad X \succeq 0, \quad \langle B_i, X \rangle \le b_i \text{ for } i = 1, \ldots, N. \tag{3.26}$$

Here A and B_i are given $n \times n$ matrices, and b_i are given numbers. The "variable" X is an $n \times n$ symmetric positive semidefinite matrix, indicated by the notation $X \succeq 0$. The inner product is the standard one on the space of $n \times n$ matrices:[9]

$$\langle A, X \rangle = \operatorname{tr}(A^\mathsf{T} X) = \sum_{i,j=1}^{n} A_{ij} X_{ij}. \tag{3.27}$$

Note that if we *minimize* instead of maximize in (3.26), we still get a semidefinite program. Same goes for replacing any "$\le$" signs by "$\ge$" or "$=$". (Why?)

Remark 3.5.5 (An SDP program is a convex program)**.** Every semidefinite program is a *convex program* because it involves maximizing a linear function

[9] Think of matrices as long vectors in $\mathbb{R}^{n^2}$ to see why (3.27) makes sense.

$\langle A, X\rangle$ over a convex set of matrices. (The set of positive semidefinite matrices is indeed convex (check!), and so is its intersection with the half-spaces defined by the constraints $\langle B_i, X\rangle \le b_i$.) This is good news because convex programs are generally *algorithmically tractable*. There are efficient solvers for general convex programs, and specifically for semidefinite programs.

Semidefinite relaxations

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

$$\text{maximize} \sum_{i,j=1}^{n} A_{ij} x_i x_j : \quad x_i = \pm 1 \text{ for } i = 1, \ldots, n \tag{3.28}$$

where A is a given $n \times n$ symmetric matrix. This is an *quadratic integer optimization problem*, whose feasible set consists of 2^n vectors $x = (x_i) \in \{-1, 1\}^n$. Finding the maximum by exhaustive search takes exponential time. Is there a smarter way? Not likely: (3.28) is a computationally hard problem (NP-hard).

Still, we can relax problem (3.28) into a semidefinite program that approximates the maximum within a constant factor. To do this, we replace in (3.28) the numbers $x_i = \pm 1$ by their higher-dimensional analogs – unit vectors X_i in $\mathbb{R}^n$. This leads to the following optimization problem:

$$\text{maximize} \sum_{i,j=1}^{n} A_{ij} \langle X_i, X_j\rangle : \quad \|X_i\|_2 = 1 \text{ for } i = 1, \ldots, n. \tag{3.29}$$

Proposition 3.5.6 (The relaxation is an SDP). *The optimization problem* (3.29) *is equivalent to the following semidefinite program:*

$$\textit{maximize } \langle A, Z\rangle : \quad Z \succeq 0, \quad Z_{ii} = 1 \textit{ for } i = 1, \ldots, n. \tag{3.30}$$

Proof Recall that the *Gram matrix* of vectors $X_1, \ldots, X_n$ is the $n \times n$ matrix Z with entries $Z_{ij} = \langle X_i, X_j\rangle$. Then the two problems are equivalent thanks to two linear algebra facts: (a) the Gram matrix of any set of vectors is symmetric and positive-semidefinite, and (b) conversely, any symmetric and positive-semidefinite matrix is a Gram matrix of some set of vectors (see Exercise 3.51). □

The guarantee of relaxation

Let us check that the semidefinite relaxation is accurate by showing that SDP (3.29) approximates (3.28) within a constant factor:

Theorem 3.5.7 (The guarantee of relaxation). *Let A be an $n \times n$ symmetric, positive-semidefinite matrix. Let* $\text{int}(A)$ *denote the maximum in the integer optimization problem* (3.28) *and* $\text{sdp}(A)$ *denote the maximum in the semidefinite problem* (3.29). *Then*

$$\text{int}(A) \le \text{sdp}(A) \le 2K \cdot \text{int}(A)$$

where $K \le 1.783$ is the constant in Grothendieck inequality.

Proof The first bound follows with $X_i = (x_i, 0, 0, \ldots, 0)^\mathsf{T}$. The second comes from the quadratic Grothendieck inequality in Remark 3.5.3. (Why can we drop the absolute values?) □

Although Theorem 3.5.7 helps us approximate the maximum value in (3.28), it is not obvious how to find the actual solution $x_1, \ldots, x_n$ that attain this approximate value. Can we convert the vectors X_i that give a solution of SDP (3.29) into labels $x_i = \pm 1$ that approximately solve (3.28)?

Yes, we can! But first, let us do this for another remarkable NP-hard problem – the max cut problem. Once we have done that, you will be ready to handle (3.30) on your own, achieving an even better approximation constant than $2K$ (Exercise 3.58.)

3.6 Application: Maximum cut for graphs

Let us see how semidefinite relaxations can be useful for tackling one of the well known NP-hard problems: finding the *maximum cut* of a graph.

An undirected *graph* $G = (V, E)$ is defined as a set of vertices V together with a set of edges E; each edge is an unordered pair of vertices. We focus on finite, *simple* graphs – finite, with no loops or multiple edges. For convenience, we label the vertices by integers, setting $V = \{1, \ldots, n\}$.

Definition 3.6.1 (Maximum cut)**.** If we partition the vertices of a graph G into two disjoint subsets, the *cut* is the number of edges between them. The *maximum cut* of G, denoted $\operatorname{maxcut}(G)$, is the largest possible cut over all partitions of vertices. See Figure 3.8 for illustration.

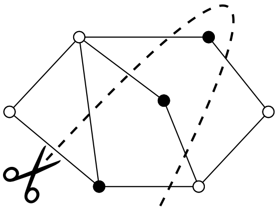


Figure 3.8 The dashed line shows the maximal cut of this graph, splitting the vertices into black and white and giving $\operatorname{maxcut}(G) = 7$.

Finding the maximum cut is generally a computationally hard problem (NP-hard).

3.6.1 A simple 0.5*-approximation algorithm*

Let us relax the maximum cut problem to a semidefinite program, using the approach from Section 3.5.1. To do this, we first translate the problem into the language of linear algebra.

Definition 3.6.2 (Adjacency matrix)**.** The *adjacency matrix* A of a graph G with vertices $V = \{1, \dots, n\}$ is a symmetric $n \times n$ matrix where $A_{ij} = 1$ if vertices i and j are connected by an edge, and $A_{ij} = 0$ otherwise.

A partition of the vertices into two sets can be described by a vector of labels

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

where the sign of x_i shows which subset vertex i belongs to. For example, in Figure 3.8, the three black vertices might have $x_i = 1$ and the four white vertices $x_i = -1$. The cut of G for this partition is simply the number of edges between vertices with opposite labels:[10]

$$\operatorname{cut}(G, x) = \frac{1}{2} \sum_{i,j:\, x_i x_j = -1} A_{ij} = \frac{1}{4} \sum_{i,j=1}^{n} A_{ij}(1 - x_i x_j). \tag{3.31}$$

The maximum cut is found by maximizing $\operatorname{cut}(G, x)$ over all partitions x:

$$\operatorname{maxcut}(G) = \frac{1}{4} \max \Big\{ \sum_{i,j=1}^{n} A_{ij}(1 - x_i x_j) :\ x_i = \pm 1 \ \forall i \Big\}. \tag{3.32}$$

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

Proposition 3.6.3 (0.5-approximation algorithm for maximum cut)**.** *If we split the vertices of G into two sets at random, uniformly over all 2^n partitions, the expected cut is at least $0.5 \operatorname{maxcut}(G)$.*

Proof A random cut is generated by a Rademacher random vector x, a vector whose coordinates are independent Rademacher random variables (recall Example 3.3.14). Then, in (3.31) we have $\mathbb{E}\, x_i x_j = 0$ for $i \neq j$ and $A_{ij} = 0$ for $i = j$ since the graph has no loops. Thus, by linearity of expectation, we get

$$\mathbb{E} \operatorname{cut}(G, x) = \frac{1}{4} \sum_{i,j=1}^{n} A_{ij} = \frac{1}{2} E \geq \frac{1}{2} \operatorname{maxcut}(G),$$

where E denotes the number of edges of G. This completes the proof. □

3.6.2 Semidefinite relaxation

Now we will do better and give a 0.878-approximation algorithm for max cut, due to Goemans and Williamson. It is based on a semidefinite relaxation of the NP-hard problem (3.32). Given (3.29), the relaxation should be easy to guess – we consider the semidefinite problem

$$\operatorname{sdp}(G) := \frac{1}{4} \max \left\{ \sum_{i,j=1}^{n} A_{ij}(1 - \langle X_i, X_j \rangle) :\ X_i \in \mathbb{R}^n,\ \|X_i\|_2 = 1 \ \forall i \right\}. \tag{3.33}$$

[10] The factor $\frac{1}{2}$ in (3.31) prevents double counting the edges (i, j) and (j, i).

(Again – why is this a semidefinite program?)

We will show that the sdp(G) approximates maxcut(G) within a 0.878 factor, and how to turn the solution (X_i) into labels $x_i = \pm 1$ for an actual partition of the graph. We do this by *randomized rounding*: pick a random hyperplane through the origin in $\mathbb{R}^n$ and assign $x_i = 1$ to the vectors X_i on one side, $x_i = -1$ to the other (see Figure 3.9). More formally, consider a standard normal random vector[11] $g \sim N(0, I_n)$ and define

$$x_i := \operatorname{sign}\langle X_i, g\rangle, \quad i = 1, \dots, n. \tag{3.34}$$

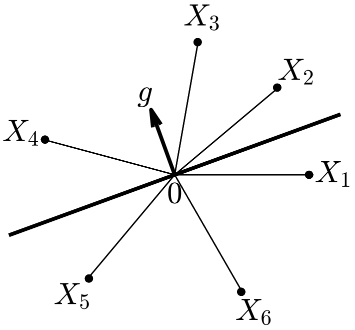


Figure 3.9 We do randomized rounding of these vectors $X_i \in \mathbb{R}^n$ into labels $x_i = \pm 1$ by choosing a random hyperplane with normal vector g (shown in bold) and assigning $x_2 = x_3 = x_4 = 1$ and $x_1 = x_5 = x_6 = -1$.

Theorem 3.6.4 (0.878-approximation algorithm for maximum cut)**.** *Let G be a graph with adjacency matrix A. Let (X_i) be a solution of the semidefinite program (3.33), and $x = (x_i)$ be the result of a randomized rounding of (X_i). Then*

$$\mathbb{E}\operatorname{cut}(G, x) \geq 0.878 \operatorname{sdp}(G) \geq 0.878 \operatorname{maxcut}(G).$$

The proof is based on an elementary inequality. In proving Grothendieck inequality (Theorem 3.5.1), we relied on the fact that if $g \sim N(0, I_n)$, then

$$\mathbb{E}\langle g, u\rangle\langle g, v\rangle = \langle u, v\rangle$$

for any fixed vectors $u, v \in \mathbb{R}^n$. (Exercise 3.9). We will need a slightly more advanced version of this identity, which you will prove in Exercise 3.53:

Lemma 3.6.5 (Grothendieck identity)**.** *Consider a random vector $g \sim N(0, I_n)$. Then, for any fixed vectors $u, v \in S^{n-1}$, we have*

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

[11] Instead of the normal distribution, we could use any other rotation invariant distribution in $\mathbb{R}^n$, like the uniform distribution on the sphere S^{n-1}.

A downside of the Grothendieck identity is the nonlinear function arcsin, which is hard to work with. We can replace it with a linear bound using the inequality

$$1 - \frac{2}{\pi}\arcsin t = \frac{2}{\pi}\arccos t \geq 0.878(1-t), \quad t \in [-1,1], \tag{3.35}$$

which can be checked easily with software (see Figure 3.10).

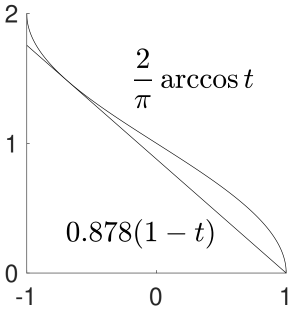


Figure 3.10 The inequality $\frac{2}{\pi}\arccos t \geq 0.878(1-t)$ holds for all $t \in [-1,1]$.

Proof of Theorem 3.6.4 By (3.31) and linearity of expectation, we have

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

The definition of labels x_i in the rounding step (3.34) gives

$$\begin{aligned} 1 - \mathbb{E}\,x_i x_j &= 1 - \mathbb{E}\,\mathrm{sign}\langle X_i, g\rangle\,\mathrm{sign}\langle X_j, g\rangle \\ &= 1 - \frac{2}{\pi}\arcsin\langle X_i, X_j\rangle \quad \text{(by Grothendieck identity, Lemma 3.6.5)} \\ &\geq 0.878(1 - \langle X_i, X_j\rangle) \quad \text{(by (3.35))}. \end{aligned}$$

Therefore

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

This proves the first inequality in the theorem. The second inequality is trivial since $\mathrm{sdp}(G) \geq \mathrm{maxcut}(G)$. (Why?) □

Try Exercises 3.56–3.58 now to get a better feel of randomized rounding, Grothendieck identity, and semidefinite relaxations.

3.7 Kernel trick, and tightening of Grothendieck inequality

Our proof of Grothendieck inequality given in Section 3.5 yields a very loose bound on the absolute constant K. Now, we will take a different approach to get (almost) the best known bound: $K \leq 1.783$.

Our new argument will build on Grothendieck identity (Lemma 3.6.5), but the non-linearity of the function $\arcsin(x)$ presents a challenge. If, hypothetically, there were no nonlinearity, and we had the ideal identity $\mathbb{E}\,\mathrm{sign}\langle g,u\rangle\,\mathrm{sign}\langle g,v\rangle = \frac{2}{\pi}\langle u,v\rangle$, then Grothendieck inequality (Theorem 3.5.1) would easily follow:

$$\frac{2}{\pi}\sum_{i,j} a_{ij}\langle u_i, v_j\rangle = \mathbb{E}\sum_{i,j} a_{ij}\,\underbrace{\mathrm{sign}\langle g,u_i\rangle}_{x_i}\,\underbrace{\mathrm{sign}\langle g,v_j\rangle}_{y_j} \le 1.$$

(The bound follows from the assumption of Grothendieck inequality since $x_i, y_j \in \{-1,1\}$.) This would give Grothendieck inequality with $K \le \pi/2 \approx 1.57$.

This argument is, of course, wrong. To properly handle the nonlinear function of an inner product $\langle u,v\rangle$, we can use a remarkably powerful trick: rewrite it as a (linear) inner product $\langle u',v'\rangle$ of some other vectors u', v' in some Hilbert space H. In the literature on machine learning, this is known as the *kernel trick*.

We will explicitly construct the non-linear transformations $u' = \Phi(u)$, $v' = \Psi(v)$ that will do the job. The best way to describe them is using *tensors*, which generalize matrices to higher dimensions.

3.7.1 Tensors

A tensor can be described as a multidimensional array. A matrix has two dimensions (rows and columns), while a tensor can have any number of dimensions.

Definition 3.7.1 (Tensors)**.** An order k tensor $(a_{i_1\dots i_k})$ is an $n_1 \times n_2 \times \cdots \times n_k$ array of real numbers $a_{i_1\dots i_k}$. The canonical inner product on $\mathbb{R}^{n_1\times\cdots\times n_k}$ defines the inner product of tensors: for $A = (a_{i_1\dots i_k})$ and $B = (b_{i_1\dots i_k})$, we set

$$\langle A, B\rangle := \sum_{i_1,\dots,i_k} a_{i_1\dots i_k} b_{i_1\dots i_k}. \tag{3.36}$$

Example 3.7.2 (Vectors and matrices)**.** Vectors are order 1 tensors, and matrices as order 2 tensors. The inner product for tensors (3.36) generalizes the inner product for vectors (1.2) and matrices (3.27).

Example 3.7.3 (Rank-one tensors)**.** For a vector $u \in \mathbb{R}^n$, the *order k tensor product* $u\otimes\cdots\otimes u$ is the tensor whose entries are the products of all k-tuples of the entries of u:

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

For example, if $k = 2$, the tensor product $u\otimes u$ is the $n\times n$ matrix

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

Lemma 3.7.4 (Powers)**.** *For any vectors $u, v \in \mathbb{R}^n$ and $k \in \mathbb{N}$, we have*

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

Proof Let's check this for $n = 3$:

$$\langle u^{\otimes 3}, v^{\otimes 3}\rangle = \sum_{i,j,\ell=1}^{n} (u_i u_j u_k)(v_i v_j v_k) = \Big(\sum_{i=1}^{n} u_i v_i\Big)\Big(\sum_{j=1}^{n} u_j v_j\Big)\Big(\sum_{\ell=1}^{n} u_\ell v_\ell\Big) = \langle u, v\rangle^3.$$

The general case is similar (write it!). □

Lemma 3.7.4 reveals something interesting: non-linear expressions like $\langle u, v\rangle^k$ can be written as a standard *linear* inner product in a different space. Specifically, there is a Hilbert space H and a transformation $\Phi : \mathbb{R}^n \to H$ such that

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

In fact we can take $H = \mathbb{R}^{n^k}$, the space of k-th order tensors and $\Phi(u) = u^{\otimes k}$.

Now that we learned how to handle the power function, we can move on to more general non-linearities:

Example 3.7.5 (Polynomials with nonnegative coefficients)**.** There exists a Hilbert space H and a transformation $\Phi : \mathbb{R}^n \to H$ such that

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

We can take

$$\Phi(u) = (\sqrt{2}u \otimes u) \oplus (\sqrt{5}u \otimes u \otimes u)$$

where $\oplus$ denotes concatenation.[12] So, the target space is $H = \mathbb{R}^{n^2+n^3}$.

Example 3.7.6 (General polynomials)**.** Polynomials with negative coefficients can make our task impossible since $\langle \Phi(u), \Phi(u)\rangle$ is always nonnegative. But here is a neat workaround: we can find *two* transformations $\Phi, \Psi : \mathbb{R}^n \to H$, possibly different, such that

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

We can take

$$\Phi(u) = (\sqrt{2}u \otimes u) \oplus (\sqrt{5}u \otimes u \otimes u),$$
$$\Psi(v) = (\sqrt{2}u \otimes u) \oplus (-\sqrt{5}u \otimes u \otimes u).$$

Note that the transformations keep the lengths of vectors under control. For any unit vector u, we have $\|\Phi(u)\|_2^2 = \|\Psi(u)\|_2^2 = 2\langle u, u\rangle^2 + 5\langle u, u\rangle^3 = 2 + 5 = 7$, which is just the sum of absolute values of the coefficients.

Following this approach, we can handle any polynomial $f(x) = \sum_{k=1}^{N} a_k x^k$. And by taking limits of polynomials, we can handle even more functions:

Lemma 3.7.7 (Real analytic functions)**.** *Consider a function* $f(x) = \sum_{k=0}^{\infty} a_k x^k$

[12] Just think of $\Phi(u)$ as a long vector obtained by listing the entries of the tensor $\sqrt{2}u \otimes u$ as a vector in $\mathbb{R}^{n^2}$, followed by those of $\sqrt{5}u \otimes u \otimes u \in \mathbb{R}^{n^3}$.

where the series converges for all $x \in \mathbb{R}$. There exists a Hilbert space H and transformations $\Phi, \Psi : \mathbb{R}^n \to H$ such that

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

Also, for any unit vector u, we have $\|\Phi(u)\|_2^2 = \|\Psi(u)\|_2^2 = \sum_{k=0}^{\infty} |a_k|$.

You will formally check this step in Exercise 3.55 – do it now!

Example 3.7.8 (Sine)**.** Let $c > 0$. The function $f(x) = \sin(cx)$ is real analytic:

$$\sin(cx) = cx - \frac{(cx)^3}{3!} + \frac{(cx)^5}{5!} - \frac{(cx)^7}{7!} + \cdots$$

Thus, there exists a Hilbert space H and transformations $\Phi, \Psi : \mathbb{R}^n \to H$ such that

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

Also, Φ and Ψ map unit vectors to unit vectors if

$$1 = c + \frac{c^3}{3!} + \frac{c^5}{5!} + \frac{c^7}{7!} + \cdots = \frac{e^c + e^{-c}}{2}.$$

Solving this equation yields $c = \ln(1 + \sqrt{2})$.

3.7.2 Proof of Theorem 3.5.1

We are ready to prove Grothendieck inequality (Theorem 3.5.1) with constant

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

We can assume without loss of generality that $u_i, v_j \in \mathbb{R}^N$ with $N = n + m$, just like in our first proof of Grothendieck inequality in Section 3.5. Then, by Example 3.7.8 with $c = \ln(1 + \sqrt{2})$, we can find unit vectors u_i', v_j' in some Hilbert space H that satisfy

$$\langle u_i', v_j' \rangle = \sin(c\langle u_i, v_j \rangle) \quad \text{for all } i, j. \tag{3.37}$$

Again, we can assume $H = \mathbb{R}^N$ without loss of generality. Applying Grothendieck identity (Lemma 3.6.5), we get

$$\mathbb{E} \operatorname{sign}\langle g, u_i' \rangle \operatorname{sign}\langle g, v_j' \rangle = \frac{2}{\pi} \arcsin\langle u_i', v_j' \rangle = \frac{2c}{\pi} \langle u_i, v_j \rangle.$$

Thus

$$\frac{2c}{\pi} \sum_{ij} a_{ij} \langle u_i, v_j \rangle = \mathbb{E} \sum_{ij} a_{ij} \underbrace{\operatorname{sign}\langle g, u_i' \rangle}_{x_i} \underbrace{\operatorname{sign}\langle g, v_j' \rangle}_{y_j} \le 1.$$

The last step follows from the assumption of Grothendieck inequality since $X_i, Y_j \in \{-1, 1\}$. The proof is complete since $c = \ln(1 + \sqrt{2})$. □

Remark 3.7.9 (An algorithmic viewpoint)**.** This proof gives a randomized algorithm that takes a matrix (a_{ij}) and unit vectors u_i, v_j and finds labels $x_i, y_j \in \{-1, 1\}$ satisfying

$$\mathbb{E} \sum_{i,j} a_{ij} x_i y_j \geq \frac{1}{K} \sum_{i,j} a_{ij} \langle u_i, v_j \rangle.$$

Here is how it works: First, find unit vectors $u'_i, v'_j \in \mathbb{R}^{n+m}$ with prescribed inner products (3.37) (which can be set up as a semidefinite program – how?). Then, use randomized rounding: pick $g \sim N(0, I_n)$ and set $x_i = \operatorname{sign}\langle g, u'_i \rangle$ and $y_i = \operatorname{sign}\langle g, v'_i \rangle$.

3.7.3 Kernels and feature maps

Since the kernel trick worked so well for Grothendieck inequality, we might wonder – what other nonlinearities can it handle? Given a function of two variables $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ on some set $\mathcal{X}$, when can we find a Hilbert space H and a transformation $\Phi : \mathcal{X} \to H$ so that

$$\langle \Phi(u), \Phi(v) \rangle = K(u, v) \quad \text{for all } u, v \in \mathcal{X}? \tag{3.38}$$

The answer is given by Mercer theorem and, more precisely, Moore-Aronszajn theorem. The necessary and sufficient condition is that K be a *positive semidefinite kernel*, meaning that for any points $u_1, \ldots, u_N \in \mathcal{X}$, the matrix $(K(u_i, u_j))_{i,j=1}^N$ is symmetric and positive semidefinite. The transformation Φ is called a *feature map*, and the Hilbert space H is called a *reproducing kernel Hilbert space* (RKHS).

Popular positive semidefinite kernels in machine learning include the *Gaussian* and *polynomial* kernels, given by:

$$K(u, v) = \exp\Big(-\frac{\|u - v\|_2^2}{2\sigma^2} \Big), \quad K(u, v) = \big(\langle u, v \rangle + r \big)^k, \quad u, v \in \mathbb{R}^n,$$

where $\sigma > 0$, $r > 0$, and $k \in \mathbb{N}$ are parameters. The kernel trick (3.38), which expresses a kernel $K(u, v)$ as an inner product, is widely used in machine learning because it lets us handle non-linear models (determined by K) with techniques designed for linear models. The exact details of the Hilbert space H and feature map Φ are not usually needed. To compute the inner product $\langle \Phi(u), \Phi(v) \rangle$ in H, you don't even need to know Φ – the identity (3.38) lets you calculate $K(u, v)$ directly.

3.8 Notes

Theorem 3.1.1 and its proof come from [213], though some versions of this result are folklore. The dependence on K in this theorem is not optimal. Jeong, Li, Plan, and Yilmaz [176] improved it from K^2 to $K\sqrt{\log K}$ and showed this is the best possible in general. This also sharpens several other results in the book that use Theorem 3.1.1.

A natural question related to Theorem 3.1.1 is how the norm $\|X\|_2$ concentrates when the coordinates of a random vector X are not independent. For X uniformly distributed in a convex set K, this is a key problem in geometric functional analysis; see [150, Section2] and [61, Chapter 12].

The “projective central limit theorem” (Theorem 3.3.9) goes back to Borel in some form [49, Chapter V]; see [102] for its history, quantitative versions, and extensions for higher-dimensional marginals.

Section 3.3.4 and Example 3.4.6 discuss random vectors uniformly distributed in convex sets. The books [21, 61] study this topic in detail, and the surveys [303, 336] explore algorithmic aspects of computing convex set volumes in high dimensions.

For more on frames introduced in Section 3.3.5, see [78, 193].

There are a few proofs of Theorem 3.4.5 and related results; for a simple geometric one, see [23]. The bound can be refined a bit based on an explicit density formula (Exercise 3.27); see [52, Exercise 7.9] based on [62].

Grothendieck inequality (Theorem 3.5.1) was originally proved by A. Grothendieck in 1953 [149] with a bound $K \le \sinh(\pi/2) \approx 2.30$; a version of this original argument appears in [215, Section 2]. Many alternative proofs exist, some with better or worse bounds on K ; see [60] for a historical overview. Surveys [185, 273] cover the impact of Grothendieck inequality across mathematics and computer science.

Our first proof (Section 3.5) follows [14, Section8.1] and was kindly brought to my attention by Mark Rudelson. Our second proof (Section 3.7) is due to J.-L. Krivine [195]; versions of this argument can be found in [16] and [205]. Krivine’s approach gives the best known *explicit* bound $K \le \frac{\pi}{2\ln(1+\sqrt{2})} \approx 1.783$. The actual optimal bound is strictly smaller [60], but an explicit expression remains unknown.

In Section 3.5.1, we explored semidefinite relaxations of combinatorial optimization problems. For an introduction to convex optimization and semidefinite programming, see [58, 64, 205, 51]. For the use of Grothendieck inequality in analyzing semidefinite relaxations, see [185, 16, 151].

Our presentation of the maximum cut problem in Section 3.6 follows [64, Section 6.6] and [205, Chapter 7]. The semidefinite approach (Section 3.6.2) was pioneered by M. Goemans and D. Williamson [139]. The approximation ratio $\frac{2}{\pi}\min_{0\le\theta\le\pi}\frac{\theta}{1-\cos(\theta)} \approx 0.878$ is still the best known. If the Unique Games Conjecture holds, no better ratio is possible: any better approximation would be NP-hard to compute [184].

In Section 3.7, we presented Krivine’s proof of Grothendieck’s inequality [195] following its exposition by Alon and Naor [16], and touched on kernel methods. For more on kernels, reproducing kernel Hilbert spaces, and their role in machine learning, see [162].

Exercise 3.13 can be improved for the standard normal distribution: $\mathbb{E}\max_{i\le N}\|X_i\|_2 \le \sqrt{n}+\sqrt{2\ln n}$, which generalizes Exercise 2.38; see [127, Proposition 8.2].

The expected ℓ^p norm of a gaussian random vector (Exercises 3.5 and 3.6) was first computed by G. Schechtman and J. Zinn [296]; they attribute the lower bound (Exercise 3.6) to J. Bourgain.

The probability that random points are in convex position (Exercise 3.23) traces back to Sylvester’s problem in stochastic geometry [29]. Related questions include the probability that the convex hull contains the origin [348] (see [297, Theorem 8.2.1]) and the broader study of the geometry of random polytopes [170, 297].

Examples 3.18 and 3.19 introduce two classical ensembles of random matrices: Ginibre and GOE. For these and other invariant ensembles, see the classic [231] and a thorough introduction to random matrix theory [20]. We will dive deeper into random matrices in Chapter 4.

In Exercise 3.57, we prove Grothendieck inequality for positive semidefinite matrices with constant $K \le \pi/2$. This result is originally due to Rietz [283], and Grothendieck showed that the constant $\pi/2$ is optimal [149]. Exercises 3.56, 3.57 and 3.58, follow the exposition from the seminal paper [16].

Exercises

3.1 ☕☕ (Thin shell) Let $X = (X_1, \ldots, X_n) \in \mathbb{R}^n$ be a random vector with independent, subgaussian coordinates X_i that satisfy $\mathbb{E}\,X_i^2 = 1$. Deduce from Theorem 3.1.1 that

$$\operatorname{Var}\left(\|X\|_2\right) \le CK^4.$$

3.2 ☕☕☕ (Thin shell, generalized) Let us relax the assumptions in Exercise 3.1 and prove the variance bound for any distribution with a bounded fourth moment. Let $X = (X_1, \dots, X_n) \in \mathbb{R}^n$ be a random vector with independent coordinates X_i that satisfy $\mathbb{E}\, X_i^2 = 1$ and $\mathbb{E}\, X_i^4 \le K^4$. Show that

$$\operatorname{Var}\big(\|X\|_2\big) \le K^4 \quad \text{and} \quad \sqrt{n} - \frac{K^4}{\sqrt{n}} \le \mathbb{E}\|X\|_2 \le \sqrt{n}.$$

3.3 ☕☕☕☕ (Thin shell, reversed) Let us show reverse bounds in Exercise 3.2. Let $X = (X_1, \dots, X_n) \in \mathbb{R}^n$ be a random vector with independent coordinates X_i that satisfy $\mathbb{E}\, X_i^2 = 1$, $\operatorname{Var}(X_i^2) > \alpha$ and $\mathbb{E}\, X_i^6 \le \beta$ for some $\alpha, \beta > 0$. Prove that if n is large enough (depending on α and β) then

$$\operatorname{Var}\big(\|X\|_2\big) \ge c\alpha \quad \text{and} \quad \mathbb{E}\|X\|_2 \le \sqrt{n} - \frac{c\alpha}{\sqrt{n}}.$$

Explain why a lower bound on the variance is an essential assumption.

3.4 ☕☕☕ (PCA maximizes the explained variance) Let X be a mean-zero random vector X in $\mathbb{R}^n$ whose covariance matrix has eigenvalues $\lambda_1 \ge \lambda_2 \ge \dots \ge \lambda_n$ and eigenvectors $v_1, \dots, v_n$.

(a) Prove that the orthogonal projection P_k onto the span of $v_1, \dots, v_k$ satisfies

$$\mathbb{E}\|P_k X\|_2^2 = \sum_{i=1}^{k} \operatorname{Var}\left(\langle X, v_i\rangle\right) = \sum_{i=1}^{k} \lambda_i.$$

(b) Prove that any rank-k orthogonal projection P in $\mathbb{R}^n$ satisfies

$$\mathbb{E}\|PX\|_2^2 \le \mathbb{E}\|P_k X\|_2^2.$$

Interpret these results as follows: PCA maximizes the explained variance of the data, with the explained fraction given by s_k/s_n, where $s_k = \lambda_1 + \ldots + \lambda_k$.

3.5 ☕☕☕ (Expected ℓ^p norm of a random vector) Let $n \in \mathbb{N}$ and $p \in [1, \infty]$. Consider a random vector $X = (X_1, \dots, X_n)$ whose coordinates are independent subgaussian random variables. Show that

$$\mathbb{E}\|X\|_p \le \begin{cases} CK\sqrt{p}\, n^{1/p}, & p \le \log n \\ CK\sqrt{\log n}, & p \ge \log n \end{cases}$$

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

3.6 ☕☕☕☕ (Expected ℓ^p norm of a Gaussian vector) Show that the bounds in Exercise 3.5 can be reversed for the standard gaussian random vector. Specifically, let $n \in \mathbb{N}$ and $p \in [1, \infty]$. Consider a random vector $X = (X_1, \dots, X_n)$ whose coordinates are independent $N(0,1)$ random variables. Show that

$$\mathbb{E}\|X\|_p \ge \begin{cases} c\sqrt{p}\, n^{1/p}, & p \le \log n \\ c\sqrt{\log n}, & p \ge \log n \end{cases}$$

where $c > 0$ is an absolute constant.

3.7 ☕☕ (Small ball probability) Let $X = (X_1, \ldots, X_n) \in \mathbb{R}^n$ be a random with independent coordinates with continuous distributions, whose densities are all bounded by K.

(a) Let us show that X is very unlikely to land in any given ball of radius $o(\sqrt{n})$. Specifically, for any fixed $a \in \mathbb{R}^n$ and $\varepsilon > 0$, prove that

$$\mathbb{P}\{\|X - a\|_2 \leq \varepsilon\sqrt{n}\} \leq \left(\sqrt{2\pi e}K\varepsilon\right)^n.$$

We will see in Exercises 4.28–4.29 that this is asymptotically sharp for small ε.

(b) Deduce for $n \geq 2$ that

$$\mathbb{E}\left[\frac{1}{\|X\|_2}\right] \leq \frac{CK}{\sqrt{n}}.$$

Explain why for $n = 1$ the expected value can be infinite.

3.8 ☕☕ (Expectation of an isotropic random vector)

(a) Prove that any isotropic random vector X in $\mathbb{R}^n$ satisfies $\|\mathbb{E}\, X\|_2 \leq 1$.

(b) Demonstrate that this bound is optimal in general: for each $n \in \mathbb{N}$, find an isotropic random vector X in $\mathbb{R}^n$ that satisfies $\|\mathbb{E}\, X\|_2 = 1$.

3.9 ☕☕ (1D marginals of an isotropic distribution) Let X be an isotropic random vector in $\mathbb{R}^n$. For any vector $u \in \mathbb{R}^n$, consider the random variable $X_u := \langle X, u\rangle$. Check that

$$\mathbb{E}\, X_u X_v = \langle u, v\rangle \quad \text{and} \quad \|X_u - X_v\|_{L^2} = \|u - v\|_2.$$

3.10 ☕☕☕ (The standard score of a random vector)

(a) Let $\mu \in \mathbb{R}^n$ be a fixed vector, Σ be a fixed $n \times n$ symmetric positive semidefinite matrix, and Z be a mean-zero, isotropic random vector in $\mathbb{R}^n$. Check that the random vector $X = \mu + \Sigma^{1/2}Z$ has mean μ and covariance matrix Σ.

(b) Conversely, let X be a random vector with mean μ and covariance matrix Σ. Find a mean-zero, isotropic random vector Z that satisfies $X = \mu + \Sigma^{1/2}Z$ almost surely.

3.11 ☕☕☕ (A randomly permuted vector) Let X be a random vector obtained by randomly permuting the coordinates of a given vector $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ with $x_1 + \cdots + x_n = 0$ and $x_1^2 + \cdots + x_n^2 = 1$. Specifically, let σ be a random permutation of $\{1, \ldots, n\}$ chosen uniformly from the set of all $n!$ permutations, and set $X_i = x_{\sigma(i)}$ for each $i = 1, \ldots, n$.

(a) Compute the mean and covariance matrix of X.

(b) You will find that the coordinates of X are negatively correlated. Can you heuristically explain why?

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

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

3.13 ☕☕ (Maximum norm of random vectors) Let's prove is a high-dimensional version of (2.22).

(a) Let $X_1, \dots, X_N \in \mathbb{R}^n$ be random vectors, not necessarily independent. Assume that the coordinates X_{ij} of each random vector X_i are independent, subgaussian, and satisfy $\mathbb{E}\, X_{ij}^2 = 1$. Prove that

$$\mathbb{E} \max_{i \le N} \|X_i\|_2 \le \sqrt{n} + CK^2 \sqrt{\log N}$$

where $K = \max_{ij} \left\|X_{ij}\right\|_{\psi_2}$.

(b) Show that this bound is tight for i.i.d. standard normal random vectors. Specifically, if $X_1, \dots, X_N \sim N(0, I_n)$ are independent, prove that

$$\mathbb{E} \max_{i \le N} \|X_i\|_2 \ge c \left(\sqrt{n} + \sqrt{\log N} \right).$$

3.14 ☕ (1D marginals of a normal distribution) Show that if $X \sim N(\mu, \Sigma)$, then

$$\langle X, v \rangle \sim N\big(\langle \mu, v \rangle, v^\mathsf{T} \Sigma v\big)$$

for any fixed vector $v \in \mathbb{R}^n$.

3.15 ☕☕ (General normal density) Prove Proposition 3.3.6 by doing the following steps.

(a) Recall the definition of the density of a random vector X: it is a function f_X satisfying $\mathbb{P}\{X \in B\} = \int_B f_X(x)\, dx$ for any (Borel) subset $B \subset \mathbb{R}^n$.

(b) Argue that we can assume that $\mu = 0$ and $X = \Sigma^{1/2} Z$ where $Z \sim N(0, I_n)$.

(c) Compute $\mathbb{P}\{X \in B\} = \mathbb{P}\{Z \in \Sigma^{-1/2}(B)\}$ by the change of variable $x = \Sigma^{1/2} z$ to express the result as an integral over B, and use the formula for the standard normal density (3.11). By part (a), you can read off the density of X from the integrand.

3.16 ☕☕☕ (A characterization of a joint normal distribution)

(a) Show that a random vector X in $\mathbb{R}^n$ is normally distributed if and only if all one-dimensional marginals $\langle X, u \rangle$, $u \in \mathbb{R}^n$, are normally distributed.

(b) Deduce that random variables $X_1, \dots, X_n$ are jointly normal if and only if all linear combination of them $a_1 X_1 + \cdots + a_n X_n$ with fixed coefficients a_i are normally distributed.

3.17 ☕ (Normal, uncorrelated, but dependent) Find normal random variables X and Y that are uncorrelated but not independent.

3.18 ☕☕ (Ginibre random matrices) A basic model for random matrices is the *Ginibre ensemble*. A Ginibre random matrix is an $n \times n$ matrix G with independent $N(0, 1)$ entries. Prove that the Ginibre distribution is invariant under multiplication by orthogonal matrices, that is, for any fixed $n \times n$ orthogonal matrix U, both UG and GU have the same distribution as G.

3.19 ☕☕ (GOE random matrices) A basic model for symmetric random matrices is the *Gaussian orthogonal ensemble* (GOE). A GOE random matrix is an $n \times n$ symmetric matrix A with independent normal entries on and above the diagonal, where the diagonal entries are $N(0, 2)$ and the off-diagonal entries are $N(0, 1)$.

(a) Check that A has the same distribution as $(G+G^{\mathsf{T}})/\sqrt{2}$ where G is a Ginibre random matrix from Exercise 3.18. This explains the different scaling of the diagonal and off-diagonal entries of A in the definition.

(b) Prove that the GOE distribution is invariant under orthogonal conjugation, that is, for any fixed $n \times n$ orthogonal matrix U, the matrix UAU^{T} has the same distribution as A. This explains the name "Gaussian orthogonal(-invariant) ensemble".

3.20 ☕☕ (A Gaussian random matrix makes orthogonal vectors independent) Let G be an $m \times n$ random matrix with independent $N(0,1)$ entries. Let $u, v \in \mathbb{R}^n$ be fixed unit orthogonal vectors. Prove that Gu and Gv are independent $N(0, I_m)$ random vectors.

3.21 ☕ Let X and Y be independent $N(0, I_n)$ random vectors. Prove that $\frac{X+Y}{\sqrt{2}}$ and $\frac{X-Y}{\sqrt{2}}$ are independent $N(0, I_n)$ random vectors.

3.22 ☕ (Length and direction of a normal random vector) Prove that the length $\|g\|_2$ and the direction $g/\|g\|_2$ of a standard normal random vector $g \sim N(0, I_n)$ are independent.

3.23 ☕☕☕ (Many random points are in convex position) A finite set of points in $\mathbb{R}^n$ is in *convex position* if no point lies in the convex hull[13] of the others (see Figure 3.11). Let $N \le e^{cn}$. Following the steps below, prove that with probability at least $1 - e^{-cn}$, independent Gaussian random vectors $g_1, \ldots, g_N \sim N(0, I_n)$ are in convex position.[14] Surprisingly, this shows that exponentially many random points are in convex position!

(a) Show that, with high probability, g_1 is linearly separated[15] from the other points, i.e. there exists $x \in \mathbb{R}^n$ (possibly random) satisfying $\langle g_1, x \rangle > \max_{j=2,\ldots,N} \langle g_j, x \rangle$.

(b) Deduce that, with high probability, g_1 is not in the convex hull of the other points.

(c) Conclude by the union bound.

Figure 3.11 The five points on the left are in convex position, while the five on the right are not.

3.24 ☕☕ (Polar decomposition of the uniform distribution on a ball) Let X be uniformly distributed on the unit Euclidean ball centered at the origin in $\mathbb{R}^n$. Decompose X as

$$X = rZ$$

[13] We introduced the convex hull in the Appetizer, see (0.1).

[14] Here, as usual, $c > 0$ is an absolute constant of your choice.

[15] A picture in $\mathbb{R}^2$ might help visualize this condition, which means there exists a line separating g_1 from the other points. In higher dimensions, a line becomes a hyperplane.

where $Z \sim \text{Unif}(S^{n-1})$ and r is a random variable in $[0,1]$ with density nx^{n-1}, and r and Z are independent.

3.25 ☕ (A ball is isotropic) Let Y be a random vector uniformly distributed on the Euclidean ball in $\mathbb{R}^n$ with center at the origin and radius $\sqrt{n+2}$. Check that Y is isotropic.

3.26 ☕☕☕ (Delocalization of random vector on the sphere) The most "spread out" or "delocalized" unit vector x in $\mathbb{R}^n$ has all coordinates with the same magnitude, i.e. $|x_i| = 1/\sqrt{n}$ for all i. It turns out that a random vector is almost delocalized. Show that a random vector X uniformly distributed on the unit Euclidean sphere in $\mathbb{R}^n$ satisfies

$$\mathbb{E}\|X\|_\infty \asymp \sqrt{\frac{\log n}{n}}.$$

As always, "$\asymp$" means equivalence up to absolute constant factors.[16]

3.27 ☕☕☕ (1D marginals of the sphere) Let X and Y be random vectors uniformly distributed on the Euclidean unit sphere and the Euclidean unit ball in $\mathbb{R}^n$.

(a) Prove that the density of any 1D marginal of Y is proportional to $(1-x^2)^{\frac{n-1}{2}}$ for $x \in [-1,1]$.

(b) Prove that the density of any 1D marginal of X is proportional to $(1-x^2)^{\frac{n-3}{2}}$ for $x \in [-1,1]$. Surprisingly, the 1D marginals of a 3D sphere are uniformly distributed!

3.28 ☕☕☕ (Large cube probability) Let us establish a surprising fact: a standard normal random vector in $\mathbb{R}^n$ is likely to be very close to a cube with constant side length, yet it is highly unlikely to land in a cube even a hundred times larger!

(a) Let π_a denote a *metric projection* onto the cube $[-a,a]^n$, which returns a point $y = \pi_a(x)$ in the cube closest to the input point x in the Euclidean distance. Show that the metric projection can be obtained by *soft thresholding* the coordinates of x as follows: $y_i = x_i$ if $|x_i| \le a$ and $y_i = a\,\text{sign}(x_i)$ otherwise, for each $i = 1,\dots,n$.

(b) Show that there exists an absolute constants n_0 and a such that for every $n > n_0$, a random vector $g \sim N(0, I_n)$ satisfies the following with probability at least 0.99:

$$\text{dist}\left(g, [-a,a]^n\right) < 0.01\|g\|_2 \quad \text{and} \quad g \notin [-100a, 100a]^n.$$

Here $\text{dist}(x,T) = \inf_{y\in T}\|x-y\|_2$ denotes the distance from a point x to a set T.

3.29 ☕ (Frames and orthogonal matrices) Show that vectors $u_1,\dots,u_N \in \mathbb{R}^n$ form a Parseval frame if and only if the $n \times N$ matrix with columns u_i has orthonormal rows.

3.30 ☕☕ (The Mercedes-Benz frame) Consider N equispaced points on a circle of radius $\sqrt{2/N}$ (see Figure 3.7 for $N = 3$). Show that they form a Parseval frame in $\mathbb{R}^2$.

3.31 ☕ (Turning any isotropic distribution into a frame) Consider an isotropic random vector

[16] Recall that $a \asymp b$ means $c_1 b \le a \le c_2 b$ for some absolute constants $c_1, c_2 > 0$.

X in $\mathbb{R}^n$ that takes values $x_1, \ldots, x_N$ with probabilities $p_1, \ldots, p_N$. Check that the vectors $\sqrt{p_1}x_1, \ldots, \sqrt{p_N}x_N$ form a Parseval frame in $\mathbb{R}^n$.

3.32 ☕ (Subgaussian norm of 1D marginals) Let X be a subgaussian random vector in $\mathbb{R}^n$. Check that

$$\|\langle X, v\rangle\|_{\psi_2} \le \|X\|_{\psi_2}\|v\|_2 \quad \text{for any vector } v \in \mathbb{R}^n.$$

3.33 ☕☕ (Subgaussian rotation invariance) Proposition 3.3.1 showed that the standard normal distribution is rotation invariant. Subgaussian distributions are not necessarily rotation invariant, but the subgaussian norm is. For any orthogonal $n \times n$ matrix U and subgaussian vector $X \in \mathbb{R}^n$, show that UX is also subgaussian, and

$$\|UX\|_{\psi_2} = \|X\|_{\psi_2}.$$

3.34 ☕☕ (Subgaussian quadratic forms) Let A be an $m \times n$ random matrix with independent, mean-zero, subgaussian rows A_i. For any fixed $u \in \mathbb{R}^m$ and $v \in \mathbb{R}^n$, show that $u^\mathsf{T} Av$ is a subgaussian random variable, and

$$\|u^\mathsf{T} Av\|_{\psi_2} \le C\|u\|_2\|v\|_2 \max_i \|A_i\|_{\psi_2}.$$

Explain why the result still holds if the same assumptions are made about the *columns* of A instead of the rows.

3.35 ☕☕ (Subgaussian norm of a sum) Proposition 2.7.1 extends to higher dimensions. Let $X_1, \ldots, X_N$ be independent, zero-mean subgaussian vectors in $\mathbb{R}^n$. Show that

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

where C is an absolute constant.

3.36 ☕☕ (Vectors with subgaussian coordinates) What happens if we drop the independence assumption in Lemma 3.4.2? X stays subgaussian, but the bound gets way worse:

(a) Let $X = (X_1, \ldots, X_n)$ be a random vector with subgaussian (not necessarily independent) coordinates X_i. Show that X is subgaussian, and

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

(b) Show by example that both bounds in part (a) are optimal in all dimensions n.

3.37 ☕☕ (The norm of subgaussian vectors needs not concentrate) It is tempting to conjecture that the norm concentration (Theorem 3.1.1) might apply more generally to any isotropic subgaussian random vector X with $K = \|X\|_{\psi_2}$, so that coordinate independence would not be necessary. Show with an example that this is false.[17]

[17] However, we will prove an upper bound for the norm in Proposition 6.2.1.

3.38 ☕☕ (A normal random vector is subgaussian) In Exercise 2.24(c), we showed that the subgaussian norm of a normal random variable $N(0, \sigma^2)$ equals $\sqrt{8/3}\,\sigma$. Extend this result to higher dimensions: show that a random vector $X \sim N(0, \Sigma)$ satisfies

$$\|X\|_{\psi_2} = \sqrt{\frac{8}{3}}\,\|\Sigma\|^{1/2}$$

where $\|\Sigma\|$ denotes the largest eigenvalue[18] of Σ.

3.39 ☕ (The Euclidean norm of a subgaussian random vector)

(a) Let $X_1, \ldots, X_n$ be subgaussian random variables, not necessarily independent. Prove that

$$\Big\| \Big(\sum_{i=1}^n X_i^2 \Big)^{1/2} \Big\|_{\psi_2} \le \Big(\sum_{i=1}^n \|X_i\|_{\psi_2}^2 \Big)^{1/2}.$$

(b) Let X be a subgaussian random vector in $\mathbb{R}^n$. Show that $\|X\|_2$ is a subgaussian random variable, and

$$\Big\| \|X\|_2 \Big\|_{\psi_2} \le \sqrt{n} \|X\|_{\psi_2}.$$

Explain why $\sqrt{n}$ is the best scaling factor in general. In Proposition 6.2.1, we will prove a more informative bound.

3.40 ☕☕☕☕ (Random vectors are almost orthogonal – sometimes.) In (3.14), we saw that two independent random vectors from the uniform distribution on the sphere S^{n-1} are likely almost orthogonal. Let's generalize this observation.

(a) Let X be a random vector in $\mathbb{R}^n$ with independent, zero-mean, unit-variance coordinates, all with subgaussian norms bounded by K. Define $\theta = X/\|X\|_2$ if $X \ne 0$, and $\theta = 0$ otherwise. Let η be an independent copy of θ. Prove that

$$|\langle \theta, \eta \rangle| \le \frac{C(K)}{\sqrt{n}} \quad \text{with probability at least } 0.99,$$

where $C(K)$ may depend only on K.

(b) Show that the independence of coordinates is crucial in part (a). Find independent, identically distributed, isotropic random vectors X and Y in $\mathbb{R}^n$, whose subgaussian norms bounded by an absolute constant, and which satisfy $X = Y$ with probability at least 0.99.

3.41 ☕☕ (Exponentially many almost orthogonal vectors!) From linear algebra, we know there cannot be more than n orthogonal vectors in $\mathbb{R}^n$. But surprisingly, there are exponentially many that are almost orthogonal! For any integer $N \le e^{cn}$, follow the steps below to find points $X_1, \ldots, X_N \in S^{n-1}$ that satisfy

$$\big| \langle X_i, X_j \rangle \big| \le 0.01 \quad \text{for all distinct } i, j. \tag{3.39}$$

(a) Argue that independent random vectors $X, Y \sim \mathrm{Unif}(S^{n-1})$ satisfy $|\langle X, Y \rangle| \le 0.01$ with probability at least $1 - 4\exp(-c_0 n)$, where c_0 is an absolute constant.

[18] $\|\Sigma\|$ is known as the *operator norm* of Σ. We will formally introduce this notion in Definition 4.1.8, and it will play a prominent role in many results later.

(b) Take a union bound over all pairs (i, j), $i \neq j$, and conclude that the event in (3.39) holds with positive probability.

3.42 ☕ (The uniform distribution on a ball is subgaussian) Let Y be a random vector uniformly distributed in the Euclidean unit ball centered at the origin in $\mathbb{R}^n$. Argue that Y satisfies the conclusion of Theorem 3.4.5.

3.43 ☕☕ (Subgaussian norm of the coordinate distribution) Let X be a random vector uniformly distributed on the standard vector basis of $\mathbb{R}^n$ for $n \geq 2$. Prove that

$$\|X\|_{\psi_2} \asymp \frac{1}{\sqrt{\log n}}$$

As always, "$\asymp$" means equivalence up to absolute constant factors.[19]

3.44 ☕☕☕ (The ℓ^1 ball is poorly subgaussian) Consider a ball in the ℓ^1 norm[20] in $\mathbb{R}^n$:

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

Let X be a random vector uniformly distributed in K.

(a) Compute the density of X_1, the first coordinate of X.
(b) Show that the random vector X is isotropic for some $r \asymp n$.
(c) For this value of r, show that $\|X\|_{\psi_2} > c\sqrt{n}$ where $c > 0$ is an absolute constant.

3.45 ☕ (The probability mass function of a subgaussian random vector) Let X be a subgaussian random vector in $\mathbb{R}^n$. Prove that

$$\mathbb{P}\{X = x\} \leq 2\exp\left(-\frac{\|x\|_2^2}{K^2}\right) \quad \text{for any } x \in \mathbb{R}^n,$$

where $K = \|X\|_{\psi_2}$.

3.46 ☕☕☕ (Isotropic subgaussian distributions have high entropy) The entropy of a discrete random vector X that takes values x_i with probabilities p_i is defined as

$$H(X) = -\sum_i p_i \ln p_i.$$

Heuristically, $H(X)$ is proportional to the expected number of bits needed to encode[21] X.

(a) (Very discrete distributions have low entropy) For any random vector X that takes N values, prove that $H(X) \leq \ln N$.
(b) (Isotropic subgaussian distributions have high entropy) For any discrete, isotropic, subgaussian random vector X in $\mathbb{R}^n$, prove that

$$H(X) \geq \frac{n}{K^2} - \ln 2, \quad \text{where} \quad K = \|X\|_{\psi_2}.$$

(c) (Isotropic subgaussian distributions have huge support) Conclude that any random vector as in (b) must take at least $\frac{1}{2}\exp(K^{-2}n)$ different values.

[19] Recall that $a \asymp b$ means $c_1 b \leq a \leq c_2 b$ for some absolute constants $c_1, c_2 > 0$.
[20] For a refresher on the unit ℓ^1 ball, check Section 1.2; it is the cross-polytope shown in Figure 1.2.
[21] This is formally established by the Shannon source coding theorem. Look it up if interested!

3.47 ☕☕ (Rewriting the assumption of Grothendieck inequality) Let A be an $m \times n$ matrix. Prove that the following properties are equivalent:

(a) $x^{\mathsf{T}} A y \le 1$ for any vectors $x \in \{-1,1\}^n$ and $y \in \{-1,1\}^m$.
(b) $x^{\mathsf{T}} A y \le 1$ for any vectors $x \in [-1,1]^n$ and $y \in [-1,1]^m$.
(c) $x^{\mathsf{T}} A y \le \|x\|_\infty \|y\|_\infty$ for any vectors $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$.
(d) Any of the three properties above but with $|x^{\mathsf{T}} A y|$ replacing $x^{\mathsf{T}} A y$.

3.48 ☕ (Tightening the probabilistic proof of Grothendieck inequality) Tighten the proof of Theorem 3.5.1 by skipping the rough $4/R^2$ bound in (3.25) and optimizing R numerically at the end. Conclude that $K \le 14.1$.

3.49 ☕☕☕ (Quadratic Grothendieck for PSD matrices) In this and next exercise, we relax the assumption of Grothendieck inequality by assuming $x_i = y_i$, as announced in Remark 3.5.3.

(a) (Polarization identity) Show that for any $n \times n$ symmetric matrix A and vectors $x, y \in \mathbb{R}^n$,we have[22]

$$x^{\mathsf{T}} A y = \bar{x}^{\mathsf{T}} A \bar{x} - \bar{y}^{\mathsf{T}} A \bar{y} \quad \text{where} \quad \bar{x} = \frac{x+y}{2}, \quad \bar{y} = \frac{x-y}{2}.$$

(b) (Quadratic Grothendieck) Let $A = (a_{ij})$ be an $n \times n$ symmetric positive-semidefinite matrix. Assume that $\sum_{i,j} a_{ij} x_i x_j \le 1$ for any numbers $x_i \in \{-1,1\}$. Prove that for any Hilbert space H, we have $|\sum_{i,j} a_{ij} \langle u_i, v_j \rangle| \le 2K$ for any unit vectors $u_i, v_j \in H$, where K is the absolute constant from Grothendieck inequality.
(c) Show the result in part (b) may fail without the positive-semidefinite assumption.

3.50 ☕☕ (Quadratic Grothendieck for diagonal-free matrices)

(a) (Separate convexity) A function $f(x_1, \ldots, x_n)$ is called *separately convex* if it is convex in each variable x_i. (Such function might not be convex – example?) Show that a separately convex function $f : [-1,1]^n \to \mathbb{R}$ attains its maximum at some extremal point of the cube, i.e. at a point in $\{-1,1\}^n$.
(b) (Quadratic Grothendieck) Let $A = (a_{ij})$ be an $n \times n$ diagonal-free matrix. Assume that $|\sum_{i,j} a_{ij} x_i x_j| \le 1$ for any numbers $x_i \in \{-1,1\}$. Prove that for any Hilbert space H, we have $|\sum_{i,j} a_{ij} \langle u_i, v_j \rangle| \le 2K$ for any unit vectors $u_i, v_j \in H$, where K is the absolute constant from Grothendieck inequality.

3.51 ☕ (Gram matrices) The *Gram matrix* of vectors $v_1, \ldots, v_n$ is the $n \times n$ matrix $(\langle v_i, v_j \rangle)$. Prove the following.

(a) The Gram matrix of any set of vectors is symmetric and positive-semidefinite.
(b) Conversely, any symmetric and positive-semidefinite matrix is a Gram matrix of some set of vectors.

[22] Note that if $A = I_n$, the polarization identity just becomes the familiar parallelogram law.

3.52 ☕☕ (SDP relaxation of a bilinear integer program) Consider the following integer optimization problem:

$$\text{maximize} \sum_{i,j=1}^{n} A_{ij} x_i y_j : \quad x_i, y_j = \pm 1 \text{ for } i, j \tag{3.40}$$

where A is an $m \times n$ matrix. Just like in (3.29), let us relax it to the following problem:

$$\text{maximize} \sum_{i,j} A_{ij} \langle X_i, Y_j \rangle : \quad \|X_i\|_2 = \|Y_j\|_2 = 1 \text{ for all } i, j, \tag{3.41}$$

where the maximum is over all unit vectors X_i, Y_j in $\mathbb{R}^k$ with some fixed dimension $k \in \mathbb{N}$.

(a) Express (3.41) as a semidefinite program.

(b) Show that the relaxation approximates the original problem (3.40) within an absolute constant factor.

3.53 ☕☕ (Grothendieck identity) Prove Grothendieck identity (Lemma 3.6.5).

3.54 ☕☕☕ (Approximating a max cut: deterministic outcome, random runtime) Goemans-Williamson Theorem 3.6.4 finds a cut that is large only *in expectation*. For any $\varepsilon > 0$, give an $(0.878 - \varepsilon)$-approximation algorithm for the maximum cut that *always* finds a valid cut but may have random running time. Bound the expected runtime.

3.55 ☕ (Kernel trick for real analytic functions) Let's prove Lemma 3.7.7. For that, we need an infinite-dimensional version of $\mathbb{R}^n$. A perfect choice is ℓ^2, the space of all square-summable sequences $x = (x_1, x_2, \ldots)$. It comes with the same norm and inner product as $\mathbb{R}^n$, but with finite sums replaced by (convergent) series:

$$\|x\|_2 = \Big(\sum_{i=1}^{\infty} x_i^2 \Big)^{1/2} \quad \text{and} \quad \langle x, y \rangle = \sum_{i=1}^{\infty} x_i y_i.$$

Now define $\Phi, \Psi : \mathbb{R}^n \to \ell^2$ that make Lemma 3.7.7 work.

3.56 ☕☕☕ (Linearizing Grothendieck identity) Let's get rid of the nonlinearity "arcsin" in Grothendieck identity (Lemma 3.6.5) by adding a correction term. Prove the following for a random vector $g \sim N(0, I_n)$ and any fixed unit vectors $u, v \in \mathbb{R}^{n-1}$:

(a) $\mathbb{E} \langle g, u \rangle \operatorname{sign} \langle g, v \rangle = \sqrt{\frac{2}{\pi}} \langle u, v \rangle$.

(b) $\mathbb{E} \operatorname{sign} \langle g, u \rangle \operatorname{sign} \langle g, v \rangle = \frac{2}{\pi} \langle u, v \rangle + \mathbb{E} [Z_u Z_v]$, where $Z_w = \sqrt{\frac{2}{\pi}} \langle g, w \rangle - \operatorname{sign} \langle g, w \rangle$.

3.57 ☕☕ (A sharper PSD Grothendieck) For positive semidefinite matrices, Grothendieck inequality (Theorem 3.5.1) holds with a better constant $\pi/2 \approx 1.571$, with a simpler proof. Let's discover it. Let $A = (a_{ij})$ be a $n \times n$ symmetric positive-semidefinite matrix. Suppose $\sum_{i,j} a_{ij} x_i x_j \leq 1$ for any numbers $x_i \in \{-1, 1\}$. Use Exercise 3.56 to show that for any Hilbert space H, we have $\sum_{i,j} a_{ij} \langle u_i, u_j \rangle \leq \pi/2$ for any unit vectors $u_i \in H$.

3.58 ☕☕ (SDP relaxation using Grothendieck inequality) Let's use the solution of Exercise 3.57 to improve Theorem 3.5.7. Let A be an $n \times n$ symmetric, positive-semidefinite matrix. Let $\operatorname{int}(A)$ denote the maximum in the integer optimization problem (3.28) and $\operatorname{sdp}(A)$ denote the maximum in the semidefinite problem (3.29).

(a) Show that $\operatorname{int}(A) \le \operatorname{sdp}(A) \le \frac{\pi}{2} \cdot \operatorname{int}(A)$.

(b) Design a randomized algorithm to turn a solution (X_i) of (3.29) into labels $x_i = \pm 1$ that approximately solve (3.28) in expectation.

4

Random Matrices

We begin to study the non-asymptotic theory of random matrices, with more to come in later chapters. We start with matrix refresher in Section 4.1 – skip what you know (like SVD) and focus on new topics like perturbation theory.

Section 4.2 introduces key geometric concepts – nets, covering and packing numbers – and links them to volume and error correction codes (Section 4.3). In Sections 4.4 and 4.6, we develop the ε-*net argument* for the analysis of random matrices, first proving bounds on the operator norm (Theorem 4.4.3) and then giving a stronger, two-sided bound on all singular values (Theorem 4.6.1).

We explore three applications: community detection in networks (Section 4.5), covariance estimation (Section 4.7) and spectral clustering for point sets (Section 4.7.1).

Don't miss the end-of-chapter exercises – you will explore the power method to compute the top singular value (Exercise 4.6), give Schur bound on the operator norm (Exercise 4.8), introduce Hermitian dilation (Exercise 4.14), construct Walsh matrices (Exercise 4.9), develop matrix perturbation theory (Exercises 4.13, 4.15, 4.16), compute more general norms for for random matrices (Exercise 4.18–4.20, 4.44), estimate the cut norm with an SDP relaxation (Exercise 4.21, 4.22), compute the volume of high-dimensional balls in three ways (Exercises 4.27–4.30), find covering numbers for low-rank matrices (Exercise 4.50), analyze Gaussian mixture models (Exercise 4.51), and more!

4.1 A quick refresher on linear algebra

Let's review some basic material about matrices. You have likely covered much this in linear algebra – so skip what you already know – but later parts might be new to you.

4.1.1 Singular value decomposition

Spectral decomposition (3.7) is a great tool, but it only works for symmetric matrices. *Singular value decomposition* (SVD) extends this idea to all matrices:

[0] This material will be published by Cambridge University Press as *High-Dimensional Probability, 2nd Edition* by Roman Vershynin. This pre-publication version is free to view and download for personal use only. Not for re-distribution, re-sale or use in derivative works. ©Roman Vershynin 202X.
If you want to be notified when the printed version is available, let me know: rvershyn@uci.edu

Theorem 4.1.1 (Singular value decomposition)**.** *Any $m \times n$ matrix A with real entries can be written as*

$$A = \sum_{i=1}^{r} s_i u_i v_i^\mathsf{T} \quad \textit{where } r = \min(m, n). \tag{4.1}$$

Here s_i are nonnegative numbers called the singular values *of A, $u_i \in \mathbb{R}^m$ are orthonormal vectors called the* left singular vectors *of A, and $v_i \in \mathbb{R}^n$ are orthonormal vectors called the* right singular vectors *of A.*

Proof Without loss of generality, we can assume that $m \geq n$. (Why?) Since $A^\mathsf{T} A$ is an $n \times n$ symmetric positive-semidefinite matrix, the spectral theorem tells us that it has real, nonnegative eigenvalues $s_1^2, \ldots, s_n^2$ and orthonormal eigenvectors $v_1, \ldots, v_n \in \mathbb{R}^n$, so that $A^\mathsf{T} A v_i = s_i^2 v_i$. The vectors Av_i are orthogonal:

$$\langle Av_i, Av_j \rangle = \langle A^\mathsf{T} A v_i, v_j \rangle = s_i^2 \langle v_i, v_j \rangle = s_i^2 \delta_{ij}, \quad i = 1, \ldots, n. \tag{4.2}$$

Therefore, there exist orthonormal vectors $u_1, \ldots, u_n \in \mathbb{R}^m$ such that

$$Av_i = s_i u_i, \quad i = 1, \ldots, n. \tag{4.3}$$

(For all i with $s_i \neq 0$, (4.3) uniquely defines the vectors u_i and (4.2) ensures they are orthonormal. If $s_i = 0$, then (4.2) (for $j = i$) gives $Av_i = 0$, so (4.3) holds trivially. In this case, we can pick any u_i while keeping orthonormality.)

Since $v_1, \ldots, v_n$ form an orthonormal basis of $\mathbb{R}^n$, we can write $I_n = \sum_{i=1}^n v_i v_i^\mathsf{T}$. Multiplying by A on the left and using (4.3) gives the desired SVD:

$$A = \sum_{i=1}^{n} (Av_i) v_i^\mathsf{T} = \sum_{i=1}^{n} s_i u_i v_j^\mathsf{T}.$$

Ask yourself: at what point in this argument did we need $m \geq n$? □

Remark 4.1.2 (What does a matrix do?)**.** From (4.3) we see how SVD gives a geometric view of matrices: A first "stretches" each orthogonal direction v_i by s_i and then "rotates" the space, mapping the orthonormal basis (v_i) into (u_i).

Remark 4.1.3 (SVD in matrix form)**.** For convenience, we often set $s_i = 0$ for $i > r$ and arrange s_i in (weakly) decreasing order. We can also extend (u_i) and (v_i) to orthonormal bases in $\mathbb{R}^m$ and $\mathbb{R}^n$, letting us rewrite the SVD (4.1) as

$$A = U \Sigma V^\mathsf{T} \tag{4.4}$$

where U is the $m \times m$ orthogonal matrix with left singular vectors u_i as columns, V is the $n \times n$ orthogonal matrix with right singular vectors v_i as columns, and Σ is the $m \times n$ diagonal matrix with singular values s_i on the diagonal. We can similarly express the spectral decomposition of a *symmetric* $n \times n$ matrix A:

$$A = \sum_{i=1}^{n} \lambda_i u_i u_i^\mathsf{T} = U \Lambda U^\mathsf{T}$$

where U is the $n \times n$ orthogonal matrix with eigenvectors u_i as columns and Λ is the $n \times n$ diagonal matrix with eigenvalues λ_i on the diagonal.

Remark 4.1.4 (Spectral decomposition vs. SVD)**.** The spectral and singular value decompositions are tightly connected. Since

$$AA^{\mathsf{T}} = \sum_{i=1}^{r} s_i^2 u_i u_i^{\mathsf{T}} \quad \text{and} \quad A^{\mathsf{T}} A = \sum_{i=1}^{r} s_i^2 v_i v_i^{\mathsf{T}}$$

(check!), the left singular vectors u_i of A are eigenvectors of AA^{T}, the right singular vectors v_i of A are eigenvectors of $A^{\mathsf{T}}A$, and the singular values s_i of A are the square roots of the eigenvalues λ_i of both AA^{T} and $A^{\mathsf{T}}A$:

$$s_i(A) = \sqrt{\lambda_i(AA^{\mathsf{T}})} = \sqrt{\lambda_i(A^{\mathsf{T}}A)}. \tag{4.5}$$

Example 4.1.5 (Orthogonal projections)**.** Consider the orthogonal projection P in $\mathbb{R}^n$ onto a k-dimensional subspace E. The projection of a vector x onto E is given by $Px = \sum_{i=1}^{k} \langle u_i, x\rangle u_i$ where $u_1, \ldots, u_k$ is an orthonormal basis of E. We can rewrite this as a spectral decomposition of P:

$$P = \sum_{i=1}^{k} u_i u_i^{\mathsf{T}} = UU^{\mathsf{T}},$$

where U is the $n \times k$ matrix with orthonormal columns u_i. In particular, P is a symmetric matrix with eigenvalues $\underbrace{1, \ldots, 1}_{k}, \underbrace{0, \ldots, 0}_{n-k}$.

4.1.2 Min-max theorem

We saw one optimization-based description of eigenvalues (Proposition 3.2.2). Here is another, very useful one:

Theorem 4.1.6 (Min-max theorem for eigenvalues)**.** *The k-th largest eigenvalue of an $n \times n$ symmetric matrix A can be written as*

$$\lambda_k(A) = \max_{\dim E = k} \min_{x \in S(E)} x^{\mathsf{T}} A x = \min_{\dim E = n-k+1} \max_{x \in S(E)} x^{\mathsf{T}} A x, \tag{4.6}$$

where the maximum (resp. minimum) is over all subspaces E of dimension k (resp. $n-k+1$) and $S(E)$ denotes the Euclidean unit sphere of E, i.e. the set of all unit vectors in E.

Proof Let us focus on the first equation. To prove the *upper bound* on $\lambda_k = \lambda_k(A)$, we need to find a k-dimensional subspace E such that

$$x^{\mathsf{T}} A x \ge \lambda_k \quad \text{for all } x \in S(E).$$

How do we find such E? Take the spectral decomposition $A = \sum_{i=1}^{n} \lambda_i u_i u_i^{\mathsf{T}}$ and pick the subspace where A is "largest": $E = \operatorname{span}(u_1, \ldots, u_k)$. The eigenvectors $u_1, \ldots, u_k$ form an orthonormal basis of E, so any vector $x \in S(E)$ can be written as $x = \sum_{i=1}^{k} a_i u_i$. Orthonormality of u_i and monotonicity of λ_i give

$$x^{\mathsf{T}} A x = \sum_{i=1}^{k} \lambda_i a_i^2 \ge \lambda_k \sum_{i=1}^{k} a_i^2 = \lambda_k,$$

and we are done with the upper bound. To prove the *lower bound* on λ_k, we need to find a vector $x \in S(E)$ in any k-dimensional subspace E such that $x^\mathsf{T} Ax \le \lambda_k$. How do we find such x? We look where A is "smallest" – inside the subspace $F = \operatorname{span}(u_k, \dots, u_n)$. Since $\dim(E) + \dim(F) = n + 1$, the intersection is nontrivial, so there is a unit vector $x \in E \cap F$. Writing $x = \sum_{i=k}^n a_i u_i$, we get

$$x^\mathsf{T} Ax = \sum_{i=k}^n \lambda_i a_i^2 \le \lambda_k \sum_{i=k}^n a_i^2 = \lambda_k.$$

So we've got the lower bound and thus the first equation in (4.6). The second equation follows if we apply the first for $-A$ instead of A, with eigenvalues reordered in the opposite way. (Check!) □

Applying Theorem 4.1.6 to $A^\mathsf{T} A$ and using (4.5), we immediately get:

Corollary 4.1.7 (Min-max theorem for singular values)**.** *Let A be an $m \times n$ matrix with singular values $s_1 \ge s_2 \ge \cdots \ge s_n \ge 0$. Then*

$$s_k(A) = \max_{\dim E = k} \min_{x \in S(E)} \|Ax\|_2 = \min_{\dim E = n-k+1} \max_{x \in S(E)} \|Ax\|_2,$$

where the maximum (resp. minimum) is over all subspaces E of dimension k (resp. $n - k + 1$) and $S(E)$ denotes the Euclidean unit sphere of E.

4.1.3 Frobenius and operator norms

The linear space of $m \times n$ matrices has several classic norms. The simplest is the *Frobenius norm*, or *Hilbert-Schmidt* norm, which is just the ℓ^2 norm on $\mathbb{R}^{m\times n}$:

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

In short, the Frobenius norm is just the Euclidean norm of the vectorized matrix. Likewise, the inner product for matrices is just the usual dot product on $\mathbb{R}^{m\times n}$:

$$\langle A, B \rangle = \sum_{i=1}^m \sum_{j=1}^n A_{ij} B_{ij} = \operatorname{tr}(A^\mathsf{T} B). \tag{4.7}$$

The inner product of a matrix with itself gives the norm squared:

$$\|A\|_F^2 = \langle A, A \rangle = \operatorname{tr}(A^\mathsf{T} A). \tag{4.8}$$

Another key matrix norm is the *operator norm* (or *spectral norm*), which measures how much the linear transformation A stretches vectors at most:

Definition 4.1.8 (Operator norm)**.** The *operator norm* of an $m \times n$ matrix A is the smallest number K such that

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

Equivalently,

$$\|A\| = \max_{x\neq 0} \frac{\|Ax\|_2}{\|x\|_2} = \max_{\|x\|_2\leq 1} \|Ax\|_2 = \max_{\|x\|_2=1} \|Ax\|_2 = \max_{\|x\|_2=\|y\|_2=1} |y^\mathsf{T} Ax|. \tag{4.9}$$

The first three equations follow by rescaling (check them!), and the last one follows because from the duality formula (1.6): $\|Ax\| = \max_{\|y\|_2=1}\langle Ax, y\rangle$. You can drop the absolute value in (4.9) or keep it; the result is the same (why?) In Exercise 4.2, you will check that $\|A\|$ is indeed a norm.

Remark 4.1.9 (Other operator norms)**.** We can replace the ℓ^2 norm in Definition 4.1.8 with other norms to get a more general concept of the operator norm. Discover it in Exercises 4.18–4.22.

4.1.4 The matrix norms and the spectrum

Lemma 4.1.10 (Orthogonal invariance)**.** *The Frobenius and spectral norms are orthogonal invariant, meaning that for any matrix A and orthogonal matrices Q and R (of proper dimensions) we have*

$$\|QAR\|_F = \|A\|_F \quad \textit{and} \quad \|QAR\| = \|A\|.$$

Proof The first part follows from (4.8) and the cyclic property of trace:

$$\begin{aligned}\|QAR\|_F^2 &= \operatorname{tr}(R^\mathsf{T} A^\mathsf{T} Q^\mathsf{T} QAR) = \operatorname{tr}(R^\mathsf{T} A^\mathsf{T} AR) = \operatorname{tr}(RR^\mathsf{T} A^\mathsf{T} A) \\ &= \operatorname{tr}(A^\mathsf{T} A) = \|A\|_F^2.\end{aligned}$$

For the second part, (4.9) says that $\|QAR\|$ is obtained by maximizing the bilinear form $y^\mathsf{T} QARx = (Qy)^\mathsf{T} A(Rx)$ over all unit vectors x and y. But since Q and R are orthogonal matrices, $v = Qy$ and $u = Rx$ range over all unit vectors, so we maximize $v^\mathsf{T} Au$ over all unit vectors and get $\|A\|$. □

Lemma 4.1.11 (Matrix norms in terms of singular values)**.** *For any $m\times n$ matrix A with singular values $s_1(A) \geq \ldots \geq s_n(A)$, we have*

$$\|A\|_F = \Big(\sum_{i=1}^n s_i(A)^2\Big)^{1/2} \quad \textit{and} \quad \|A\| = s_1(A).$$

Proof To prove the part about the Frobenius norm, let's use SVD (4.4) and orthogonal invariance (Lemma 4.1.10): $\|A\|_F = \|U\Sigma V^\mathsf{T}\|_F = \|\Sigma\|_F$; now recall that the only nonzero entries of Σ are $s_1, \ldots, s_n$. The part about the operator norm follows directly from the min-max theorem (Corollary 4.1.7) for $k = 1$. □

Remark 4.1.12 (Symmetric matrices)**.** For a symmetric matrix A with eigenvalues $\lambda_k(A)$, we have

$$\|A\| = \max_k |\lambda_k(A)| = \max_{\|x\|=1} |x^\mathsf{T} Ax|, \tag{4.10}$$

so we can take $x = y$ in the operator norm definition (4.9). The first equation in (4.10) comes from Lemma 4.1.11 since the singular values of A are $|\lambda_k(A)|$. The min-max theorem (Theorem 4.1.6) gives $|\lambda_k(A)| \le \max_{\|x\|=1} |x^\mathsf{T} Ax|$, proving the upper bound in (4.10). The lower bound is clear by taking $x = y$ in (4.9).

Take a moment to sharpen your skills with Frobenius and operator norms in Exercises 4.2–4.10.

4.1.5 Low-rank approximation

Suppose we want to approximate a given matrix A by a matrix B with a given rank k. How do we construct the best B, and what is the smallest approximation error measured in the operator norm? The answer is: get B by truncating the SVD of A; the approximation error is the $(k+1)$-th singular value of A:

Theorem 4.1.13 (Eckart-Young-Mirsky theorem). *Let A be an $m \times n$ matrix with SVD $A = \sum_{i=1}^n s_i u_i v_i^\mathsf{T}$. Then for any $1 \le k \le n$ we have:*

$$\min_{\operatorname{rank}(B)=k} \|A - B\| = s_{k+1},$$

and the minimum is attained for $B = \sum_{i=1}^k s_i u_i v_i^\mathsf{T}$.

Proof If B is any $m \times n$ matrix with rank k, its kernel $E = \ker(B)$ has dimension $n - k$. Then the min-max theorem (Corollary 4.1.7) for $k+1$ instead of k gives

$$\|A - B\| \ge \max_{x \in S(E)} \|(A - B)x\|_2 = \max_{x \in S(E)} \|Ax\|_2 \ge s_{k+1}(A).$$

In the opposite direction, setting $B = \sum_{i=1}^k s_i u_i v_i^\mathsf{T}$ gives $A - B = \sum_{i=k+1}^n s_i u_i v_i^\mathsf{T}$. The maximal singular value of this matrix is s_{k+1}, which is the same as its operator norm by Lemma 4.1.11. □

4.1.6 Perturbation theory

Perturbation theory strudies how eigenvalues and eigenvectors change under matrix perturbations. For eigenvalues, an application of min-max theorem (Theorem 4.1.6 and Corollary 4.1.7) gives the following (check!):

Lemma 4.1.14 (Weyl inequality). *The k-th largest eigenvalues of symmetric matrices A and B (of the same dimensions) satisfy*

$$|\lambda_k(A) - \lambda_k(B)| \le \|A - B\|.$$

Similarly, the k-th largest singular values of general rectangular matrices satisfy

$$|s_k(A) - s_k(B)| \le \|A - B\|.$$

A similar result holds for eigenvectors, but we need to be careful to track the same eigenvector before and after perturbation. If the eigenvalues are too close, a small perturbation can swap them, leading to huge error since their eigenvectors are orthogonal and thus far apart. To avoid this, we assume that the eigenvalues of A are well separated.

Theorem 4.1.15 (Davis-Kahan inequality)**.** *Consider two symmetric matrices A and B with spectral decompositions $A = \sum_{i=1}^n \lambda_i u_i u_i^\mathsf{T}$ and $B = \sum_{i=1}^n \mu_i v_i v_i^\mathsf{T}$, where the eigenvalues are (weakly) decreasing. Assume the k-th largest eigenvalue of A is δ-separated from the rest:*

$$\min_{i:i\neq k} |\lambda_k - \lambda_i| = \delta > 0.$$

Then the angle between the eigenvectors u_k and v_k (as a number between 0 and $\pi/2$) satisfies

$$\sin\angle(u_k,\, v_k) \le \frac{2\|A-B\|}{\delta}.$$

We will derive this from a stronger result that lets us target multiple eigenvectors at once. We will be looking at *spectral projections* – the orthogonal projections onto the span of some subsets of eigenvectors.

Lemma 4.1.16 (Davis-Kahan inequality for spectral projections)**.** *Consider two symmetric matrices A and B with spectral decompositions $A = \sum_{i=1}^n \lambda_i u_i u_i^\mathsf{T}$ and $B = \sum_{j=1}^n \mu_j v_j v_j^\mathsf{T}$. Let I, J be two δ-separated*[1] *subsets of $\mathbb{R}$, with I being an interval. Then the spectral projections*

$$P = \sum_{i:\lambda_i \in I} u_i u_i^\mathsf{T} \quad \textit{and} \quad Q = \sum_{j:\mu_j \in J} v_j v_j^\mathsf{T} \quad \textit{satisfy} \quad \|QP\| \le \frac{\|A-B\|}{\delta}.$$

Proof Without loss of generality, assume that the interval I is finite and closed. Adding the same multiple of identity to A and B, we can center I as $I = [-r, r]$, so that $|\lambda_i| \le r$ for $\lambda_i \in I$ and $|\mu_j| \ge r + \delta$ for $\mu_j \in J$. The idea is to study how P and Q interact through $H := B - A$:

$$\|H\| \ge \|QHP\| = \|QBP - QAP\| \ge \|QBP\| - \|QAP\|. \tag{4.11}$$

The spectral projection Q commutes with B, so

$$\|QBP\| = \|BQP\| \ge (r+\delta)\|QP\|. \tag{4.12}$$

To see the last inequality, note that the image of Q is spanned by orthogonal vectors v_j with $|\mu_j| \ge r + \delta$. The matrix B maps each such vector v_j to $\mu_j v_j$, thus scaling it by at least $r + \delta$. Thus, B expands the norm of any vector in the image of Q by at least $r + \delta$, giving $\|BQPx\|_2 \ge (r+\delta)\|QPx\|_2$ for any x. Taking the supremum over unit vectors x gives the inequality in (4.12). Also, $AP = PAP = \sum_{i:\lambda_i \in I} \lambda_i u_i u_i^\mathsf{T}$, so

$$\|QAP\| = \|QPAP\| \le \|QP\| \cdot \|AP\| \le r\|QP\|, \tag{4.13}$$

because $\|AP\| = \max_{i:\lambda_i \in I} |\lambda_i| \le r$. Putting the bounds (4.12) and (4.13) into (4.11), we get $\|H\| \ge \delta\|QP\|$, which completes the proof. □

[1] Subsets I and J are δ-separated if $|x - y| \ge \delta$ for all $x \in I$ and $y \in J$.

Proof of Theorem 4.1.15 We can assume that $\varepsilon := \|A - B\| \le \delta/2$, otherwise the conclusion is trivial. By Weyl inequality (Lemma 4.1.14), $|\lambda_j - \mu_j| \le \varepsilon$ for each j, so

$$\min_{j:j\neq k} |\lambda_k - \mu_j| \ge \min_{j:j\neq k} |\lambda_k - \lambda_j| - \varepsilon = \delta - \varepsilon \ge \delta/2.$$

Apply Lemma 4.1.16 for the $\delta/2$-separated subsets $I = \{\lambda_k\}$ and $J = \{\mu_j : j \neq k\}$ to get $\|QP\| \le 2\varepsilon/\delta$. To finish the proof, recall that P and $I_n - Q$ are the orthogonal projections on the directions of u_k and v_k respectively, so a little computation gives $\|QP\| = \|Qu_k\|_2 = \sin\angle(u_k, v_k)$. (Check!) □

Explore more versions of Davis-Kahan inequality in Exercises 4.11–4.16.

4.1.7 Isometries

The extreme singular values s_1 and s_n of an $m \times n$ matrix A have key geometric meaning. By the min-max theorem (Corollary 4.1.7), they can be expressed as

$$s_1(A) = \max_{\|x\|_2=1} \|Ax\|_2 \quad \text{and} \quad s_n(A) = \min_{\|x\|_2=1} \|Ax\|_2. \tag{4.14}$$

Applying this inequality for $x - y$ instead of x gives

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

So the extreme singular values set the limits on how much the linear map A distorts space. The *condition number* $\kappa(A) = s_1(A)/s_n(A)$ measures the worst-case distortion factor and is key in numerical algorithms.

A matrix that exactly preserve distances, meaning

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

is called an *isometry*. A basic example is an $m \times m$ orthogonal matrix U. More generally, any $m \times n$ matrix made of n columns of U is an isometry, giving an *"isometric embedding"* of $\mathbb{R}^n$ into $\mathbb{R}^m$. In fact, all isometries are of this form, because for $m \times n$ matrix A with[2] $m \ge n$, the following properties are equivalent:

(a) the columns of A are orthonormal, i.e. $A^\mathsf{T} A = I_n$;
(b) A is an isometry;
(c) all singular values of A equal 1.

Let's prove a stronger result where these properties hold approximately instead of exactly; this will come handy when dealing with random matrices:

Lemma 4.1.17 (Approximate isometries). *For an $m \times n$ matrix A with $m \ge n$ and a number $\varepsilon \ge 0$, the following properties are equivalent:*

(a) $\|A^\mathsf{T} A - I_n\| \le \varepsilon$.
(b) $(1-\varepsilon)\|x\|_2^2 \le \|Ax\|_2^2 \le (1+\varepsilon)\|x\|_2^2$ *for any* $x \in \mathbb{R}^n$.
(c) $1 - \varepsilon \le s_n(A)^2 \le s_1(A)^2 \le 1 + \varepsilon$.

[2] Note that an $m \times n$ matrix can be an isometry only if it is "tall", meaning $m \ge n$. Why?

Proof (a)⇔(b) By rescaling, we can assume that $\|x\|_2 = 1$ in (b). We have

$$\|A^\mathsf{T} A - I_n\| \overset{(4.10)}{=} \max_{\|x\|_2=1} \left| x^\mathsf{T}(A^\mathsf{T} A - I_n)x \right| = \max_{\|x\|_2=1} \left| \|Ax\|_2^2 - 1 \right|,$$

and this being bounded by ε is equivalent to (b) for all unit vectors x.

(b)⇔(c) follows from (4.14). □

Remark 4.1.18. Here is a more handy version of the key implication (a)⇒(c) in Lemma 4.1.17. For any numbers $z, \delta \geq 0$, we have

$$|z^2 - 1| \leq \max(\delta, \delta^2) \quad \Longrightarrow \quad |z - 1| \leq \delta$$

(check!) Then, substituting $\varepsilon = \max(\delta, \delta^2)$, we get:

$$\|A^\mathsf{T} A - I_n\| \leq \max(\delta, \delta^2) \quad \Longrightarrow \quad 1 - \delta \leq s_n(A) \leq s_1(A) \leq 1 + \delta. \tag{4.15}$$

Try Exercise 4.17 to add another handy equivalent property to Lemma 4.1.17.

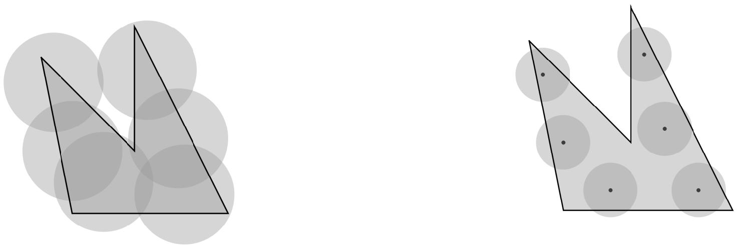
4.2 Nets, covering and packing

We are going to develop a simple but powerful method – an ε-net argument – and illustrate its usefulness for the analysis of random matrices. In this section, we recall the concept of an *ε-net*, something you might have come across in real analysis, and connect it to other basic notions – covering, packing, entropy, volume, and coding.

Definition 4.2.1 (ε-net)**.** Let (T, d) be a metric space. Consider a set $K \subset T$ and a number $\varepsilon > 0$. A subset $\mathcal{N} \subset K$ is called an *ε-net of K* if every point in K is within distance ε of some point of $\mathcal{N}$, i.e.

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

Equivalently, $\mathcal{N}$ is an ε-net of K if balls of radius ε centered at points in $\mathcal{N}$ cover K, like in Figure 4.1a.



(a) This covering of a polygon K by six ε-balls shows that $\mathcal{N}(K, \varepsilon) \leq 6$.

(b) $\mathcal{P}(K, \varepsilon) \geq 6$ means that there exist six ε-separated points in K; the $\varepsilon/2$-balls centered at these points are disjoint.

Figure 4.1 Covering and packing

If this generality is feeling overwhelming, keep in mind a key example. Let $T = \mathbb{R}^n$ with d being the Euclidean distance:

$$d(x, y) = \|x - y\|_2, \quad x, y \in \mathbb{R}^n. \tag{4.16}$$

In this case, we cover a subset $K \subset \mathbb{R}^n$ with *round balls*, like in Figure 4.1a. We already saw an example of this kind of covering in Corollary 0.0.3 where K was a polytope.

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

Remark 4.2.3 (Compactness)**.** An important result in real analysis says that a subset K of a complete metric space (T, d) is *precompact* (i.e. the closure of K is compact) if and only if

$$\mathcal{N}(K, d, \varepsilon) < \infty \quad \text{for every } \varepsilon > 0.$$

Thus we can think about the covering number $\mathcal{N}(K, d, \varepsilon)$ as a quantitative measure of compactness of K.

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

Definition 4.2.4 (Packing numbers)**.** A subset $\mathcal{N}$ of a metric space (T, d) is *ε-separated* if

$$d(x, y) > \varepsilon \quad \text{for any distinct points } x, y \in \mathcal{N}.$$

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

Remark 4.2.5 (Packing balls into K)**.** If $\mathcal{N}$ is ε-separated, the closed $\varepsilon/2$-balls centered at points in $\mathcal{N}$ are disjoint by the triangle inequality.[3] So, we can always "pack" into K at least $\mathcal{P}(K, d, \varepsilon)$ disjoint $\varepsilon/2$-balls like in Figure 4.1b.

Lemma 4.2.6 (Nets from separated sets)**.** *Let $\mathcal{N}$ be a* maximal[4] *ε-separated subset of K. Then $\mathcal{N}$ is an ε-net of K.*

Proof Let $x \in K$; we want to show that there exists $x_0 \in \mathcal{N}$ such that $d(x, x_0) \le \varepsilon$. If $x \in \mathcal{N}$, the conclusion is trivial by choosing $x_0 = x$. Suppose now $x \notin \mathcal{N}$. The maximality assumption implies that $\mathcal{N} \cup \{x\}$ is not ε-separated. But this means precisely that $d(x, x_0) \le \varepsilon$ for some $x_0 \in \mathcal{N}$. □

Remark 4.2.7 (Constructing a net)**.** Lemma 4.2.6 gives an iterative algorithm to construct an ε-net of a given set K. Pick a point $x_1 \in K$ arbitrarily, then pick $x_2 \in K$ that is farther than ε from x_1, then pick x_3 that it is farther than ε

[3] Otherwise we could find distinct $x, y \in \mathcal{N}$ and $z \in T$ such that $d(x, z) \le \varepsilon$ and $d(y, z) \le \varepsilon$. Triangle inequality would give $d(x, y) \le \varepsilon$, contradicting the separability assumption.

[4] Here by "maximal" we mean that adding any new point to $\mathcal{N}$ destroys the separation property.

from both x_1 and x_2, and so on. If K is compact, this process stops in finite time (why?) and gives an ε-net of K.

The covering and packing numbers are essentially equivalent:

Lemma 4.2.8 (Equivalence of covering and packing numbers)**.** *For any set $K \subset T$ and any $\varepsilon > 0$, we have*

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

Proof The upper bound follows from Lemma 4.2.6. (How?)

To prove the lower bound, take any 2ε-separated subset $\mathcal{P} = \{x_i\}$ in K and any ε-net $\mathcal{N} = \{y_j\}$ of K. By definition, each point x_i is in the ε-ball centered at some point y_j. Since any closed ε-ball cannot contain two 2ε-separated points, each ε-ball centered at y_j may contain at most one x_i. The pigeonhole principle gives $|\mathcal{P}| \leq |\mathcal{N}|$. Since $\mathcal{P}$ and $\mathcal{N}$ are arbitrary, the lower bound follows. □

Get some practice with covering and packing in Exercises 4.23–4.26.

4.2.1 Covering numbers and volume

Let's now study covering numbers in the most important example: $T = \mathbb{R}^n$ with the Euclidean metric

$$d(x, y) = \|x - y\|_2$$

as in (4.16). To ease the notation, we often omit the metric when it is understood, thus writing

$$\mathcal{N}(K, \varepsilon) = \mathcal{N}(K, d, \varepsilon).$$

If the covering numbers measure the size of K, how are they related to the most classical measure of size, the volume of K in $\mathbb{R}^n$? There could not be a full equivalence between these two quantities, since "flat" sets have zero volume but non-zero covering numbers. Still, there is a useful partial equivalence, which is often quite sharp. It is based on the notion of *Minkowski sum* of sets in $\mathbb{R}^n$.

Definition 4.2.9 (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 shows an example of Minkowski sum of two sets on the plane.

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

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

Here B_2^n denotes the unit Euclidean ball[5] in $\mathbb{R}^n$, so εB_2^n is a Euclidean ball with radius ε.

[5] Thus $B_2^n = \{x \in \mathbb{R}^n : \ \|x\|_2 \leq 1\}$.

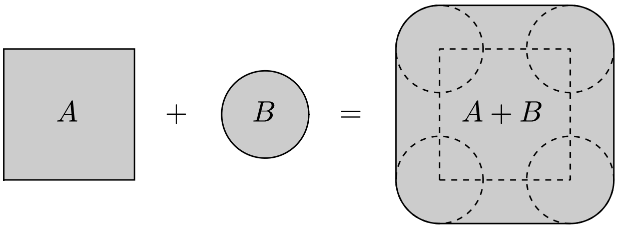


Figure 4.2 Minkowski sum of a square and a circle is a square with rounded corners.

Proof The middle inequality follows from Lemma 4.2.8, so all we need to prove is the left and right bounds.

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

$$\operatorname{Vol}(K) \le N \cdot \operatorname{Vol}(\varepsilon B_2^n),$$

proving the lower bound.

(Upper bound) Let $N := \mathcal{P}(K, \varepsilon)$. Then we can find N disjoint closed $\varepsilon/2$-balls with centers $x_i \in K$ (Remark 4.2.5). While these balls may not fit entirely in K (see Figure 4.1b), they do fit in a slightly inflated set, namely $K + (\varepsilon/2)B_2^n$. (Why?) Comparing the volumes gives

$$N \cdot \operatorname{Vol}((\varepsilon/2)B_2^n) \le \operatorname{Vol}(K + (\varepsilon/2)B_2^n).$$

which leads to the upper bound in the proposition. □

An important consequence of the volumetric bound (4.18) is that the covering (and thus packing) numbers are typically *exponential* in the dimension n:

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 $\varepsilon > 0$:*

$$\Big(\frac{1}{\varepsilon}\Big)^n \le \mathcal{N}(B_2^n, \varepsilon) \le \Big(\frac{2}{\varepsilon} + 1\Big)^n. \tag{4.17}$$

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 in $\mathbb{R}^n$ scales as $\operatorname{Vol}(\varepsilon B_2^n) = \varepsilon^n \operatorname{Vol}(B_2^n)$. The upper bound follows from Proposition 4.2.10, too:

$$\mathcal{N}(B_2^n, \varepsilon) \le \frac{\operatorname{Vol}((1 + \varepsilon/2)B_2^n)}{\operatorname{Vol}((\varepsilon/2)B_2^n)} = \frac{(1 + \varepsilon/2)^n}{(\varepsilon/2)^n} = \Big(\frac{2}{\varepsilon} + 1\Big)^n.$$

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

To simplify (4.17), note that in the non-trivial range $\varepsilon \in (0, 1]$ we have

$$\Big(\frac{1}{\varepsilon}\Big)^n \le \mathcal{N}(B_2^n, \varepsilon) \le \Big(\frac{3}{\varepsilon}\Big)^n. \tag{4.18}$$

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

Remark 4.2.12 (The volume of the ball)**.** The proof of Corollary 4.2.11 works with the volume of Euclidean ball, but it cleverly avoids calculating the actual volume! Try computing it yourself in three ways–geometric, probabilistic, and analytic–in Exercises 4.27–4.29, extend it to ℓ^p balls in Exercise 4.30.

Remark 4.2.13 (How to construct a net?)**.** Remark 4.2.7 describes a general iterative algorithm to building an ε-net, but for the Euclidean ball, you can just use a scaled integer lattice (Exercise 4.31) or even random points (Exercise 4.39).

The volumetric argument we just gave is pretty flexible and works in many other settings. Here is an important example.

Definition 4.2.14 (Hamming distance)**.** The Hamming cube $\{0,1\}^n$ consists of all binary strings of length n. To turn it into a metric space, we define the *Hamming distance* $d_H(x, y)$ as the number of bits where the strings x and y differ:

$$d_H(x,y) := \#\big\{i : \ x(i) \neq y(i)\big\}, \quad x, y \in \{0,1\}^n.$$

(Check that this is actually a metric!)

Proposition 4.2.15 (Covering and packing numbers of the Hamming cube)**.** *The covering and packing numbers of the Hamming cube $K = \{0,1\}^n$ satisfy the following for any integer $m \in [0, n]$:*

$$\frac{2^n}{\sum_{k=0}^{m} \binom{n}{k}} \le \mathcal{N}(K, d_H, m) \le \mathcal{P}(K, d_H, m) \le \frac{2^n}{\sum_{k=0}^{\lfloor m/2 \rfloor} \binom{n}{k}}.$$

The proof is just a tweak of the volumetric argument, replacing volume with cardinality – give it a try in Exercise 4.32. For more practice with the volumetric argument, compute the covering numbers for low-rank matrices in Exercise 4.50.

4.3 Application: error correcting codes

Covering and packing arguments frequently appear in applications to *coding theory*. Here we give two examples that relate covering and packing numbers to complexity and error correction.

4.3.1 Metric entropy and complexity

Intuitively, the covering and packing numbers measure the *complexity* of a set K. The logarithm of the covering numbers $\log_2 \mathcal{N}(K, \varepsilon)$ is often called the *metric entropy* of K. As we will see now, the metric entropy is equivalent to the number of bits needed to encode points in K:

Proposition 4.3.1 (Metric entropy and coding)**.** *Let (T, d) be a metric space,* *and consider a subset $K \subset T$. Let $\mathcal{C}(K, d, \varepsilon)$ denote the smallest number of bits sufficient to specify every point $x \in K$ with accuracy ε in the metric d. Then*

$$\log_2 \mathcal{N}(K, d, \varepsilon) \leq \mathcal{C}(K, d, \varepsilon) \leq \lceil \log_2 \mathcal{N}(K, d, \varepsilon/2) \rceil .$$

Proof **(Lower bound)** Assume $\mathcal{C}(K, d, \varepsilon) \leq N$. This means that there exists a mapping ("encoding") of points $x \in K$ into bit strings of length N, which specifies every point with accuracy ε. This gives a partition of the domain K into at most 2^N subsets, each consisting of the points represented by the same string (see Figure 4.3). Each subset has diameter[6] at most ε, so it can be covered by a ball centered in K and with radius ε. (Why?) Therefore, K can be covered by at most 2^N balls with radius ε, meaning $\mathcal{N}(K, d, \varepsilon) \leq 2^N$. Taking logarithms gives the lower bound.

(Upper bound) Assume $\log_2 \mathcal{N}(K, d, \varepsilon/2) \leq N$ for some integer N. This means that there exists an $(\varepsilon/2)$-net $\mathcal{N}$ of K with at most 2^N points. For each point $x \in K$, assign a closest point $x_0 \in \mathcal{N}$. Since there are at most 2^N such points, N bits are sufficient to specify x_0. Let's show that the encoding $x \mapsto x_0$ represents points in K with accuracy ε. If both x and y are encoded by the same x_0 then by triangle inequality,

$$d(x, y) \leq d(x, x_0) + d(y, x_0) \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon$$

This shows that $\mathcal{C}(K, d, \varepsilon) \leq N$. This completes the proof. □

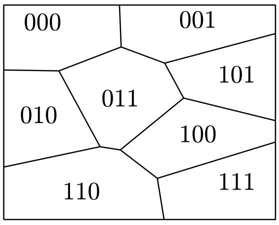


Figure 4.3 Encoding points in K as N-bit strings induces a partition of K into at most 2^N subsets.

4.3.2 Error correcting codes

Suppose Alice wants to send Bob a message with k letters, such as

$$x := \text{"fill the glass"}.$$

Suppose further that an adversary can corrupt Alice's message by changing up to r letters. For example, Bob may receive

$$y := \text{"bill the class"}$$

[6] The diameter of a subset K of a metric space is defined as $\operatorname{diam}(K) := \sup\{d(x, y) : x, y \in K\}$.

if $r = 2$. Is there a way to protect the communication channel between Alice and Bob, a method that can correct adversarial errors?

A common approach uses *redundancy.* Alice encodes her k-letter message into a longer n-letter message ($n > k$), hoping that the extra information helps Bob recover the message, even if up to r errors occur.

Example 4.3.2 (Repetition code)**.** Alice may just repeat her message several times, thus sending to Bob

$E(x) :=$ *"fill the glass fill the glass fill the glass fill the glass fill the glass".*

Bob can use *majority decoding*: he checks the received copies of each letter in $E(x)$ and picks the one that appears most often. If the original message x is repeated $2r + 1$ times, majority decoding will recover x correctly, even if r letters in $E(x)$ are corrupted. (Why?)

The problem with majority decoding is that it is inefficient: it uses

$$n = (2r + 1)k \tag{4.19}$$

letters to encode a k-letter message. As we will see shortly, there exist error correcting codes with much smaller n.

But first let us formalize the notion of an error correcting code – an encoding that turns k-letter strings into n-letter strings capable of correcting r errors. For simplicity, we will use the binary alphabet, consisting of just the letters 0 and 1, instead of the English alphabet.

Definition 4.3.3 (Error correcting code)**.** An *error correcting code* that encodes k-bit strings into n-bit strings and can correct r errors consists of encoding map $E : \{0, 1\}^k \to \{0, 1\}^n$ and decoding map $D : \{0, 1\}^n \to \{0, 1\}^k$ that satisfy

$$D(y) = x$$

for any word $x \in \{0, 1\}^k$ and any string $y \in \{0, 1\}^n$ that differs from $E(x)$ in at most r bits.

We now relate error correction to packing numbers of the Hamming cube $\{0, 1\}^n$ equipped with the Hamming metric introduced in Definition 4.2.14.

Lemma 4.3.4 (Error correction and packing)**.** *Assume that positive integers k, n and r are such that*

$$\log_2 \mathcal{P}(\{0, 1\}^n, d_H, 2r) \geq k.$$

Then there exists an error correcting code that encodes k-bit strings into n-bit strings and can correct r errors.

Proof By assumption, there exists a subset $\mathcal{N} \subset \{0, 1\}^n$ with $|\mathcal{N}| = 2^k$, where the closed balls of radius r centered at the points in $\mathcal{N}$ are disjoint (see Remark 4.2.5). Let the encoder $E : \{0, 1\}^k \to \mathcal{N}$ be any one-to-one mapping, and let $D : \{0, 1\}^n \to \{0, 1\}^k$ be a nearest neighbor decoder.[7]

[7] Formally, we set $D(y) = x_0$ where $E(x_0)$ is a closest codeword in $\mathcal{N}$ to y; break ties arbitrarily.

If $y \in \{0,1\}^n$ differs from $E(x)$ in at most r bits, it lies in the closed ball of radius r centered at $E(x)$. Since such balls are disjoint by construction, y is strictly closer to $E(x)$ than to any other codeword $E(x')$. So, nearest-neighbor decoding correctly decodes y, meaning $D(y) = x$. □

Let us substitute into Lemma 4.3.4 the bounds on the packing numbers of the Hamming cube from Proposition 4.2.15.

Theorem 4.3.5 (Guarantees for an error correcting code)**.** *Assume that positive integers k, n and r are such that*

$$n - k \geq 2r \log_2 \left(\frac{en}{2r} \right).$$

Then there exists an error correcting code that encodes k-bit strings into n-bit strings and can correct r errors.

Proof Passing from packing to covering numbers using Lemma 4.2.8 and then using the bounds on the covering numbers from Proposition 4.2.15 (and simplifying using Exercise 0.6), we get

$$\mathcal{P}(\{0,1\}^n, d_H, 2r) \geq \mathcal{N}(\{0,1\}^n, d_H, 2r) \geq \frac{2^n}{\sum_{i=0}^{2r} \binom{n}{i}} \geq 2^n \left(\frac{2r}{en} \right)^{2r}.$$

By assumption, this quantity is further bounded below by 2^k. An application of Lemma 4.3.4 completes the proof. □

Remark 4.3.6 (Extra bits grow nearly linearly with errors)**.** Theorem 4.3.5 shows that we can correct r errors with $n-k$ growing nearly linear in r (ignoring a logarithmic term). This is way more efficient than the repetition code (4.19), and is optimal (Exercise 4.33).

4.4 Upper bounds on subgaussian random matrices

We are now ready to enter the non-asymptotic theory of random matrices. Random matrix theory is concerned with $m \times n$ matrices A with random entries. The central questions here are about the distributions of singular values, eigenvalues (if A is symmetric), and eigenvectors.

Theorem 4.4.3 will give us the first bound on the operator norm (or the largest singular value) of a random matrix with independent subgaussian entries. It is neither the sharpest nor the most general result; we will improve and extend it in Sections 4.6 and 6.4.

But before that, let us take a moment to learn how ε-nets can help us compute the operator norm of a matrix.

4.4.1 Computing the norm on an ε-net

The operator norm of an $m \times n$ matrix A, introduced in Definition 4.1.8, is

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

To evaluate $\|A\|$, we need to control $\|Ax\|_2$ uniformly over the sphere S^{n-1} – sometimes a daunting task. We will show that instead of the entire sphere, it is enough to control just an ε-net of the sphere (in the Euclidean metric).

Lemma 4.4.1 (Computing the operator norm on a net)**.** *Let A be an $m \times n$ matrix and $\varepsilon \in [0, 1)$. Then, for any 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 is trivial since $\mathcal{N} \subset S^{n-1}$. To prove the upper bound, fix a vector $x \in S^{n-1}$ for which $\|A\| = \|Ax\|_2$ and choose $x_0 \in \mathcal{N}$ that approximates x so that $\|x - x_0\|_2 \leq \varepsilon$. By the definition of the operator norm, this implies

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

Using triangle inequality, we find that

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

Dividing by $1 - \varepsilon$, we complete the proof. □

Lemma 4.4.1 is quite flexible; here is a useful version of it. From (4.9), we know that the operator norm of an $m \times n$ matrix A can be found by maximizing a quadratic form:

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

and if A is symmetric, we can take $x = y$, see (4.10). We can replace the two spheres by their nets:

Lemma 4.4.2 (Maximizing quadratic forms on a net)**.** *Let A be an $m \times n$ matrix and $\varepsilon \in [0, 1/2)$. Then, 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*

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

Moreover, if $m = n$, A is symmetric and $\mathcal{N} = \mathcal{M}$, then we can take $x = y$.

One can prove this by tweaking the proof of Lemma 4.4.1 – try Exercise 4.36. For a different method, see Exercises 4.34, and for more practice, do Exercise 4.37.

4.4.2 The norms of subgaussian random matrices

We are ready for our first result on random matrices. It says that the operator norm of an $m \times n$ random matrix A with independent subgaussian entries satisfies

$$\|A\| \lesssim \sqrt{m} + \sqrt{n} \quad \text{with high probability.}$$

Theorem 4.4.3 (Norm of matrices with subgaussian entries)**.** *Let A be an $m \times n$ random matrix with independent, mean-zero, subgaussian entries A_{ij}. Then, for any $t > 0$ we have*[8]

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

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

Proof This proof is an 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

$$|\mathcal{N}| \le 9^n \quad \text{and} \quad |\mathcal{M}| \le 9^m. \tag{4.20}$$

By Lemma 4.4.2, the norm of A can be bounded using these nets as follows:

$$\|A\| \le 2 \max_{x \in \mathcal{N},\, y \in \mathcal{M}} |\langle Ax, y\rangle|. \tag{4.21}$$

Step 2: Concentration. Fix $x \in \mathcal{N}$ and $y \in \mathcal{M}$. 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, subgaussian random variables. Proposition 2.7.1 states that the sum is subgaussian, and

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

Recalling Proposition 2.6.6(i), we can restate this as the tail bound

$$\mathbb{P}\{|\langle Ax, y\rangle| \ge u\} \le 2\exp(-cu^2/K^2), \quad u \ge 0. \tag{4.22}$$

Step 3: Union bound. Next, we unfix x and y using a union bound. The event $\max_{x \in \mathcal{N},\, y \in \mathcal{M}} |\langle Ax, y\rangle| \ge u$ means that there exist $x \in \mathcal{N}$ and $y \in \mathcal{M}$ such that $|\langle Ax, y\rangle| \ge u$, so the union bound gives

$$\mathbb{P}\Big\{\max_{x \in \mathcal{N},\, y \in \mathcal{M}} |\langle Ax, y\rangle| \ge u\Big\} \le \sum_{x \in \mathcal{N},\, y \in \mathcal{M}} \mathbb{P}\{|\langle Ax, y\rangle| \ge u\}.$$

Using the tail bound (4.22) and the estimate (4.20) on the sizes of $\mathcal{N}$ and $\mathcal{M}$, we bound the probability above by

$$9^{n+m} \cdot 2\exp(-cu^2/K^2). \tag{4.23}$$

[8] In results like this, C and c will always denote some positive absolute constants.

Choose

$$u = CK(\sqrt{n} + \sqrt{m} + t). \tag{4.24}$$

Then $u^2 \geq C^2K^2(n + m + t^2)$, and if the constant C is chosen sufficiently large, the exponent in (4.23) is large enough, say $cu^2/K^2 \geq 3(n + m) + t^2$. Thus

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

Finally, combining this with (4.21), we conclude that

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

Recalling our choice of u in (4.24), we complete the proof. □

Remark 4.4.4 (Expectation)**.** *High-probability bounds* like Theorem 4.4.3 can usually be turned into simpler (but less informative) *expectation bounds* using the integrated tail formula (Lemma 1.6.1). Try Exercise 4.41 to get

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

Remark 4.4.5 (Optimality)**.** Theorem 4.4.3 is typically tight since the matrix's operator norm is bounded below by the Euclidean norm of any row or column (Exercise 4.7). For example, if A has Rademacher entries, its columns have norm $\sqrt{m}$ and rows $\sqrt{n}$, so

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

with probability 1. For a fully general lower bound, try Exercise 4.42.

Remark 4.4.6 (Relaxing independence)**.** The independence assumption in Theorem 4.4.3 can be relaxed: we just need the rows (or columns) of A to be independent, even with dependent entries (Exercise 4.43).

4.4.3 Symmetric matrices

Theorem 4.4.3 extends easily for symmetric matrices, giving the bound

$$\|A\| \lesssim \sqrt{n} \quad \text{with high probability.}$$

Corollary 4.4.7 (Norm of symmetric matrices with subgaussian entries)**.** *Let A be an $n \times n$ symmetric random matrix with independent, mean-zero, subgaussian entries A_{ij} on and above the diagonal. Then, for any $t > 0$ we have*

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

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

Proof Split A into the upper-triangular part A^+ and lower-triangular part A^-. The diagonal can go either way; let us include it in A^+ to be specific. Then

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

Theorem 4.4.3 applies for each part A^+ and A^- separately. By a union bound, with probability at least $1-4\exp(-t^2)$ both of these bounds hold simultaneously:

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

Since by the triangle inequality $\|A\| \le \|A^+\| + \|A^-\|$, the proof is complete. □

To practice more with norms of random matrices, try Exercise 4.44.

4.5 Application: community detection in networks

Random matrix theory has many applications. Here we give an example in network analysis.

Real-world networks often have *communities*, or clusters of tightly connected nodes. Identifying them accurately and efficiently is a main challenge known as the *community detection problem.*

4.5.1 Stochastic Block Model

Let's explore community detection in a simple probabilistic network model with two communities. It's a straightforward extension of the Erdős-Rényi model random graph model, which we described in Section 2.5.

Definition 4.5.1 (Stochastic block model)**.** Split n vertices into two groups ("communities") of size $n/2$ each. Build a random graph G by connecting each pair of vertices independently with probability p if they are in the same community and q if they are in different communities. Self-loops are included under this rule. This random graph model is called the *stochastic block model*,[9] denoted $G(n,p,q)$.

When $p=q$, we get a version of the Erdös-Rényi model $G(n,p)$ with self-loops. But when $p>q$, edges are more likely to form within communities than between them, creating a community structure (see Figure 4.4).

4.5.2 The expected adjacency matrix holds the key

A graph G can be conveniently represented by its adjacency matrix A, which we introduced in Definition 3.6.2. For a random graph $G \sim G(n,p,q)$, the adjacency matrix A is a *random matrix*, and we can analyze A using the tools we developed earlier in this chapter.

Let's split A into deterministic and random parts:

$$A = D + R \quad \text{where} \quad D = \mathbb{E}\,A$$

and think of D as the *signal* (informative part) and R as "noise".

To see why D is informative, let us compute its eigenstructure. The entries

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

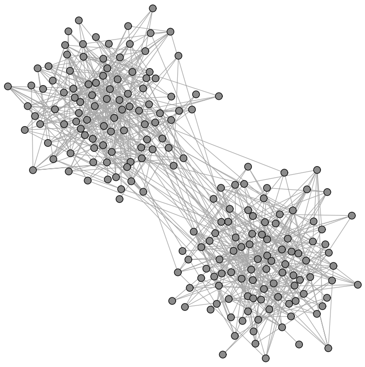


Figure 4.4 A random graph following the stochastic block model $G(n, p, q)$ with $n = 200$, $p = 1/20$ and $q = 1/200$.

A_{ij} have a Bernoulli distribution, either $\mathrm{Ber}(p)$ or $\mathrm{Ber}(q)$, depending on the community membership of vertices i and j. So, the entries of D are either p or q, depending on the membership. For example, if we arrange vertices by community, then for $n = 4$, the matrix D look like:

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

Take a look at the simpler, 2×2 matrix $\left[\begin{smallmatrix} p & q \\ q & p \end{smallmatrix}\right]$: it has eigenvalues $\frac{p+q}{2}$ and $\frac{p-q}{2}$ with eigenvectors $\left[\begin{smallmatrix} 1 \\ 1 \end{smallmatrix}\right]$ and $\left[\begin{smallmatrix} 1 \\ -1 \end{smallmatrix}\right]$ (check!). The matrix D is similar but with p and q replaced by $n/2 \times n/2$ blocks of the same values. So, D also has rank 2, and its nonzero eigenvalues and eigenvectors are:

$$\lambda_1(D) = \Big(\frac{p+q}{2}\Big)n, \; u_1(D) = \left[\begin{array}{c} 1 \\ \vdots \\ 1 \\ \hline 1 \\ \vdots \\ 1 \end{array}\right]; \quad \lambda_2(D) = \Big(\frac{p-q}{2}\Big)n, \; u_2(D) = \left[\begin{array}{c} 1 \\ \vdots \\ 1 \\ \hline -1 \\ \vdots \\ -1 \end{array}\right].$$

(Check!) The key object here is the *second* eigenvector $u_2(D)$, which holds all the information about the community structure. If we knew $u_2(D)$, we could identify the communities based on the signs of its coefficients.

4.5.3 The actual adjacency matrix is a good approximation

But we don't know the expected adjacency matrix $D = \mathbb{E}\, A$, so can't access $u_2(D)$. Instead, we know the actual adjacency matrix $A = D + R$, which is a noisy version of D. The level of the signal D is

$$\|D\| = \lambda_1 \asymp n$$

while the level of the noise R can be estimated using Corollary 4.4.7:

$$\|R\| \le C\sqrt{n} \quad \text{with probability at least } 1 - 4e^{-n}. \tag{4.25}$$

So for large n, the noise R is much smaller than the signal D. It means that A is close to D, so we can use A instead of D to extract the community information. Let's justify this using matrix perturbation theory developed in Section 4.1.6.

4.5.4 Perturbation theory

Apply Davis-Kahan inequality (Theorem 4.1.15) to D and A, focusing on the second-largest eigenvalue. We need to check that $\lambda_2(D)$ is well separated from the rest of the spectrum of D, that is from 0 and $\lambda_1(D)$. The distance is

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

Recalling the bound on $R = A - D$ from (4.25), the Davis-Kahan inequality gives a bound on the angle between the *unit* eigenvectors of D and A (indicated here by bars on top of the vectors):

$$\sin \angle\left(\bar{u}_2(D),\, \bar{u}_2(A)\right) \le \frac{2\|R\|}{\delta} \lesssim \frac{\sqrt{n}}{\mu n} \lesssim \frac{1}{\mu\sqrt{n}}.$$

If the sine of the angle between two unit vectors is small, the vectors are close up to a sign (Exercise 4.16), so there exists $\theta \in \{-1, 1\}$ such that

$$\|\bar{u}_2(D) - \theta \bar{u}_2(A)\|_2 \lesssim \frac{1}{\mu\sqrt{n}}.$$

We already computed the eigenvector $u_2(D)$ of D, but it wasn't a unit vector – it had norm $\sqrt{n}$ since its coefficients were ±1. Multiplying both sides by $\sqrt{n}$, we get

$$\|u_2(D) - \theta u_2(A)\|_2 \lesssim \frac{1}{\mu}.$$

This implies that that the *signs* of most coefficients of $v_2(D)$ and $\theta v_2(A)$ must agree. Indeed, rewriting the bound as

$$\sum_{j=1}^{n} |u_2(D)_j - \theta u_2(A)_j|^2 \lesssim \frac{1}{\mu^2}$$

and noting that all coefficients of $u_2(D)$ are ±1, we see that each disagreement between the signs of $u_2(D)_j$ and $\theta u_2(A)_j$ contributes at least 1 to the sum. Therefore, the number of disagreeing signs is $\lesssim 1/\mu^2$.

4.5.5 Spectral Clustering

In summary, we can use the vector $u_2(A)$ to accurately estimate the vector $u_2 = u_2(D)$, whose coefficients are ±1 and identify the two communities. This method for community detection is usually called *spectral clustering*. Let's now state the method and the guarantees we have just derived.

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 $v_2(A)$ for the second largest eigenvalue of $A$.
3: Split vertices into two communities based on the signs of $v_2(A)$'s coefficients.[10]
```

We have proved:

Theorem 4.5.2 (Spectral clustering for the stochastic block model)**.** *Let $G \sim G(n,p,q)$ and $\min(q, p-q) = \mu > 0$. Then, with probability at least $1-4e^{-n}$, the spectral clustering algorithm identifies the communities of G with at most C/μ^2 misclassified vertices.*

Summarizing, the spectral clustering algorithm correctly classifies all but $O(1)$ number of vertices, as long as the random graph is dense enough ($q \ge \text{const}$) and the probabilities of within- and across-community edges are well separated ($p-q \ge \text{const}$). The condition $q \ge \text{const}$ is not essential; try removing it by doing Exercise 4.45.

Remark 4.5.3 (Sparsity)**.** The sparsest graphs for which Theorem 4.5.2 is nontrivial, meaning $C/\mu^2 \le n$, have expected average degree

$$\frac{n(p+q)}{2} \asymp \sqrt{n}.$$

With more tools, we will handle much sparser graphs in Section 5.5.

4.6 Two-sided bounds on subgaussian matrices

Let us revisit Theorem 4.4.3, which gives an upper bound on the singular values of an $m \times n$ subgaussian random matrix A. It tells us that

$$s_1(A) = \|A\| \le C(\sqrt{m} + \sqrt{n})$$

with high probability. Now, we will prove sharper *two-sided* bounds on the entire spectrum of A:

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

In other words, we will show that a tall random matrix $\frac{1}{\sqrt{m}}A$ (with $m \gg n$) is an *approximate isometry* (see Section 4.1.7).

[10] To be specific, if $v_2(A)_j > 0$ put vertex j into first community, otherwise in the second.

Theorem 4.6.1 (Two-sided bound on subgaussian matrices)**.** *Let A be an $m \times n$ random matrix with independent, mean-zero, subgaussian, isotropic rows A_i. Then for any $t \geq 0$ we have*

$$\sqrt{m} - CK^2(\sqrt{n} + t) \leq s_n(A) \leq s_1(A) \leq \sqrt{m} + CK^2(\sqrt{n} + t) \tag{4.26}$$

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

We will prove a slightly stronger conclusion than (4.26), namely that

$$\Big\| \frac{1}{m} A^\mathsf{T} A - I_n \Big\| \leq K^2 \max(\delta, \delta^2) \quad \text{where} \quad \delta = C\Big(\sqrt{\frac{n}{m}} + \frac{t}{\sqrt{m}} \Big). \tag{4.27}$$

Using (4.15), one can check that (4.27) indeed implies (4.26). (Do this!)

Proof We will prove (4.27) using an *ε-net argument*, like in Theorem 4.4.3, but with Bernstein concentration inequality instead of Hoeffding.

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

$$|\mathcal{N}| \leq 9^n.$$

Using Lemma 4.4.2, we can evaluate the operator norm in (4.27) on the $\mathcal{N}$:

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

So to prove (4.27), it's enough to show that, with the required probability,

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

Step 2: Concentration. Fix $x \in \mathcal{N}$ and express $\|Ax\|_2^2$ as a sum of independent random variables:

$$\|Ax\|_2^2 = \sum_{i=1}^m \langle A_i, x \rangle^2 =: \sum_{i=1}^m X_i^2. \tag{4.28}$$

By assumption, the rows A_i are independent, isotropic, and subgaussian random vectors with $\|A_i\|_{\psi_2} \leq K$. Thus $X_i = \langle A_i, x \rangle$ are independent subgaussian random variables with $\mathbb{E}\, X_i^2 = 1$ and $\|X_i\|_{\psi_2} \leq K$. This makes $X_i^2 - 1$ independent, mean-zero, and subexponential random variables[11] with

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

[11] This reasoning mirrors a step in the proof of Theorem 3.1.1.

Thus we can use Bernstein inequality (Corollary 2.9.2) and obtain

$$\begin{aligned}\mathbb{P}\Big\{\Big|\frac{1}{m}\|Ax\|_2^2 - 1\Big| \geq \frac{\varepsilon}{2}\Big\} &= \mathbb{P}\Big\{\Big|\frac{1}{m}\sum_{i=1}^m X_i^2 - 1\Big| \geq \frac{\varepsilon}{2}\Big\} \\ &\leq 2\exp\Big[-c_1 \min\Big(\frac{\varepsilon^2}{K^4}, \frac{\varepsilon}{K^2}\Big)m\Big] \\ &= 2\exp\Big[-c_1\delta^2 m\Big] \quad (\text{since } \frac{\varepsilon}{K^2} = \max(\delta, \delta^2)) \\ &\leq 2\exp\big[-c_1 C^2(n+t^2)\big].\end{aligned}$$

The last bound follows from the definition of δ in (4.27) and using the inequality $(a+b)^2 \geq a^2 + b^2$ for $a, b \geq 0$.

Step 3: Union bound. Now we can unfix $x \in \mathcal{N}$ using a union bound. Recalling that $|\mathcal{N}| \leq 9^n$, we get

$$\mathbb{P}\Big\{\max_{x\in\mathcal{N}}\Big|\frac{1}{m}\|Ax\|_2^2 - 1\Big| \geq \frac{\varepsilon}{2}\Big\} \leq 9^n \cdot 2\exp\big[-c_1C^2(n+t^2)\big] \leq 2\exp(-t^2)$$

if we chose the absolute constant C in (4.27) large enough. As we noted in Step 1, this completes the proof of the theorem. □

Remark 4.6.2 (Expectation)**.** As we noted in Remark 4.4.4, high-probability bounds can be converted into expectation bounds. Do Exercise 4.41 to get the following expected form of Theorem 4.6.1:

$$\mathbb{E}\Big\|\frac{1}{m}A^{\mathsf{T}}A - I_n\Big\| \leq CK^2\Big(\sqrt{\frac{n}{m}} + \frac{n}{m}\Big).$$

Try Exercise 4.46 for an alternative proof of Theorem 4.6.1.

4.7 Application: covariance estimation and clustering

Suppose we want to analyze some high-dimensional data, given as points $X_1, \ldots, X_m$ sampled from an unknown distribution in $\mathbb{R}^n$. A basic tool for exploring such data is principal component analysis (PCA), which we touched on in Section 3.2.2.

PCA finds the "principal components" as top eigenvectors of the data's covariance matrix. Although we do not know the covariance matrix of the underlying distribution (the "population covariance matrix"), we can approximately estimate it using the sample $X_1, \ldots, X_m$. The Davis-Kahan theorem 4.1.15 then helps us estimate the principal components of the underlying distribution.

How do we estimate the covariance matrix from the data? Let X denote the random vector from the unknown distribution. For simplicity, assume X has zero mean,[12] and let's denote its covariance matrix by

$$\Sigma = \mathbb{E}\, XX^{\mathsf{T}}.$$

[12] Our analysis doesn't actually require zero mean, in which case Σ is simply the second moment matrix of X, as we explained in Section 3.2.

To estimate Σ, we use the *sample covariance* matrix Σ_m, computed from the sample $X_1, \ldots, X_m$ as follows:

$$\Sigma_m = \frac{1}{m} \sum_{i=1}^m X_i X_i^\mathsf{T}.$$

We just replaced the population average with the sample average here.

Since X_i and X are identically distributed, our estimate is unbiased:

$$\mathbb{E}\, \Sigma_m = \Sigma.$$

By the law of large numbers (Theorem 1.7.1) applied to each entry of Σ, we get

$$\Sigma_m \to \Sigma \quad \text{almost surely}$$

as the sample size m increases to infinity. This leads to the quantitative question: how big does the sample size m need to be to make sure that

$$\Sigma_m \approx \Sigma \quad \text{with high probability?}$$

For dimension reasons, we need at least $m \gtrsim n$ sample points. (Why?) Let's show that $m \asymp n$ is enough.

Theorem 4.7.1 (Covariance estimation)**.** *Let X be a subgaussian random vector in $\mathbb{R}^n$. More specifically, assume that there exists $K \ge 1$ such that*[13]

$$\|\langle X, x\rangle\|_{\psi_2} \le K \|\langle X, x\rangle\|_{L^2} \quad \textit{for any } x \in \mathbb{R}^n. \tag{4.29}$$

Then, for every positive integer m, we have

$$\mathbb{E}\|\Sigma_m - \Sigma\| \le CK^2 \Big(\sqrt{\frac{n}{m}} + \frac{n}{m} \Big) \|\Sigma\|.$$

Proof Let's first bring the random vectors $X, X_1, \ldots, X_m$ to the isotropic position. For simplicity, assume that Σ is invertible (you can drop this condition later, like in Exercise 3.10). Setting $Z = \Sigma^{-1/2} X$ and $Z_i = \Sigma^{-1/2} X_i$, we see that $Z, Z_1, \ldots, Z_m$ are independent and isotropic random vectors satisfying

$$X = \Sigma^{1/2} Z \quad \text{and} \quad X_i = \Sigma^{1/2} Z_i.$$

The assumption (4.29) then implies that

$$\|Z\|_{\psi_2} \le K \quad \text{and} \quad \|Z_i\|_{\psi_2} \le K.$$

(Check!) Then

$$\|\Sigma_m - \Sigma\| = \|\Sigma^{1/2} R_m \Sigma^{1/2}\| \le \|R_m\|\, \|\Sigma\| \quad \text{where } R_m := \frac{1}{m} \sum_{i=1}^m Z_i Z_i^\mathsf{T} - I_n. \tag{4.30}$$

[13] Here we used the notation for the L^2 norm of random variables from (1.10), so that $\|\langle X, x\rangle\|_{L^2}^2 = \mathbb{E}\langle X, x\rangle^2 = x^\mathsf{T} \Sigma x$ due to (3.6).

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

$$\frac{1}{m}A^{\mathsf{T}}A - I_n = \frac{1}{m}\sum_{i=1}^{m} Z_i Z_i^{\mathsf{T}} - I_n = R_m.$$

Applying Theorem 4.6.1, we get

$$\mathbb{E}\|R_m\| \leq CK^2\Big(\sqrt{\frac{n}{m}} + \frac{n}{m}\Big)$$

(see Remark 4.6.2). Substituting this into (4.30), we complete the proof. □

Remark 4.7.2 (Sample complexity)**.** Theorem 4.7.1 shows that for any $\varepsilon \in (0,1)$, we can estimate the covariance matrix with a small relative error:

$$\mathbb{E}\|\Sigma_m - \Sigma\| \leq \varepsilon\|\Sigma\|,$$

as long as the sample size is

$$m \asymp \varepsilon^{-2}n.$$

So, the sample covariance matrix gives a good estimate for the population covariance matrix *if the sample size m is proportional to the dimension n.*

Remark 4.7.3 (High-probability bound)**.** Our argument also gives a high-probability bound: for any $u \geq 0$, we have

$$\|\Sigma_m - \Sigma\| \leq CK^2\Big(\sqrt{\frac{n+u}{m}} + \frac{n+u}{m}\Big)\,\|\Sigma\|$$

with probability at least $1 - 2e^{-u}$. Check this! (Exercise 4.49).

We will revisit covariance estimation later, handling heavy-tailed and approximately low-dimensional distributions in Sections 5.6 and 9.2.3. For now, try Exercise 4.48 to capture all 1D marginals with small relative error.

4.7.1 Application: clustering of point sets

Let us illustrate Theorem 4.7.1 with an application to clustering. This follows a similar idea to Section 4.5, but this time we will find clusters in point sets in $\mathbb{R}^n$ instead of networks. Although the notion of a cluster is not strictly defined, common sense says points in the same cluster should be closer together than those in different ones.

As we did for networks, we begin by building a simple probabilistic model for point sets in $\mathbb{R}^n$ with two clusters:

Definition 4.7.4 (Gaussian mixture model)**.** Generate m random points in $\mathbb{R}^n$ like this. Flip a fair coin; if it comes up heads, draw a point from $N(\mu, I_n)$, and if it comes up tails, from $N(-\mu, I_n)$. This distribution is called the Gaussian mixture model with means $\pm\mu$.

Equivalently, consider a random vector

$$X = \theta\mu + g$$

where θ is a Rademacher random variable, $g \sim N(0, I_n)$, and θ and g are independent. Draw a sample $X_1, \ldots, X_m$ of independent random vectors identically distributed with X. Then the sample is distributed according to the Gaussian mixture model; see Figure 4.5 for illustration.

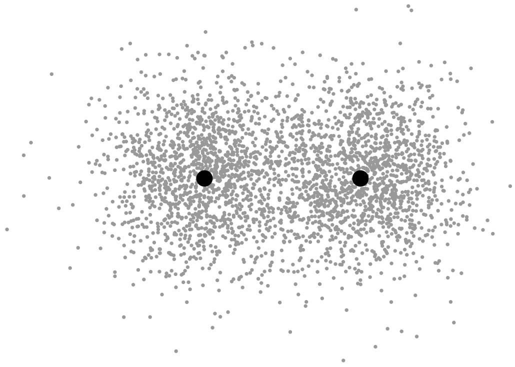


Figure 4.5 $m = 3000$ points drawn from the Gaussian mixture model with means $-\mu$ and μ shown as two big black dots, for $\mu = (-1.6, 0)$.

Given m sample points from the Gaussian mixture model, our goal is to figure out which points belong to which cluster. To do this, we can use a variant of the *spectral clustering* algorithm we introduced for networks in Section 3.2.2.

To see why a spectral method might work here, notice that the distribution of X is not isotropic, but rather stretched along μ (the horizontal direction in Figure 4.5.) Thus, we can approximately find μ by computing the first principal component of the data — the eigenvector corresponding to the largest eigenvalue of the sample covariance matrix. Next, we can project the data points onto the principal component and classify them based on which side of the origin the lie on. Here is the algorithm:

Spectral Clustering Algorithm

```
Input: points $X_1, \ldots, X_m$ in $\mathbb{R}^n$
Output: a partition of the points into two clusters
Compute the sample covariance matrix $\Sigma_m = \frac{1}{m} \sum_{i=1}^m X_i X_i^\mathsf{T}$.
Compute the top eigenvector $v = v_1(\Sigma_m)$.
Split the points $X_i$ into two communities based on the sign of $\langle X_i, v \rangle$.
```

Theorem 4.7.5 (Spectral clustering for the Gaussian mixture model)**.** *Take points $X_1, \ldots, X_m \in \mathbb{R}^n$ from the Gaussian mixture model with means μ and $-\mu$. If $m \geq Cn$ and $\|\mu\|_2 \geq C$, then with probability at least* 0.99*, the spectral* *clustering algorithm identifies the communities, with at most* 1% *of misclassified points.*

Now it's your turn! Prove Theorem 4.7.5 in Exercise 4.51.

It is remarkable that accurate classification is possible even when the cluster separation $\|\mu\|_2$ is much smaller than their diameter, which is $\asymp \sqrt{n}$.

4.8 Notes

The min-max theorem (Theorem 4.1.6) is also known as *Courant-Fischer-Weyl min-max principle.*

The Davis-Kahan theorem (a form of Theorem 4.1.15, Lemma 4.1.16) was originally proved in [95] and is now an indispensable tool in numerical analysis and statistics. There are numerous extensions, variants, and different proofs of this theorem, see in particular [346, 352], [41, Section VII.3], [306, Chapter V]. For random perturbations, Davis-Kahan can be improved, see [343, 120, 262, 263, 345].

Section 4.2 introduced covering numbers, packing numbers, and metric entropy. For a deeper dive, see [21, Chapter 4] and [272]; for an exposition with a view toward applications, see [280].

Section 4.3.2 covers some basic results about error correcting codes. The book [334] offers a more systematic introduction. Theorem 4.3.5 is a simplified version of the landmark *Gilbert-Varshamov bound* on the rate of error correcting codes, and the converse result (Exercise 4.33) is a simplified version of the *Hamming bound*.

In Section 4.5 we gave an application of random matrix theory to networks. For a comprehensive introduction into the interdisciplinary area of network analysis, see e.g. the book [257]. *Stochastic block models* (Definition 4.5.1) were introduced in [163]. The *community detection problem* in stochastic block models has attracted a lot of attention: see the book [257], the surveys [1, 126], papers including [229, 353, 254, 153, 2, 48, 86, 206, 151, 175, 120] and many others.

In Section 4.7 we discussed the *covariance estimation problem*; more general results will appear in Sections 5.6 and 9.2.3. This topic has been extensively studied in high-dimensional statistics, see e.g. [340, 284, 190, 69, 211, 82, 320, 238, 4, 261].

In Section 4.7.1 we gave an application to the clustering of *Gaussian mixture models*. This problem has been studied in statistics and computer science communities, see [250, Chapter 6] and [181, 251, 33, 165, 19, 146, 212].

The *power method* (Exercise 4.6), also called *von Mises iteration*, can also be used to find the top eigenvector. *Schur bound*, also known as Schur test (Exercise 4.8), was originally proved by I. Schur [298, p. 6].

The *Walsh matrices* (Exercise 4.9) are a special case of *Hadamard matrices* – orthogonal matrices with ±1 entries. The exercisebuilds Hadamard matrices for all sizes 2^k, and it is conjectured they exist for all $4k$.

The equation that connects the norms of the difference and the product of orthogonal projections (Exercise 4.12) is sometimes called *Krein-Krasnoselskii-Milman formula*, attributing it to the work [194]. This formula is explicitly proved in the work of Wedin [347]. A geometric proof is given in [264] and a linear-algebraic proof can be found in [239, p.454]; the hint for Exercise 4.12 points to this latter proof.

Wedin theorem (a form of Exercise 4.15) was established in [346].

Hermitian dilation (Exercise 4.14) is a simple but powerful trick with many applications [306, Chapter I, Section 4].

The *cut norm* (Exercise 4.21) is important in theoretical computer science [16], combinatorics [173], numerical approximation [130], etc. The equivalence of the cut norm to the $\infty \to 1$ norm (Exercise 4.21) and an SDP relaxation for computing the cut norm (Exercise 4.22) are borrowed from the seminal paper by N. Alon and A. Naor [16].

The idea of the *ε-net expansion* (Exercises 4.34, 4.35) traces back to [246].

Exercise 4.48 is about direction-aware covariance estimation; for deeper results on this problem, see [3].

The bound on the covering numbers of low-rank matrices in Exercise 4.50 is due to E. Candes and Y. Plan [71].

Exercises

4.1 ☕ (SVD of the inverse) Let A be an $n \times n$ matrix with singular value decomposition $A = \sum_{i=1}^n s_i u_i v_i^\mathsf{T}$. Show that A is invertible if and only if all singular values s_i are nonzero. In this case, check that $A^{-1} = \sum_{i=1}^n s_i^{-1} v_i u_i^\mathsf{T}$.

4.2 ☕☕ (Basic properties of the operator norm)

(a) Prove that the operator norm $\|A\|$ introduced in Definition 4.1.8 is indeed a norm on the space of $m \times n$ matrices.

(b) For any matrix, show that

$$\|A^\mathsf{T}\| = \|A\|.$$

(c) For any two matrices of appropriate dimensions, show that

$$\|AB\| \le \|A\| \cdot \|B\|.$$

Find an example where the left side is zero, but the right side is not.

(d) Prove that the operator norm of any submatrix is bounded by the operator norm of the matrix.

4.3 ☕ (Operator norm: simple examples) The operator norm is usually tricky to express in terms of entries, but here are couple of exceptions.

(a) (Rank-one) For any vectors $u \in \mathbb{R}^m$ and $v \in \mathbb{R}^n$, check that

$$\|uv^\mathsf{T}\| = \|uv^\mathsf{T}\|_F = \|u\|_2 \, \|v\|_2.$$

(b) (Diagonal) For any diagonal matrix A with diagonal entries $a_1, \ldots, a_n$, check that

$$\|A\| = \max_i |a_i|.$$

4.4 ☕☕ (Operator norm vs. Frobenius norms) Let A be an $m \times n$ matrix.

(a) If A has rank r, prove that

$$\|A\| \le \|A\|_F \le \sqrt{r}\|A\|.$$

Show that both bounds are achievable for any m, n and $1 \le r \le \min(m, n)$.

(b) If Z is an isotropic random vector in $\mathbb{R}^n$, show that

$$\mathbb{E}\|AZ\|_2^2 = \|A\|_F^2.$$

(c) If B is an $k \times m$ matrix, prove that

$$\|BA\|_F \le \|B\| \, \|A\|_F.$$

(d) If A is a diagonal matrix, improve (c) to

$$\|BA\|_F \le \|B\|_{1\to 2}\,\|A\|_F,$$

where the $\|B\|_{1\to 2}$ denotes the maximal Euclidean norm of columns[14] of B.

4.5 ☕☕ (A bound on the singular values) Prove that the singular values $s_i(A)$ of any matrix A satisfy

$$s_k(A) \le \frac{\|A\|_F}{\sqrt{k}}, \quad k = 1, 2, \dots.$$

4.6 ☕☕☕ (Power method) Here is a computationally friendly way to approximate the operator norm of a matrix without computing the spectrum. Consider an $m \times n$ matrix A and a random vector $x \sim N(0, I_n)$. Show that, with probability 1, we have

$$\sqrt[k]{\|A^k x\|_2} \to \|A\| \quad \text{as } k \to \infty.$$

4.7 ☕☕☕ (Operator norm vs. the norms of columns) The operator norm of a matrix is hard to express in terms of the entries, but in this and next exercises, you will prove some useful bounds. Let A be an $m \times n$ matrix with columns $A_1, \dots, A_n$.

(a) Show that

$$\|A\| \ge \max_i \|A_i\|_2,$$

with equality when A_i are orthogonal.

(b) Prove that if $\|A\| = \|A_i\|_2$ for some i, then A_i is orthogonal to all other columns.

(c) Argue that the same is true for the rows.

4.8 ☕☕☕ (Schur bound) For an $m \times n$ matrix A with rows $A_{i:}$ and columns $A_{:j}$, prove that

$$\|A\| \le \sqrt{\max_i \|A_{i:}\|_1 \cdot \max_j \|A_{:j}\|_1}.$$

4.9 ☕☕ (Walsh matrices) Let's construct the "most spread" orthogonal matrices, whose all entries have the same absolute value. Start with the 2×2 matrix $W_1 := \left[\begin{smallmatrix} 1 & 1 \\ 1 & -1 \end{smallmatrix}\right]$ and then iteratively define the block matrix

$$W_{k+1} := \begin{bmatrix} W_k & W_k \\ W_k & -W_k \end{bmatrix}, \quad k = 1, 2, \dots$$

Thus W_2 is a 4×4 matrix, W_3 is a 8×8 matrix, and so on. These matrices, all with ± 1 entries, are called Walsh matrices. Prove that $\frac{1}{\sqrt{n}} W_k$ is an orthogonal matrix for each k.

4.10 ☕☕ (Matrices with ± 1 entries) Prove that any $n \times n$ matrix A with ± 1 entries satisfies

$$\sqrt{n} \le \|A\| \le n,$$

and both bounds are achievable for infinitely many n.

[14] We will explore the $1 \to 2$ norm and general $p \to q$ norms in Exercise 4.18, 4.19.

4.11 ☕☕ (Difference of orthogonal projections) Prove that any two orthogonal projections P and Q in $\mathbb{R}^n$ satisfy $\|P - Q\| \le 1$. And if you are up for a challenge, try the next exercise for a refined bound.

4.12 ☕☕☕☕ (Difference vs. products of orthogonal projections) For any two orthogonal projections P and Q in $\mathbb{R}^n$, prove the following.

(a) $\|P - Q\| = \max(\|(P_\perp Q\|, \|PQ_\perp\|)$, where $P_\perp = I_n - P$ and $Q_\perp = I_n - Q$.

(b) If the ranks of P and Q are different, then $\|P - Q\| = 1$.

(c) If the ranks of P and Q are the same, then $\|P - Q\| = \|P_\perp Q\| = \|PQ_\perp\|$.

4.13 ☕☕ (Davis-Kahan for spectral projections) Let us prove a version of Davis-Kahan theorem (Theorem 4.1.15) for projections on the top k eigenvectors. Consider two symmetric matrices A and B with spectral decompositions $A = \sum_i \lambda_i u_i u_i^\mathsf{T}$ and $B = \sum_i \mu_i v_i v_i^\mathsf{T}$, in which the eigenvalues are (weakly) decreasing. Prove that the spectral projections $P_A = \sum_{i=1}^k u_i u_i^\mathsf{T}$ and $P_B = \sum_{i=1}^k v_i v_i^\mathsf{T}$ satisfy

$$\|P_A - P_B\| \le \frac{2\|A - B\|}{\lambda_k - \lambda_{k+1}}.$$

4.14 ☕☕ (Hermitian dilation) Two common ways to turn any $m \times n$ matrix A into a symmetric one are: (1) use $A^\mathsf{T} A$ or AA^T, or (2) take its *Hermitian dilation*, defined as the $(m+n) \times (m+n)$ block matrix

$$H = \begin{bmatrix} 0 & A \\ A^\mathsf{T} & 0 \end{bmatrix}.$$

Hermitian dilation has several advantages: it is a linear transformation and it preserves sparsity. Let $A = \sum_i s_i u_i v_i^\mathsf{T}$ be a singular value decomposition. Prove that the only nonzero eigenvalues of H are of the form $\pm s_i$, with corresponding eigenvalues $\left[\begin{smallmatrix} u_i \\ \pm v_i \end{smallmatrix}\right]$.

4.15 ☕☕ (Wedin theorem) Davis-Kahan inequality (Theorem 4.1.15) applies for symmetric matrices. Let's prove a version for general rectangular matrices. Let A and B be $m \times n$ matrices with singular value decompositions $A = \sum_i s_i u_i v_i^\mathsf{T}$ and $B = \sum_i t_i w_i z_i^\mathsf{T}$, where the singular values are (weakly) decreasing. Prove that the projections on the top k left singular vectors $P_A = \sum_{i=1}^k u_i u_i^\mathsf{T}$ and $P_B = \sum_{i=1}^k w_i w_i^\mathsf{T}$ satisfy

$$\|P_A - P_B\| \le \frac{2\|A - B\|}{s_k - s_{k+1}}.$$

Get a similar bound for the projections on the top k right singular vectors. Also, prove a version of Davis-Kahan inequality (Theorem 4.1.15) on the k-th singular vectors.

4.16 ☕ (Angle and distance between vectors) Let's simplify the conclusion of Davis-Kahan inequality (Theorem 4.1.15). Suppose the angle between two unit vectors $u, v \in \mathbb{R}^n$ (as a number between 0 and $\pi/2$) is small, i.e.

$$\sin \angle(u, v) \le \varepsilon \quad \text{for some } \varepsilon > 0.$$

Show that u and v are close up to a sign, i.e.

$$\|u - \theta v\|_2 \le \sqrt{2}\,\varepsilon \quad \text{for some } \theta \in \{-1, 1\}.$$

4.17 ☕☕☕ (Approximate projections) Let's add one more equivalent property to Lemma 4.1.17. Let A be an $m \times n$ matrix A with $m \geq n$, and $\varepsilon \geq 0$.

(a) Prove that A is an isometry if and only if AA^{T} is an orthogonal projection in $\mathbb{R}^m$ onto a subspace of dimension n.

(b) More generally, prove that that A is an ε-approximate isometry (i.e. the three equivalent properties in Lemma 4.1.17 hold) if and only if

$$\|AA^{\mathsf{T}} - P\| \leq \varepsilon$$

for some orthogonal projection P in $\mathbb{R}^m$ with $\operatorname{rank}(P) = n$.

4.18 ☕ ($p \to q$ norm) The operator norm of an $m \times n$ matrix A, defined in (4.9), measures how much A can stretch vectors. To measure the "stretch", we used the ℓ^2 norm, any two norms work, like ℓ^p and ℓ^q for $p, q \in [1, \infty]$. Define the $\ell^p \to \ell^q$ operator norm of A, or simply the *$p \to q$ norm*, as

$$\|A\|_{p \to q} = \max_{x \neq 0} \frac{\|Ax\|_q}{\|x\|_p} = \max_{\|x\|_p = \|y\|_{q'} = 1} |y^{\mathsf{T}} A x|, \tag{4.31}$$

where q' is the conjugate exponent of q defined in (1.5).

(a) Extend the equations in (4.9) for the $p \to q$ norm and justify them.

(b) Verify that this indeed defines a norm on the space of $m \times n$ matrices.

(c) (Duality) Prove that $\|A^{\mathsf{T}}\|_{p \to q} = \|A\|_{q' \to p'}$ where p' and q' are the conjugate exponents of p and q.

4.19 ☕☕ ($1 \to \infty$ and $1 \to 2$ norms) Let A be an $m \times n$ matrix. For most p and q, it is hard to express $\|A\|_{p \to q}$ in terms of the entries, but here are some exceptions.

(a) Prove that the $1 \to \infty$ norm equals the maximal absolute value of the entries:

$$\|A\|_{1 \to \infty} = \max_{i,j} |A_{ij}|.$$

(b) Prove that the $1 \to 2$ norm equals the maximal Euclidean norm of the columns $A_{:j}$, and the $2 \to \infty$ norm equals the maximal Euclidean norm of the rows $A_{i:}$:

$$\|A\|_{1 \to 2} = \max_{j \leq n} \left\|A_{:j}\right\|_2 \quad \text{and} \quad \|A\|_{2 \to \infty} = \max_{i \leq m} \|A_{i:}\|_2.$$

4.20 ☕☕☕ ($\infty \to 1$ norm) We have seen this norm before! It was implicit in the assumption of Grothendieck inequality (Theorem 3.5.1). Let A be an $m \times n$ matrix.

(a) Prove that

$$\|A\|_{\infty \to 1} = \max_{\substack{x \in \{-1,1\}^m \\ y \in \{-1,1\}^n}} |x^{\mathsf{T}} A y|. \tag{4.32}$$

(b) (Duality) Prove that

$$\|A\|_{\infty \to 1} = \sup_{\|Z\|_\infty \leq 1,\, \operatorname{rank}(Z) = 1} |\langle A, Z \rangle|$$

where the supremum is over all rank-one $m \times n$ matrices Z whose all entries are bounded by 1 in absolute value, and the matrix inner product is defined in (4.7).

(c) Demonstrate that the identity in (4.32) does not generalize to quadratic forms: find an $n \times n$ symmetric matrix A for which

$$\max_{x\in[-1,1]^n} |x^\mathsf{T} Ax| \neq \max_{x\in\{-1,1\}^n} |x^\mathsf{T} Ax|.$$

4.21 ☕☕☕ (Cut norm) This norm is important in theoretical computer science and graph theory. To find the cut norm of an $m \times n$ matrix A, we sum the entries of each submatrix of A, take the absolute value, and maximize over all submatrices. Formally, we define

$$\|A\|_{\text{cut}} = \max_{I,J} \Big| \sum_{i\in I,\, j\in J} A_{ij} \Big|,$$

where the maximum is over all subsets $I \subset \{1,\dots,m\}$ and $J \subset \{1,\dots,n\}$. Prove that the cut norm is equivalent to the $\infty \to 1$ norm up to a constant factor:

$$\|A\|_{\text{cut}} \leq \|A\|_{\infty\to1} \leq 4\|A\|_{\text{cut}}.$$

4.22 ☕ (SDP relaxations) Let A be an $m \times n$ matrix. Argue that $\|A\|_{\infty\to1}$, the cut norm, $\|A\|_{\infty\to2}$, and the cut norm of A (Exercise 4.21) can be approximated up to an absolute constant factor by solving a semidefinite program.

4.23 ☕ (Transitivity of ε-nets) If $\mathcal{N}$ is an ε-net of $\mathcal{M}$ and $\mathcal{M}$ is an δ-net of K, show that $\mathcal{N}$ is an $(\varepsilon+\delta)$-net of K.

4.24 ☕☕☕ (Packing balls) In Remark 4.2.5, we noted that if $\mathcal{N}$ is an ε-separated subset of a metric space (T,d), then the closed $\varepsilon/2$-balls centered at points in $\mathcal{N}$ are disjoint.

(a) Show that the converse is false in a general metric space (T,d),

(b) but is true in any normed space (T,d).

4.25 ☕☕☕ (External covering numbers) In defining covering numbers of K, we required that the ball centers x_i lie in K. Relaxing this, the *external covering number* $\overline{\mathcal{N}}(K,d,\varepsilon)$ allows centers outside K. Prove that the usual and external covering numbers are essentially equivalent:

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

4.26 ☕☕☕ (Monotonicity of covering numbers) Although the covering numbers are monotone in ε (why?), they are not monotone in K.

(a) Show by example that, in general,

$$L \subset K \quad \not\Longrightarrow \quad \mathcal{N}(L,d,\varepsilon) \leq \mathcal{N}(K,d,\varepsilon).$$

(b) Prove an approximate version of monotonicity:

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

4.27 ☕☕ (Volume of the Euclidean ball: a geometric argument) In this and next two exercises, we compute the volume of the unit Euclidean ball in $\mathbb{R}^n$ in three different ways: geometric, probabilistic, analytic. Let's start with geometric.

(a) The *canonical simplex* Δ_n, shaded in Figure 4.6, consists of all points in $\mathbb{R}^n$ with nonnegative coordinates that sum to at most 1. Show that

$$\mathrm{Vol}(\Delta_n) = \frac{1}{n!}$$

(b) Recall that the unit ℓ^1 ball in $\mathbb{R}^n$ is $B_1^n \coloneqq \{x \in \mathbb{R}^n : \|x\|_1 \le 1\}$. Partition it into simplices as in Figure 4.6 to show that

$$\mathrm{Vol}(B_1^n) = \frac{2^n}{n!} \le \left(\frac{2e}{n}\right)^n.$$

(c) Sandwich the unit Euclidean ball B_2^n between the appropriately scaled cube and the ℓ^1 ball as in Figure 4.6 to conclude that

$$\left(\frac{2}{\sqrt{n}}\right)^n \le \mathrm{Vol}(B_2^n) \le \left(\frac{2e}{\sqrt{n}}\right)^n. \tag{4.33}$$

So, the volume of the unit ball in high dimensions is *exponentially small*! Even though there is some slack in the bounds, they are good enough for most purposes.

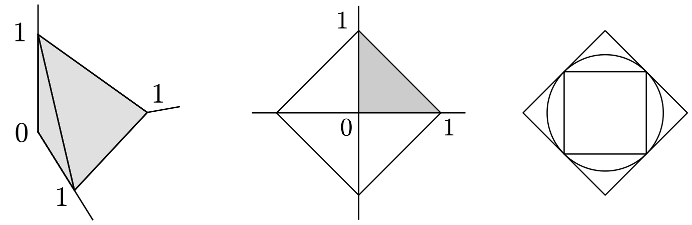


Figure 4.6 The shaded shapes are the canonical simplices in $\mathbb{R}^3$ and $\mathbb{R}^2$.

4.28 ☕☕ (Volume of the Euclidean ball: a probabilistic argument) Let us tighten up the upper bound in (4.33). Deduce from the small ball probability (Exercise 3.7) that

$$\mathrm{Vol}(B_2^n) \le \left(\sqrt{\frac{2\pi e}{n}}\right)^n.$$

This bound is asymptotically sharp, as we will see from the next exercise.

4.29 ☕☕☕ (Volume of the Euclidean ball: an analytic argument) Finally, let us compute the volume of the Euclidean ball *exactly*.

(a) Let $\|\cdot\|$ be a norm on $\mathbb{R}^n$. Consider the unit ball of this normed space, $B \coloneqq \{x \in \mathbb{R}^n : \|x\| \le 1\}$. Let $f : \mathbb{R}^+ \to \mathbb{R}^+$ be a decreasing, differentiable function satisfying $\lim_{x\to+\infty} f(x) = 0$. Check the following identity:

$$\int_{\mathbb{R}^n} f\left(\|x\|\right) dx = -\,\mathrm{Vol}(B) \int_0^\infty t^n f'(t) dt.$$

(b) Substitute $f(t) = e^{-t^2/2}$ and conclude that

$$\text{Vol}\left(B_2^n\right) = \frac{\pi^{n/2}}{\Gamma(n/2+1)}.$$

(c) Denote by R_n the radius of a Euclidean ball with volume 1. Show that

$$R_n = \sqrt{\frac{n}{2\pi e}}\,(1+o(1)) \quad \text{as } n \to \infty.$$

The fact that R_n is large indicates that the volume of the unit ball B_2^n is small.

4.30 ☕☕☕ (Volume of the ℓ^p ball) Let $p \in [1, \infty]$.

(a) Repeating the geometric argument from Exercise 4.27, deduce the following bounds:

$$\left(\frac{2}{n^{1/p}}\right)^n \le \text{Vol}(B_p^n) \le \left(\frac{2e}{n^{1/p}}\right)^n.$$

(b) Using the analytic argument from Exercise 4.29, compute the volume exactly:

$$\text{Vol}(B_p^n) = \frac{(2\Gamma(1/p+1))^n}{\Gamma(n/p+1)}.$$

4.31 ☕☕☕ (A lattice is an ε-net) The construction of an ε-net (Remark 4.2.7) is not very efficient, since nets tend to be exponentially large and hard to store. Instead, let $\mathcal{N}$ be the set of points in the scaled integer lattice $\frac{\varepsilon}{\sqrt{n}}\mathbb{Z}^n$ that lie inside the unit ball B_2^n. Check that $\mathcal{N}$ forms an ε-net of B_2^n in the Euclidean metric, and

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

This bound nearly matches the one in Corollary 4.2.11. Also mention how to quickly approximate any given vector in the unit ball by a vector in $\mathcal{N}$.

4.32 ☕ (Covering and packing numbers of the Hamming cube) Prove Proposition 4.2.15 by adapting the volumetric method.

4.33 ☕☕☕ (A limitation of error correcting codes) Let us prove a converse to Theorem 4.3.5. For any error correcting code that encodes k-bit strings into n-bit strings and can correct r errors, prove that the

$$n - k \ge r \log_2\left(\frac{n}{r}\right).$$

4.34 ☕☕ (An ε-net expansion) ε-nets help approximate vectors – here is how to use them for *exact* representation.

(a) Let $\mathcal{N}$ be an ε-net of the unit sphere S^{n-1} of $\mathbb{R}^n$. Show that any vector $x \in S^{n-1}$ can be written as a convergent series

$$x = \sum_{k=0}^{\infty} \lambda_k x_k \quad \text{for some coefficients } 0 \le \lambda_k \le \varepsilon^k.$$

(b) Use the ε-net expansion to give an alternative proof of Lemma 4.4.1.

4.35 ☕ (Computing the norm on an ε-net) Let $x \in \mathbb{R}^n$ and $\mathcal{N}$ be an ε-net of the sphere S^{n-1}. Show that

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

4.36 ☕☕☕ (Maximizing quadratic forms on an ε-net) Prove Lemma 4.4.2 by modifying the proof of Lemma 4.4.1.

4.37 ☕☕☕ (Deviation of the norm on an ε-net) 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

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

4.38 ☕☕ (The convex hull of an ε-net) Let's show that an ε-net of a sphere "absorbs" a slightly smaller sphere, see Figure 4.7. Let $\mathcal{N}$ be an ε-net of the unit sphere S^{n-1} for some $\varepsilon \in (0, 1)$. Prove that

$$(1-\varepsilon) B_2^n \subset \operatorname{conv}(\mathcal{N}).$$

Here as ususal B_2^n denotes the unit Euclidean ball in $\mathbb{R}^n$, and $\operatorname{conv}(\cdot)$ denotes the convex hull introduced in the Appetizer.

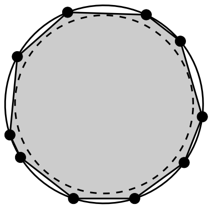


Figure 4.7 The convex hull of an ε-net of a sphere contains a slightly smaller (dashed) sphere (Exercise 4.38).

4.39 ☕☕☕☕ (Random points form an ε-net) Let $g_1, \ldots, g_N$ be independent $N(0, I_n)$ random vectors. For any $R, \varepsilon > 0$, show that with probability at least $1 - e^{-cn}$, the set

$$\left\{ \frac{g_1}{\sqrt{n}}, \ldots, \frac{g_N}{\sqrt{n}} \right\} \cap RB_2^n$$

forms an ε-net of RB_2^n (the centered ball of radius R), as long as $N \ge e^{C(R,\varepsilon)n}$. Here $C(R, \varepsilon)$ is allowed to depend only on R and ε.

4.40 ☕☕ (Too many random points are not in convex position) In Exercise 3.23 we saw that independent Gaussian random vectors $g_1, \ldots, g_N \sim N(0, I_n)$ are in convex position with high probability if $N \le e^{cn}$. Show the converse: with probability at least $1 - e^{-cn}$, these vectors are *not* in convex position if $N \ge e^{Cn}$.

4.41 ☕☕☕ (Expected operator norm of a random matrix)

(a) Deduce from Theorem 4.4.3 that

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

(b) Deduce from Theorem 4.6.1 that

$$\mathbb{E}\Big\|\frac{1}{m}A^{\mathsf{T}}A - I_n\Big\| \le CK^2\Big(\sqrt{\frac{n}{m}}+\frac{n}{m}\Big).$$

and

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

4.42 ☕☕ (Norm of random matrices: a lower bound) Let's prove a matching lower bound for Theorem 4.4.3. Let A be an $m \times n$ random matrix whose entries A_{ij} are independent, subgaussian random variables satisfying $\mathbb{E}\, A_{ij}^2 = 1$. For any $t > 0$, show that

$$\|A\| \ge \frac{1}{2}\left(\sqrt{m}+\sqrt{n}-t\right)$$

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

4.43 ☕☕ (Upper bounds on subgaussian matrices: relaxing independence)

(a) Show that Theorem 4.4.3 holds without any independence assumptions if A is a subgaussian matrix (meaning that A is a subgaussian random vector in $\mathbb{R}^{m\times n}$, see Definition 3.4.1), with $K = \|A\|_{\psi_2}$.

(b) In particular, show that Theorem 4.4.3 holds if A has independent, mean-zero, subgaussian rows (or columns) A_i, with $K = \max_i \|A_i\|_{\psi_2}$.

4.44 ☕☕ (Some $p \to q$ norms of random matrices) In Theorem 4.4.3, we looked at the $2 \to 2$ operator norm of random matrices. But what about other $p \to q$ operator norms? (We defined them in Exercise 4.18 and gave some examples in Exercises 4.19, 4.20, 4.21.) Let A be an $m\times n$ matrix with independent, mean-zero, variance-one subgaussian entries A_{ij}. Denote $K = \max_{ij}\|A_{ij}\|_{\psi_2}$.

(a) Show that

$$\|A\|_{1\to\infty} \le CK\left(\sqrt{\log m}+\sqrt{\log n}\right).$$

(b) Show that

$$\|A\|_{1\to 2} = \|A^{\mathsf{T}}\|_{2\to\infty} \le \sqrt{m} + CK^2\sqrt{\log n}.$$

Later, with more advanced tools, we will be able to handle all $p \to q$ norms (see Exercise 8.41).

(c) Let's check that the bounds in (a) and (b) are essentially optimal. If A has independent standard normal random entries, check that

$$\|A\|_{1\to\infty} \ge c\left(\sqrt{\log m}+\sqrt{\log n}\right) \quad \text{and} \quad \|A\|_{1\to 2} \ge c\left(\sqrt{m}+\sqrt{\log n}\right).$$

4.45 ☕☕☕☕ (Community detection) The guarantee of the spectral clustering algorithm for community detection (Theorem 4.5.2) assumes not only that $p - q$ is not too small, but also q, the probability of across-community edges, is not too small. Let's remove the latter assumption. Design a version of the spectral clustering algorithm that results in at most $C/(p-q)^2$ misclassified vertices.

4.46 ☕☕ (An alternative proof of two-sided bound on random matrices) Give a simpler proof of Theorem 4.6.1, using Theorem 3.1.1 to obtain a concentration bound for $\|Ax\|_2$ and Exercise 4.37 to reduce to a union bound over a net.

4.47 ☕☕ (Intermediate singular values of random matrices) Deduce from Theorem 4.6.1 the following bound on intermediate singular values. For any fixed $1 \le k \le n$ and $t \ge 0$,

$$s_k(A) \ge \sqrt{m} - CK^2(\sqrt{k} + t)$$

with probability at least $1 - 2\exp(-t^2)$.

4.48 ☕☕ (Covariance estimation with a relative error) Let's give a more sensitive version of covariance estimation (Theorem 4.7.1) with relative, rather than absolute, error. Show that

$$\mathbb{E} \sup_{v \in \mathbb{R}^n} \left| \frac{v^\mathsf{T} \Sigma_m v}{v^\mathsf{T} \Sigma v} - 1 \right| \le CK^2 \Big(\sqrt{\frac{n}{m}} + \frac{n}{m} \Big),$$

assuming Σ is invertible. In other words, we get a uniform, *relative* approximation of the variance of 1D marginals of X (recall (3.6)).

4.49 ☕ (Covariance estimation with high probability) Check the high-probability guarantee on the covariance estimation mentioned in Remark 4.7.3.

4.50 ☕☕☕ (Covering numbers of low-rank matrices) An $m \times n$ matrix of rank r can be described with $(m+n+1)r$ parameters via its SVD: mr for the left singular vectors, nr for the right singular vectors, and r for the singular values.[15] Since covering numbers are typically exponential in the dimension (see Corollary 4.2.11), we can guess that the covering numbers of the set

$$M_{m,n,r} = \{m \times n \text{ matrices } A \text{ of rank } r \text{ and } \|A\|_F \le 1\}$$

is exponential in $(m+n+1)r$. This is indeed the case!

(a) Consider the set $O_{m,r}$ consisting of $m \times r$ matrices with orthonormal columns. Prove[16] for every $\varepsilon > 0$:

$$\mathcal{N}(O_{m,r}, \|\cdot\|_{1\to2}, \varepsilon) \le \Big(\frac{C}{\varepsilon}\Big)^{mr}.$$

(b) Prove for every $\varepsilon > 0$:

$$\mathcal{N}(M_{m,n,r}, \|\cdot\|_F, \varepsilon) \le \Big(\frac{C}{\varepsilon}\Big)^{(m+n+1)r}.$$

[15] This is only an upper bound; what is the exact number of parameters determining such a matrix?

[16] Recall that the 1 → 2 norm of a matrix is the maximal Euclidean norm of columns (Exercise 4.19(b)).

(c) Conversely, show for every $\varepsilon > 0$:

$$\mathcal{N}(M_{m,n,r}, \|\cdot\|_F, \varepsilon) \geq \left(\frac{c}{\varepsilon}\right)^{\frac{1}{2}(m+n)r}.$$

4.51 ☕☕☕☕ (Spectral clustering of the Gaussian mixture model) Prove Theorem 4.7.5, which guarantees that spectral clustering learns the Gaussian mixture model. Here is how:

(a) Compute the covariance matrix Σ of X and note that the top eigenvector u is collinear with μ.

(b) Conclude that the signs of $\langle X_i, u\rangle$ classify most points X_i correctly.

(c) Use results about covariance estimation to show that the sample covariance matrix Σ_m approximates Σ.

(d) Use the Davis-Kahan inequality (Theorem 4.1.15) to deduce that the top eigenvector $v = v_1(\Sigma_m)$ approximates u.

(e) Using (b), conclude that the signs of $\langle X_i, v\rangle$ classify most points X_i correctly.

5

Concentration Without Independence

So far, our approach to concentration inequalities relied crucially on independence of random variables. Now, we will explore other approaches to concentration that do not rely on independence. In Section 5.1, we introduce an isoperimetric approach using the Euclidean sphere as an example, then cover other settings in Section 5.2.

In Section 5.3, we use concentration on the sphere to derive the classical Johnson-Lindenstrauss Lemma, a key result on dimension reduction for high-dimensional data.

Section 5.4 introduces matrix concentration inequalities, focusing on the matrix Bernstein inequality, which extends the classical Bernstein inequality to random matrices. We then apply it in Sections 5.5 and 5.6, extending our analysis of community detection and covariance estimation problems to sparse networks and more general distributions in $\mathbb{R}^n$.

Don't miss the exercises! We explore dimension reduction with binary coins in Exercise 5.14, matrix calculus in Exercise 5.16–5.19, build various matrix concentration inequalities in Exercises 5.20–5.24 and apply them for matrix sketching (Exercise 5.32), community detection (Exercise 5.25), and more.

5.1 Concentration of Lipschitz functions on the sphere

For a random vector X in $\mathbb{R}^n$ and a function $f : \mathbb{R}^n \to \mathbb{R}$, when does the random variable $f(X)$ concentrate, i.e.

$$f(X) \approx \mathbb{E}\, f(X) \quad \text{with high probability?}$$

If X is normal and f is *linear*, this is easy: $f(X)$ is normal (Corollary 3.3.2) and concentrates well (Proposition 2.1.2).

What about *non-linear* functions f? We can't expect good concentration for just any f (why?), but if f does not oscillate too wildly, we might expect concentration. To make this precise, we introduce Lipschitz functions – they help rule out extreme oscillations.

[0] This material will be published by Cambridge University Press as *High-Dimensional Probability, 2nd Edition* by Roman Vershynin. This pre-publication version is free to view and download for personal use only. Not for re-distribution, re-sale or use in derivative works. ©Roman Vershynin 202X.
If you want to be notified when the printed version is available, let me know: rvershyn@uci.edu

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

$$d_Y(f(u), f(v)) \le L \cdot d_X(u, v) \quad \text{for every } u, v \in X.$$

The infimum of all L in this definition is called the *Lipschitz norm*[1] because of f and is denoted $\|f\|_{\mathrm{Lip}}$.

In other words, Lipschitz functions don't stretch distances too much. When $\|f\|_{\mathrm{Lip}} \le 1$, they are *contractions* since they can only shrink distances. The class of Lipschitz functions sits between differentiable and uniformly continuous:

$$f \text{ is differentiable} \Rightarrow f \text{ is Lipschitz} \Rightarrow f \text{ is uniformly continuous},$$

and in Exercise 5.1 you will even quantify the first implication for $f : \mathbb{R}^n \to \mathbb{R}$:

$$\|f\|_{\mathrm{Lip}} \le \sup_{x \in \mathbb{R}^n} \|\nabla f(x)\|_2.$$

Example 5.1.2. Vectors, matrices and norms define natural Lipschitz functions:

(a) For a fixed vector $\theta \in \mathbb{R}^n$, the linear functional

$$f(x) = \langle x, \theta \rangle \quad \text{has Lipschitz norm} \quad \|f\|_{\mathrm{Lip}} = \|\theta\|_2.$$

(b) More generally, any $m \times n$ matrix A, the linear operator[2]

$$f(x) = Ax \quad \text{has Lipschitz norm} \quad \|f\|_{\mathrm{Lip}} = \|A\|.$$

(c) For any norm $\|\cdot\|$ on $\mathbb{R}^n$, the function

$$f(x) = \|x\|$$

has Lipschitz norm equal to the smallest L such that

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

These claims should be easy to verify – try it in Exercise 5.2.

5.1.2 Concentration via isoperimetric inequalities

We will now prove that any Lipschitz function on the Euclidean sphere $S^{n-1} = \{x \in \mathbb{R}^n : \|x\|_2 = 1\}$ concentrates:

Theorem 5.1.3 (Concentration of Lipschitz functions on the sphere)**.** *Let X be a random vector uniformly distributed on the Euclidean sphere of radius $\sqrt{n}$, i.e. $X \sim \mathrm{Unif}(\sqrt{n}S^{n-1})$. Then for any Lipschitz function*[3] *$f : \sqrt{n}S^{n-1} \to \mathbb{R}$ we have*

$$\|f(X) - \mathbb{E} f(X)\|_{\psi_2} \le C\|f\|_{\mathrm{Lip}}.$$

[1] Technically, $\|f\|_{\mathrm{Lip}}$ is only a seminorm, since it vanishes on nonzero constant functions – but we will call it a norm for brevity.

[2] Here we consider the linear operator as a map from $(\mathbb{R}^m, \|\cdot\|_2)$ to $(\mathbb{R}^n, \|\cdot\|_2)$

[3] This theorem works for both the geodesic metric (shortest arc length) and the Euclidean metric $d(x, y) = \|x - y\|_2$. We will prove it for the Euclidean case – try Exercise 5.4 to extend it to geodesic distance.

By the definition of the subgaussian norm, Theorem 5.1.3 can be written as

$$\mathbb{P}\{|f(X) - \mathbb{E}\, f(X)| \geq t\} \leq 2\exp\Big(-\frac{ct^2}{\|f\|_{\mathrm{Lip}}^2}\Big) \quad \text{for any } t \geq 0.$$

We already proved Theorem 5.1.3 for *linear* functions. Theorem 3.4.5 tells us that X is a subgaussian random vector, and this by definition means that any linear function of X is a subgaussian random variable.

To fully prove Theorem 5.1.3, we need to argue that any Lipschitz function concentrates at least as well as a linear function. Instead of comparing them directly, we will compare the areas of their *sublevel sets* – regions of the sphere where $f(x) \leq a$ for a given level a. For linear functions, these regions are just spherical caps. To compare the areas of general sets and spherical caps, we can use a remarkable geometric principle – the *isoperimetric inequality.*

The most familiar form of the isoperimetric inequality is for subsets of $\mathbb{R}^3$ (and also in $\mathbb{R}^n$):

Theorem 5.1.4 (Isoperimetric inequality on $\mathbb{R}^n$)**.** *Among all subsets $A \subset \mathbb{R}^n$ with given volume, the Euclidean balls have minimal area. Moreover, for any $\varepsilon > 0$, the Euclidean balls minimize the volume of the ε-neighborhood of A, defined as*[4]

$$A_\varepsilon := \{x \in \mathbb{R}^n : \exists y \in A \text{ such that } \|x - y\|_2 \leq \varepsilon\} = A + \varepsilon B_2^n.$$

Figure 5.1 illustrates the isoperimetric inequality. Note that the "moreover" part of Theorem 5.1.4 implies the first part; to see this, let $\varepsilon \to 0$.

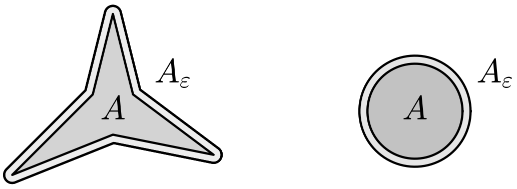


Figure 5.1 The isoperimetric inequality says that among all sets A with a given volume, Euclidean balls minimize the volume of their ε-neighborhood A_ε.

A similar isoperimetric inequality holds for subsets of the sphere S^{n-1}, and in this case the minimizers are the *spherical caps* – neighborhoods of a single point.[5] To state this principle, we use σ_{n-1} to denote the normalized area on the sphere S^{n-1} (the $n-1$-dimensional Lebesgue measure).

Theorem 5.1.5 (Isoperimetric inequality on the sphere)**.** *Let $\varepsilon > 0$. Then, among all sets $A \subset S^{n-1}$ with given area $\sigma_{n-1}(A)$, the spherical caps minimize the area of the neighborhood $\sigma_{n-1}(A_\varepsilon)$, where*

$$A_\varepsilon := \left\{x \in S^{n-1} : \exists y \in A \text{ such that } \|x - y\|_2 \leq \varepsilon\right\}.$$

[4] Here we used the notation for Minkowski sum introduced in Definintion 4.2.9.

[5] More formally, a closed spherical cap centered at a point $a \in S^{n-1}$ and with radius ε is $C(a, \varepsilon) = \{x \in S^{n-1} : \|x - a\|_2 \leq \varepsilon\}$.

We do not prove isoperimetric inequalities (Theorems 5.1.4 ans 5.1.5) in this book; the bibliography notes for this chapter refer to several known proofs of these results.

5.1.3 Blow-up of sets on the sphere

The isoperimetric inequality leads to a remarkable and counter-intuitive result: if a set A covers at least *half* of the sphere in area, its ε-neighborhood A_ε will cover *most* of the sphere. We will state and prove this "blow-up" phenomenon and then try to explain it intuitively. To simplify things in view of Theorem 5.1.3, we will work with the sphere of radius $\sqrt{n}$ instead of the unit sphere.

Lemma 5.1.6 (Blow-up)**.** *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,*[6] *for every $t \geq 0$,*

$$\sigma(A_t) \geq 1 - 2\exp(-ct^2).$$

Proof Consider the hemisphere defined by the first coordinate:

$$H := \left\{x \in \sqrt{n}S^{n-1} :\ x_1 \leq 0\right\}.$$

By assumption, $\sigma(A) \geq 1/2 = \sigma(H)$, so the isoperimetric inequality (Theorem 5.1.5) implies that

$$\sigma(A_t) \geq \sigma(H_t). \tag{5.1}$$

The neighborhood H_t of the hemisphere H is a spherical cap, and we could compute its area directly. It is, however, easier to use Theorem 3.4.5 instead, which states a random vector

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

is subgaussian, and $\|X\|_{\psi_2} \leq C$. Since σ is the uniform probability measure on the sphere, it follows that

$$\sigma(H_t) = \mathbb{P}\{X \in H_t\}.$$

Now, the definition of the neighborhood implies that

$$H_t \supset \left\{x \in \sqrt{n}S^{n-1} :\ x_1 \leq t/\sqrt{2}\right\}. \tag{5.2}$$

(Check this – a picture would help.) Thus

$$\sigma(H_t) \geq \mathbb{P}\{X_1 \leq t/\sqrt{2}\} \geq 1 - 2\exp(-ct^2).$$

The last inequality holds because $\|X_1\|_{\psi_2} \leq \|X\|_{\psi_2} \leq C$. In view of (5.1), the lemma is proved. □

Remark 5.1.7 (An even more dramatic blow-up)**.** The 1/2 value for the area in Lemma 5.1.6 was arbitrary and can be replaced with any constant, or even an exponentially small quantity! Verify this is Exercise 5.3.

[6] Here the neighborhood A_t of a set A is defined in the same way as before, that is $A_t := \left\{x \in \sqrt{n}S^{n-1} :\ \exists y \in A \text{ such that } \|x-y\|_2 \leq t\right\}$.

Remark 5.1.8 (A zero-one law)**.** The blow-up phenomenon we just saw may be quite counter-intuitive at first sight. How can an exponentially small set A change so dramatically to an exponentially large set A_{2t} under just a tiny perturbation $2t$? (Remember t can be much smaller than the radius $\sqrt{n}$ of the sphere.) However perplexing it seems, this is a typical phenomenon in high dimensions. It is similar to *zero-one laws* in probability theory, which basically say that events influenced by many random variables tend to have probabilities either zero or one.

5.1.4 Proof of Theorem 5.1.3

Without loss of generality, we can assume that $\|f\|_{\mathrm{Lip}} = 1$. (Why?) Let M denote a median of $f(X)$, which by definition is a number satisfying[7]

$$\mathbb{P}\{f(X) \le M\} \ge \frac{1}{2} \quad \text{and} \quad \mathbb{P}\{f(X) \ge M\} \ge \frac{1}{2}.$$

Consider the sublevel set

$$A := \left\{x \in \sqrt{n}S^{n-1} : \; f(x) \le M\right\}.$$

Since $\mathbb{P}\{X \in A\} \ge 1/2$, Lemma 5.1.6 implies that

$$\mathbb{P}\{X \in A_t\} \ge 1 - 2\exp(-ct^2). \tag{5.3}$$

On the other hand, we claim that

$$\mathbb{P}\{X \in A_t\} \le \mathbb{P}\{f(X) \le M + t\}. \tag{5.4}$$

Indeed, if $X \in A_t$ then $\|X - y\|_2 \le t$ for some point $y \in A$. By definition, $f(y) \le M$. Since f Lipschitz with $\|f\|_{\mathrm{Lip}} = 1$, it follows that

$$f(X) \le f(y) + \|X - y\|_2 \le M + t.$$

This proves our claim (5.4).

Combining (5.3) and (5.4), we conclude that

$$\mathbb{P}\{f(X) \le M + t\} \ge 1 - 2\exp(-ct^2).$$

Repeating the argument for $-f$, we obtain a similar bound for the probability that $f(X) \ge M - t$. (Do this!) Combining the two, we obtain a similar bound for the probability that $|f(X) - M| \le t$, and conclude that

$$\|f(X) - M\|_{\psi_2} \le C.$$

It remains to replace the median M by the mean $\mathbb{E}\, f$, which follows by centering (use Lemma 2.7.8 – check!) The proof of Theorem 5.1.3 is complete. □

We just derived concentration from a blow-up phenomenon. Try Exercise 5.7 to see that these two are generally equivalent.

[7] If the median is not unique, just take any median.

5.2 Concentration on other metric measure spaces

We'll now extend concentration from the sphere to other spaces. Our proof of Theorem 5.1.3 relied on two ingredients:

(a) an isoperimetric inequality,
(b) a blow-up of its minimizers.

These are not unique to the sphere – many spaces satisfy them, leading to similar concentration results. We'll cover two key examples: Gaussian concentration in $\mathbb{R}^n$ and concentration on the Hamming cube, then briefly mention other cases.

Remark 5.2.1 (Mean, median, L^p norm – take your pick)**.** Concentration keeps the mean, median, and L^p norms close. So, we can always replace the mean $\mathbb{E}\, f(X)$ with the median (Exercise 5.6) or, if the mean is nonnegative, with the L^p norm for any $p \geq 1$, though the constant may depend on p (see Exercise 5.10).

5.2.1 Gaussian concentration

The classical isoperimetric inequality in $\mathbb{R}^n$, Theorem 5.1.4, holds not only with respect to the volume but also with respect to the *Gaussian measure* on $\mathbb{R}^n$. The Gaussian measure of a (Borel) set $A \subset \mathbb{R}^n$ is defined as[8]

$$\gamma_n(A) := \mathbb{P}\{X \in A\} = \frac{1}{(2\pi)^{n/2}} \int_A e^{-\|x\|_2^2/2}\, dx$$

where $X \sim N(0, I_n)$ is the standard normal random vector in $\mathbb{R}^n$.

Theorem 5.2.2 (Gaussian isoperimetric inequality)**.** *Let $\varepsilon > 0$. Then, among all sets $A \subset \mathbb{R}^n$ with given Gaussian measure $\gamma_n(A)$, the half-spaces minimize the Gaussian measure of the neighborhood $\gamma_n(A_\varepsilon)$.*

With the same method as we developed for the sphere, we can then deduce the following Gaussian concentration inequality (see Exercise 5.8):

Theorem 5.2.3 (Gaussian concentration)**.** *Consider a random vector $X \sim N(0, I_n)$ and a Lipschitz function $f : \mathbb{R}^n \to \mathbb{R}$ (with respect to the Euclidean metric). Then*

$$\|f(X) - \mathbb{E}\, f(X)\|_{\psi_2} \leq C\|f\|_{\mathrm{Lip}}. \tag{5.5}$$

Example 5.2.4. Two special cases of Theorem 5.2.3 should already be familiar:

(a) For linear functions f, it follows since $X \sim N(0, I_n)$ is subgaussian.
(b) For the Euclidean norm $f(x) = \|x\|_2$, it follows from norm concentration (Theorem 3.1.1).

To practice more with Gaussian concentration, try Exercise 5.9 to prove concentration for the maximum of n gaussians.

[8] Recall the definition of the standard normal distribution in $\mathbb{R}^n$ from Section 3.3.1.

5.2.2 Hamming cube

The method based on isoperimetry also works for concentration for the Hamming cube

$$(\{0,1\}^n, d, \mathbb{P}),$$

(recall Definition 4.2.14), where $d(x,y)$ is the *normalized* Hamming distance, measuring the fraction of the digits where the binary strings x and y differ:

$$d(x,y) = \frac{1}{n}|\{i : x_i \neq y_i\}|.$$

The measure $\mathbb{P}$ is the uniform probability measure on the cube:

$$\mathbb{P}(A) = \frac{|A|}{2^n} \quad \text{for any } A \subset \{0,1\}^n.$$

Theorem 5.2.5 (Concentration on the Hamming cube)**.** *Consider a random vector $X \sim \mathrm{Unif}\{0,1\}^n$. (Thus, the coordinates of X are independent* Ber(1/2) *random variables.) Then for any function $f : \{0,1\}^n \to \mathbb{R}$ we have*

$$\|f(X) - \mathbb{E}\, f(X)\|_{\psi_2} \leq \frac{C\|f\|_{\mathrm{Lip}}}{\sqrt{n}}. \tag{5.6}$$

This result follows from the isoperimetric inequality on the Hamming cube, whose minimizers are known to be the *Hamming balls* – neighborhoods of single points with respect to the Hamming distance. (Try deducing it!)

5.2.3 Symmetric group

A similar results holds for the symmetric group S_n, a set of all $n!$ permutations of n symbols $\{1,\dots,n\}$. We can view the symmetric group as a metric measure space

$$(S_n, d, \mathbb{P}).$$

Here $d(\pi,\rho)$ is the normalized Hamming distance – the fraction of the symbols on which permutations π and ρ differ:

$$d(\pi,\rho) = \frac{1}{n}|\{i : \pi(i) \neq \rho(i)\}|.$$

The measure $\mathbb{P}$ is the uniform probability measure on S_n, i.e.

$$\mathbb{P}(A) = \frac{|A|}{n!} \quad \text{for any } A \subset S_n.$$

Theorem 5.2.6 (Concentration on the symmetric group)**.** *Consider a random permutation $X \sim \mathrm{Unif}(S_n)$ and a function $f : S_n \to \mathbb{R}$. Then the concentration inequality* (5.6) *holds.*

5.2.4 Riemannian manifolds with strictly positive curvature

Riemannian manifolds provide many examples of spaces with concentration. If you are not into differential geometry, feel free to skip the rest of this section.

A compact connected Riemannian manifold (M, g) comes with the geodesic distance $d(x, y)$, the shortest length of a curve connecting the points. It can be seen as a metric measure space

$$(M, d, \mathbb{P}),$$

where $\mathbb{P}$ is the uniform probability measure derived by normalizing the Riemannian volume. 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

$$\|f(X) - \mathbb{E}\, f(X)\|_{\psi_2} \le \frac{C\|f\|_{\mathrm{Lip}}}{\sqrt{c(M)}} \tag{5.7}$$

for any Lipschitz function $f : M \to \mathbb{R}$.

To give an example, it is known that $c(S^{n-1}) = n - 1$. Thus (5.7) gives an alternative approach to concentration inequality (5.29) for the sphere S^{n-1}. We give several other examples next.

5.2.5 Special orthogonal group

The special orthogonal group $\mathrm{SO}(n)$ consists of all rotations in $\mathbb{R}^n$ or equivalently, $n \times n$ orthogonal matrices with determinant 1. We can treat it as a metric measure space

$$(\mathrm{SO}(n), \|\cdot\|_F, \mathbb{P}),$$

with distance given by the Frobenius norm $\|A - B\|_F$ and $\mathbb{P}$ as the uniform measure.

Theorem 5.2.7 (Concentration on the special orthogonal group)**.** *Consider a random orthogonal matrix $X \sim \mathrm{Unif}(\mathrm{SO}(n))$ and a function $f : \mathrm{SO}(n) \to \mathbb{R}$. Then the concentration inequality* (5.6) *holds.*

This result can be deduced from concentration on general Riemannian manifolds from Section 5.2.4.

Remark 5.2.8 (Haar measure)**.** To generate a random orthogonal matrix $X \sim \mathrm{Unif}(SO(n))$, one way is to start with an $n \times n$ Gaussian random matrix G with $N(0, 1)$ independent entries, and compute its singular value decomposition $G = U\Sigma V^\mathsf{T}$. Then the matrix $X := UV^\mathsf{T}$ is uniformly distributed in $\mathrm{SO}(n)$.

The uniform probability distribution on $\mathrm{SO}(n)$ is given by

$$\mu(A) := \mathbb{P}\{X \in A\} \quad \text{for } A \subset \mathrm{SO}(n).$$

This is the unique rotation-invariant probability measure[9] on $\mathrm{SO}(n)$, called the *Haar measure*. (Check the rotation invariance!)

[9] That is, for any measurable set $E \subset \mathrm{SO}(n)$ and any $T \in \mathrm{SO}(n)$, we have $\mu(E) = \mu(T(E))$.

5.2.6 Grassmannian

The Grassmann manifold $G_{n,m}$ consists of all m-dimensional subspaces of $\mathbb{R}^n$. When $m = 1$, it can be identified with the sphere S^{n-1} (can you see how?), so the concentration result on the Grassmanian includes concentration on the sphere. We treat $G_{n,m}$ as a metric measure space

$$(G_{n,m}, d, \mathbb{P}),$$

where distance between subspaces E and F is given by the operator norm[10]

$$d(E, F) = \|P_E - P_F\|$$

where P_E and P_F are the orthogonal projections onto the subspaces. (To practice with this distance, try the tricky Exercise 4.12.)

The probability distribution $\mathbb{P}$ is, like before, the uniform (Haar) probability measure.. A random subspace $E \sim \mathrm{Unif}(G_{n,m})$, and thus the Haar measure on the Grassmannian, can be constructed by computing the image of the random $n \times m$ Gaussian random matrix G with i.i.d. $N(0, 1)$ entries. (The rotation invariance should be straightforward – check it!)

Theorem 5.2.9 (Concentration on the Grassmannian)**.** *Consider a random subspace $X \sim \mathrm{Unif}(G_{n,m})$ and a function $f : G_{n,m} \to \mathbb{R}$. Then the concentration inequality (5.6) holds.*

This follows from concentration on the special orthogonal group from Section 5.2.5. (For the interested reader, here is how: express the Grassmannian as the quotient $G_{n,m} = SO(n)/(SO_m \times SO_{n-m})$ and use the fact that concentration carries over to quotients.)

5.2.7 Continuous cube and Euclidean ball

You can prove similar concentration inequalities for the unit Euclidean cube $[0, 1]^n$ and the Euclidean ball[11] $\sqrt{n}B_2^n$ (with Euclidean distance and the uniform probability measures). This follows from Gaussian concentration by "pushing forward" the Gaussian measure to the uniform measures on the cube and ball. We will state the result here and leave the proof for Exercises 5.12, 5.13.

Theorem 5.2.10 (Concentration on the continuous cube and ball)**.** *Let T be either the cube $[0, 1]^n$ or the ball $\sqrt{n}B_2^n$. Consider a random vector $X \sim \mathrm{Unif}(T)$ and a Lipschitz function $f : T \to \mathbb{R}$, where the Lipschitz norm is with respect to the Euclidean distance. Then the concentration inequality (5.5) holds.*

[10] The operator norm was introduced in Section 4.1.3.

[11] Recall that B_2^n denotes the unit Euclidean ball, i.e. $B_2^n = \{x \in \mathbb{R}^n : \|x\|_2 \le 1\}$, and $\sqrt{n}B_2^n$ is the Euclidean ball of radius $\sqrt{n}$.

5.2.8 Densities $e^{-U(x)}$

The push forward method from the previous section can be applied to many other distributions in $\mathbb{R}^n$. For example, suppose a random vector X has a density of the form

$$f(x) = e^{-U(x)}$$

for some function $U : \mathbb{R}^n \to \mathbb{R}$. As an example, if $X \sim N(0, I_n)$ the normal density (3.11) gives $U(x) = \|x\|_2^2 + c$ where c is a constant (that depends on n but not x), and Gaussian concentration holds for X.

Now, if U is a general function with curvature at least like $\|x\|_2^2$, then we should expect at least Gaussian concentration, as the next theorem shows. The curvature of U is measured by its *Hessian* $\operatorname{Hess} U(x)$, which is the $n \times n$ symmetric matrix with second derivatives: its (i, j)-th entry equals $\partial^2 U / \partial x_i \partial x_j$.

Theorem 5.2.11. *Consider a random vector X in $\mathbb{R}^n$ whose density has the form $f(x) = e^{-U(x)}$ for some function $U : \mathbb{R}^n \to \mathbb{R}$. Assume there exists $\kappa > 0$ such that*[12] $\operatorname{Hess} U(x) \succeq \kappa I_n$ *for all $x \in \mathbb{R}^n$. Then any Lipschitz function $f : \mathbb{R}^n \to \mathbb{R}$ satisfies*

$$\|f(X) - \mathbb{E} f(X)\|_{\psi_2} \le \frac{C\|f\|_{\mathrm{Lip}}}{\sqrt{\kappa}}.$$

Notice the similarity between this theorem and the concentration inequality (5.7) for Riemannian manifolds. Both can be proved using semigroup methods, which are not covered in this book.

5.2.9 Random vectors with independent bounded coordinates

There is a remarkable partial generalization of Theorem 5.2.10 for random vectors $X = (X_1, \ldots, X_n)$ with independent coordinates that have arbitrary bounded distributions (not just uniform). By scaling, we can assume without loss of generality that $|X_i| \le 1$.

Theorem 5.2.12 (Talagrand concentration inequality)**.** *Consider a random vector $X = (X_1, \ldots, X_n)$ whose coordinates are independent and satisfy $|X_i| \le 1$ almost surely. Then concentration inequality (5.5) holds for any* convex *Lipschitz function $f : [-1, 1]^n \to \mathbb{R}$.*

In particular, Talagrand concentration inequality holds for any *norm* on $\mathbb{R}^n$. We do not prove this result here.

5.3 Application: Johnson-Lindenstrauss lemma

Suppose we have N data points in $\mathbb{R}^n$ where the dimension n is very large. Can we reduce dimension without sacrificing too much of the data's geometry? The

[12] The matrix inequality here means $\operatorname{Hess} U(x) - \kappa I_n$ is a symmetric positive semidefinite matrix.

simplest way is to project the data points onto a low-dimensional subspace

$$E \subset \mathbb{R}^n, \quad \dim(E) := m \ll n,$$

see Figure 5.2 for illustration. How shall we choose the subspace E, and how small its dimension m can be?

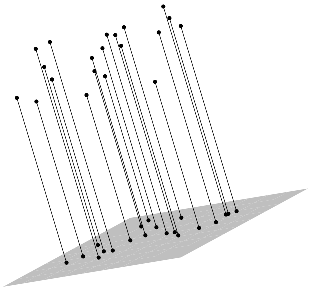


Figure 5.2 Johnson-Lindenstrauss Lemma reduces dimension of the data by random projection onto a low-dimensional subspace.

Johnson-Lindenstrauss Lemma below states that the geometry of data is well preserved if we choose E to be a *random subspace* of dimension

$$m \asymp \log N.$$

We already came across the notion of a random subspace in Section 5.2.6; let us recall it here. We say that E is a random m-dimensional subspace in $\mathbb{R}^n$ uniformly distributed in $G_{n,m}$, i.e.

$$E \sim \mathrm{Unif}(G_{n,m}),$$

if E is a random m-dimensional subspace of $\mathbb{R}^n$ whose distribution is rotation invariant, i.e.

$$\mathbb{P}\{E \in \mathcal{E}\} = \mathbb{P}\{U(E) \in \mathcal{E}\}$$

for any fixed subset $\mathcal{E} \subset G_{n,m}$ and $n \times n$ orthogonal matrix U.

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*

$$m \geq C\varepsilon^{-2} \log N.$$

Let P be the orthogonal projection in $\mathbb{R}^n$ onto a random m-dimensional subspace $E \sim \mathrm{Unif}(G_{n,m})$. Then, with probability at least $1 - 2\exp(-c\varepsilon^2 m)$, the scaled projection $Q = \sqrt{\frac{n}{m}} P$ is an approximate isometry on $\mathcal{X}$:

$$(1-\varepsilon)\|x-y\|_2 \leq \|Qx - Qy\|_2 \leq (1+\varepsilon)\|x-y\|_2 \quad \textit{for all } x, y \in \mathcal{X}. \tag{5.8}$$

The proof of Johnson-Lindenstrauss Lemma will be based on concentration of Lipschitz functions on the sphere, which we studied in Section 5.1. We use it to first examine how the random projection P acts on a *fixed* vector $x - y$, and then take *union bound* over all N^2 differences $x - y$.

Lemma 5.3.2 (Random projection). *Let P be a projection in $\mathbb{R}^n$ onto a random m-dimensional subspace $E \sim \mathrm{Unif}(G_{n,m})$. Fix any $z \in \mathbb{R}^n$ and $\varepsilon > 0$. Then:*

(a) $\left(\mathbb{E}\|Pz\|_2^2\right)^{1/2} = \sqrt{\dfrac{m}{n}}\,\|z\|_2$.

(b) With probability at least $1 - 2\exp(-c\varepsilon^2 m)$, we have

$$(1-\varepsilon)\sqrt{\frac{m}{n}}\,\|z\|_2 \le \|Pz\|_2 \le (1+\varepsilon)\sqrt{\frac{m}{n}}\,\|z\|_2.$$

Proof Without loss of generality, we may assume that $\|z\|_2 = 1$. Now switch the view. A random m-dimensional subspace E can be obtained by randomly rotating some fixed subspace, such as the coordinate subspace $\mathbb{R}^m$. But instead of fixing z and randomly rotating $\mathbb{R}^m$, we can fix the subspace $E = \mathbb{R}^m$ and randomly rotate z, which makes z uniformly distributed on the sphere: $z \sim \mathrm{Unif}(S^{n-1})$. By rotation invariance, $\|Pz\|_2$ has the same distribution (check!).

(a) Since P is the projection onto the first m coordinates in $\mathbb{R}^n$,

$$\mathbb{E}\|Pz\|_2^2 = \mathbb{E}\sum_{i=1}^m z_i^2 = \sum_{i=1}^m \mathbb{E}\, z_i^2 = m\,\mathbb{E}\, z_1^2, \tag{5.9}$$

because 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 $\sum_{i=1}^n z_i^2 = 1$. Taking expectations of both sides, we obtain $\sum_{i=1}^n \mathbb{E}\, z_i^2 = 1$ which yields $\mathbb{E}\, z_1^2 = 1/n$ because all terms in the sum are the same. Putting this into (5.9), we get $\mathbb{E}\|Pz\|_2^2 = m/n$.

(b) follows from concentration of the sphere. Indeed, $x \mapsto \|Px\|_2$ is a Lipschitz function on S^{n-1} with Lipschitz norm bounded by 1 (check!). Then concentration inequality (5.30) gives

$$\mathbb{P}\left\{\left|\|Px\|_2 - \sqrt{\frac{m}{n}}\right| \ge t\right\} \le 2\exp(-cnt^2).$$

(We replaced $\mathbb{E}\|x\|_2$ by the $(\mathbb{E}\|x\|_2^2)^{1/2}$ in the concentration inequality using Remark 5.2.1.) 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

$$(1-\varepsilon)\|z\|_2 \le \|Qz\|_2 \le (1+\varepsilon)\|z\|_2$$

holds for all $z \in \mathcal{X} - \mathcal{X}$. Since $Q = \sqrt{n/m}\,P$, this is inequality is equivalent to

$$(1-\varepsilon)\sqrt{\frac{m}{n}}\,\|z\|_2 \le \|Pz\|_2 \le (1+\varepsilon)\sqrt{\frac{m}{n}}\,\|z\|_2. \tag{5.10}$$

For any fixed z, Lemma 5.3.2 states that (5.10) holds with probability at least $1 - 2\exp(-c\varepsilon^2 m)$. It remains to take a union bound over $z \in \mathcal{X} - \mathcal{X}$. It follows that inequality (5.10) holds simultaneously for all $z \in \mathcal{X} - \mathcal{X}$, 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 $m \geq C\varepsilon^{-2}\log N$ then this probability is at least $1-2\exp(-c\varepsilon^2 m/2)$, as claimed. Johnson-Lindenstrauss Lemma is proved. □

Remark 5.3.3 (Non-adaptive, dimension-free)**.** A remarkable feature of Johnson-Lindenstrauss lemma is dimension reduction map A is *non-adaptive*, it does not depend on the data. Note also that the ambient dimension n of the data plays no role. With more tools, we will develop more advanced versions of Johnson-Lindenstrauss lemma in Sections 9.2.4 and 9.7 (see Exercises 9.37–9.39).

Remark 5.3.4 (Optimality)**.** Johnson-Lindenstrauss lemma makes such a striking dimension reduction from N to $n = O(\log N)$. Can we go even smaller, say $n = o(\log N)$? Exercise 5.15 shows we can't – the log dimension is the best we can do, even with nonlinear maps.

For more practice with Johnson-Lindenstrauss lemma, prove its subgaussian version (Exercise 5.14).

5.4 Matrix Bernstein inequality

Here, we generalize concentration inequalities from sums of independent random variables $\sum X_i$ to sums of independent *random matrices*. We will make a matrix version of Bernstein inequality (Theorem 2.9.5) by replacing random variables X_i by random matrices and absolute value $|\cdot|$ by the operator norm $\|\cdot\|$. No need for independence of entries, rows, or columns within each random matrix X_i – a remarkably general assumption!

Theorem 5.4.1 (Matrix Bernstein inequality)**.** *Let $X_1, \ldots, X_N$ be independent, mean-zero, $n \times n$ symmetric random matrices, such that $\|X_i\| \leq K$ almost surely for all i. Then, for every $t \geq 0$, we have*

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

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

We can rewrite the right-hand side as the mixture of subgaussian and subexponential tail, just like in the scalar Bernstein inequality:

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

The proof of matrix Bernstein inequality follows a simple idea: repeat the MGF argument (Section 2.9), swapping scalars for matrices. Most of it works, except for one major challenge: matrix multiplication is not commutative. Before tackling that, let us build some *matrix calculus* to treat matrices as numbers.

5.4.1 Matrix calculus

For an $n \times n$ symmetric matrix X, operations such as inversion or squaring only affect eigenvalues, keeping eigenvectors the same. If the spectral decomposition[13] of X is $X = \sum_{i=1}^n \lambda_i u_i u_i^\mathsf{T}$, then

$$X^{-1} = \sum_{i=1}^n \frac{1}{\lambda_i} u_i u_i^\mathsf{T}, \quad X^2 = \sum_{i=1}^n \lambda_i^2 u_i u_i^\mathsf{T}, \quad 2I_n - 5X^3 = \sum_{i=1}^n (2 - 5\lambda_i^3) u_i u_i^\mathsf{T}. \tag{5.11}$$

(Check it!) This suggests how to define arbitrary functions of matrices – just apply the function to eigenvalues, keeping eigenvectors the same:

Definition 5.4.2 (Functions of matrices)**.** For a function $f : \mathbb{R} \to \mathbb{R}$ and an $n \times n$ symmetric matrix X with spectral decomposition

$$X = \sum_{i=1}^n \lambda_i u_i u_i^\mathsf{T}, \quad \text{define} \quad f(X) := \sum_{i=1}^n f(\lambda_i) u_i u_i^\mathsf{T}.$$

As we saw in (5.11), this definition agrees with matrix addition and multiplication, and (by a limiting argument) with Taylor series (Exercise 5.16).

Just like numbers, matrices can be compared to each other if we define a *partial order* on the set of $n \times n$ symmetric matrices like this:

Definition 5.4.3 (Loewner order)**.** We write $X \succeq 0$ if X is a symmetric positive semidefinite matrix.[14] Now write $X \succeq Y$ and $Y \preceq X$ if $X - Y \succeq 0$.

Note that $\succeq$ is a partial, not total, order, because there are matrices for which neither $X \succeq Y$ nor $Y \succeq X$ holds. (Example?)

Proposition 5.4.4 (Simple properties of Loewner order)**.** *We have:*

(a) (Eigenvalue monotonicity) $X \preceq Y$ implies $\lambda_i(X) \le \lambda_i(Y)$ for all i.

(b) (Trace monotonicity) For a (weakly) increasing function $f : \mathbb{R} \to \mathbb{R}$,

$$X \preceq Y \quad \Longrightarrow \quad \operatorname{tr} f(X) \le \operatorname{tr} f(Y).$$

(c) (Operator norm) For any $a \ge 0$,

$$\|X\| \le a \quad \Longleftrightarrow \quad -aI_n \preceq X \preceq aI_n. \tag{5.12}$$

(d) (Upgrading scalar to matrix inequalities) For functions $f, g : \mathbb{R} \to \mathbb{R}$,

$$f(x) \le g(x) \ \forall x \text{ with } |x| \le a \quad \Longrightarrow \quad f(X) \preceq g(X) \ \forall X \text{ with } \|X\| \le a.$$

Proof (a) If $X \preceq Y$ then $Y - X \ge 0$ and so $u^\mathsf{T}(Y - X)u \ge 0$ for all u, meaning $u^\mathsf{T} X u \le u^\mathsf{T} Y u$ for all u. Now use min-max theorem (Theorem 4.1.6).

(b) The eigenvalues of $f(X)$ are $f(\lambda_i(X))$, and similarly for $f(Y)$. By part (a) and assumption, $f(\lambda_i(X)) \le f(\lambda_i(Y))$. Summing these gives the result, since the trace is the sum of eigenvalues.

(c) Recalling (4.10) we see that $\|X\| \le t$ implies $u^\mathsf{T} X u \le a$ for all unit u, so

[13] Spectral decomposition was discussed in Section 3.2.2.

[14] Equivalently, $X \succeq 0$ if X is symmetric with nonnegative eigenvalues of X – why?

$u^\mathsf{T}(aI_n - X)u \ge 0$ for all u, meaning $aI_n - X \succeq 0$, thus $X \preceq aI_n$. A similar argument gives $X \succeq -aI_n$, and also proves the opposite direction (write it!).

(d) By considering $g - f$, we can assume that $f = 0$. If $\|X\| \le a$, then all eigenvalues of X satisfy $|\lambda_i| \le a$ (by (4.10)), which implies $g(\lambda_i) \ge 0$ by assumption. So, by definition, $g(X)$ has nonnegative eigenvalues $g(\lambda_i)$ and so $g(X) \succeq 0$. □

Remark 5.4.5 (Operator norm as matrix "absolute value")**.** Does (5.12) look familiar? It is a matrix version of the basic fact about absolute values: for $x \in \mathbb{R}$,

$$|x| \le a \quad \Longleftrightarrow \quad -a \le x \le a.$$

This makes the operator norm $\|\cdot\|$ a natural matrix version of absolute value $|\cdot|$, and that's why it appears in matrix Bernstein inequality 5.4.1.

Remark 5.4.6 (Matrix monotonicity)**.** Can we strenghten trace monotonicity (Proposition 5.4.4(b)) to matrix monotonicity, that is

$$X \preceq Y \quad \Longrightarrow \quad f(X) \preceq f(Y) \quad \text{for any weakly increasing } f : \mathbb{R} \to \mathbb{R}? \tag{5.13}$$

If X and Y commute, yes – but in general, no (Exercise 5.17). However, some functions, like $1/x$ and $\log x$ on $[0, \infty)$, are *matrix monotone*, meaning that (5.13) holds even for non-commuting matrices:

$$0 \preceq X \preceq Y \quad \Longrightarrow \quad X^{-1} \succeq Y^{-1} \succeq 0 \quad \text{and} \quad \log X \preceq \log Y$$

whenever X is invertible. Check this in Exercise 5.18.

5.4.2 Trace inequalities

So far, extending scalar concepts to matrices has been pretty smooth. But it doesn't always work. We already saw in Remark 5.4.6 how non-commutativity ($AB \ne BA$) may cause scalar properties to fail for matrices. Here is another example: the identity is $e^{x+y} = e^x\, e^y$ holds for scalars, but in Exercise 5.19 you will find $n \times n$ symmetric matrices X and Y such that

$$e^{X+Y} \ne e^X e^Y.$$

This is unfortunate, because the identity $e^{x+y} = e^x\, e^y$ was crucial to our approach to concentration of sums of random variables: it let us split the MGF $\mathbb{E}\exp(\lambda S)$ of the sum into the product of exponentials, see (2.6).

Nevertheless, there exists useful substitutes for the missing identity $e^{X+Y} = e^X e^Y$. We state two of them here without proof; they belong to the rich family of *trace inequalities.*

Theorem 5.4.7 (Golden-Thompson inequality)**.** *For any $n \times n$ symmetric matrices A and B, we have*

$$\operatorname{tr}(e^{A+B}) \le \operatorname{tr}(e^A e^B).$$

Unfortunately, Golden-Thompson inequality does not hold for three or more matrices: in general, the inequality $\operatorname{tr}(e^{A+B+C}) \le \operatorname{tr}(e^A e^B e^C)$ may fail.

Theorem 5.4.8 (Lieb inequality). *Let H be an $n \times n$ symmetric matrix. Define the function on matrices*

$$f(X) := \operatorname{tr} \exp(H + \log X).$$

Then f is concave on the space on positive definite $n \times n$ symmetric matrices.[15]

In the scalar case ($n = 1$), f is linear and Lieb inequality holds trivially.

A proof of matrix Bernstein inequality can be based on either Golden-Thompson or Lieb inequalities. We use Lieb inequality, which we will now restate for random matrices. If X is a random matrix, then Lieb and Jensen inequalities imply that

$$\mathbb{E} f(X) \le f(\mathbb{E} X).$$

(Why does Jensen inequality hold for random matrices?) Applying this with $X = e^Z$, we obtain the following.

Lemma 5.4.9 (Lieb 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*

$$\mathbb{E} \operatorname{tr} \exp(H + Z) \le \operatorname{tr} \exp(H + \log \mathbb{E} e^Z).$$

5.4.3 Proof of matrix Bernstein inequality

We are now ready to prove matrix Bernstein inequality, Theorem 5.4.1, using Lieb 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

$$\|S\| = \max_i |\lambda_i(S)| = \max\left(\lambda_{\max}(S),\, \lambda_{\max}(-S)\right). \tag{5.14}$$

To bound $\lambda_{\max}(S)$, we proceed with the MGF method like in Section 2.2. Fix $\lambda \ge 0$ and use Markov inequality to obtain

$$\mathbb{P}\{\lambda_{\max}(S) \ge t\} = \mathbb{P}\{e^{\lambda \cdot \lambda_{\max}(S)} \ge e^{\lambda t}\} \le e^{-\lambda t}\, \mathbb{E}\, e^{\lambda \cdot \lambda_{\max}(S)}. \tag{5.15}$$

Since by Definition 5.4.2 the eigenvalues of $e^{\lambda S}$ are $e^{\lambda \cdot \lambda_i(S)}$, we have

$$E := \mathbb{E}\, e^{\lambda \cdot \lambda_{\max}(S)} = \mathbb{E}\, \lambda_{\max}(e^{\lambda S}).$$

[15] Concavity means that the inequality $f(\lambda X + (1-\lambda)Y) \ge \lambda f(X) + (1-\lambda) f(Y)$ holds for matrices X and Y, and for $\lambda \in [0,1]$.

Since the eigenvalues of $e^{\lambda S}$ are all positive, the maximal eigenvalue of $e^{\lambda S}$ is bounded by the sum of all eigenvalues, the trace of $e^{\lambda S}$, which leads to

$$E \le \mathbb{E}\,\mathrm{tr}\, e^{\lambda S}.$$

Step 2: Application of Lieb inequality. To prepare for an application of Lieb inequality (Lemma 5.4.9), let us separate the last term from the sum S:

$$E \le \mathbb{E}\,\mathrm{tr}\exp\Big[\sum_{i=1}^{N-1} \lambda X_i + \lambda X_N\Big].$$

Condition on $(X_i)_{i=1}^{N-1}$ and apply Lemma 5.4.9 for the fixed matrix $H := \sum_{i=1}^{N-1} \lambda X_i$ and the random matrix $Z := \lambda X_N$. We obtain

$$E \le \mathbb{E}\,\mathrm{tr}\exp\Big[\sum_{i=1}^{N-1} \lambda X_i + \log \mathbb{E}\, e^{\lambda X_N}\Big].$$

(To be more specific here, we first apply Lemma 5.4.9 for the conditional expectation, and then take expectation of both sides using the law of total expectation.)

We continue in a similar way: separate the next term λX_{N-1} from the sum $\sum_{i=1}^{N-1} \lambda X_i$ and apply Lemma 5.4.9 again for $Z = \lambda X_{N-1}$. Repeating N times, we obtain

$$E \le \mathrm{tr}\exp\Big[\sum_{i=1}^{N} \log \mathbb{E}\, e^{\lambda X_i}\Big]. \tag{5.16}$$

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 5.4.10 (Moment generating function). *Let X be an $n \times n$ symmetric mean-zero 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 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 $e^z = 1 + z + z^2 \cdot \sum_{p=2}^{\infty} z^{p-2}/p!$ 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 upgrade this to a matrix inequality using Proposition 5.4.4(d). 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 get

$$\mathbb{E}\, e^{\lambda X} \preceq I + g(\lambda)\, \mathbb{E}\, X^2.$$

To complete the proof of the lemma, let's use the inequality $1 + z \le e^z$ that holds for all scalars z. Thus, using again Proposition 5.4.4(d), we see that $I + Z \preceq e^Z$ holds for all matrices Z, and in particular for $Z = g(\lambda)\, \mathbb{E}\, X^2$. □

Step 4: Completion of the proof. Let us return to bounding the quantity in (5.16). Using Lemma 5.4.10, we obtain

$$E \le \operatorname{tr} \exp \Big[\sum_{i=1}^N \log \mathbb{E}\, e^{\lambda X_i} \Big] \le \operatorname{tr} \exp \left[g(\lambda) Z \right], \quad \text{where} \quad Z := \sum_{i=1}^N \mathbb{E}\, X_i^2.$$

Here we used matrix monotonicity of $\ln x$ (see Remark 5.4.6) to take logarithms on both sides, summed up the results, and then used trace monotonicity (Proposition 5.4.4(b)) to take traces of the exponential on both sides.

Since the trace of $\exp\,[g(\lambda)Z]$ 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 (5.15), we obtain

$$\mathbb{P}\{\lambda_{\max}(S) \ge t\} \le n \cdot \exp\left[-\lambda t + g(\lambda) \sigma^2 \right].$$

We obtained a bound that holds for any $\lambda > 0$ such that $|\lambda| < 3/K$, so we can minimize it in λ. Better yet, instead of computing the exact minimum (which might be a little too ugly), we can choose the following value: $\lambda = t/(\sigma^2 + Kt/3)$. Substituting it into the bound above and simplifying yields

$$\mathbb{P}\{\lambda_{\max}(S) \ge t\} \le n \cdot \exp\Big(- \frac{t^2/2}{\sigma^2 + Kt/3} \Big).$$

Repeating the argument for $-S$ and combining the two bounds via (5.14), we complete the proof of Theorem 5.4.1. (Do this!) □

Remark 5.4.11 (Matrix Bernstein inequality: expectation)**.** Matrix Bernstein inequality gives a high-probability bound. It can be turned into a simpler (but less informative) expectation bound in a standard way – using the integrated tail formula (Lemma 1.6.1). Try Exercise 5.20 to deduce from Theorem 5.4.1 that

$$\mathbb{E} \Big\| \sum_{i=1}^N X_i \Big\| \lesssim \Big\| \sum_{i=1}^N \mathbb{E}\, X_i^2 \Big\|^{1/2} \sqrt{\log(2n)} + K \log(2n), \tag{5.17}$$

where "$\lesssim$" hides an absolute constant factor. Note that in the scalar case ($n = 1$), an expectation bound is trivial: the variance of sum formula (1.8) gives

$$\mathbb{E}\Big|\sum_{i=1}^{N} X_i\Big| \le \Big(\mathbb{E}\Big|\sum_{i=1}^{N} X_i\Big|^2\Big)^{1/2} = \Big(\sum_{i=1}^{N} \mathbb{E}\, X_i^2\Big)^{1/2}.$$

Remark 5.4.12 (The logarithmic price)**.** Compared to (5.17), the high-dimensional upgrade (5.17) differs by just a *logarithmic* factor. This is a surprisingly small price for high dimensions! And this price is essentially optimal–check out Exercise 5.28 for an example for why we can't get rid of it.

5.4.4 Matrix Hoeffding and Khintchine inequalities

The techniques we developed in the proof of matrix Bernstein inequality can be used to give matrix versions of other classical concentration inequalities. Here is a matrix version of Hoeffding inequality (Theorem 2.2.1):

Theorem 5.4.13 (Matrix Hoeffding inequality)**.** *Let $\varepsilon_1, \ldots, \varepsilon_N$ be independent Rademacher random variables and let $A_1, \ldots, A_N$ be any (fixed) symmetric $n \times n$ matrices. Then, for any $t \ge 0$, we have*

$$\mathbb{P}\Big\{\Big\|\sum_{i=1}^{N} \varepsilon_i A_i\Big\| \ge t\Big\} \le 2n \exp\Big(-\frac{t^2}{2\sigma^2}\Big),$$

where $\sigma^2 = \big\|\sum_{i=1}^{N} A_i^2\big\|$.

The proof is a bit simpler than of matrix Bernstein; give it a try in Exercise 5.21!

Matrix Hoeffding inequality is a high-probability bound. Like before, you can convert it into an expectation bound using the integrated tail formula (the one in Exercise 1.15(c) is helpful here) and get a matrix version of Khintchine inequality (Theorem 2.7.5):

Theorem 5.4.14 (Matrix Khintchine inequality)**.** *Let $\varepsilon_1, \ldots, \varepsilon_N$ be independent Rademacher random variables and let $A_1, \ldots, A_N$ be any (fixed) symmetric $n \times n$ matrices. Then, for every $p \in [1, \infty)$ we have*

$$\Big(\mathbb{E}\,\Big\|\sum_{i=1}^{N} \varepsilon_i A_i\Big\|^p\Big)^{1/p} \le C\sqrt{p + \log n}\,\Big\|\sum_{i=1}^{N} A_i^2\Big\|^{1/2}.$$

Deduce this result from matrix Hoeffding inequality in Exercise 5.22.

Remark 5.4.15 (Non-symmetric, rectangular matrices)**.** Matrix concentration inequalities easily extend to rectangular matrices using the neat *Hermitian dilation* introduced in Exercise 4.14. Just replace each matrix X_i with the symmetric block-matrix $\left[\begin{smallmatrix} 0 & X_i \\ X_i^{\mathsf{T}} & 0 \end{smallmatrix}\right]$ and apply usual matrix concentration. Try out Exercises 5.23 and 5.24 to get matrix Bernstein and Khintchine inequalities for rectangular matrices this way.

5.5 Application: community detection in sparse networks

In Section 4.5, we explored spectral clustering – a basic method for community detection in networks. We analyzed its performance on the stochastic block model $G(n,p,q)$ with two communities and showed it works for relatively *dense* networks, where the expected average degree $\gtrsim \sqrt{n}$ (Remark 4.5.3). Now, using matrix Bernstein inequality, we will show that spectral clustering actually works for *much sparser* networks, with an expected average degree as low as $O(\log n)$.

Theorem 5.5.1 (Spectral clustering for sparse stochastic block model)**.** *Let $G \sim G(n,p,q)$ where*[16] *$p = a/n$, $q = b/n$ and $b < a < 3b$. Assume that*

$$(a-b)^2 \geq Ca \log n.$$

Then, with probability at least 0.99, the spectral clustering algorithm (Section 4.5.5) identifies the communities of G with 99% accuracy, i.e. misclassifying at most $0.01n$ vertices.

Proof Let's follow the argument in Section 4.5, just with a sharper error bound.

Step 1: Decomposition. Like in Section 4.5.2, we look at the adjacency matrix A of a random graph $G \sim G(n,p,q)$, and split it into deterministic and random parts:

$$A = D + R \quad \text{where} \quad D = \mathbb{E}\, A$$

In Section 4.5.2, we analyzed the expected adjacency matrix D, noting that its second top eigenvector $u_2(D)$ has ± 1 coefficients that represent community membership. The main difference now is in analyzing the random part

$$R = A - \mathbb{E}\, A.$$

Let's decompose it entry by entry, keeping symmetry in mind. Denote the standard basis vectors in $\mathbb{R}^n$ by $e_1, \ldots, e_n$ and write R as a *sum of independent, mean-zero random matrices* Z_{ij} that isolate entries (i,j) and (j,i):

$$R = \sum_{i \leq j} Z_{ij}, \quad \text{where} \quad Z_{ij} = \begin{cases} R_{ij}(e_i e_j^\mathsf{T} + e_j e_i^\mathsf{T}), & i < j \\ R_{ii} e_i e_i^\mathsf{T}, & i = j. \end{cases}$$

Step 2: Bounding the error. Since $A_{ij} \in \{0,1\}$, we have $|R_{ij}| \leq 1$, so $\|Z_{ij}\| = \|R_{ij}\| \leq 1$ (why?), so $\|R_{ij} Z_{ij}\| \leq 1$. Thus, applying matrix Bernstein inequality (5.17) combined with Markov inequality, we obtain with probability at least 0.99:

$$\|R\| \lesssim \sigma \sqrt{\log n} + \log n \quad \text{where} \quad \sigma^2 = \Big\| \mathbb{E} \sum_{i \leq j} Z_{ij}^2 \Big\|. \tag{5.18}$$

[16] With this parametrization, any node is connected to $a/2$ nodes in its own community and $b/2$ nodes in the other, on average. (Why?)

Let's compute σ^2. A quick check shows that Z_{ij}^2 is a diagonal matrix:

$$Z_{ij}^2 = \begin{cases} R_{ij}^2(e_i e_i^\mathsf{T} + e_j e_j^\mathsf{T}), & i < j \\ R_{ii}^2 e_i e_i^\mathsf{T}, & i = j. \end{cases}$$

Then, by symmetry,

$$\sum_{i \le j} Z_{ij}^2 = \sum_{i<j} R_{ij}^2(e_i e_i^\mathsf{T} + e_j e_j^\mathsf{T}) + \sum_i R_{ii}^2 e_i e_i^\mathsf{T} = \sum_{i=1}^n \Big(\sum_{j=1}^n R_{ij}^2 \Big) e_i e_i^\mathsf{T}.$$

This is a diagonal matrix, and so is its expectation. Thus

$$\sigma^2 = \Big\| \mathbb{E} \sum_{i \le j} Z_{ij}^2 \Big\| = \max_{i=1,\dots,n} \sum_{j=1}^n \mathbb{E}\, R_{ij}^2$$

since the operator norm of a diagonal matrix is the maximal absolute value of its entries (Exercise 4.3(b)). Recall that $R_{ij} = A_{ij} - \mathbb{E}\, A_{ij}$. In the stochastic block model, A_{ij} is either $\mathrm{Ber}(p)$ or $\mathrm{Ber}(q)$, so $\mathbb{E}\, R_{ij}^2 = \mathrm{Var}(A_{ij}) \le p$ since $p > q$. Thus

$$\sigma^2 \le np = a,$$

and substituting this into (5.18) we get

$$\|R\| \lesssim \sqrt{a \log n} + \log n \lesssim \sqrt{a \log n} \tag{5.19}$$

because the assumption implies that $a \gtrsim \log n$ (why?).

Step 3: Applying Davis-Kahan. Let's apply Theorem 4.1.15 (see Exercise 4.16) to D and A, focusing on the second-largest eigenvalue. As we noted in Section 4.5.4, the separation between $\lambda_2(D)$ of D and the rest of the spectrum is

$$\delta = \min(\lambda_2(D),\, \lambda_1(D) - \lambda_2(D)) = \min\left(\frac{p-q}{2},\, q \right) n = \frac{a-b}{2}$$

since $a \le 3b$ by assumption. Using the bound on $R = A - D$ from (5.19), the Davis-Kahan inequality guarantees that for some $\theta \in \{-1, 1\}$, the distance between the *unit* eigenvectors of D and A (denoted with bars) satisfies

$$\|\bar{u}_2(D) - \theta \bar{u}_2(A)\|_2 \le \frac{2\|R\|}{\delta} \le \frac{C_1 \sqrt{a \log n}}{a - b} < \frac{1}{10}$$

if we choose the constant C in the assumption of the theorem large enough. Multiplying both sides by $\sqrt{n}$, we get

$$\|u_2(D) - \theta u_2(A)\|_2 \lesssim \frac{\sqrt{n}}{10}.$$

Since all coefficients of $u_2(D)$ are ± 1 and correctly identify community membership, it follows that at least 99% of the coefficients of $\theta u_2(A)_j$ have the same sign as $u_2(D)_j$ (check!), and thus correctly identify the community membership. □

Remark 5.5.2 (Sparsity)**.** The sparsest graphs for which Theorem 5.5.1 is nontrivial have expected average degree

$$\frac{n(p+q)}{2} = \frac{a+b}{2} \asymp \log n.$$

That's way sparser than $O(\sqrt{n})$ we have achieved previously (Remark 4.5.3)!

To practice more with these ideas, try Exercise 5.25 to do community detection in the stochastic bock model without loops (a pretty natural model).

5.6 Application: covariance estimation for general distributions

In Section 4.7, we learned how to estimate the covariance matrix of a subgaussian distribution in $\mathbb{R}^n$ from a sample of size $O(n)$. Now, we drop the subgaussian assumption, making it work for much broader distributions, even discrete ones. The trade-off is just a logarithmic oversampling factor!

Like in Section 4.7, we estimate the second moment matrix $\Sigma = \mathbb{E}\, XX^\mathsf{T}$ by its sample version

$$\Sigma_m = \frac{1}{m}\sum_{i=1}^m X_i X_i^\mathsf{T}.$$

If X has zero mean, then Σ is the covariance matrix of X, and Σ_m is the sample covariance matrix.

Theorem 5.6.1 (General covariance estimation)**.** *Let X be a random vector in $\mathbb{R}^n$, $n \ge 2$. Assume that for some $K \ge 1$,*

$$\|X\|_2 \le K\,(\mathbb{E}\|X\|_2^2)^{1/2} \quad \textit{almost surely.} \tag{5.20}$$

Then, for every positive integer m, we have

$$\mathbb{E}\|\Sigma_m - \Sigma\| \le C\Big(\sqrt{\frac{K^2 n\log n}{m}} + \frac{K^2 n \log n}{m}\Big)\,\|\Sigma\|.$$

Proof By Proposition 3.2.1(b), we have $\mathbb{E}\|X\|_2^2 = \operatorname{tr}(\Sigma)$, so (5.20) becomes

$$\|X\|_2^2 \le K^2 \operatorname{tr}(\Sigma) \quad \text{almost surely.} \tag{5.21}$$

Apply the expected version of matrix Bernstein inequality (5.17) for the sum of i.i.d. mean-zero random matrices $X_iX_i^\mathsf{T} - \Sigma$ and get[17]

$$\mathbb{E}\|\Sigma_m - \Sigma\| = \frac{1}{m}\,\mathbb{E}\Big\|\sum_{i=1}^m (X_iX_i^\mathsf{T} - \Sigma)\Big\| \lesssim \frac{1}{m}\Big(\sigma\sqrt{\log n} + M\log n\Big) \tag{5.22}$$

where

$$\sigma^2 = \Big\|\sum_{i=1}^m \mathbb{E}(X_iX_i^\mathsf{T} - \Sigma)^2\Big\| = m\|\mathbb{E}(XX^\mathsf{T} - \Sigma)^2\|$$

[17] As usual, the notation $a \lesssim b$ hides absolute constant factors, i.e. it means that $a \le Cb$ where C is an absolute constant.

and M is any number chosen so that

$$\|XX^\mathsf{T} - \Sigma\| \le M \quad \text{almost surely.}$$

To complete the proof, it remains to bound σ^2 and M.

Let us start with σ^2. Expanding the square, we find that[18]

$$\mathbb{E}(XX^\mathsf{T} - \Sigma)^2 = \mathbb{E}(XX^\mathsf{T})^2 - \Sigma^2 \preceq \mathbb{E}(XX^\mathsf{T})^2. \tag{5.23}$$

Further, the assumption (5.21) gives

$$(XX^\mathsf{T})^2 = \|X\|^2 XX^\mathsf{T} \preceq K^2 \operatorname{tr}(\Sigma) XX^\mathsf{T}.$$

Taking expectation and recalling that $\mathbb{E}\, XX^\mathsf{T} = \Sigma$, we obtain

$$\mathbb{E}(XX^\mathsf{T})^2 \preceq K^2 \operatorname{tr}(\Sigma)\Sigma.$$

Substituting this bound into (5.23), we obtain a good bound on σ, namely

$$\sigma^2 \le K^2 m \operatorname{tr}(\Sigma)\|\Sigma\|.$$

Bounding M is easy:

$$\begin{aligned}\|XX^\mathsf{T} - \Sigma\| &\le \|X\|_2^2 + \|\Sigma\| \quad \text{(by triangle inequality)}\\ &\le K^2 \operatorname{tr}(\Sigma) + \|\Sigma\| \quad \text{(by assumption (5.21))}\\ &\le 2K^2 \operatorname{tr}(\Sigma) =: M \quad (\text{since } \|\Sigma\| \le \operatorname{tr}(\Sigma) \text{ and } K \ge 1).\end{aligned}$$

Substituting our bounds for σ and M into (5.22), we get

$$\mathbb{E}\|\Sigma_m - \Sigma\| \le \frac{1}{m}\Big(\sqrt{K^2 m \operatorname{tr}(\Sigma)\|\Sigma\|} \cdot \sqrt{\log n} + 2K^2 \operatorname{tr}(\Sigma) \cdot \log n\Big).$$

To finish the proof, use $\operatorname{tr}(\Sigma) \le n\|\Sigma\|$ and simplify the bound. □

Remark 5.6.2 (Sample complexity)**.** Theorem 5.6.1 shows that for any $\varepsilon \in (0,1)$, we can estimate the covariance matrix with a small relative error:

$$\mathbb{E}\|\Sigma_m - \Sigma\| \le \varepsilon\|\Sigma\|, \tag{5.24}$$

as long as the sample size is

$$m \asymp \varepsilon^{-2} n \log n. \tag{5.25}$$

Compared to the sample complexity $m \asymp \varepsilon^{-2} n$ for subgaussian distributions (see Remark 4.7.2), dropping the subgaussian assumption costs just a small logarithmic oversampling factor! In general, this factor cannot be dropped (Exercise 5.28).

Remark 5.6.3 (Low-dimensional distributions)**.** At the end of the proof of Theorem 5.6.1, we used a rough bound $\operatorname{tr}(\Sigma) \le n\|\Sigma\|$. But instead, we can express the conclusion in terms of the *effective rank* of Σ:

$$r = r(\Sigma) = \frac{\operatorname{tr}(\Sigma)}{\|\Sigma\|} \tag{5.26}$$

[18] Recall Definition 5.4.3 of the positive semidefinite order (or Loewner order) $\preceq$ used here.

and get a sharper bound

$$\mathbb{E}\|\Sigma_m - \Sigma\| \le C\Big(\sqrt{\frac{K^2 r \log n}{m}} + \frac{K^2 r \log n}{m}\Big)\,\|\Sigma\|. \tag{5.27}$$

It shows that a sample of size

$$m \asymp \varepsilon^{-2} r \log n$$

is enough to estimate the covariance matrix as in (5.24). Since $r \le n$ (why?), this sample size is at least as small as (5.25). It is even much smaller for *approximately low-dimensional* distributions that concentrate near lower-dimensional subspaces.

Remark 5.6.4 (Effective and stable rank of a matrix)**.** What does the effective rank (5.26) really tell us about a positive-semidefinite matrix Σ? To get an idea, write it as the sum of eigenvalues divided by the biggest one:

$$r(\Sigma) = \frac{\sum_{i=1}^n \lambda_i(\Sigma)}{\max_i \lambda_i(\Sigma)}$$

(check!). This is always bounded by the actual rank (the number of nonzero eigenvalues) and can be much smaller for "approximately" low-rank matrices – ones having only a few large eigenvalues. A related idea is *stable rank*, defined for any matrix A

$$s(A) = \frac{\|A\|_F^2}{\|A\|^2} = \frac{\sum_{i=1}^n s_i^2(A)}{\max_i s_i^2(A)} = r(A^\mathsf{T} A) = r(AA^\mathsf{T}).$$

where $s_i(A)$ denote singular values. Both are "soft" versions of rank that are stable under small changes. Try Exercise 5.29 to get some intuition!

Remark 5.6.5 (High-probability guarantees)**.** We covered expectation bounds, but our argument actually gives a more informative high-probability guarantee: for any $u \ge 0$,

$$\|\Sigma_m - \Sigma\| \le C\Big(\sqrt{\frac{K^2 r(\log n + u)}{m}} + \frac{K^2 r(\log n + u)}{m}\Big)\,\|\Sigma\| \tag{5.28}$$

with probability at least $1 - 2e^{-u}$. Here $r = \operatorname{tr}(\Sigma)/\|\Sigma\| \le n$ is the effective rank. Check this in Exercise 5.26.

Remark 5.6.6 (Boundedness assumption)**.** The boundedness assumption (5.20) might seem strong, but it cannot be dropped in general: if X is isotropic but zero with high probability, the sample is likely to consist entirely of zeros, making covariance estimation impossible (formalize this argument in Exercise 5.27). However, this assumption can be relaxed (see Exercise 6.34). In practice, it is usually enforced by truncation – dropping a small percentage of samples with the largest norm.

To practice with covariance estimation, try Exercise 5.30 about sampling from frames, Exercise 5.31 to explore heavy-tailed random matrices, and Exercise 5.32 about matrix sketching.

5.7 Notes

For more on concentration inequalities, see the book [52, 209], also [21, Chapter 3], parts of [246, 136], and the tutorial [23].

The isoperimetric approach to concentration (Section 5.1) was first discovered by P.Lévy, who proved Theorems 5.1.4 and 5.1.3 (see [147]). A full proof of the isoperimetric inequality on the sphere (Theorem 5.1.3) can be found in [136, Section 2.2.1].

When V. Milman realized the power and generality of Lévy approach in 1970's, it led to far-reaching extensions of the *concentration of measure* principle, some of which we surveyed in Section 5.2. To keep this book concise, we skipped many key methods, including bounded differences, martingales, semigroups, transport, Poincaré and log-Sobolev inequalities, hypercontractivity, Stein's method, and Talagrand inequalities (see [330, 209, 52]). Most of what we covered in Sections 5.1 and 5.2 can be found in [21, Chapter 3], [246, 209].

The Gaussian isoperimetric inequality (Theorem 5.2.2) was first proved by B. Tsirelson, I. Ibragimov and V. Sudakov [87] and C. Borell [50]. There are several other proofs of Gaussian isoperimetric inequality, see [45, 22, 30]. There is also an elementary derivation of Gaussian concentration (Theorem 5.2.3) from Gaussian interpolation instead of isoperimetry, see [272].

Concentration on the Hamming cube (Theorem 5.2.5) is a consequence of Harper theorem, which is an isoperimetric inequality for the Hamming cube [155], see [46]. Concentration on the symmetric group (Theorem 5.2.6) is due to B. Maurey [227]. Both Theorems 5.2.5 and 5.2.6 can be also proved using martingales, see [246, Chapter 7].

The proof of concentration on Riemannian manifolds with positive curvature (Section 5.2.4) can be found e.g. in [209, Proposition 2.17]. Many interesting special cases follow from this general result, including Theorem 5.2.7 for the special orthogonal group [246, Section 6.5.1] and, consequently, Theorem 5.2.9 for the Grassmannian [246, Section 6.7.2]. A construction of Haar measure we mentioned in Remark 5.2.8 can be found e.g. in [246, Chapter 1] and [125, Chapter 2]; the survey [240] discusses numerically stable ways to generate random unitary matrices.

Concentration on the continuous cube and ball (Theorem 5.2.10) can be found in [209, Propositions 2.8, 2.9]. Theorem 5.2.11 on concentration for exponential densities is borrowed from [209, Proposition 2.18]. The proof of Talagrand concentration inequality (Theorem 5.2.12) can be found in [314, Theorem 6.6], [209, Corollary 4.10], [52, Section 6.6]; extensions for unbounded distributions can be found in [167].

The original version of Johnson-Lindenstrauss Lemma (Theorem 5.3.1) was proved in [178]. For various versions of this lemma, related results, applications, and bibliographic notes, see [225, Section 15.2]. The condition $m \gtrsim \varepsilon^{-2} \log N$ is optimal [197].

The approach to matrix concentration inequalities we follow in Section 5.4 originates in the work of R. Ahlswede and A. Winter [12]. A short proof of Golden-Thompson inequality (Theorem 5.4.7), a result on which Ahlswede-Winter approach rests, can be found e.g. in [41, Theorem 9.3.7] and [338]. While the work of R. Ahlswede and A. Winter was originally motivated by quantum information theory, their approach later found use in other areas, with early work including [349, 339, 148, 260].

The original argument of R. Ahlswede and A. Winter [12] yields a version of matrix Bernstein inequality that is somewhat weaker than Theorem 5.4.1, namely with $\sum_{i=1}^N \|\mathbb{E}\, X_i^2\|$ instead of σ. This quantity was later tightened by R. Oliveira [259] by a modification of Ahlswede-Winter's method and independently by J. Tropp [324] using Lieb inequality (Theorem 5.4.8) instead of Golden-Thompson. In this book, we mainly follow J. Tropp's proof of Theorem 5.4.1. J. Tropp's book [327] presents a self-contained proof of Lieb inequality (Theorem 5.4.8), matrix Hoeffding inequality from Exercise 5.21, matrix Chernoff inequality, and many other matrix analogs of classical scalar concentration inequalities. In matrix Bernstein inequality (Theorem 5.4.1) and thus in general covariance estimation (5.28), the dimension factor n can be replaced with effective rank [248], making these results dimension-free. For further results on matrix concentration, see [248, 188, 25, 26, 59].

The survey [270] discusses several useful trace inequalities and outlines proofs of Golden-Thompson inequality in Section 3 and Lieb inequality (embedded in the proof of Proposition 7). The book [127] also contains a detailed exposition of matrix Bernstein inequality and some of its variants (Section 8.5) and a proof of Lieb inequality (Appendix B.6).

We state in Remark 5.4.6 and prove in Exercise 5.18 that $1/x$ and $\ln x$ are matrix monotone functions. A general theory of matrix monotone functions was developed by K Loewner [218].

Matrix Khintchine inequality (Theorem 5.4.14) can alternatively be deduced from non-commutative Khintchine inequality due to F. Lust-Piquard [220]; see also [221, 65, 66, 281]. This derivation was first observed and used by M. Rudelson [287].

For community detection in networks (Section 5.5), see the notes at the end of Chapter 4. The approach to concentration of random graphs using matrix Bernstein inequality outlined in Section 5.5 was first proposed by R. Oliveira [259].

In Section 5.6 we discussed covariance estimation for general high-dimensional distributions following [340]. An alternative and earlier approach to covariance estimation, which gives similar results, relies on non-commutative Khintchine inequalities; it was developed earlier by M. Rudelson [287]. For more references on covariance estimation problem, see the notes at the end of Chapter 4.

The notion of effective rank, introduced in Remark 5.6.3 and explored in Exercise 5.29, is also known as the intrinsic dimension [327, Chapter 7].

The result of Exercise 5.31 is from [340, Section 5.4.2]. Exercise 5.32 gives an example of sketching – a popular technique in numerical linear algebra, see for example [113, 98, 289, 5, 57, 214, 326, 325] and the surveys [223, 255].

We omitted a lot of concentration inequalities from this chapter; one particularly useful is the *bounded differences inequality* (or *McDiarmid inequality*), which works not only for sums but for general functions of independent random variables. It is a generalization of Hoeffding inequality (Theorem 2.2.6).

Theorem 5.7.1 (Bounded differences inequality)**.** *Let $X = (X_1, \dots, X_N)$ be a random vector with independent entries.*[19] *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a measurable function. Assume that the value of $f(x)$ can change by at most $c_i > 0$ under an arbitrary change*[20] *of a single coordinate of $x \in \mathbb{R}^n$. Then, for any $t > 0$, we have*

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

Exercises

5.1 ☕☕ (Continuous, differentiable, and Lipschitz functions)

(a) Show that every Lipschitz function is uniformly continuous.

(b) Show that every differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is Lipschitz, and

$$\|f\|_{\text{Lip}} \leq \sup_{x \in \mathbb{R}^n} \|\nabla f(x)\|_2.$$

(c) Find a non-Lipschitz but uniformly continuous function $f : [-1, 1] \to \mathbb{R}$.

(d) Find a non-differentiable but Lipschitz function $f : [-1, 1] \to \mathbb{R}$.

5.2 ☕☕ Check all claims in Example 5.1.2.

5.3 ☕☕☕ (Blow-up of exponentially small sets) Let A be a subset of the sphere $\sqrt{n}S^{n-1}$ such that $\sigma(A) > 2\exp(-cs^2)$.

(a) Prove that $\sigma(A_s) > 1/2$.

(b) Deduce that $\sigma(A_{2t}) \geq 1 - 2\exp(-ct^2)$ for any $t \geq s$.

[19] The theorem remains valid if the random variables X_i take values in an abstract set $\mathcal{X}$ and $f : \mathcal{X} \to \mathbb{R}$.

[20] This means that for any index i and any $x_1, \dots, x_n, x_i'$, we have $|f(x_1, \dots, x_{i-1}, x_i, x_{i+1}, \dots, x_n) - f(x_1, \dots, x_{i-1}, x_i', x_{i+1}, \dots, x_n)| \leq c_i$.

5.4 ☕☕ (Concentration in the geodesic metric) We stated Theorem 5.1.3 for functions that are Lipschitz with respect to the Euclidean metric on the sphere. Show that it also works with the geodesic metric, which measures the shortest arc between points.

5.5 ☕ (Concentration on the unit sphere) We stated Theorem 5.1.3 for the scaled sphere $\sqrt{n}S^{n-1}$. Deduce that a Lipschitz function f on the *unit* sphere S^{n-1} satisfies

$$\|f(X) - \mathbb{E}\, f(X)\|_{\psi_2} \le \frac{C\|f\|_{\mathrm{Lip}}}{\sqrt{n}}. \tag{5.29}$$

where $X \sim \mathrm{Unif}(S^{n-1})$. Equivalently, show that for every $t \ge 0$,

$$\mathbb{P}\{|f(X) - \mathbb{E}\, f(X)| \ge t\} \le 2\exp\Big(-\frac{cnt^2}{\|f\|_{\mathrm{Lip}}^2}\Big) \tag{5.30}$$

5.6 ☕☕ (Concentration about the mean and median are equivalent) For a random variable Z with median M, show that

$$c\|Z - \mathbb{E}\, Z\|_{\psi_2} \le \|Z - M\|_{\psi_2} \le C\|Z - \mathbb{E}\, Z\|_{\psi_2},$$

This allows us to swap the mean and median in concentration inequalities.

5.7 ☕☕☕ (Concentration is equivalent to blow-up) In Section 5.1.4, we derived Lipschitz concentration on the sphere from the blow-up phenomenon (Lemma 5.1.6). Let's reverse this logic and show that Lipschitz concentration is always equivalent to blow-up. Consider a random vector X taking values in a metric space (T, d). Suppose there exists $K > 0$ such that

$$\|f(X) - \mathbb{E}\, f(X)\|_{\psi_2} \le K\|f\|_{\mathrm{Lip}}$$

for every Lipschitz function $f : T \to \mathbb{R}$. For a subset $A \subset T$, define $\sigma(A) := \mathbb{P}\{X \in A\}$, thus making σ is a probability measure on T. Show that if $\sigma(A) \ge 1/2$, then[21] for every $t \ge 0$ we have

$$\sigma(A_t) \ge 1 - 2\exp(-ct^2/K^2).$$

5.8 ☕☕ (Gaussian concentration) Deduce the Gaussian concentration inequality (Theorem 5.2.3) from Gaussian isoperimetric inequality (Theorem 5.2.2).

5.9 ☕☕☕ (Concentration of maximum) In Proposition 2.7.6, we estimated the expected maximum of n subgaussians. Now, let's show that the maximum concentrates.

(a) Let $X_1, \ldots, X_n$ be independent $N(0, 1)$ random variables. Prove that[22]

$$\|\max_i X_i - \mathbb{E}\max_i X_i\|_{\psi_2} \le C.$$

[21] Here the neighborhood A_t of a set A is defined in the same way as before, that is $A_t := \{x \in T :\ \exists y \in A \text{ such that } d(x, y) \le t\}$.

[22] This result is often used together with the fact $\mathbb{E}\max_i X_i \approx \sqrt{2\ln n}$, see Exercise 2.38(b).

(b) More generally, let $X_1, \dots, X_n$ be jointly normal random variables (recall Section 3.3.2), not necessarily independent. Prove that

$$\|\max_i X_i - \mathbb{E}\max_i X_i\|_{\psi_2} \le C \max_i \sqrt{\mathrm{Var}(X_i)}.$$

5.10 ☕☕☕ (Concentration around the L^p norm) In Exercise 5.6 we saw how we can always swap the mean and median in concentration inequalities. Let's prove the same for the L^p norm. For a random variable Z with $\mathbb{E}\, Z \ge 0$ and $p \ge 1$, show that

$$\big\|Z - \|Z\|_{L^p}\big\|_{\psi_2} \le C\sqrt{p}\big\|Z - \mathbb{E}\, Z\big\|_{\psi_2}.$$

5.11 ☕☕ (Pushing forward Gaussian to uniform measures) Let $\Phi(x)$ denote the cumulative distribution function of the standard normal distribution $N(0,1)$. Consider a random vector $Z = (Z_1, \dots, Z_n) \sim N(0, I_n)$. Check that

$$\phi(Z) := \big(\Phi(Z_1), \dots, \Phi(Z_n)\big) \sim \mathrm{Unif}([0,1]^n).$$

5.12 ☕☕ (Concentration on the continuous cube) Prove Theorem 5.2.10 for $T = [0,1]^n$ as follows:

(a) Express $X = \phi(Z)$ using Exercise 5.11. Then, use Gaussian concentration to control the deviation of $f(\phi(Z))$ in terms of $\|f \circ \phi\|_{\mathrm{Lip}} \le \|f\|_{\mathrm{Lip}}\,\|\phi\|_{\mathrm{Lip}}$.

(b) Show that $\|\phi\|_{\mathrm{Lip}}$ is bounded by a constant, and then finish the proof of the theorem.

5.13 ☕☕☕ (Concentration on the continuous ball) Prove Theorem 5.2.10 for $T = B_2^n$ by a strategy similar to Exercise 5.12:

(a) Define a function $\phi : \mathbb{R}^n \to \sqrt{n}B_2^n$ that pushes forward the Gaussian measure on $\mathbb{R}^n$ into the uniform measure on $\sqrt{n}B_2^n$.

(b) Check that ϕ has bounded Lipschitz norm.

5.14 ☕☕☕ (Johnson-Lindenstrauss with subgaussian matrices) Let A be an $m \times n$ random matrix with mean-zero, subgaussian, isotropic rows. Show that Johnson-Lindenstrauss lemma (Theorem 5.3.1) holds for $Q = (1/\sqrt{m})A$. This covers the important *binary* Johnson-Lindenstrauss case, where A has ± 1 entries (how?)

5.15 ☕☕☕ (Optimality of Johnson-Lindenstrauss lemma) Let's show that the dimension $n = O(\log N)$ in Johnson-Lindenstrauss lemma (Theorem 5.3.1) is optimal – even if we allow nonlinear dimension reduction maps. Here is how:

(a) Let $z_1, \dots, z_N$ be vectors in $\mathbb{R}^n$ that satisfy $1 < \|z_i - z_j\|_2 \le 2$ for all distinct i, j. Show that $N \le 5^n$.

(b) Let $n < \frac{1}{2}\log N$. Find vectors $x_1, \dots, x_N$ in $\mathbb{R}^N$ for which there does not exist any map $T : \mathbb{R}^N \to \mathbb{R}^n$ that satisfies

$$0.99\|x_i - x_j\|_2 \le \|T(x_i) - T(x_j)\|_2 \le 1.01\|x_i - x_j\|_2 \quad \text{for all } i, j = 1, \dots, N.$$

5.16 ☕☕ (Matrix Taylor series) Let's practice with Definition 5.4.2 of matrix functions. Let X be an symmetric matrix.

(a) For any polynomial $f(x) = a_0 + a_1 x + \cdots + a_p x^p$, check that $f(X) = a_0 I_n + a_1 X + \cdots + a_p X^p$. On the right hand side, we use matrix addition and multiplication, so X^p there is interpreted as X matrix-multiplied by itself p times.

(b) More generally, if $f(x) = \sum_{k=1}^{\infty} a_k (x - x_0)^k$ is a convergent power series in a neighborhood of x_0, check that $f(X) = \sum_{k=1}^{\infty} a_k (X - x_0 I_n)^k$, where this matrix series converges in the Frobenius norm (where?), and thus also in the operator norm.

(c) Argue that

$$e^X = I_n + X + \frac{X^2}{2!} + \frac{X^3}{3!} + \cdots$$

5.17 ☕☕ (Matrix monotonicity)

(a) Check that for commuting matrices X, Y,

$$X \preceq Y \quad \Longrightarrow \quad f(X) \preceq f(Y) \quad \text{for any weakly increasing } f : \mathbb{R} \to \mathbb{R}.$$

(b) Give an example showing that property (a) may fail for non-commuting matrices.

5.18 ☕☕☕ (Matrix monotonicity of $1/x$ and $\ln x$) Let X and Y be symmetric $n \times n$ matrices such that X is invertible and

$$0 \preceq X \preceq Y.$$

(a) Prove that Y is also invertible, and

$$X^{-1} \succeq Y^{-1} \succeq 0.$$

(b) Check the identity

$$\ln x = \int_0^\infty \Big(\frac{1}{1+t} - \frac{1}{x+t} \Big)\, dt.$$

(c) Using the formula in (b), deduce from (a) that

$$\ln X \preceq \ln Y.$$

5.19 ☕☕ (Matrix exponentiation) Let X and Y be $n \times n$ symmetric matrices.

(a) Show that if the matrices commute, i.e. $XY = YX$, then $e^{X+Y} = e^X e^Y$.

(b) Find an example of matrices X and Y such that $e^{X+Y} \neq e^X e^Y$.

5.20 ☕☕ Deduce from matrix Bernstein inequality (Theorem 5.4.1) the following expectation bound:

$$\mathbb{E} \Big\| \sum_{i=1}^N X_i \Big\| \lesssim \Big\| \sum_{i=1}^N \mathbb{E}\, X_i^2 \Big\|^{1/2} \sqrt{1 + \log n} + K(1 + \log n).$$

5.21 ☕☕☕ (Matrix Hoeffding inequality) Prove matrix Hoeffding inequality (Theorem 5.4.13) following the proof of matrix Bernstein inequality (Theorem 5.4.1).

5.22 ☕☕☕ (Matrix Khintchine inequality) Deduce matrix Khintchine inequality (Theorem 5.4.14) from matrix Hoeffding inequality (Theorem 5.4.13).

5.23 ☕☕ (Matrix Bernstein inequality for rectangular matrices) Let $X_1, \dots, X_N$ be independent, mean-zero, $m \times n$ random matrices, such that $\|X_i\| \le K$ almost surely for all i. Prove that for any $t \ge 0$, we have

$$\mathbb{P}\Big\{\Big\|\sum_{i=1}^N X_i\Big\| \ge t\Big\} \le 2(m+n)\exp\Big(-\frac{t^2/2}{\sigma^2 + Kt/3}\Big),$$

where

$$\sigma^2 = \Big\|\sum_{i=1}^N \mathbb{E}\, X_i^\mathsf{T} X_i\Big\| + \Big\|\sum_{i=1}^N \mathbb{E}\, X_i X_i^\mathsf{T}\Big\|.$$

5.24 ☕☕ (Matrix Khintchine for rectangular matrices) Let's prove a version of matrix Khintchine inequality (Theorem 5.4.14) for rectangular matrices. Let $\varepsilon_1, \dots, \varepsilon_N$ be independent Rademacher random variables and let $A_1, \dots, A_N$ be any (fixed) $m \times n$ matrices. Show that, for every $p \in [1, \infty)$ we have

$$\Big(\mathbb{E}\Big\|\sum_{i=1}^N \varepsilon_i A_i\Big\|^p\Big)^{1/p} \le C\sigma\sqrt{p + \log(m+n)}, \quad \text{where } \sigma^2 = \Big\|\sum_{i=1}^N A_i^\mathsf{T} A_i\Big\| + \Big\|\sum_{i=1}^N A_i A_i^\mathsf{T}\Big\|.$$

5.25 ☕☕☕ (Stochastic block model without loops) Definition 4.5.1 of the stochastic block model $G(n, p, q)$ allows loops, meaning every vertex connects to itself with probability p. Modify this definition to disallow loops, then prove a version of Theorem 5.5.1 under this adjustment.

5.26 ☕☕ (Covariance estimation: a high-probability guarantee) Prove the high-probability version of the covariance estimation result stated in Remark 5.6.5.

5.27 ☕☕ (Covariance estimation: the boundedness assumption) Show that the boundedness assumption (5.20) cannot be dropped from Theorem 5.6.1 in general. (See Remark 5.6.6.)

5.28 ☕☕☕ (Logarithmic factor in covariance estimation and matrix Bernstein) Let's show that a logarithmic factor is unavoidable in general covariance estimation (5.24) and matrix Bernstein inequality (5.17).

(a) Give an example of a probability distribution where the covariance estimation bound $\|\Sigma_m - \Sigma\| < \|\Sigma\|$ fails with high probability unless $m \gtrsim n \log n$.

(b) Conclude that the logarithmic factors in matrix Bernstein inequality (5.17) cannot be dropped.

5.29 ☕☕ (Effective rank) For an $n \times n$ symmetric positive-semidefinite matrix Σ, we defined the *effective rank* of Σ in (5.26) as

$$r = \frac{\operatorname{tr}(\Sigma)}{\|\Sigma\|}.$$

Unlike exact rank, which counts the nonzero eigenvalues, the effective rank is a more robust sense of how many eigenvalues actually matter for the matrix's structure.

(a) Show that $1 \le r(\Sigma) \le \operatorname{rank}(\Sigma) \le n$.

(b) Show that this inequality is optimal: there are matrices of full rank but whose effective rank is arbitrarily close to 1.

(c) (Stability) Show that, unlike the (algebraic) rank, the effective rank is a continuous function of the matrix (e.g. with respect to the operator norm).

(d) Show that if the random vector X takes values in a k-dimensional subspace of $\mathbb{R}^n$, then $\Sigma = \mathbb{E}\, XX^\mathsf{T}$ satisfies $\operatorname{rank}(\Sigma) \le k$ and thus $r(\Sigma) \le k$.

5.30 ☕☕ (Sampling from frames) Consider an equal-norm Parseval frame[23] $(u_1, \ldots, u_m)$ in $\mathbb{R}^n$. Show that, with high probability, a random sample of

$$m \gtrsim n \log n$$

frame elements still forms an "approximate" frame (formalize this notion).

5.31 ☕☕☕ (Random matrices with general independent rows) Let's prove a version of Theorem 4.6.1 for random matrices with arbitrary, not necessarily subgaussian, distributions of rows. Let A be an $m \times n$ random matrix with independent, isotropic rows A_i. Assume that for some $K \ge 0$,

$$\|A_i\|_2 \le K\sqrt{n} \quad \text{almost surely for every } i.$$

Prove that, for every $t \ge 1$, one has

$$\sqrt{m} - Kt\sqrt{n \log n} \le s_n(A) \le s_1(A) \le \sqrt{m} + Kt\sqrt{n \log n}$$

with probability at least $1 - 2n^{-ct^2}$.

5.32 ☕☕☕ (Matrix sketching) Some matrices are too big to compute eigenvalues and eigenvectors directly. A trick to handle this is subsampling–picking a random set of rows or columns to make a smaller matrix. Let's show that we can approximate the singular values of a tall $N \times n$ matrix A using a smaller matrix B, made by randomly selecting $m = O(n \log n)$ rows (with replacement, uniform probability). Assuming all rows of A have the same norm, prove that if $m \ge Cn \log n$, then the singular values satisfy

$$\max_{i=1,\ldots,n} \left| s_i(A)^2 - \frac{N}{n} s_i(B)^2 \right| \le 0.1\, s_1(A)^2$$

with probability at least 0.9.

[23] The concept of frames was introduced in Section 3.3.5. By equal-norm frame we mean that $\|u_i\|_2 = \|u_j\|_2$ for all i and j.

6

Quadratic Forms, Symmetrization and Contraction

In this chapter, we introduce a number of basic tools of high-dimensional probability: decoupling in Section 6.1, concentration for quadratic forms (the Hanson-Wright inequality) in Section 6.2, symmetrization in Section 6.3 and contraction in Section 6.6.

We illustrate these tools with a few applications. In Section 6.4 (and Exercise 6.28), we show that the operator norm of a random matrix is essentially equivalent to the maximal Euclidean norm of its rows and columns. We use this result in Section 6.5 for *matrix completion*, where we recover a low-rank matrix from a random sample of its entries.

Many more topics are explored in exercises. You will bound the norm of a subgaussian random vector (Exercise 6.10) and apply it to derive Hanson-Wright inequality for subgaussian random vectors (Exercise 6.11) and for *mean estimation* of a subgaussian distribution (Exercise 6.12); extend the concentration of norm (Theorem 3.1.1) to anisotropic random vectors (Exercise 6.13) and apply it for the distance from a random vector to a subspace (Exercise 6.14) and graph cutting (Exercise 6.15), explore the notion of type of normed spaces (Exercises 6.23, 6.24) and apply it to extend the approximate Caratheodory theorem (Theorem 0.0.2) for the ℓ^p norm, and extend covariance estimation for unbounded distributions (Exercise 6.34).

6.1 Decoupling

In Chapter 2, we studied sums of independent random variables like

$$\sum_{i=1}^{n} a_i X_i \tag{6.1}$$

where $X_1, \dots, X_n$ are independent random variables and a_i are fixed coefficients. Now, let's look at *quadratic forms* like

$$\sum_{i,j=1}^{n} a_{ij} X_i X_j = X^\mathsf{T} A X = \langle X, AX \rangle \tag{6.2}$$

[0] This material will be published by Cambridge University Press as *High-Dimensional Probability, 2nd Edition* by Roman Vershynin. This pre-publication version is free to view and download for personal use only. Not for re-distribution, re-sale or use in derivative works. ©Roman Vershynin 202X.
If you want to be notified when the printed version is available, let me know: rvershyn@uci.edu

where $A = (a_{ij})$ is an $n \times n$ coefficient matrix and $X = (X_1, \dots, X_n)$ is a random vector with independent coordinates. Such a quadratic forms are known as *chaos*.

Computing the expectation of a chaos is easy. For simplicity, let us assume that X_i have zero means and unit variances. Then

$$\mathbb{E}\, X^\mathsf{T} A X = \sum_{i,j=1}^{n} a_{ij}\, \mathbb{E}\, X_i X_j = \sum_{i=1}^{n} a_{ii} = \operatorname{tr} A.$$

It is harder to establish a concentration for a chaos, because the terms of the sum in (6.2) are not independent. We can overcome this difficulty by the *decoupling* technique, which we will introduce now.

The goal of decoupling is to replace the quadratic form (6.2) with the *bilinear* form

$$\sum_{i,j=1}^{n} a_{ij} X_i X'_j = X^\mathsf{T} A X' = \langle X, AX' \rangle,$$

where $X' = (X'_1, \dots, X'_n)$ is an independent copy of X – a random vector independent of X and with the same distribution as X. Bilinear forms are easier to analyze than the quadratic forms, since they are linear in X. If we condition on X', we may treat the bilinear form as a sum of independent random variables

$$\sum_{i=1}^{n} \Big(\sum_{j=1}^{n} a_{ij} X'_j \Big) X_i = \sum_{i=1}^{n} b_i X_i$$

with fixed coefficients b_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.*[1] *Let X be a random vector in $\mathbb{R}^n$ with independent mean-zero coordinates, and let X' be an independent copy. Then, for every convex function $F : \mathbb{R} \to \mathbb{R}$,*

$$\mathbb{E}\, F(X^\mathsf{T} A X) \le \mathbb{E}\, F(4 X^\mathsf{T} A X'). \tag{6.3}$$

Here is the proof idea in a nutshell. We start by replacing the chaos $X^\mathsf{T} A X = \sum_{i,j} a_{ij} X_i X_j$ by the "partial chaos"

$$\sum_{(i,j) \in I \times I^c} a_{ij} X_i X_j$$

where $I \subset \{1, \dots, n\}$ is a randomly chosen subset of indices. The benefit is that sums run over disjoint sets for i and j, so we can swap X_j with an independent copy X'_j without changing the distribution. Then, we use Jensen's inequality to extend this partial chaos back to the full sum $X^\mathsf{T} A X' = \sum_{i,j} a_{ij} X_i X'_j$. Now we pass to a detailed proof.

Proof **Step 1: Randomly selecting a partial sum.** To specify a random subset of indices I, let us introduce *selectors* – independent Bernoulli random variables $\delta_1, \dots, \delta_n \in \{0, 1\}$ with $\mathbb{P}\{\delta_i = 0\} = \mathbb{P}\{\delta_i = 1\} = 1/2$, and define

$$I := \{i : \ \delta_i = 1\}.$$

[1] Diagonal-free means that all diagonal entries of A equal zero.

Condition on X. Since by assumption $a_{ii} = 0$ and

$$\mathbb{E}\,\delta_i(1-\delta_j) = \frac{1}{2}\cdot\frac{1}{2} = \frac{1}{4} \quad \text{for all } i \neq j,$$

we may express the chaos as

$$X^{\mathsf{T}}AX = \sum_{i\neq j} a_{ij}X_iX_j = 4\,\mathbb{E}_\delta \sum_{i\neq j} \delta_i(1-\delta_j)a_{ij}X_iX_j = 4\,\mathbb{E}_I \sum_{(i,j)\in I\times I^c} a_{ij}X_iX_j.$$

(The subscripts δ and I indicate the sources of randomness in these conditional expectations. Since X is fixed, the expectations are taken over the random selectors $\delta = (\delta_1,\ldots,\delta_n)$, or equivalently, the random index set I. We will keep using this notation later.)

Step 2. Applying F. Apply the function F to both sides and take expectation over X. Using Jensen inequality and Fubini theorem, 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 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 of I until the end of the proof (and drop the subscripts X in the expectation for convenience.) 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 get

$$\mathbb{E}\,F(X^{\mathsf{T}}AX) \le \mathbb{E}\,F\Big(4\sum_{(i,j)\in I\times I^c} a_{ij}X_iX'_j\Big).$$

Step 3. Completing the partial sum. It remains to complete the sum in the right side to the sum over all pairs of indices. We want to show that

$$\mathbb{E}\,F\Big(4\sum_{(i,j)\in I\times I^c} a_{ij}X_iX'_j\Big) \le \mathbb{E}\,F\Big(4\sum_{(i,j)\in [n]\times[n]} a_{ij}X_iX'_j\Big), \tag{6.4}$$

where we use the notation $[n] = \{1,\ldots,n\}$. To do this, we decompose the sum in the right side as

$$\sum_{(i,j)\in[n]\times[n]} a_{ij}X_iX'_j = \underbrace{\sum_{(i,j)\in I\times I^c} a_{ij}X_iX'_j}_{Y} + \underbrace{\sum_{(i,j)\in I\times I} a_{ij}X_iX'_j + \sum_{(i,j)\in I^c\times[n]} a_{ij}X_iX'_j}_{Z}.$$

Condition on all $(X_i)_{i\in I}$ and $(X'_j)_{j\in I^c}$, and denote this conditional expectation by $\mathbb{E}'$. This fixes Y, while Z has zero conditional expectation (check!). Thus, by Jensen inequality, we get

$$F(4Y) = F(4Y + \mathbb{E}'[4Z]) = F(\mathbb{E}'[4Y+4Z]) \le \mathbb{E}'\,F(4Y+4Z).$$

Finally, taking expectation over all remaining random variables, we get

$$\mathbb{E}\, F(4Y) \le \mathbb{E}\, F(4Y + 4Z).$$

This proves (6.4) and finishes the argument. □

Remark 6.1.2 (Diagonal-free assumption)**.** This assumption is essential in Theorem 6.1.1, since the conclusion fails for diagonal matrices when $F(x) = x$ (why?). But we can include the diagonal on the right hand side: for any $n \times n$ matrix $A = (a_{ij})$, we get

$$\mathbb{E}\, F\Big(\sum_{i,j:\, i \ne j} a_{ij} X_i X_j\Big) \le \mathbb{E}\, F\Big(4 \sum_{i,j} a_{ij} X_i X'_j\Big) \tag{6.5}$$

Check this in Exercise 6.1 and explore other variants of decoupling in Exercises 6.2–6.4.

6.2 Hanson-Wright inequality

As a warm-up, let's ask: if X is a subgaussian random vector in $\mathbb{R}^n$, what can we say about its norm? If X has independent coordinates, the norm concentrates (Theorem 3.1.1). But in general, it does not have to – it can be too small with high probability (Exercise 3.37). However, it can't be too large:

Proposition 6.2.1 (The norm of a subgaussian random vector)**.** *Let X be a mean-zero subgaussian random vector in $\mathbb{R}^n$ with $\|X\|_{\psi_2} \le K$. Then, for every $t \ge 0$,*

$$\mathbb{P}\big\{\|X\|_2 \ge CK(\sqrt{n} + t)\big\} \le e^{-t^2}.$$

Proof Without loss of generality, we may assume that $K = 1$. Squaring and exponentiating both sides and using Markov inequality, we get

$$\mathbb{P}\big\{c\|X\|_2 \ge \sqrt{n} + t\big\} \le e^{-(n+t^2)}\, \mathbb{E} \exp(c^2 \|X\|_2^2). \tag{6.6}$$

Now we use a **Gaussian replacement** trick: for some absolute constant $c > 0$, we claim that

$$\mathbb{E} \exp(c^2 \|X\|_2^2) \le \mathbb{E} \exp(\|g\|_2^2/4) \quad \text{where} \quad g \sim N(0, I_n). \tag{6.7}$$

To see this, condition on X (this treating it as a fixed vector); then $\langle g, X\rangle \sim N(0, \|X\|_2^2)$ by Corollary 3.3.2, so[2]

$$\exp(c^2 \|X\|_2^2) = \mathbb{E}_g \exp(\sqrt{2} c \langle g, X\rangle),$$

where $\mathbb{E}_g$ denotes the conditional expectation over g (conditional on X). Now take expectation over X on both sides and apply Fubini:

$$\mathbb{E}_X \exp(c^2 \|X\|_2^2) = \mathbb{E}_X\, \mathbb{E}_g \exp(\sqrt{2} c \langle g, X\rangle) = \mathbb{E}_g\, \mathbb{E}_X \exp(\sqrt{2} c \langle X, g\rangle). \tag{6.8}$$

[2] Here we used that the MGF of $h = N(0, \sigma^2)$ is $\mathbb{E} \exp(\lambda h) = \exp(\lambda^2 \sigma^2 / 2)$, see (2.16).

When we condition on g (thus treating g as a fixed vector), the subgaussian norm of $\langle X, g\rangle$ is at most 1 by assumption[3] $\|X\|_{\psi_2} = 1$, so Proposition 2.6.6(iv) gives

$$\mathbb{E}_X \exp(\sqrt{2}c\langle X, g\rangle) \le \exp(\|g\|_2^2/4)$$

for some absolute constant $c > 0$. Substitute this into (6.8), and (6.7) is proved.

Since $\|g\|_2^2 = g_1^2 + \cdots + g_n^2$ where $g_i \sim N(0,1)$ are i.i.d., we have[4]

$$\mathbb{E}_g \exp\left(\|g\|_2^2/4\right) = \left(\mathbb{E}\exp(g_1^2/4)\right)^n = (\sqrt{2})^n \le e^n.$$

Substitute into (6.7) and (6.6) to complete the proof. □

To practice Gaussian replacement, prove an *anisotropic* version of Proposition 6.2.1 in Ecercises 6.9, 6.10.

The Gaussian replacement trick we just learned will be handy when proving concentration of a chaos – a quadratic version of Bernstein inequalty:

Theorem 6.2.2 (Hanson-Wright inequality)**.** *Let A be an $n\times n$ matrix. Let $X = (X_1,\ldots,X_n) \in \mathbb{R}^n$ be a random vector with independent, mean-zero, subgaussian coordinates. Then, for every $t \ge 0$, we have*

$$\mathbb{P}\left\{|X^\mathsf{T} AX - \mathbb{E}\, X^\mathsf{T} AX| \ge t\right\} \le 2\exp\left[-c\min\left(\frac{t^2}{K^4\|A\|_F^2}, \frac{t}{K^2\|A\|}\right)\right],$$

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

Our proof will be based on bounding the MGF of $X^\mathsf{T} AX$. Here is the plan:

(a) replace $X^\mathsf{T} AX$ by $X^\mathsf{T} AX'$ by decoupling;
(b) replace $X^\mathsf{T} AX'$ by $g^\mathsf{T} Ag'$ using Gaussian replacement, for $g \sim N(0, I_n)$;
(c) compute $g^\mathsf{T} Ag'$ by diagonalizing A using rotation invariance of $N(0, I_n)$.

We start with (b), a version of Gaussian replacement we learned in the proof of Proposition 6.2.1. Can you prove it on your own without peeking at the proof?

Lemma 6.2.3 (Gaussian replacement)**.** *Let A be an $n \times n$ matrix. Let X be a mean-zero, subgaussian random vector in $\mathbb{R}^n$ with $\|X\|_{\psi_2} \le K$, and X' be its independent copy. Let $g, g' \sim N(0, I_n)$ be independent. Then, for any $\lambda \in \mathbb{R}$,*

$$\mathbb{E}\exp(\lambda X^\mathsf{T} AX') \le \mathbb{E}\exp(CK^2\lambda g^\mathsf{T} Ag').$$

Proof Condition on X' and take expectation over X, which we denote $\mathbb{E}_X$. Then the random variable $X^\mathsf{T} AX' = \langle X, AX'\rangle$ is (conditionally) subgaussian, with subgaussian bounded by $K\|AX'\|_2$. Then Proposition 2.6.6(iv) gives

$$\mathbb{E}_X \exp(\lambda X^\mathsf{T} AX') \le \exp(C\lambda^2K^2\|AX'\|_2^2), \quad \lambda \in \mathbb{R}. \tag{6.9}$$

[3] Recall Definition 3.4.1.

[4] Here we used the MGF of square of $h = N(0,1)$: $\mathbb{E}\exp(\lambda h^2) = (1-2\lambda)^{-1/2}$ for $\lambda < \frac{1}{2}$ (check!).

Compare this to the normal MGF formula (2.16). Applied to the normal random variable $g^{\mathsf{T}} AX' = \langle g, AX' \rangle$ (still conditionally on X'), it gives

$$\mathbb{E}_g \exp(\mu g^{\mathsf{T}} AX') = \exp(\mu^2 \|AX'\|_2^2/2), \quad \mu \in \mathbb{R}. \tag{6.10}$$

Setting $\mu = \sqrt{2C}K\lambda$, we match the right sides of (6.9) and (6.10) and obtain

$$\mathbb{E}_X \exp(\lambda X^{\mathsf{T}} AX') \leq \mathbb{E}_g \exp(\sqrt{2C}K\lambda g^{\mathsf{T}} AX').$$

Taking expectation over X' of both sides, we see that we have replaced X by g in the chaos, at a cost of the factor $\sqrt{2C}K$. Repeating for X', we can replace it with g', paying another factor $\sqrt{2C}K$. (Try it yourself!) □

Now we move on to step (c) of our plan:

Lemma 6.2.4 (MGF of a Gaussian quadratic form)**.** *Let $A = (a_{ij})$ be an $n \times n$ matrix, and let $g, g' \sim N(0, I_n)$ be independent. Then*

$$\mathbb{E} \exp(\lambda g^{\mathsf{T}} Ag') \leq \exp\left(\lambda^2 \|A\|_F^2\right) \quad \textit{whenever} \quad |\lambda| \leq \frac{1}{2\|A\|}.$$

Proof Let's use rotation invariance of the normal distribution to diagonalize A. With its singular value decomposition (4.4), $A = U\Sigma V^{\mathsf{T}}$, we can write

$$g^{\mathsf{T}} Ag' = (U^{\mathsf{T}} g)^{\mathsf{T}} \Sigma (V^{\mathsf{T}} g').$$

By rotation invariance of the normal distribution (Proposition 3.3.1), $U^{\mathsf{T}} g$ and $V^{\mathsf{T}} g'$ are independent standard normal random vectors in $\mathbb{R}^n$. So,

$$g^{\mathsf{T}} Ag' \overset{\text{dist}}{=} g^{\mathsf{T}} \Sigma g' = \sum_{i=1}^n s_i g_i g_i'$$

where "dist" indicates equality in distribution and s_i are the singular values of A. This is a sum of independent random variables, so

$$\mathbb{E} \exp(\lambda g^{\mathsf{T}} Ag') = \mathbb{E} \prod_i \mathbb{E} \exp(\lambda s_i g_i g_i') = \prod_i \mathbb{E} \exp(\lambda s_i g_i g_i'). \tag{6.11}$$

Now, for each i and $t \in \mathbb{R}$, we have

$$\mathbb{E} \exp(t g_i g_i') = \mathbb{E} \exp\left(\frac{t^2 g_i^2}{2}\right) = \frac{1}{\sqrt{1-t^2}} \leq \exp(t^2) \quad \text{if } t^2 \leq \frac{1}{2}.$$

The first identity here follows by conditioning on g_i and using the MGF formula (2.16) for the normal random variable g_i'; the other steps are direct calculations (check them!).[5] Substituting this bound with $t = \lambda s_i$ into (6.11), we get

$$\mathbb{E} \exp(\lambda g^{\mathsf{T}} Ag') \leq \exp\left(\lambda^2 \sum_i s_i^2\right) \quad \text{if} \quad \lambda^2 \leq \frac{1}{2 \max s_i^2}.$$

Since s_i are the singular values of A, we have $\sum_i s_i^2 = \|A\|_F^2$ and $\max_i s_i = \|A\|$ (recall Lemma 4.1.11). The lemma is proved. □

[5] Or use (2.31) for the subexponential random variable g_i^2 to get the same result up to constants.

Proof of Hanson-Wright inequality (Theorem 6.2.2) Without loss of generality, we may assume that $K = 1$. (Why?) As usual, it is enough to bound the one-sided tail

$$p := \mathbb{P}\{X^\mathsf{T} AX - \mathbb{E}\, X^\mathsf{T} AX \ge t\}.$$

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 would complete the proof.

In terms of the entries of $A = (a_{ij})_{i,j=1}^n$, we have

$$X^\mathsf{T} AX = \sum_{i,j} a_{ij} X_i X_j \quad \text{and} \quad \mathbb{E}\, X^\mathsf{T} AX = \sum_i a_{ii}\, \mathbb{E}\, X_i^2,$$

where we used the mean-zero assumption and independence. So

$$X^\mathsf{T} AX - \mathbb{E}\, X^\mathsf{T} AX = \sum_i a_{ii}(X_i^2 - \mathbb{E}\, X_i^2) + \sum_{i,j:\, i \ne j} a_{ij} X_i X_j.$$

The problem reduces to estimating the diagonal and off-diagonal sums:

$$p \le \mathbb{P}\Big\{ \sum_i a_{ii}(X_i^2 - \mathbb{E}\, X_i^2) \ge t/2 \Big\} + \mathbb{P}\Big\{ \sum_{i,j:\, i \ne j} a_{ij} X_i X_j \ge t/2 \Big\} =: p_1 + p_2.$$

Step 1: Diagonal sum. Since X_i are independent and subgaussian, $X_i^2 - \mathbb{E}\, X_i^2$ are independent, mean-zero, and subexponential, 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 centering (2.26) and Lemma 2.8.5.) Then Bernstein inequality (Corollary 2.9.2) gives

$$p_1 \le \exp\Big[-c \min\Big(\frac{t^2}{\sum_i a_{ii}^2}, \frac{t}{\max_i |a_{ii}|} \Big) \Big] \le \exp\Big[-c \min\Big(\frac{t^2}{\|A\|_F^2}, \frac{t}{\|A\|} \Big) \Big].$$

Step 2: Off-diagonal sum. It remains to bound the off-diagonal sum

$$S := \sum_{i,j:\, i \ne j} a_{ij} X_i X_j.$$

Let $\lambda > 0$ be a parameter to be determined later. By Markov inequality, we have

$$p_2 = \mathbb{P}\{S \ge t/2\} = \mathbb{P}\{\lambda S \ge \lambda t/2\} \le \exp(-\lambda t/2)\, \mathbb{E} \exp(\lambda S). \tag{6.12}$$

Now,

$$\begin{aligned}
\mathbb{E} \exp(\lambda S) &\le \mathbb{E} \exp(4\lambda X^\mathsf{T} AX') \quad \text{(by decoupling (6.5))} \\
&\le \mathbb{E} \exp(C_1 \lambda g^\mathsf{T} Ag') \quad \text{(by Gaussian replacement, Lemma 6.2.3)} \\
&\le \exp(C\lambda^2 \|A\|_F^2) \quad \text{(by Gaussian MGF bound, Lemma 6.2.4)},
\end{aligned}$$

whenever $|\lambda| \le \frac{1}{2\|A\|}$. Putting this bound into (6.12), we obtain

$$p_2 \le \exp\big(-\lambda t/2 + C\lambda^2 \|A\|_F^2 \big).$$

Optimizing over $0 \le \lambda \le \frac{1}{2\|A\|}$, we conclude that

$$p_2 \le \exp\Big[-c\min\Big(\frac{t^2}{\|A\|_F^2}, \frac{t}{\|A\|}\Big)\Big].$$

(Check!) Summarizing, we obtained the desired bounds for the probabilities of diagonal deviation p_1 and off-diagonal deviation p_2. Putting them together,[6] we complete the proof of Theorem 6.2.2. □

Practice these techniques! Find a direct proof of Hanson-Wright for the Gaussian distribution (Exercise 6.7), prove versions where X_i are random vectors (Exercise 6.8) or have dependent entries (Exercise 6.11), and apply them to mean estimation (Exercise 6.12), concentration of norm for anisotropic distributions (Exercise 6.13), distance from a random vector to a subspace (Exercise 6.14), and graph cutting (Exercise 6.15).

6.3 Symmetrization

A random variable X is called *symmetric* if it has the same distribution as $-X$. A basic is a *Rademacher* random variable, which takes values -1 and 1 with equal probabilities. Mean-zero normal random variables $X \sim N(0, \sigma^2)$ are also symmetric, while Poisson or exponential random variables are not.

This section introduces *symmetrization*, a. useful trick for reducing problems to symmetric distributions – and sometimes even to the Rademacher distribution. It is based on the following elementary observation:

Lemma 6.3.1 (Constructing symmetric distributions)**.** *Let X be a random variable and ξ be an independent Rademacher random variable.*

(a) ξX and $\xi|X|$ are identically distributed and symmetric.
(b) If X is symmetric, both ξX and $\xi|X|$ have the same distribution as X.
(c) If X' is an independent copy of X, then $X - X'$ is symmetric.

Proof Let's just check that ξX is symmetric – you will prove the rest in Exercise 6.16. For any interval $A \subset \mathbb{R}$, the law of total probability (1.17) gives

$$\begin{aligned}\mathbb{P}\{\xi X \in A\} &= \mathbb{P}\{\xi X \in A \mid \xi = 1\} \cdot \frac{1}{2} + \mathbb{P}\{\xi X \in A \mid \xi = -1\} \cdot \frac{1}{2}\\ &= \frac{1}{2}\Big(\mathbb{P}\{X \in A\} + \mathbb{P}\{-X \in A\}\Big).\end{aligned}$$

Doing the same for $-\xi X$ gives the same result (check!). So ξX and $-\xi X$ have the same CDF, meaning they have the same distribution. □

For practice, try making symmetric versions of distributions like Bernoulli and exponential (see Exercise 6.17).

[6] You might notice we get a factor of 4 instead of 2 in Theorem 6.2.2 – this happens because we bound both upper and lower tails separately. But we can replace 4 with 2 by adjusting the constant c in the exponent. (How?)

Lemma 6.3.2 (Symmetrization)**.** *Let $X_1, \ldots, X_N$ be independent, mean-zero random vectors in a normed space, and let $\varepsilon_1, \ldots, \varepsilon_N$ be independent*[7] *Rademacher random variables. Then*

$$\frac{1}{2}\,\mathbb{E}\,\Big\|\sum_{i=1}^N \varepsilon_i X_i\Big\| \le \mathbb{E}\,\Big\|\sum_{i=1}^N X_i\Big\| \le 2\,\mathbb{E}\,\Big\|\sum_{i=1}^N \varepsilon_i X_i\Big\|.$$

This lemma lets us replace any random variables X_i by symmetric ones $\varepsilon_i X_i$.

Proof **Upper bound.** Let (X_i') be an independent copy of (X_i). Since $\sum_i X_i'$ has zero mean, we have

$$p := \mathbb{E}\,\Big\|\sum_i X_i\Big\| \le \mathbb{E}\,\Big\|\sum_i X_i - \sum_i X_i'\Big\| = \mathbb{E}\,\Big\|\sum_i (X_i - X_i')\Big\|.$$

The inequality comes this fact: for independent random vectors Y and Z,

$$\mathbb{E}\,Z = 0 \quad \text{implies} \quad \mathbb{E}\,\|Y\| \le \mathbb{E}\,\|Y + Z\|. \tag{6.13}$$

(Check it!) Next, since $(X_i - X_i')$ are symmetric random vectors, they have the same distribution as $\varepsilon_i(X_i - X_i')$ (Exercise 6.16(b)). Then

$$\begin{aligned} p &\le \mathbb{E}\,\Big\|\sum_i \varepsilon_i (X_i - X_i')\Big\| \\ &\le \mathbb{E}\,\Big\|\sum_i \varepsilon_i X_i\Big\| + \mathbb{E}\,\Big\|\sum_i \varepsilon_i X_i'\Big\| \quad \text{(by triangle inequality)} \\ &= 2\,\mathbb{E}\,\Big\|\sum_i \varepsilon_i X_i\Big\| \quad \text{(since the two terms are identically distributed).} \end{aligned}$$

Lower bound. The argument here is similar:

$$\begin{aligned} \mathbb{E}\,\Big\|\sum_i \varepsilon_i X_i\Big\| &\le \mathbb{E}\,\Big\|\sum_i \varepsilon_i (X_i - X_i')\Big\| \quad \text{(condition on } (\varepsilon_i) \text{ and use (6.13))} \\ &= \mathbb{E}\,\Big\|\sum_i (X_i - X_i')\Big\| \quad \text{(the distribution is the same)} \\ &\le \mathbb{E}\,\Big\|\sum_i X_i\Big\| + \mathbb{E}\,\Big\|\sum_i X_i'\Big\| \quad \text{(by triangle inequality)} \\ &\le 2\,\mathbb{E}\,\Big\|\sum_i X_i\Big\| \quad \text{(by identical distribution).} \end{aligned}$$

□

Now, ask yourself – where did we use X_i's independence? Do we need mean zero for both upper and lower bound?

Try proving a few versions of symmetrization lemma yourself (see Exercises 6.19–6.21).

[7] We assume without saying that (ε_i) are independent of (X_i), so technically $X_1, \ldots, \varepsilon_1, \ldots, \varepsilon_N$ are jointly independent.

6.4 Random matrices with non-i.i.d. entries

A typical application of symmetrization consists of two steps: first, replace random variables X_i with symmetric ones $\varepsilon_i X_i$, then condition on X_i so that all randomness comes from signs ε_i. This reduces the problem to Rademacher random variables. To illustrate this technique, let's bound the operator norm of a random matrix with independent, non-identically distributed entries.

Theorem 6.4.1 (Norm of random matrices with non-i.i.d. entries)**.** *Let A be an $n \times n$ symmetric random matrix with independent, mean-zero entries above and on the diagonal. Then*

$$\mathbb{E} \max_i \|A_i\|_2 \le \mathbb{E}\|A\| \le C\sqrt{\log n} \cdot \mathbb{E} \max_i \|A_i\|_2,$$

where A_i denote the rows of A.

Although the operator norm of a matrix is usually hard to express in terms of its entries, but for *random* matrices, we get pretty close: Theorem 6.4.1 shows it is roughly the maximal Euclidean norm of the rows, up to a log factor. And unlike earlier results, we require *no moment assumptions* on the entries at all!

Proof The lower bound should already be familiar to you (recall Exercise 4.7). For the upper bound, we will use symmetrization and the matrix Khintchine inequality (Theorem 5.4.14).

Let's decompose A entry by entry, keeping symmetry in mind – just like we did in the proof of Theorem 5.5.1. Denote the standard canonical basis vectors in $\mathbb{R}^n$ by $e_1, \dots, e_n$ and write A as a sum of independent, mean-zero random matrices:

$$A = \sum_{i \le j} Z_{ij}, \quad \text{where} \quad Z_{ij} = \begin{cases} A_{ij}(e_i e_j^\mathsf{T} + e_j e_i^\mathsf{T}), & i < j \\ A_{ii} e_i e_i^\mathsf{T}, & i = j \end{cases}$$

Apply symmetrization (Lemma 6.3.2):

$$\mathbb{E}\|A\| = \mathbb{E}\Big\| \sum_{i \le j} Z_{ij} \Big\| \le 2\,\mathbb{E}\Big\| \sum_{i \le j} \varepsilon_{ij} Z_{ij} \Big\|, \tag{6.14}$$

where (ε_{ij}) are independent Rademacher random variables.

Condition on (Z_{ij}), apply the matrix Khintchine inequality (Theorem 5.4.14) for $p = 1$, then take expectation over (Z_{ij}) using the law of total expectation. This gives

$$\mathbb{E}\Big\| \sum_{i \le j} \varepsilon_{ij} Z_{ij} \Big\| \le C\sqrt{\log n}\; \mathbb{E}\left(\Big\| \sum_{i \le j} Z_{ij}^2 \Big\|^{1/2} \right). \tag{6.15}$$

Now, just like in the proof of Theorem 5.5.1, each Z_{ij} is a diagonal matrix:

$$Z_{ij}^2 = \begin{cases} A_{ij}^2 (e_i e_i^\mathsf{T} + e_j e_j^\mathsf{T}), & i < j \\ A_{ii}^2 e_i e_i^\mathsf{T}, & i = j, \end{cases}$$

so

$$\sum_{i\le j} Z_{ij}^2 = \sum_{i=1}^n \Big(\sum_{j=1}^n A_{ij}^2 \Big) e_i e_i^\mathsf{T} = \sum_{i=1}^n \|A_i\|_2^2 e_i e_i^\mathsf{T}.$$

In other words, $\sum_{i\le j} Z_{ij}^2$ is a diagonal matrix with diagonal entries equal $\|A_i\|_2^2$. Since the operator norm of a diagonal matrix is the maximal absolute value of its entries, we get

$$\Big\| \sum_{i\le j} Z_{ij}^2 \Big\| = \max_i \|A_i\|_2^2.$$

Substitute this into (6.15) and then into (6.14) to complete the proof. □

To practice the symmetrization technique, try Exercises 6.22–6.29 now.

6.5 Application: matrix completion

One exciting application of the methods we learned is matrix completion – recovering missing entries from a partially observed matrix. Of course, this is not possible without knowing something extra about the matrix. Let's show that for low-rank matrices, we can recovering the missing entries algorithmically.

To describe the problem mathematically, consider an $n \times n$ 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 observe the $n \times n$ matrix Y with entries

$$Y_{ij} := \delta_{ij} X_{ij} \quad \text{where} \quad \delta_{ij} \sim \operatorname{Ber}(p) \text{ are independent.}$$

These δ_{ij} are *selectors* – Bernoulli random variables that select the entries we observe (and all unobserved entries are replaced with zeros). If

$$p = \frac{m}{n^2} \tag{6.16}$$

then we *observe m entries of X on average.*

How can we recover X from Y? Although X has small rank r, Y may not have small rank. (Why?) To fix this, we can pick the *best rank r approximation* to Y (see Section 4.1.5). Properly scaled, this gives a good estimate of X:

Theorem 6.5.1 (Matrix completion)**.** *Let $\hat{X}$ be a best rank r approximation to $p^{-1}Y$. Then*

$$\mathbb{E} \frac{1}{n} \|\hat{X} - X\|_F \le C \sqrt{\frac{rn \log n}{m}} \, \|X\|_\infty,$$

as long as $m \ge n \log n$. Here $\|X\|_\infty = \max_{i,j} |X_{ij}|$ is the largest entry.

Before we prove Theorem 6.5.1, note that the recovery error

$$\frac{1}{n}\|\hat{X}-X\|_F = \Big(\frac{1}{n^2}\sum_{i,j=1}^n |\hat{X}_{ij}-X_{ij}|^2\Big)^{1/2}.$$

represents the average error per entry (in the L^2 sense). If we choose the average number of observed entries m so that

$$m \ge C' rn \log n$$

with large constant C', then Theorem 6.5.1 guarantees that the average error is much smaller than $\|X\|_\infty$. So, *matrix completion is possible if the number of observed entries exceeds rn by a logarithmic margin.*

Proof We 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 the triangle inequality, we can 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 rank r approximation to $p^{-1}Y$, the second summand dominates, i.e. $\|\hat{X}-p^{-1}Y\| \le \|p^{-1}Y-X\|$, so we have

$$\|\hat{X}-X\| \le 2\|p^{-1}Y-X\| = \frac{2}{p}\|Y-pX\|. \tag{6.17}$$

Note that the matrix $\hat{X}$, which is tricky to handle, is gone the bound. Instead, we get $Y-pX$, which is easier to understand since its entries,

$$(Y-pX)_{ij} = (\delta_{ij}-p)X_{ij},$$

are independent, mean-zero random variables. Using Theorem 6.4.1 (more precisely, its non-symmetric version, see Exercise 6.28), we get

$$\mathbb{E}\|Y-pX\| \le C\sqrt{\log n}\,\Big(\mathbb{E}\max_{i\le n}\|(Y-pX)_{i:}\|_2 + \mathbb{E}\max_{j\le n}\|(Y-pX)_{:j}\|_2\Big). \tag{6.18}$$

To bound the norms of the rows of $Y-pX$, we write them as

$$\|(Y-pX)_{i:}\|_2^2 = \sum_{j=1}^n (\delta_{ij}-p)^2 X_{ij}^2 \le \sum_{j=1}^n (\delta_{ij}-p)^2 \cdot \|X\|_\infty^2,$$

and similarly for columns. These sums of independent random variables can be easily bounded using Bernstein (or Chernoff) inequality, which yields

$$\mathbb{E}\max_{i\in[n]}\sum_{j=1}^n (\delta_{ij}-p)^2 \lesssim pn.$$

(Do this calculation yourself in Exercise 6.30!) Combining with a similar bound for the columns and substituting into (6.18), we obtain

$$\mathbb{E}\,\|Y-pX\| \lesssim \sqrt{pn\log n}\;\|X\|_\infty.$$

Then, by (6.17), we get

$$\mathbb{E}\,\|\hat{X}-X\| \lesssim \sqrt{\frac{n\log n}{p}}\,\|X\|_\infty. \tag{6.19}$$

Step 2: Passing to Frobenius norm. We have not used the low rank assumption yet, and will do this now. Since $\operatorname{rank}(X)\le r$ by assumption and $\operatorname{rank}(\hat{X})\le r$ by construction, we have $\operatorname{rank}(\hat{X}-X)\le 2r$. This implies

$$\|\hat{X}-X\|_F \le \sqrt{2r}\|\hat{X}-X\|.$$

(This follows from Lemma 4.1.11, see Exercise 4.4.) Taking expectations and using the bound on the error in the operator norm (6.19), we get

$$\mathbb{E}\,\|\hat{X}-X\|_F \lesssim \sqrt{\frac{rn\log n}{p}}\,\|X\|_\infty.$$

Dividing both sides by n, we can rewrite this bound as

$$\mathbb{E}\,\frac{1}{n}\|\hat{X}-X\|_F \lesssim \sqrt{\frac{rn\log n}{pn^2}}\,\|X\|_\infty.$$

To finish the proof, recall that $pn^2=m$ by the definition (6.16) of p. □

Remark 6.5.2 (Extensions)**.** Theorem 6.5.1 can be extended and improved in many ways. Try to generalize it to rectangular matrices (Exercise 6.31) and noisy observations (Exercise 6.32). It is less trivial but possible to remove the logarithmic factor from the error bound, and achieve zero error for noiseless observations – see the notes after this chapter for details.

6.6 Contraction principle

We conclude this chapter with one more useful inequality.

Theorem 6.6.1 (Contraction principle)**.** *Let $x_1,\ldots,x_N$ be any vectors in a normed space, let $a=(a_1,\ldots,a_N)\in\mathbb{R}^N$, and let $\varepsilon_1,\ldots,\varepsilon_N$ be independent Rademacher random variables. Then*

$$\mathbb{E}\Big\|\sum_{i=1}^N a_i\varepsilon_i x_i\Big\| \le \|a\|_\infty\cdot\mathbb{E}\Big\|\sum_{i=1}^N \varepsilon_i x_i\Big\|.$$

Proof Without loss of generality, assume $\|a\|_\infty\le 1$. (Why?) Define the function

$$f(a):=\mathbb{E}\Big\|\sum_{i=1}^N a_i\varepsilon_i x_i\Big\|. \tag{6.20}$$

Then $f:\mathbb{R}^N\to\mathbb{R}$ is convex – check this in Exercise 6.35!

We want to bound for f on the set of points a satisfying $\|a\|_\infty\le 1$, i.e. on the unit cube $[-1,1]^N$. By the maximum principle (Exercises 1.4 and 1.5), the maximum of the convex function f on the cube is attained at a vertex, where all $a_i = \pm 1$.

For such a, the random variables $(\varepsilon_i a_i)$ have the same distribution as (ε_i) by symmetry. Thus

$$\mathbb{E} \Big\| \sum_{i=1}^N a_i \varepsilon_i x_i \Big\| = \mathbb{E} \Big\| \sum_{i=1}^N \varepsilon_i x_i \Big\|,$$

Thus, $f(a) \le \mathbb{E} \left\| \sum_{i=1}^N \varepsilon_i x_i \right\|$ whenever $\|a\|_\infty \le 1$, completing the proof. □

As an application, let us prove a version of symmetrization (Lemma 6.3.2) but with Gaussian random variables $g_i \sim N(0,1)$ instead of Rademachers.

Lemma 6.6.2 (Symmetrization with Gaussians)**.** *Let $X_1, \ldots, X_N$ be independent, mean-zero random vectors in a normed space. Let $g_1, \ldots, g_N \sim N(0,1)$ 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 **Upper bound.** By symmetrization (Lemma 6.3.2), we have

$$E := \mathbb{E} \Big\| \sum_{i=1}^N X_i \Big\| \le 2 \, \mathbb{E} \Big\| \sum_{i=1}^N \varepsilon_i X_i \Big\|.$$

To replace the Rademacher random variables with Gaussians, recall that $\mathbb{E}\,|g_i| = \sqrt{2/\pi}$. Thus we can continue our bound as follows:[8]

$$\begin{aligned} E &\le 2\sqrt{\frac{\pi}{2}} \, \mathbb{E}_X \Big\| \sum_{i=1}^N \varepsilon_i \, \mathbb{E}_g \, |g_i| X_i \Big\| \\ &\le 2\sqrt{\frac{\pi}{2}} \, \mathbb{E} \Big\| \sum_{i=1}^N \varepsilon_i |g_i| X_i \Big\| \quad \text{(by Jensen inequality)} \\ &= 2\sqrt{\frac{\pi}{2}} \, \mathbb{E} \Big\| \sum_{i=1}^N g_i X_i \Big\|. \end{aligned}$$

The last equality holds since the random variables $(\varepsilon_i |g_i|)$ have the same joint distribution as (g_i) (Lemma 6.3.1(b)).

Lower bound can be proved by using contraction principle (Theorem 6.6.1)

[8] Here we use index g in $\mathbb{E}_g$ to indicate that this is an expectation "over (g_i)", i.e. conditional on (X_i). Similarly, E_X denotes the expectation over (X_i).

and symmetrization (Lemma 6.3.2). We have

$$\begin{aligned}
\mathbb{E}\Big\|\sum_{i=1}^N g_i X_i\Big\| &= \mathbb{E}\Big\|\sum_{i=1}^N \varepsilon_i g_i X_i\Big\| \quad \text{(by symmetry of } g_i\text{)} \\
&\le \mathbb{E}_g\,\mathbb{E}_X \Big(\|g\|_\infty \,\mathbb{E}_\varepsilon \Big\|\sum_{i=1}^N \varepsilon_i X_i\Big\| \Big) \quad \text{(by Theorem 6.6.1)} \\
&= \mathbb{E}_g \Big(\|g\|_\infty \,\mathbb{E}_\varepsilon\,\mathbb{E}_X \Big\|\sum_{i=1}^N \varepsilon_i X_i\Big\| \Big) \quad \text{(by independence)} \\
&\le 2\,\mathbb{E}_g \Big(\|g\|_\infty \,\mathbb{E}_X \Big\|\sum_{i=1}^N X_i\Big\| \Big) \quad \text{(by Lemma 6.3.2)} \\
&= 2\Big(\mathbb{E}\,\|g\|_\infty\Big)\Big(\mathbb{E}\Big\|\sum_{i=1}^N X_i\Big\|\Big) \quad \text{(by independence).}
\end{aligned}$$

It remains to recall that

$$\mathbb{E}\,\|g\|_\infty \le C\sqrt{\log N}.$$

(see Proposition 2.7.6 or Exercise 2.38(a)). The proof is complete. □

Remark 6.6.3 (The log factor is unavoidable)**.** The logarithmic factor in Lemma 6.6.2 is necessary and optimal in general (Exercise 6.37), making Gaussian symmetrization weaker than Rademacher's.

Get more practice with contraction: prove it for general distributions (Exercise 6.36) and convex functions of norms (Exercise 6.38).

6.7 Notes

The decoupling inequality in Theorem 6.1.1 was originally proved by J. Bourgain and L. Tzafriri [55]. For related results and extensions, see the paper [97] and the books [96], [127, Section 8.4].

The original Hanson-Wright inequality, somewhat weaker than Theorem 6.2.2, goes back to [154, 350]. The modern version of Hanson-Wright inequality inequality (Theorem 6.2.2) and its proof in Section 6.2 are from [292]. Several special cases appeared earlier in [127, Proposition 8.13] for Bernoulli random variables, in [315, Lemma 2.5.1] for Gaussian random variables, and in [31] for diagonal-free matrices.

In [176], the dependence on K in Hanson-Wright inequality (Theorem 6.2.2) was improved, assuming that X_i have variance one. For some extensions Hanson-Wright inequality, see [6, 34, 354, 8, 188, 145, 293, 159].

Concentration for anisotropic random vectors (Exercise 6.13) and the bound on the distance from a random vector to a subspace (Exercise 6.14) are from [292].

Symmetrization Lemma 6.3.2 and its proof can be found e.g. in [210, Lemma 6.3], [127, Section 8.2].

Although the precise statement of Theorem 6.4.1 is difficult to locate in existing literature, it can be deduced, for example, from the inequalities in [324, 327]. The factor $\sqrt{\log n}$ in Theorem 6.4.1 can be improved to $\log^{1/4} n$ by combining a result of Y. Seginer [299, Theorem 3.1] with symmetrization (Lemma 6.3.2); see [27, Corollary 4.7] for an alternative approach to Seginer theorem. This improved factor is optimal as is demonstrated by the result of Exercise 6.29, which is due to Y. Seginer [299, Theorem 3.2]. Moreover, for many classes of matrices the factor $\log^{1/4} n$ can be removed completely; this happens, in particular, for matrices with i.i.d. entries

[299] and matrices with Gaussian entries [203]. Refer to [331, Section 4], [203, 25, 26, 59, 204] for more elaborate results, whose goal is to describe the operator norm of a random matrix A in terms of the variances of its entries.

Theorem 6.5.1 on matrix completion and its proof are from [278, Section 2.5], although versions of it may have appeared before. In particular, Keshavan, Montanari and Oh [182] showed how to obtain a slightly better bound–one without the logarithmic factor–by "trimming" the random matrix Y, where one removes the rows and columns of Y that have, say, twice more many nonzero entries than expected. E. Candes and B. Recht [72] demonstrated that under some additional incoherence assumptions, *exact* matrix completion (with zero error) is possible with $m \asymp rn \log^2(n)$ randomly sampled entries. For further developments on matrix completion, see [74, 282, 148, 92].

The contraction principle (Theorem 6.6.1) is from [210, Section 4.2]; see also [210, Corollary 3.17, Theorem 4.12] for different versions. Gaussian symmetrization (Lemma 6.6.2) can be found in [210, inequality (4.9)]. While the logarithmic factor is in general needed there, it can be removed if the normed space has non-trivial cotype, see [210, Proposition 9.14].

In Exercise 6.4, the constant 4 can be improved to 2 for symmetric matrices [323].

A version of Exercise 6.10 on the norm of anisotropic subgaussian random vectors was originally proved by D. Hsu, S. Kakade and T. Zhang [166]; they obtained sharp bounds working with exact subgaussian norm (introduced in Exercise 2.40).

The symmetry assumption in Exercise 6.22 is essential and cannot be skipped [135, Example 2.8].

Exercises

6.1 ☕☕ (Decoupling for matrices with diagonal terms) As we noted in Remark 6.1.2, the diagonal-free assumption cannot be removed from Theorem 6.1.1. But we can include the diagonal on the right hand side: show that for any real numbers $(a_{ij})_{i,j=1}^n$, we have

$$\mathbb{E}\, F\Big(\sum_{i,j:\, i\neq j} a_{ij} X_i X_j \Big) \le \mathbb{E}\, F\Big(4 \sum_{i,j} a_{ij} X_i X'_j \Big)$$

6.2 ☕☕ (L^p and subgaussian decoupling) Let $(a_{ij})_{i,j=1}^n$ be real numbers. Let $X_1, \ldots, X_n$ be independent, mean-zero random variables, and (X'_i) be an independent copy of (X_i).

(a) For any $p \in [1, \infty)$, show that

$$\Big\| \sum_{i,j:\, i\neq j} a_{ij} X_i X_j \Big\|_{L^p} \le 4 \Big\| \sum_{i,j} a_{ij} X_i X'_j \Big\|_{L^p}.$$

(b) Show that

$$\Big\| \sum_{i,j:\, i\neq j} a_{ij} X_i X_j \Big\|_{\psi_2} \le 4 \Big\| \sum_{i,j} a_{ij} X_i X'_j \Big\|_{\psi_2}.$$

6.3 ☕ (Vector decoupling)

(a) Let $(v_{ij})_{i,j=1}^n$ be vectors in some vector space V. Let $X_1, \ldots, X_n$ be independent, mean-zero random variables, and (X'_i) be an independent copy of (X_i). Prove that for every convex function $F : V \to \mathbb{R}$,

$$\mathbb{E}\, F\Big(\sum_{i,j:\, i\neq j} v_{ij} X_i X_j \Big) \le \mathbb{E}\, F\Big(4 \sum_{i,j} v_{ij} X_i X'_j \Big).$$

(b) Let $(a_{ij})_{i,j=1}^n$ be real numbers. Let $X_1, \dots, X_n$ be independent, mean-zero random vectors in $\mathbb{R}^N$. Show that for every convex function $F : \mathbb{R} \to \mathbb{R}$, one has

$$\mathbb{E}\, F\Big(\sum_{i,j:\, i \neq j} a_{ij} \langle X_i, X_j \rangle \Big) \le \mathbb{E}\, F\Big(4 \sum_{i,j} a_{ij} \langle X_i, X_j' \rangle \Big).$$

(c) Let $(a_{ij})_{i,j=1}^n$ be real numbers. Let $X_1, \dots, X_n$ be independent, mean-zero random vectors in $\mathbb{R}^N$. Show that for every convex function $F : \mathbb{R}^{N \times N} \to \mathbb{R}$, one has

$$\mathbb{E}\, F\Big(\sum_{i,j:\, i \neq j} a_{ij} X_i X_j^{\mathsf{T}} \Big) \le \mathbb{E}\, F\Big(4 \sum_{i,j} a_{ij} X_i (X_j')^{\mathsf{T}} \Big).$$

6.4 ☕☕☕ (Decoupling the norm of a random submatrix) Let A be an $n \times n$ diagonal-free matrix. Pick a random subset $J \subset \{1, \dots, n\}$ by including each index independently with probability $p \in (0, 1)$. Let J' be an independent copy of J. Show that

$$\mathbb{E}\|A_{J \times J}\| \le 4\, \mathbb{E}\|A_{J \times J'}\|.$$

Here $A_{I \times J}$ denotes the submatrix of A with rows from I and columns from J.

6.5 ☕☕☕ (No deviation bound from the mean) It is tempting to conjecture that the first-order term in Proposition 6.2.1 should be the mean of $\|X\|$, which is approximately $\sqrt{n}$, instead of $CK\sqrt{n}$. But that is false! Give an example showing that the bound

$$\mathbb{P}\Big\{ \|X\|_2 \ge C(\sqrt{n} + Kt) \Big\} \le e^{-t^2} \quad \text{for all } t \ge 0$$

does not always hold.

6.6 ☕☕ (Norms of random matrices with subgaussian columns)

(a) Extend the maximal inequality (Exercise 3.13) to subgaussian random vectors, even without independent entries.

(b) Extend the bounds on the $1 \to \infty$ and $1 \to 2$ norms bound (Exercise 4.44(a),(b)) to random matrices with independent subgaussian columns, even if entries are not independent.

6.7 ☕☕☕ (Gaussian Hanson-Wright) Give an alternative proof of Hanson-Wright inequality for normal distributions, without separating the diagonal part or decoupling.

6.8 ☕☕☕ (Higher-dimensional Hanson-Wright) Let $A = (a_{ij})$ be an $n \times n$ matrix. Let $X_1, \dots, X_n$ be independent, mean-zero, subgaussian random vectors in $\mathbb{R}^d$. Prove that for every $t \ge 0$, we have

$$\mathbb{P}\Big\{ \Big| \sum_{i,j:\, i \neq j} a_{ij} \langle X_i, X_j \rangle \Big| \ge t \Big\} \le 2 \exp\Big[-c \min\Big(\frac{t^2}{K^4 d \|A\|_F^2}, \frac{t}{K^2 \|A\|} \Big) \Big]$$

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

6.9 ☕☕☕ (MGF of the squared norm) Let B be an $m \times n$ matrix, and let X be a mean-zero subgaussian random vector in $\mathbb{R}^n$ with $\|X\|_{\psi_2} \le K$. Show that

$$\mathbb{E}\exp\left(\lambda^2 \|BX\|_2^2\right) \le \exp\left(CK^2\lambda^2\|B\|_F^2\right) \quad \text{whenever} \quad |\lambda| \le \frac{c}{K\|B\|}.$$

6.10 ☕☕ (The norm of an anisotropic random vector) Let us extend Proposition 6.2.1 for anisotropic random vectors. Let B be an $m \times n$ matrix, and let X be a mean-zero subgaussian random vector in $\mathbb{R}^n$ with $\|X\|_{\psi_2} \le K$. For every $t \ge 0$, show that

$$\mathbb{P}\big\{\|BX\|_2 \ge CK\big(\|B\|_F + t\|B\|\big)\big\} \le e^{-t^2}.$$

6.11 ☕☕ (Hanson-Wright for subgaussian vectors) Here is a useful version of Hanson-Wright inequality (Theorem 6.2.2) without assuming independence of entries. Let A be a symmetric positive-semidefinite $n \times n$ matrix, and let X be a mean-zero subgaussian random vector in $\mathbb{R}^n$ with $\|X\|_{\psi_2} \le K$. For any $s \ge 0$, show that

$$\mathbb{P}\Big\{X^\mathsf{T} AX \ge CK^2\big(\operatorname{tr} A + s\|A\|\big)\Big\} \le e^{-s}.$$

(To compare with Theorem 6.2.2, set $t = CK^2 s\|A\|$ and note that $\operatorname{tr} A = \mathbb{E}\, X^\mathsf{T} AX$ for isotropic X.) Explain why there can be only one-sided bound here, even if X is isotropic.

6.12 ☕☕☕ (Mean estimation) Let's revisit the mean estimation problem (Section 2.4), where we sample N i.i.d. points $X_1, \ldots, X_N$ from some unknown distribution in $\mathbb{R}^n$ with mean μ and covariance Σ, and we want to estimate μ. Assume the distribution is subgaussian:

$$\|\langle X_i - \mu, u\rangle\|_{\psi_2} \le K\|\langle X_i - \mu, u\rangle\|_{L^2} \quad \text{for all } u \in \mathbb{R}^n.$$

For any $\alpha \in (0,1)$, show the sample mean $\mu_n := \frac{1}{N}\sum_{i=1}^N X_i$ satisfies

$$\|\mu_n - \mu\|_2 \le CK\sqrt{\frac{\operatorname{tr}\Sigma}{N}} + CK\sqrt{\frac{\|\Sigma\|\log(1/\alpha)}{N}}$$

with probability at least $1 - \alpha$.

6.13 ☕☕☕☕ (Anisotropic concentration of the norm) Let's extend the concentration of norm (Theorem 3.1.1) to anisotropic distributions. Let B be an $m \times n$ matrix, and let $X = (X_1, \ldots, X_n) \in \mathbb{R}^n$ be a random vector with independent, mean-zero, unit variance, subgaussian coordinates. Show that

$$\Big\|\|BX\|_2 - \|B\|_F\Big\|_{\psi_2} \le CK^2\|B\|,$$

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

6.14 ☕☕ (Distance from a random vector to a subspace) Let E be a subspace of $\mathbb{R}^n$ of dimension d. Let $X = (X_1, \ldots, X_n) \in \mathbb{R}^n$ be a random vector with independent, mean-zero, unit variance, subgaussian coordinates.

(a) Check that $\big(\mathbb{E}\operatorname{dist}(X, E)^2\big)^{1/2} = \sqrt{n-d}$.

(b) Deduce from Exercise 6.13 that for any $t \geq 0$, the distance nicely concentrates:

$$\mathbb{P}\Big\{\big|\operatorname{dist}(X,E) - \sqrt{n-d}\big| > t\Big\} \leq 2\exp\big(-ct^2/K^4\big),$$

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

6.15 ☕☕ (Randomly cutting a graph) Take any graph with E edges and randomly split its vertices into two groups (as in Section 3.6.3). As we saw in Proposition 3.6.3, the expected "cut" (the number of crossing edges) is $E/2$. Now show that it is tightly concentrated around that value:

$$\mathbb{P}\left\{\left|\text{cut} - \frac{E}{2}\right| \geq s\sqrt{E}\right\} \leq 2e^{-cs} \quad \text{for any } s \geq 1.$$

6.16 ☕☕ (Constructing symmetric distributions) Prove all statements in Lemma 6.3.1.

6.17 ☕☕ (Symmetrizing Bernoulli and exponential distributions) Recall Lemma 6.3.1(c): if X' is an independent copy of a random variable X, then $X - X'$ is symmetric. Let's compute the distribution of $X - X'$ for two examples of X: Bernoulli and exponential.

(a) Let $X \sim \operatorname{Ber}(p)$. Compute the probability mass function of $X - X'$.

(b) Let $X \sim \operatorname{Exp}(1)$. Show that $X - X' \sim \operatorname{Lap}(0,1)$, meaning that the density is $\frac{1}{2}e^{-|x|}$ for $x \in \mathbb{R}$.

6.18 ☕☕ (Norm of a shifted random vector) Let X be a symmetric random vector (meaning X and $-X$ have the same distribution) taking values in some normed space, and v be a fixed vector in that space. Show that

$$\mathbb{E}\|X + v\| \asymp \mathbb{E}\|X\| + \|v\|,$$

where the notation $\asymp$ hides positive absolute constant factors.

6.19 ☕☕ (Symmetrization without zero mean)

(a) Prove the following generalization of Symmetrization Lemma 6.3.2 for random vectors X_i that do not necessarily have zero means:

$$\mathbb{E}\Big\|\sum_{i=1}^N X_i - \sum_{i=1}^N \mathbb{E}\, X_i\Big\| \leq 2\,\mathbb{E}\Big\|\sum_{i=1}^N \varepsilon_i X_i\Big\|.$$

(b) Argue that there can not be any non-trivial reverse inequality.

6.20 ☕ (Symmetrization with a convex function) 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

$$\mathbb{E}\, F\Big(\frac{1}{2}\Big\|\sum_{i=1}^N \varepsilon_i X_i\Big\|\Big) \leq \mathbb{E}\, F\Big(\Big\|\sum_{i=1}^N X_i\Big\|\Big) \leq \mathbb{E}\, F\Big(2\Big\|\sum_{i=1}^N \varepsilon_i X_i\Big\|\Big).$$

6.21 ☕☕ (Symmetrizing subgaussian sums) Let $X_1, \ldots, X_N$ be independent, mean-zero random variables. Show that their sum $\sum_i X_i$ is subgaussian if and only if $\sum_i \varepsilon_i X_i$ is subgaussian, and

$$c\Big\|\sum_{i=1}^N \varepsilon_i X_i\Big\|_{\psi_2} \le \Big\|\sum_{i=1}^N X_i\Big\|_{\psi_2} \le C\Big\|\sum_{i=1}^N \varepsilon_i X_i\Big\|_{\psi_2}.$$

6.22 ☕☕ (Self-normalized sums) Let $X_1, \ldots, X_N$ be independent symmetric random variables. Show that, for any $t > 0$,

$$\mathbb{P}\left\{\Big|\sum_{i=1}^N X_i\Big| \ge t\Big(\sum_{i=1}^N X_i^2\Big)^{1/2}\right\} \le 2e^{-t^2/2}$$

This is very general –no subgaussian or moment assumptions needed!

6.23 ☕☕☕ (Type[9] p). Let $p \in [1,2]$.

(a) Let $x_1, \ldots, x_N \in \mathbb{R}^n$ be any (fixed) vectors. Check that

$$\mathbb{E}\Big\|\sum_{i=1}^N \varepsilon_i x_i\Big\|_p^p \le \sum_{i=1}^N \|x_i\|_p^p.$$

(b) Generalize part (a) as follows. Let $X_1, \ldots, X_N$ be mean-zero, independent random vectors in $\mathbb{R}^n$. Show that

$$\mathbb{E}\Big\|\sum_{i=1}^N X_i\Big\|_p^p \le \sum_{i=1}^N \mathbb{E}\|X_i\|_p^p.$$

(c) Show by example that the results in parts (a)–(b) fail for each $p > 2$.

6.24 ☕☕☕ (Type 2)

(a) Let $x_1, \ldots, x_N \in \mathbb{R}^n$ be any (fixed) vectors. Check that

$$\mathbb{E}\Big\|\sum_{i=1}^N \varepsilon_i x_i\Big\|_p^2 \le Cp\sum_{i=1}^N \|x_i\|_p^2.$$

(b) Generalize part (a) as follows. Let $X_1, \ldots, X_N$ be mean-zero, independent random vectors in $\mathbb{R}^n$. Show that

$$\mathbb{E}\Big\|\sum_{i=1}^N X_i\Big\|_p^2 \le Cp\sum_{i=1}^N \mathbb{E}\|X_i\|_p^2.$$

(c) Show by example that the results in parts (a)–(b) fail for each $p < 2$.

6.25 ☕☕ (Approximate ℓ^p-Caratheodory theorem) State and prove a version of the approximate Caratheodory theorem (Theorem 0.0.2) for the ℓ^p norm, for any $p \in [1, \infty)$.

[9] Here is a broader perspective: a normed space X is said to be of *type* p if there exists K such that $\mathbb{E}\|\sum_{i=1}^N \varepsilon_i x_i\|^p \le K^p \sum_{i=1}^N \|x_i\|^p$ for any N and vectors $x_1, \ldots, x_N \in X$. In this and next exercises, we show that ℓ^p is of type p if and only if $p \in [1,2]$, and type 2 if and only if $p \in [2, \infty)$.

6.26 ☕☕☕ (Marcinkiewicz-Zygmund inequality) For independent, mean-zero random variables $X_1, \dots, X_N$ and $p \in [2, \infty)$, show that[10]

$$\Big\| \sum_{i=1}^N X_i \Big\|_{L^p}^2 \le Cp \sum_{i=1}^N \|X_i\|_{L^p}^2.$$

6.27 ☕☕☕ (Concentration of the norm under finite moments) Let's prove a version of Theorem 3.1.1 under weaker, finite-moment conditions. Let $X = (X_1, \dots, X_n) \in \mathbb{R}^n$ be a random vector with coordinates X_i that satisfy $\mathbb{E}\, X_i^2 = 1$ and $\|X_i\|_{L^{2p}} \le K$ for some $p \ge 2$ and $K \ge 0$. Show that

$$\Big\| \|X\|_2 - \sqrt{n} \Big\|_{L^p} \le C\sqrt{p} K^2.$$

6.28 ☕☕ (Norm of rectangular random matrices) Extend Theorem 6.4.1 to non-symmetric, rectangular matrices. For an $m \times n$ random matrix A with independent, mean-zero entries, show that its expected operator norm is roughly the expected max Euclidean norm of its rows and columns, up to a logarithmic factor:

$$\mathbb{E} \max_{i,j} \big(\|A_{i:}\|_2,\, \|A_{:j}\|_2 \big) \le \mathbb{E}\|A\| \le C\sqrt{\log(m+n)} \max_{i,j} \big(\|A_{i:}\|_2,\, \|A_{:j}\|_2 \big).$$

6.29 ☕☕☕ (The log factor is unremovable) Show that the logarithmic factor in Theorem 6.4.1 cannot be completely removed in general. Construct a random matrix A satisfying the assumptions of the theorem and for which

$$\mathbb{E}\|A\| \ge c \log^{1/4}(n) \cdot \mathbb{E} \max_i \|A_i\|_2.$$

6.30 ☕☕ (Norms of the rows of random matrices) Consider i.i.d. random variables $\delta_{ij} \sim \mathrm{Ber}(p)$, where $i, j = 1, \dots, n$. Assuming that $pn \ge \log n$, show that

$$\mathbb{E} \max_{i \le n} \sum_{j=1}^n (\delta_{ij} - p)^2 \le Cpn.$$

6.31 ☕ (Matrix completion for rectangular matrices) State and prove a version of matrix completion (Theorem 6.5.1) for general $m \times n$ matrices.

6.32 ☕☕ (Matrix completion with 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 subgaussian random variables representing noise.

6.33 ☕☕☕☕ (Sums of independent unbounded random matrices) There are not many results that don't make any assumptions on the distribution of random matrices. Let's prove one

[10] If you are interested in a broader perspective, this inequality basically says that the L^p spaces for $p \ge 2$ are of type 2, just like the ℓ^p spaces we examined in Exercise 6.24.

– a version of matrix Bernstein inequality (5.17) for unbounded random matrices. Let $Z_1, \dots, Z_N$ be independent $n \times n$ positive semidefinite random matrices. Show that

$$S := \sum_{i=1}^{N} Z_i \quad \text{satisfies} \quad \mathbb{E}\|S - \mathbb{E}\, S\| \le C\Big(\sqrt{\|\mathbb{E}\, S\| \cdot L} + L\Big)$$

where $L = \log(n)\, \mathbb{E} \max_i \|Z_i\|$.

6.34 ☕☕☕ (Covariance estimation for unbounded distributions) Let's relax the boundedness assumption (5.20) in the general covariance estimation result (5.27). Let X be a random vector in $\mathbb{R}^n$ with $\Sigma = \mathbb{E}\, XX^\mathsf{T}$. Assume that for some $K \ge 1$,

$$\mathbb{E} \max_{i \le m} \|X\|_2^2 \le K^2\, \mathbb{E}\|X\|_2^2.$$

Let $X_1, \dots, X_m$ be i.i.d. copies of X, and $\Sigma_m = \frac{1}{m} \sum_{i=1}^{m} X_i X_i^\mathsf{T}$. Prove that

$$\mathbb{E}\|\Sigma_m - \Sigma\| \le C\bigg(\sqrt{\frac{K^2 r \log n}{m}} + \frac{K^2 r \log n}{m}\bigg)\, \|\Sigma\|.$$

where $r = \operatorname{tr}(\Sigma)/\|\Sigma\|$ is the effective rank of Σ.

6.35 ☕☕ Let $z_1, \dots, z_N$ be any vectors in a normed space, let $a = (a_1, \dots, a_N) \in \mathbb{R}^N$, and let $\varepsilon_1, \dots, \varepsilon_N$ be independent Rademacher random variables. Show that the following function is convex:

$$f(a) := \mathbb{E}\Big\|\sum_{i=1}^{N} a_i \varepsilon_i z_i\Big\|.$$

6.36 ☕☕ (Contraction principle for general distributions) Let's prove the following generalization of Theorem 6.6.1. Let $a = (a_1, \dots, a_n) \in \mathbb{R}^n$, and let $X_1, \dots, X_N$ be independent, mean-zero random vectors taking values in a normed space. Show that

$$\mathbb{E}\Big\|\sum_{i=1}^{N} a_i X_i\Big\| \le 4\|a\|_\infty \cdot \mathbb{E}\Big\|\sum_{i=1}^{N} X_i\Big\|.$$

6.37 ☕☕ (Logarithmic factor is unavoidable in Gaussian symmetrization) Show that, in general, the factor $\sqrt{\log N}$ in Lemma 6.6.2 is optimal.

6.38 ☕ (Contraction and symmetrization for functions of norms) Let $F : \mathbb{R}_+ \to \mathbb{R}$ be a convex increasing function. Generalize contraction (Theorem 6.6.1) and Gaussian symmetrization (Lemma 6.6.2) by replacing the norm $\|\cdot\|$ with $F(\|\cdot\|)$ throughout.

7

Random Processes

In this chapter, we turn our attention to random processes – collections of random variables $(X_t)_{t \in T}$, which may be dependent. In classical settings like Brownian motion, t represents time, so $T \subset \mathbb{R}$. But in high-dimensional probability, T can be any abstract set. A key example is the canonical Gaussian process

$$X_t = \langle g, t \rangle, \quad t \in T,$$

where $T \subset \mathbb{R}^n$ and $g \sim N(0, I_n)$. We introduce it in Section 7.1.

In Section 7.2, we explore powerful comparison inequalities for Gaussian processes – Slepian, Sudakov-Fernique, and Gordon – using a new trick: Gaussian interpolation. In Section 7.3, we use these tools to prove a sharp bound on the operator norm of an $m \times n$ Gaussian random matrices.

How does a Gaussian process $(X_t)_{t \in T}$ capture the geometry of T? In Section 7.4, we prove Sudakov inequality – a lower bound on the *Gaussian width*

$$w(T) = \mathbb{E} \sup_{t \in T} \langle g, t \rangle$$

using covering numbers. (Upper bounds come later in Chapter 8.) Gaussian width is a key concept linking probability with metric geometry; we take a closer look at it in Section 7.5 and connect it to other ideas like effective dimension.

In Section 7.6, we compute the size of a random projection of any bounded set $T \subset \mathbb{R}^n$. The crucial quantity that determines it is the Gaussian width.

You will have plenty of chances to practice these ideas: prove symmetrization and contraction inequalities for random processes (Exercises 7.2–7.4, 7.8), derive the important min-max inequality for Gaussian processes – Gordon inequality (Exercise 7.9), get sharp bounds for Gaussian matrices (Exercises 7.11 and 7.13), compute the Gaussian width for the ℓ^p balls (Exercise 7.17), explore the nuclear norm (Exercises 7.18 and 7.19), effective dimension (Exercises 7.21–7.22), random projections of general sets (Exercises 7.25 and 7.26), and matrix sketching (Exercise 7.27).

[0] This material will be published by Cambridge University Press as *High-Dimensional Probability, 2nd Edition* by Roman Vershynin. This pre-publication version is free to view and download for personal use only. Not for re-distribution, re-sale or use in derivative works. ©Roman Vershynin 202X.
If you want to be notified when the printed version is available, let me know: rvershyn@uci.edu

7.1 Basic concepts and examples

Definition 7.1.1 (Random process). A *random process* is just a collection of random variables $(X_t)_{t\in T}$ on the same probability space, which are indexed by elements t of some set T.

In some classical examples, t stands for *time*, in which case T is a subset of $\mathbb{R}$. But we primarily study processes in high-dimensional settings, where T is a subset of $\mathbb{R}^n$ and where the analogy with time is lost.

Example 7.1.2 (Discrete time). If $T = \{1,\dots,n\}$ then the random process

$$(X_1,\dots,X_n)$$

can be identified 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.

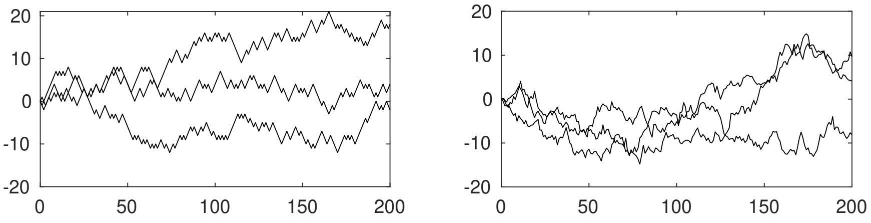


Figure 7.1 A few trials of a random walk (left) and standard Brownian motion (right).

Example 7.1.4 (Brownian motion). The most classical continuous-time random process is the standard *Brownian motion* $(X_t)_{t\geq 0}$, or the *Wiener process*. It can be characterized as follows:

(i) The process has continuous sample paths, i.e. the random function $f(t) := X_t$ is continuous almost surely;

(ii) The increments are independent and satisfy $X_t - X_s \sim N(0, t-s)$ for all $t \geq s$.

Figure 7.1 shows 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 $(X_t)_{t\in T}$ is sometimes called a spacial random process, or a *random field*. For example, the water temperature X_t at the location on Earth that is parametrized by t can be modeled as a spacial random process.

7.1.1 Covariance and increments

In Section 3.2, we introduced the notion of the covariance matrix of a random vector. We now define the *covariance function* of a random process $(X_t)_{t \in T}$ in a similar manner. For simplicity, let us assume in this section that the random process has zero mean, i.e.

$$\mathbb{E}\, X_t = 0 \quad \text{for all } t \in T.$$

(The adjustments for the general case will be obvious.) The covariance function of the process is defined as

$$\Sigma(t,s) := \operatorname{cov}(X_t, X_s) = \mathbb{E}\, X_t X_s, \quad t,s \in T.$$

The *increments* of the random process are defined as

$$d(t,s) := \|X_t - X_s\|_{L^2} = \left(\mathbb{E}(X_t - X_s)^2\right)^{1/2}, \quad t,s \in T. \tag{7.1}$$

Example 7.1.6. The increments of the standard Brownian motion satisfy

$$d(t,s) = \sqrt{t-s}, \quad t \ge s$$

by definition. The increments of a random walk of Example 7.1.3 with $\mathbb{E}\, Z_i^2 = 1$ behave similarly:

$$d(n,m) = \sqrt{n-m}, \quad n \ge m.$$

(Check!)

Remark 7.1.7 (The canonical metric)**.** Even if the index T has no geometric structure, the increments $d(t,s)$ always define a *metric* on T, thus automatically turning T into a *metric space*.[1] However, as we see in Example 7.1.6, this metric may not match the Euclidean distance on $\mathbb{R}^n$.

Remark 7.1.8 (Covariance vs. increments)**.** The covariance and the increments contain roughly the same information about the random process. Increments can be written using the covariance: just expand the square in (7.1) to see that

$$d(t,s)^2 = \Sigma(t,t) - 2\Sigma(t,s) + \Sigma(s,s).$$

Vice versa, if the zero random variable 0 belongs to the process, you can also recover the covariance from the increments (see Exercise 7.1).

7.1.2 Gaussian processes

Definition 7.1.9 (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. (To see why these two are equivalent, recall Exercise 3.16.)

[1] More precisely, $d(t,s)$ is a *pseudometric* on T since the distance between two distinct points can be zero, i.e. $d(t,s) = 0$ does not necessarily imply $t = s$.

The notion of Gaussian processes generalizes that of Gaussian random vectors in $\mathbb{R}^n$. A classical example of a Gaussian process is the standard Brownian motion.

Remark 7.1.10 (Distribution is determined by covariance, increments)**.** The distribution of a mean-zero Gaussian random vector in $\mathbb{R}^n$ is completely determined by its covariance matrix (recall Proposition 3.3.5). Then the same goes for a mean-zero Gaussian process: its distribution is determined by the covariance function $\Sigma(t,s)$, or equivalently (due to Exercise 7.1) by the increments $d(t,s)$, assuming the zero variable is part of the process.

Many tools we learned about random vectors can be applied to random processes. For example, Gaussian concentration (Theorem 5.2.3) implies:

Theorem 7.1.11 (Concentration of Gaussian processes)**.** *Let $(X_t)_{t\in T}$ be a Gaussian process with finite*[2] *T. Then*

$$\Big\| \sup_{t\in T} X_t - \mathbb{E} \sup_{t} X_t \Big\|_{\psi_2} \le C \sup_{t\in T} \sqrt{\operatorname{Var}(X_t)}.$$

This is just a restatement of Exercise 5.9(b) – if you skipped it then, do it now!

Let's look at a broad class of Gaussian processes indexed by high-dimensional sets $T \subset \mathbb{R}^n$. Take a standard normal vector $g \sim N(0, I_n)$ and define

$$X_t := \langle g, t \rangle, \quad t \in T. \tag{7.2}$$

This gives us a Gaussian process $(X_t)_{t\in T}$ called *the canonical Gaussian process.* The increments match the Euclidean distance:

$$\|X_t - X_s\|_{L^2} = \|t - s\|_2, \quad t, s \in T.$$

(Check!)

Actually, one can realize any Gaussian process as the canonical process (7.2). This follows from a simple observation about Gaussian vectors:

Lemma 7.1.12 (Gaussian random vectors)**.** *Let X 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*

$$X \overset{dist}{=} \left(\langle g, t_i \rangle\right)_{i=1}^n, \quad \textit{where } g \sim N(0, I_n).$$

Here $\overset{dist}{=}$ means the equality of distributions.

Proof If Σ denotes the covariance matrix of X, then, by (3.12), we have

$$X \equiv \Sigma^{1/2} g \quad \text{where } g \sim N(0, I_n).$$

The entries of $\Sigma^{1/2} g$ are $\langle t_i, g \rangle$ where t_i are the rows of $\Sigma^{1/2}$. Done! □

It follows that for any Gaussian process $(X_s)_{s\in S}$, all finite-dimensional marginals $(X_s)_{s\in S_0}$, $|S_0| = n$ can be represented as the canonical Gaussian process (7.2) indexed in a certain subset $T_0 \subset \mathbb{R}^n$.

[2] We are assuming T is finite just to avoid measurability issues. But in practice, you can usually extend to general T by approximation. Try it for the canonical Gaussian process (7.2)!

7.2 Slepian, Sudakov-Fernique and Gordon inequalities

In many applications, it helps to have a *uniform* bound on a random process:

$$\mathbb{E}\sup_{t\in T} X_t = ?$$

Remark 7.2.1 (Making T finite)**.** To avoid measurability issues, let's think of $\mathbb{E}\sup_{t\in T} X_t$ as shorthand for $\sup_{T_0\subset T}\mathbb{E}\max_{t\in T_0} X_t$ where T_0 runs over all finite subsets. The general case usually follows by approximation.

For some processes, this quantity can be computed exactly. For example, if (X_t) is a standard Brownian motion, the so-called reflection principle gives

$$\mathbb{E}\sup_{t\le t_0} X_t = \sqrt{\frac{2t_0}{\pi}} \quad \text{for every } t_0 \ge 0.$$

For general random processes – even Gaussian – the problem is nontrivial.

The first general bound we prove is the Slepian comparison inequality for Gaussian processes. It basically says: the faster the process grows (in terms of the increments), the farther it gets.

Theorem 7.2.2 (Slepian inequality)**.** *Let $(X_t)_{t\in T}$ and $(Y_t)_{t\in T}$ be two mean-zero Gaussian processes. Assume that for all $t,s\in T$, we have*

$$\mathbb{E}X_t^2 = \mathbb{E}Y_t^2 \quad \textit{and} \quad \mathbb{E}(X_t - X_s)^2 \le \mathbb{E}(Y_t - Y_s)^2. \tag{7.3}$$

Then $\sup_{t\in T} X_t$ is stochastically dominated by $\sup_{t\in T} Y_t$: for every $\tau\in\mathbb{R}$, we have

$$\mathbb{P}\Big\{\sup_{t\in T} X_t \ge \tau\Big\} \le \mathbb{P}\Big\{\sup_{t\in T} Y_t \ge \tau\Big\}. \tag{7.4}$$

Consequently,

$$\mathbb{E}\sup_{t\in T} X_t \le \mathbb{E}\sup_{t\in T} Y_t. \tag{7.5}$$

We now prepare for the proof of Slepian inequality.

7.2.1 Gaussian interpolation

We will prove Slepian inequality using a trick called Gaussian interpolation. Assume that T is finite; then we can look at $X = (X_t)_{t\in T}$ and $Y = (Y_t)_{t\in T}$ as Gaussian random vectors in $\mathbb{R}^n$ where $n = |T|$. We may also assume that X and Y are independent. (Why?)

Define the Gaussian random vector $Z(u)$ in $\mathbb{R}^n$ 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 $Z(u)$ continuously interpolates linearly between the covariance matrices of Y and X:

$$\Sigma(Z(u)) = u\,\Sigma(X) + (1-u)\,\Sigma(Y).$$

(Check!)

For a given function $f : \mathbb{R}^n \to \mathbb{R}$, let's study how $\mathbb{E}\, f(Z(u))$ changes as u increases from 0 to 1. Of special interest to us is the function

$$f(x) = \mathbf{1}_{\{\max_i x_i < \tau\}}.$$

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 inequality at once, since then

$$\mathbb{E}\, f(Z(1)) \geq \mathbb{E}\, f(Z(0)), \quad \text{thus} \quad \mathbb{P}\Big\{\max_i X_i < \tau\Big\} \geq \mathbb{P}\Big\{\max_i Y_i < \tau\Big\}$$

as claimed.

Now let us pass to a detailed argument. To develop Gaussian interpolation, let us start with the following useful identity.

Lemma 7.2.3 (Gaussian integration by parts)**.** *Let $X \sim N(0,1)$. Then for any differentiable function $f : \mathbb{R} \to \mathbb{R}$ we have*

$$\mathbb{E}\, X f(X) = \mathbb{E}\, f'(X),$$

assuming both expectations exist and are finite.

Proof Assume first that f has bounded support. Denoting the Gaussian density 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

$$\int_{\mathbb{R}} f(x)p(x)x\, dx = \mathbb{E}\, X f(X),$$

as claimed. The result can be extended to general functions by an approximation argument. The lemma is proved. □

By rescaling, we can extend Gaussian integration by parts for $X \sim N(0, \sigma^2)$:

$$\mathbb{E}\, X f(X) = \sigma^2\, \mathbb{E}\, f'(X).$$

(Just write $X = \sigma Z$ for $Z \sim N(0,1)$ and apply Lemma 7.2.3.) We can also extend it to high dimensions:

Lemma 7.2.4 (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*

$$\mathbb{E}\, X f(X) = \Sigma \cdot \mathbb{E}\, \nabla f(X),$$

assuming both expectations exist and are finite. In other words,

$$\mathbb{E}\, X_i f(X) = \sum_{j=1}^{n} \Sigma_{ij}\, \mathbb{E}\, \frac{\partial f}{\partial x_j}(X), \quad i = 1, \ldots, n. \tag{7.7}$$

You will prove this extension in Exercise 7.6.

Lemma 7.2.5 (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*

$$Z(u) \coloneqq \sqrt{u}\, X + \sqrt{1-u}\, Y, \quad u \in [0,1]. \tag{7.8}$$

Then for any twice-differentiable function $f : \mathbb{R}^n \to \mathbb{R}$, we have

$$\frac{d}{du}\, \mathbb{E}\, f(Z(u)) = \frac{1}{2} \sum_{i,j=1}^{n} (\Sigma^X_{ij} - \Sigma^Y_{ij})\, \mathbb{E}\Big[\frac{\partial^2 f}{\partial x_i\, \partial x_j}(Z(u))\Big], \tag{7.9}$$

assuming all expectations exist and are finite.

Proof Using the chain rule,[3] we have

$$\begin{aligned}\frac{d}{du}\, \mathbb{E}\, f(Z(u)) &= \sum_{i=1}^{n} \mathbb{E}\, \frac{\partial f}{\partial x_i}(Z(u)) \frac{dZ_i}{du} \\ &= \frac{1}{2} \sum_{i=1}^{n} \mathbb{E}\, \frac{\partial f}{\partial x_i}(Z(u)) \Big(\frac{X_i}{\sqrt{u}} - \frac{Y_i}{\sqrt{1-u}}\Big) \quad \text{(by (7.8))}. \end{aligned} \tag{7.10}$$

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

$$\sum_{i=1}^{n} \frac{1}{\sqrt{u}}\, \mathbb{E}\, X_i \frac{\partial f}{\partial x_i}(Z(u)) = \sum_{i=1}^{n} \frac{1}{\sqrt{u}}\, \mathbb{E}\, X_i g_i(X), \tag{7.11}$$

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 7.2.4). According to (7.7), we have

$$\begin{aligned}\mathbb{E}\, X_i g_i(X) &= \sum_{j=1}^{n} \Sigma^X_{ij}\, \mathbb{E}\, \frac{\partial g_i}{\partial x_j}(X) \\ &= \sum_{j=1}^{n} \Sigma^X_{ij}\, \mathbb{E}\, \frac{\partial^2 f}{\partial x_i\, \partial x_j}(\sqrt{u}\, X + \sqrt{1-u}\, Y) \cdot \sqrt{u}.\end{aligned}$$

[3] Here we use the multivariate chain rule to differentiate a function $f(g_1(u), \ldots, g_n(u))$ where $g_i : \mathbb{R} \to \mathbb{R}$ and $f : \mathbb{R}^n \to \mathbb{R}$ as follows: $\frac{df}{du} = \sum_{i=1}^n \frac{\partial f}{\partial x_i} \frac{dg_i}{du}$.

Substitute this into (7.11) to get

$$\sum_{i=1}^{n} \frac{1}{\sqrt{u}} \mathbb{E}\, X_i \frac{\partial f}{\partial x_i}(Z(u)) = \sum_{i,j=1}^{n} \Sigma^X_{ij}\, \mathbb{E}\, \frac{\partial^2 f}{\partial x_i\, \partial x_j}(Z(u)).$$

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. □

7.2.2 Proof of Slepian inequality

We are ready to establish a preliminary, functional form Slepian inequality.

Lemma 7.2.6 (Slepian inequality, functional form). *Consider two mean-zero Gaussian random vectors X and Y in $\mathbb{R}^n$. Assume that for all $i, j = 1, \ldots, n$, we have*

$$\mathbb{E}\, X_i^2 = \mathbb{E}\, Y_i^2 \quad \textit{and} \quad \mathbb{E}(X_i - X_j)^2 \le \mathbb{E}(Y_i - Y_j)^2.$$

Consider a twice-differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ such that

$$\frac{\partial^2 f}{\partial x_i\, \partial x_j} \ge 0 \quad \textit{for all } i \ne j.$$

Then

$$\mathbb{E}\, f(X) \ge \mathbb{E}\, f(Y),$$

assuming both expectations exist and are finite.

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 $i, j = 1, \ldots, n$. We can assume that X and Y are independent. (Why?) Apply Lemma 7.2.5 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 inequality, Theorem 7.2.2. Let us state and prove it in the equivalent form for Gaussian random vectors.

Theorem 7.2.7 (Slepian inequality). *Let X and Y be Gaussian random vectors as in Lemma 7.2.6. Then for every $\tau \ge 0$ we have*

$$\mathbb{P}\Big\{\max_{i \le n} X_i \ge \tau\Big\} \le \mathbb{P}\Big\{\max_{i \le n} Y_i \ge \tau\Big\}.$$

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 $(-\infty, \tau)$:

$$h(x) \approx \mathbf{1}_{(-\infty,\tau)},$$

like in Figure 7.2. Define the function $f : \mathbb{R}^n \to \mathbb{R}$ by

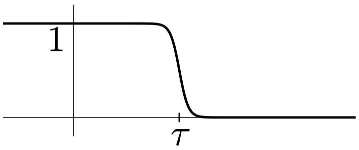


Figure 7.2 The function $h(x)$ is a smooth, non-increasing approximation to the indicator function $\mathbf{1}_{(-\infty,\tau)}$.

$$f(x) = h(x_1) \cdots h(x_n).$$

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 inequality, Lemma 7.2.6, for $f(x)$. To check the assumptions of this result, note that for $i \ne j$ 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 non-positive and the others are nonnegative by the assumption. Thus the second derivative is nonnegative, as required.

It follows that

$$\mathbb{E} f(X) \ge \mathbb{E} f(Y).$$

By approximation, this implies

$$\mathbb{P}\Big\{\max_{i \le n} X_i < \tau\Big\} \ge \mathbb{P}\Big\{\max_{i \le n} Y_i < \tau\Big\}.$$

This proves the first part of the conclusion. The second part follows by using the integrated tail formula in Exercise 1.15(b) (check!) □

7.2.3 Sudakov-Fernique and Gordon inequalities

Slepian 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 now remove the assumption on the equality of variances, and still be able to obtain (7.5).

Theorem 7.2.8 (Sudakov-Fernique inequality)**.** *Let $(X_t)_{t\in T}$ and $(Y_t)_{t\in T}$ be two mean-zero Gaussian processes. Assume that for all $t, s \in T$, we have*

$$\mathbb{E}(X_t - X_s)^2 \le \mathbb{E}(Y_t - Y_s)^2.$$

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 inequality in Theorem 7.2.7. We again deduce the result from Gaussian Interpolation Lemma 7.2.5. 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

$$f(x) \to \max_{i\le n} x_i \quad \text{as } \beta \to \infty.$$

(Do this!) Substituting $f(x)$ into the Gaussian interpolation formula (7.9) and simplifying the expression shows that

$$\frac{d}{du}\,\mathbb{E}\, f(Z(u)) \le 0 \quad \text{for all } u \in [0,1].$$

(Work through this step carefully in Exercise 7.7!) Then we can finish the proof just like in Slepian inequality. □

Some applications call for *min-max* bounds on random processes. For example, by (4.14), the smallest singular value of an $m \times n$ Gaussian matrix A with all $N(0,1)$ i.i.d. entries can be written as

$$s_n(A) = \min_{u\in S^{n-1}} \|Au\|_2 = \min_{u\in S^{n-1}} \max_{v\in S^{m-1}} X_{uv},$$

where $X_{uv} = \langle Au, v\rangle$ are normal random variables. A handy tool for this kind of problems is Gordon inequality, which extends Slepian and Sudakov-Fernique to the min-max setting:

Theorem 7.2.9 (Gordon inequality)**.** *Let $(X_{ut})_{u\in U,\, t\in T}$ and $Y = (Y_{ut})_{u\in U,\, t\in T}$ be two mean-zero Gaussian processes indexed by pairs of points (u,t) in a product set $U \times T$. Assume that*

$$\begin{aligned} &\mathbb{E}(X_{ut} - X_{us})^2 \le \mathbb{E}(Y_{ut} - Y_{us})^2 \quad \textit{for all } u, t, s;\\ &\mathbb{E}(X_{ut} - X_{vs})^2 \ge \mathbb{E}(Y_{ut} - Y_{vs})^2 \quad \textit{for all } u \ne v \textit{ and all } t, s. \end{aligned}$$

[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*.

Then, for every $\tau \geq 0$,

$$\mathbb{P}\Big\{\inf_{u\in U}\sup_{t\in T} X_{ut} \geq \tau\Big\} \leq \mathbb{P}\Big\{\inf_{u\in U}\sup_{t\in T} Y_{ut} \geq \tau\Big\},$$

and so, by the integrated tail formula,

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

Try yourself to prove Gordon inequality under an additional assumption of equal variances (Exercise 7.9). Also try your hand at a Gaussian version of Talagrand contraction principle (Exercise 7.8).

7.3 Application: sharp bounds for Gaussian matrices

Let's apply the Gaussian comparison inequalities we just proved to random matrices. In Section 4.6, we studied $m \times n$ random matrices A with independent subgaussian rows. We used the ε-net argument to bound the expected operator norm of A like this:

$$\mathbb{E}\|A\| \leq \sqrt{m} + C\sqrt{n}$$

where C is a constant (see Exercise 4.41). Now, using the Sudakov-Fernique inequality, we will tighten this bound for *Gaussian* random matrices, proving it with sharp constant $C = 1$.

Theorem 7.3.1 (Norms of Gaussian random matrices)**.** *Let A be an $m\times n$ matrix with independent $N(0,1)$ entries. Then*

$$\mathbb{E}\|A\| \leq \sqrt{m} + \sqrt{n}.$$

Proof Let's write the norm of A as a supremum of a Gaussian process: by (4.9),

$$\|A\| = \max_{u\in S^{n-1},\ v\in S^{m-1}} \langle Au, v\rangle = \max_{(u,v)\in T} X_{uv}$$

where $T = S^{n-1} \times S^{m-1}$ and

$$X_{uv} := \langle Au, v\rangle \sim N(0,1).$$

To apply Sudakov-Fernique comparison inequality (Theorem 7.2.8), let us compute the increments of the process (X_{uv}). For any $(u,v), (w,z) \in T$, we have

$$\begin{aligned}
\mathbb{E}(X_{uv} - X_{wz})^2 &= \mathbb{E}\left(\langle Au, v\rangle - \langle Aw, z\rangle\right)^2 = \mathbb{E}\Big(\sum_{i,j} A_{ij}(u_j v_i - w_j z_i)\Big)^2 \\
&= \sum_{i,j}(u_j v_i - w_j z_i)^2 \quad \text{(by independence, mean 0, variance 1)} \\
&= \|uv^\mathsf{T} - wz^\mathsf{T}\|_F^2 \\
&\leq \|u - w\|_2^2 + \|v - z\|_2^2 \quad \text{(check this in Exercise 7.10).}
\end{aligned}$$

Let's define a simpler Gaussian process (Y_{uv}) with similar increments:

$$Y_{uv} := \langle g, u\rangle + \langle h, v\rangle, \quad (u,v) \in T,$$

where $g \sim N(0, I_n)$ and $h \sim N(0, I_m)$ 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 inequality. Applying Theorem 7.2.8, we obtain

$$\begin{aligned}\mathbb{E}\,\|A\| = \mathbb{E} \sup_{(u,v)\in T} X_{uv} &\le \mathbb{E} \sup_{(u,v)\in T} Y_{uv} \\ &= \mathbb{E} \sup_{u\in S^{n-1}} \langle g, u\rangle + \mathbb{E} \sup_{v\in S^{m-1}} \langle h, v\rangle \\ &= \mathbb{E}\,\|g\|_2 + \mathbb{E}\,\|h\|_2 \\ &\le (\mathbb{E}\,\|g\|_2^2)^{1/2} + (\mathbb{E}\,\|h\|_2^2)^{1/2} \quad \text{(recall Exercise 1.11)} \\ &= \sqrt{n} + \sqrt{m} \quad \text{(see Proposition 3.2.1(b)).}\end{aligned}$$

□

Theorem 7.3.1 only gives an expectation bound, but we can boost it to a high-probability bound using the concentration tools from Section 5.2:

Corollary 7.3.2 (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*

$$\mathbb{P}\{\|A\| \ge \sqrt{m} + \sqrt{n} + t\} \le 2\exp(-ct^2).$$

Proof Let's combine the expectation bound (Theorem 7.3.1) with Gaussian concentration (Theorem 5.2.3). Think of A as a long random vector in $\mathbb{R}^{m\times n}$ by concatenating the rows. This makes A a standard normal random vector: $A \sim N(0, I_{nm})$. Consider the function

$$f(A) := \|A\|$$

that maps the vectorized matrix to the matrix's operator norm. Since the operator norm is bounded by Frobenius norm, and the Frobenius norm is just the Euclidean norm on $\mathbb{R}^{m\times n}$, f is a Lipschitz function on $\mathbb{R}^{m\times n}$ with Lipschitz norm is bounded by 1. (Why?) Then Theorem 5.2.3 yields

$$\mathbb{P}\{\|A\| \ge \mathbb{E}\|A\| + t\} \le 2\exp(-ct^2).$$

The bound on $\mathbb{E}\|A\|$ from Theorem 7.3.1 completes the proof. □

Now try Exercise 7.11 to prove that a *symmetric* Gaussian matrix satisfies

$$\mathbb{E}\|A\| \le 2\sqrt{n},$$

and Exercise 7.13 to show that the *smallest* singular value of an $m \times n$ Gaussian matrix A satisfies

$$\mathbb{E}\, s_n(A) \ge \sqrt{m} - \sqrt{n}.$$

7.4 Sudakov inequality

Let's go back to general mean-zero Gaussian processes $(X_t)_{t\in T}$ on any index set T. As we saw earlier (Remark 7.1.7), the increments

$$d(t,s) := \|X_t - X_s\|_{L^2} = \big(\mathbb{E}(X_t - X_s)^2\big)^{1/2} \tag{7.13}$$

define a metric on T, called the *canonical metric*. This metric determines the covariance function $\Sigma(t,s)$, which in turn determines the distribution of the process $(X_t)_{t\in T}$ (recall Remark 7.1.10). So, in theory, we can answer any question about the distribution of the process just by understanding the geometry of the metric space (T,d) – in other words, we can study probability via geometry.

Here is a big specific question: how can we estimate

$$\mathbb{E}\sup_{t\in T} X_t \tag{7.14}$$

in terms of the geometry of (T,d)? This is a hard problem, which we begin to study now and continue in Chapter 8.

We will start with a lower bound on (7.14) in terms of *metric entropy*, which was in introduced Section 4.2. Recall that for any $\varepsilon > 0$, the *covering number*

$$\mathcal{N}(T,d,\varepsilon)$$

is the smallest cardinality of an ε-net of T in the metric d, or equivalently the smallest number[5] of closed balls of radius ε whose union covers T. The logarithm of the covering number, $\log_2 \mathcal{N}(T,d,\varepsilon)$, is called the *metric entropy* of T.

Theorem 7.4.1 (Sudakov inequality)**.** *Let $(X_t)_{t\in T}$ be a mean-zero Gaussian process. Then, for any $\varepsilon \geq 0$, we have*

$$\mathbb{E}\sup_{t\in T} X_t \geq c\varepsilon\sqrt{\log \mathcal{N}(T,d,\varepsilon)}.$$

where d is the canonical metric defined in (7.13)*.*

Proof Let us deduce this result from Sudakov-Fernique comparison inequality (Theorem 7.2.8). Assume that

$$\mathcal{N}(T,d,\varepsilon) =: N$$

is finite; you will handle the infinite case in Exercise 7.14. Let $\mathcal{N}$ be a maximal ε-separated subset of T. Then $\mathcal{N}$ is an ε-net of T (recall Lemma 4.2.6), and thus

$$|\mathcal{N}| \geq N.$$

Restricting the process to $\mathcal{N}$, we see that it suffices to show that

$$\mathbb{E}\sup_{t\in\mathcal{N}} X_t \geq c\varepsilon\sqrt{\log N}.$$

[5] If T does not have a finite ε-net, we set $\mathcal{N}(T,d,\varepsilon) = \infty$.

Let's do it by comparing $(X_t)_{t\in\mathcal{N}}$ to a simpler Gaussian process $(Y_t)_{t\in\mathcal{N}}$, defined 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 comparison inequality (Theorem 7.2.8), 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

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

(In the last line, we use that $g_t - g_s \sim N(0,2)$.) This implies 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 7.2.8, we obtain

$$\mathbb{E}\sup_{t\in\mathcal{N}} X_t \geq \mathbb{E}\sup_{t\in\mathcal{N}} Y_t = \frac{\varepsilon}{\sqrt{2}}\, \mathbb{E}\max_{t\in\mathcal{N}} g_t \geq c\varepsilon\sqrt{\log N}.$$

In the last step, we used that the expected maximum of N i.i.d. $N(0,1)$ random variables is at least $c\sqrt{\log N}$, see Exercise 2.38(b). The proof is complete. □

7.4.1 Application for covering numbers in $\mathbb{R}^n$

Sudakov inequality can be used to bound the covering numbers of an arbitrary set $T \subset \mathbb{R}^n$:

Corollary 7.4.2 (Sudakov 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 smallest number of Euclidean balls with radius ε and centers in T that cover T, just like in Section 4.2.1.

Proof Consider the canonical Gaussian process $X_t := \langle g, t\rangle$ where $g \sim N(0, I_n)$. As we noted in Section 7.1.2, the canonical distance for this process is the Euclidean distance in $\mathbb{R}^n$, i.e. $d(t,s) = \|X_t - X_s\|_{L^2} = \|t-s\|_2$ for any $t, s \in T$. Then the corollary follows from Sudakov inequality (Theorem 7.4.1). □

In Exercise 8.5, you will show that Corollary 7.4.2 is sharp up to a log factor:

$$\mathbb{E}\sup_{t\in T}\langle g, t\rangle \leq C\log(n)\cdot\varepsilon\sqrt{\log \mathcal{N}(T,\varepsilon)}.$$

For a quick application of Sudakov inequality, let's (roughly) re-derive the bound on covering numbers of polytopes in $\mathbb{R}^n$ from Corollary 0.0.3:

Corollary 7.4.3 (Covering numbers of polytopes)**.** *Let P be a polytope in $\mathbb{R}^n$ with N vertices, contained in the unit Euclidean ball. Then, for every $\varepsilon > 0$ we have*

$$\mathcal{N}(P, \varepsilon) \le N^{C/\varepsilon^2}.$$

Proof If $x_1, \dots, x_N$ are the vertices of P, then

$$\mathbb{E} \sup_{t \in P} \langle g, t \rangle \le \mathbb{E} \sup_{i \le N} \langle g, x_i \rangle \le C\sqrt{\log N}. \tag{7.15}$$

The first bound follows from the maximal principle (Exercise 1.4): since P lies the convex hull of its vertices, for each fixed g the linear (and thus convex) function $t \mapsto \langle g, t \rangle$ attains its maximum at a vertex. The second bound in (7.15) is due to the maximal inequality (2.22), since $\langle g, x \rangle \sim N(0, \|x\|_2^2)$ and $\|x\|_2 \le 1$. Substitute this into Corollary 7.4.2 and simplify to complete the proof. □

7.5 Gaussian width

In Section 7.4.1, we saw an important quantity associated with any set $T \subset \mathbb{R}^n$: the size of the canonical Gaussian process on T. It shows up a lot in high-dimensional probability, so let's give it a name and look at its basic properties.

Definition 7.5.1. The *Gaussian width* of a subset $T \subset \mathbb{R}^n$ is defined as

$$w(T) := \mathbb{E} \sup_{x \in T} \langle g, x \rangle \quad \text{where} \quad g \sim N(0, I_n).$$

You can think of Gaussian width as a fundamental geometric measure of a set $T \subset \mathbb{R}^n$, like volume or surface area.

Proposition 7.5.2 (Simple properties of Gaussian width)**.**

(a) (Finiteness) $w(T)$ is finite if and only if T is bounded.
(b) (Invariance) $w(UT + y) = w(T)$ for any orthogonal matrix U and vector y.
(c) (Convex hulls) $w(\mathrm{conv}(T)) = w(T)$.
(d) (Minkowski addition and scaling)[6] $w(T + S) = w(T) + w(S)$ and $w(aT) = |a|\, w(T)$ for any $T, S \subset \mathbb{R}^n$ and $a \in \mathbb{R}$.
(e) (Symmetry):

$$w(T) = \frac{1}{2} w(T - T) = \frac{1}{2} \mathbb{E} \sup_{x, y \in T} \langle g, x - y \rangle.$$

(f) (Width and diameter):[7]

$$\frac{1}{\sqrt{2\pi}} \cdot \mathrm{diam}(T) \le w(T) \le \frac{\sqrt{n}}{2} \cdot \mathrm{diam}(T).$$

(g) (Linear maps) For any $m \times n$ matrix A, we have $w(AT) \le \|A\|\, w(T)$.

[6] Recall Definition 4.2.9 of Minkowski sum and difference of two sets: $T \pm S = \{t \pm s : \ t \in T,\ s \in S\}$.
[7] Recall that the diameter of a set $T \subset \mathbb{R}^n$ is defined as $\mathrm{diam}(T) := \sup\{\|x - y\|_2 : \ x, y \in T\}$.

Proof Let's just prove part (f) and leave the rest for you in Exercise 7.15.

For the lower bound, fix any $x, y \in T$. Since both $x-y$ and $y-x$ are in $T-T$, property (e) gives

$$w(T) \geq \frac{1}{2}\,\mathbb{E}\max\big(\langle x-y, g\rangle,\ \langle y-x, g\rangle\big) = \frac{1}{2}\,\mathbb{E}|\langle x-y, g\rangle| = \frac{1}{2}\sqrt{\frac{2}{\pi}}\|x-y\|_2.$$

The last equality holds since $\langle x-y, g\rangle \sim N(0, \|x-y\|_2^2)$ and $\mathbb{E}|X| = \sqrt{2/\pi}$ for $X \sim N(0,1)$. (Check!) Taking the supremum over all $x, y \in T$ gives the result.

For the upper bound in (f), use property (e) again to get

$$w(T) = \frac{1}{2}\,\mathbb{E}\sup_{x,y\in T}\langle g, x-y\rangle \leq \frac{1}{2}\,\mathbb{E}\sup_{x,y\in T}\|g\|_2\|x-y\|_2 \leq \frac{1}{2}\,\mathbb{E}\|g\|_2 \cdot \operatorname{diam}(T).$$

Since $\mathbb{E}\|g\|_2 \leq (\mathbb{E}\|g\|_2^2)^{1/2} = \sqrt{n}$, we are done. □

Remark 7.5.3 (Width and diameter)**.** Both upper and lower bounds in (f) are optimal and the $O(\sqrt{n})$ gap between them cannot be improved (see Exercise 7.16). So, diameter is not a great way to capture Gaussian width.

7.5.1 Geometric meaning of width

Gaussian width has a nice geometric meaning: it's about how wide the set $T \subset \mathbb{R}^n$ looks in random directions. The width of T in the direction $\theta \in S^{n-1}$ is the width of the smallest slab (between parallel hyperplanes orthogonal to θ) that contains T (see Figure 7.3), which can be expressed as $\sup_{x,y\in T}\langle\theta, x-y\rangle$ (check)! If we average the width over all unit directions θ, we get

$$\mathbb{E}\sup_{x,y\in T}\langle\theta, x-y\rangle = \mathbb{E}\sup_{z\in T-T}\langle\theta, z\rangle. \tag{7.16}$$

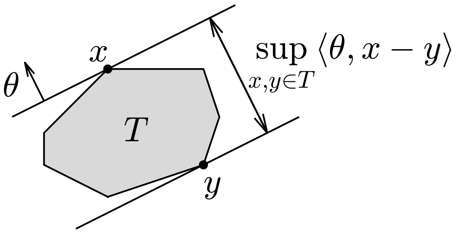


Figure 7.3 The width of a set $T \subset \mathbb{R}^n$ in the direction of a unit vector θ.

Definition 7.5.4 (Spherical width)**.** The *spherical width* of a set $T \subset \mathbb{R}^n$ is

$$w_s(T) := \mathbb{E}\sup_{x\in T}\langle\theta, x\rangle \quad \text{where} \quad \theta \sim \operatorname{Unif}(S^{n-1}).$$

The only difference between the Gaussian and spherical widths is in the random vectors we average over: $g \sim N(0, I_n)$ versus $\theta \sim \operatorname{Unif}(S^{n-1})$. Both are rotation invariant, but g is approximately $\sqrt{n}$ times longer than θ. Thus:

Lemma 7.5.5 (Gaussian vs. spherical widths)**.** *The Gaussian width is approximately $\sqrt{n}$ times the spherical width:*

$$\Big(\sqrt{n} - \frac{C}{\sqrt{n}}\Big)\, w_s(T) \le w(T) \le \sqrt{n}\, w_s(T).$$

Proof Express the Gaussian vector g through its length and direction: $g = r\theta$, where $r = \|g\|_2$ and $\theta = g/\|g\|_2$. Now, $\theta \sim \mathrm{Unif}(S^{n-1})$ is independent of r (see Exercise 3.22). 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 use concentration of the norm (see Exercise 3.2), which gives $\sqrt{n} - C/\sqrt{n} \le \mathbb{E}\|g\|_2 \le \sqrt{n}$, and finishes the proof. □

7.5.2 Examples

Example 7.5.6 (Euclidean ball and sphere)**.** The Gaussian widths of the unit ball and sphere are

$$w(S^{n-1}) = w(B_2^n) = \mathbb{E}\|g\|_2 = \sqrt{n} \pm \frac{C}{\sqrt{n}}, \tag{7.17}$$

where we used concentration of norm (Exercise 3.2) in the last step. The spherical widths of these sets of course equal 1.

Example 7.5.7 (Cube)**.** The unit ball of the ℓ^∞ norm in $\mathbb{R}^n$ is the cube $B_\infty^n = [-1, 1]^n$. So, using duality (1.6), we get

$$w(B_\infty^n) = \mathbb{E}\|g\|_1 = \mathbb{E}|g_1| \cdot n = \sqrt{\frac{2}{\pi}} \cdot n. \tag{7.18}$$

Example 7.5.8 (Cross-polytope)**.** The unit ball of the ℓ^1 norm in $\mathbb{R}^n$ is the *cross-polytope* $B_1^n = \{x \in \mathbb{R}^n : \ \|x\|_1 \le 1\}$, see (1.3). Its Gaussian width satisfies

$$w(B_1^n) \asymp \sqrt{\log n} \tag{7.19}$$

where the notation $\asymp$ hides absolute constant factors. This is because

$$w(B_1^n) = \mathbb{E}\, \|g\|_\infty = \mathbb{E} \max_{i \le n} |g_i|,$$

where the first equation uses duality (1.6). Then (7.19) follow from Exercise 2.38(b).

Example 7.5.9 (Finite point sets)**.** Any finite set of points $T \subset \mathbb{R}^n$ satisfies

$$w(T) \le C\sqrt{\log|T|} \cdot \mathrm{diam}(T).$$

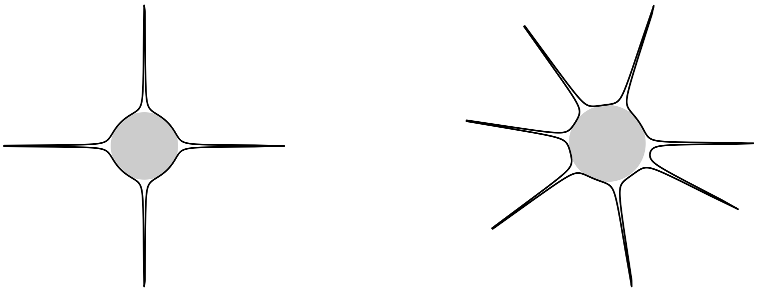
To prove this, we can assume that $\mathrm{diam}(T) = 1/2$ (by rescaling), and that T lies in the unit Euclidean ball (by translation). Then argue as in (7.15) (do it!)

Remark 7.5.10 (Surprising behavior of width in high dimensions)**.** As we can see from Examples 7.5.6–7.5.8, the Gaussian width of the cube B_∞^n is roughly (up to a constant factor) the same as that of its *circumscribed* ball $\sqrt{n}B_2^n$. But for the cross-polytope B_1^n, the width is roughly (up to a log factor) like that of its *inscribed* ball $\frac{1}{\sqrt{n}}B_2^n$, which is tiny! Why?

The cube B_∞^n has so many vertices (2^n) that in most directions it sticks out to roughly the circumscribed ball, which drives the width. But the cross-polytope B_1^n only has $2n$ vertices, so a random direction $g \sim N(0, I_n)$ likely to be far from all of them. The width is not driven by those lonely $2n$ "spikes" – it's driven by the "bulk", which is roughly the inscribed ball.

Figure 7.4a shows Milman's *hyperbolic sketch* of B_1^n, highlighting how the bulk (the inscribed ball) dominates since set has few vertices (spikes). You can make similar sketches for general convex sets, too (Figure 7.4b) – they are great for building high-dimensional intuition, even if you lose convexity in the picture.

For more practice, try Exercise 7.17 to compute the Gaussian width of any ℓ^p ball, covering all three examples above at once – the ball ($p = 2$), the cube ($p = \infty$), and the cross-polytope ($p = 1$). Also work through Exercises 7.18–7.19 to compute the Gaussian width of the set of $n \times n$ matrices with operator norm at most 1.



(a) The octahedron B_1^n

(b) General convex set

Figure 7.4 V. Milman's hyperbolic sketch of high-dimensional convex sets

7.5.3 Gaussian complexity and effective dimension

There are a couple of helpful cousins of the Gaussian width $w(T)$. Normally, we would take the expected max of $\langle g, t\rangle$, but sometimes it's easier to work with L^1 or L^2 averages:

$$w(T) = \mathbb{E}\sup_{x\in T}\langle g, x\rangle, \quad \gamma(T) := \mathbb{E}\sup_{x\in T}|\langle g, x\rangle|, \quad h(T) := \Big(\mathbb{E}\sup_{x\in T}\langle g, x\rangle^2\Big)^{1/2},$$

where $g \sim N(0, I_n)$. We call $\gamma(T)$ the *Gaussian complexity* of T. Clearly,

$$w(T) \le \gamma(T) \le h(T),$$

and the reverse bounds are basically true too:

Lemma 7.5.11 (Almost equivalent versions of Gaussian width)**.** *For any bounded set $T \subset \mathbb{R}^n$, we have:*[8]

(a) $\gamma(T - T) = 2w(T)$.

(b) $h(T) \asymp \gamma(T) \asymp w(T) + \|y\|_2$ *for any point* $y \in T$.

In particular, if T contains the origin, all three versions are equivalent:

$$h(T) \asymp \gamma(T) \asymp w(T).$$

Proof (a) follows from Proposition 7.5.2(e), since $T - T$ is origin-symmetric.

(b) Let's just prove the first part and leave the equivalence $\gamma(T) \asymp w(T) + \|y\|_2$ for you to prove in Exercise 7.20. We trivially have $\gamma(T) \le h(T)$. For the reverse, look at the function $z \mapsto \sup_{x \in T} |\langle z, x \rangle|$ on $\mathbb{R}^n$. Its Lipschitz norm is bounded by the radius $\sup_{x \in T} \|x\|_2 = r(T)$ (check!). Then, by Gaussian concentration (5.5),

$$\Big\| \sup_{x \in T} |\langle g, x \rangle| - \gamma(T) \Big\|_{\psi_2} \lesssim r(T).$$

So, by triangle inequality and Proposition 2.6.6(ii), we get

$$h(T) = \Big\| \sup_{x \in T} |\langle g, x \rangle| \Big\|_{L^2} \lesssim \Big\| \sup_{x \in T} |\langle g, x \rangle| \Big\|_{\psi_2} \lesssim \gamma(T) + r(T) \lesssim \gamma(T).$$

The last step used the fact that $\gamma(T) \gtrsim r(T)$, which comes from the second part of (b) – just take the supremum over $y \in T$. □

The Gaussian width helps us define a robust version of the notion of dimension. The usual linear-algebraic dimension of a set $T \subset \mathbb{R}^n$, which is the dimension of the smallest affine space containing it, can change a lot with tiny perturbations of T. Here is a more robust alternative:

Definition 7.5.12 (Effective dimension)**.** The *effective dimension* of a bounded set $T \subset \mathbb{R}^n$ is

$$d(T) := \frac{h(T-T)^2}{\operatorname{diam}(T)^2} \asymp \frac{w(T)^2}{\operatorname{diam}(T)^2}.$$

The equivalence follows from Lemma 7.5.11. The effective dimension is bounded by the linear-algebraic one:

$$d(T) \le \dim(T),$$

with equality when T is a Euclidean ball in some subspace (see Exercise 7.21). Unlike the usual dimension, the effective one is stable – small perturbations to T only slightly change its width and diameter.

To get some practice with effective dimension, compute it for ellipsoids (Exercise 7.23) and bound it for general finite sets (Exercise 7.22).

[8] Here, as usual, the notation $\asymp$ hides positive asbsolute constant factors.

7.6 Application: random projections of sets

What happens if we project a set $T \subset \mathbb{R}^n$ onto a random m-dimensional subspace in $\mathbb{R}^n$ (picked uniformly from the Grassmanian $G_{n,m}$), like in Figure 5.2? In practice, we might see T as data and P as a way of dimension reduction, like in Johnson-Lindenstrauss lemma. What can we say about the size (diameter) of the projected set PT?

For a finite set T, the Johnson-Lindenstrauss Lemma (Theorem 5.3.1) says that if $m \gtrsim \log|T|$, the random projection P acts essentially as a scaling of T: it shrinks all distances between points in T by a factor $\approx \sqrt{m/n}$. In particular,

$$\operatorname{diam}(PT) \approx \sqrt{\frac{m}{n}} \operatorname{diam}(T). \tag{7.20}$$

But if the cardinality of T is too large or infinite, (7.20) may fail. For instance, if $T = B_2^n$ is a Euclidean ball, no projection can shrink its size:

$$\operatorname{diam}(PT) = \operatorname{diam}(T). \tag{7.21}$$

What about general sets T? The next result states that a random projection shrinks T as in (7.20), but it cannot shrink it beyond the spherical width $w_s(T)$:

Theorem 7.6.1 (Sizes of random projections of sets). *Let $T \subset \mathbb{R}^n$ be a bounded set, and P be the orthogonal projection in $\mathbb{R}^n$ onto a random m-dimensional subspace $E \sim \operatorname{Unif}(G_{n,m})$. Then*

$$\mathbb{E} \operatorname{diam}(PT) \asymp w_s(T) + \sqrt{\frac{m}{n}} \operatorname{diam}(T),$$

where the notation $\asymp$ hides positive absolute constant factors.

Proof Let's prove the upper bound and leave the lower bound to you (see Exercise 7.26).

Step 1: Change the model. Let's switch the view just like in the proof of Proposition 5.3.2. A random subspace $E \subset \mathbb{R}^n$ can be obtained by randomly rotating some fixed subspace, such as $\mathbb{R}^m$. But instead of fixing T and randomly rotating $\mathbb{R}^m$, we can fix $E = \mathbb{R}^m$ and randomly rotate T. A random rotation of a vector $x \in T$ is Ux where $U \sim \operatorname{Unif}(O(n))$ is a random orthogonal matrix. Projecting Ux onto $E = \mathbb{R}^m$ means keeping the first m coordinates, i.e. Qx where Q is the $m \times n$ matrix consisting of the first m columns of U. So, we can work with Q instead of P.

Step 2: Approximation. Without loss of generality, $\operatorname{diam}(T) \le 1$. (Why?) We need to bound

$$\operatorname{diam}(QT) = \sup_{x \in T-T} \|Qx\|_2 = \sup_{x \in T-T} \max_{z \in S^{m-1}} \langle Qx, z \rangle.$$

We will proceed with an ε-net argument as in the proof of Theorem 4.4.3. Choose an $(1/2)$-net $\mathcal{N}$ of the sphere S^{m-1} so that

$$|\mathcal{N}| \le 5^m$$

using Corollary 4.2.11. We can replace the supremum over the sphere S^{m-1} by the supremum over the net $\mathcal{N}$ paying a factor 2:

$$\operatorname{diam}(QT) \le 2 \sup_{x\in T-T} \max_{z\in\mathcal{N}} \langle Qx, z\rangle = 2 \max_{z\in\mathcal{N}} \sup_{x\in T-T} \langle Q^\mathsf{T} z, x\rangle. \tag{7.22}$$

(see Exercise 4.35). Now, here is the plan: we will first bound

$$\sup_{x\in T-T} \langle Q^\mathsf{T} z, x\rangle \tag{7.23}$$

for a fixed $z \in \mathcal{N}$, and then take union bound over all z.

Step 3: Concentration. So, let's fix $z \in \mathcal{N}$. By construction, $Q^\mathsf{T} z$ is uniformly distributed on the sphere: $Q^\mathsf{T} z \sim \operatorname{Unif}(S^{n-1})$ (check this carefully in Exercise 7.24). The expectation of (7.23) can be expressed as the spherical width:

$$\mathbb{E} \sup_{x\in T-T} \langle Q^\mathsf{T} z, x\rangle = w_s(T-T) = 2w_s(T).$$

(The last equality is just a spherical version of Proposition 7.5.2(e).)

To check that (7.23) concentrates around its mean, we use the concentration inequality on the sphere (Theorem 5.1.3, or more specifically (5.30)). Since $\operatorname{diam}(T) \le 1$ by assumption, the function $z \mapsto \sup_{x\in T-T} \langle z, x\rangle$ on the sphere has Lipschitz norm at most 1 (check!). So (5.30) gives

$$\mathbb{P}\Big\{ \sup_{x\in T-T} \langle Q^\mathsf{T} z, x\rangle \ge 2w_s(T) + t \Big\} \le 2\exp(-cnt^2).$$

Step 4: Union bound. Now we unfix $z \in \mathcal{N}$ by taking the union bound:

$$\mathbb{P}\Big\{ \max_{z\in\mathcal{N}} \sup_{x\in T-T} \langle Q^\mathsf{T} z, x\rangle \ge 2w_s(T) + t \Big\} \le |\mathcal{N}| \cdot 2\exp(-cnt^2). \tag{7.24}$$

Recall that $|\mathcal{N}| \le 5^m$. Choosing $t = Cs\sqrt{m/n}$ with onstant C large enough, the probability in (7.24) is bounded by $2e^{-ms^2}$ for any $s \ge 1$. So, (7.24) and (7.22) give the result:

$$\mathbb{P}\Big\{ \frac{1}{2}\operatorname{diam}(QT) \ge 2w_s(T) + Cs\sqrt{\frac{m}{n}} \Big\} \le e^{-ms^2} \quad \text{for any } s \ge 1.$$

From this, one can bound the expected value of $\operatorname{diam}(QT)$ using the integrated tail formula (Lemma 1.6.1) – do this! □

Remark 7.6.2 (Phase transition)**.** Let's get more insight from Theorem 7.6.1. Since the sum of two terms is equivalent to maximum (up to factor 2), we can write:

$$\operatorname{diam}(PT) \asymp \max\Big[w_s(T),\ \sqrt{\frac{m}{n}} \operatorname{diam}(T) \Big].$$

Let's find the "phase transition" point where these two terms are equal. Set them equal and solve for m:

$$m = \frac{(\sqrt{n}\, w_s(T))^2}{\operatorname{diam}(T)^2} \asymp \frac{w(T)^2}{\operatorname{diam}(T)^2} \asymp d(T),$$

using Lemma 7.5.5 and the definition of effective dimension $d(T)$ (see Definition 7.5.12). So the takeaway is:

$$\operatorname{diam}(PT) \asymp \begin{cases} \sqrt{\frac{m}{n}}\operatorname{diam}(T), & \text{if } m \geq d(T) \\ w_s(T), & \text{if } m \leq d(T), \end{cases}$$

(see Figure 7.5). That is, as we decrease the dimension m of the random projection, initially it shrinks T roughly by $\sqrt{m/n}$ like in Johnson-Lindenstrauss (7.20). But once m dips below the effective dimension $d(T)$, the shrinking stops and the diameter stays near the spherical width $w_s(T)$ as in (7.21). This is because $\operatorname{conv}(PT)$ looks like a ball of radius $w_s(T)$, as we will see in Section 9.7.2.

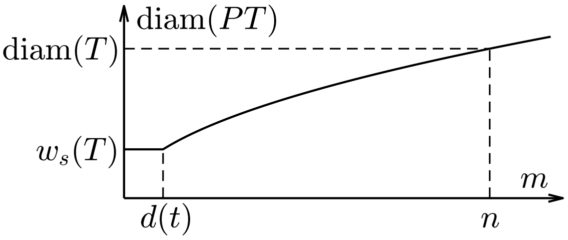


Figure 7.5 The diameter of a random m-dimensional projection of a set T as a function of m.

With more tools, we will improve on Theorem 7.6.1 by dropping the constant in front of $\sqrt{m/n}$ in Section 9.2.2.

7.7 Notes

Slepian inequality (Theorem 7.2.2) is originally due to D. Slepian [304, 305]; modern proofs can be found e.g. in [210, Corollary 3.12], [11, Section 2.2], [330, Section 6.1], [169], [180].

Sudakov-Fernique inequality (Theorem 7.2.8) is attributed to V. N. Sudakov [309, 310] and X. Fernique [124]. Our presentation of the proofs of Slepian and Sudakov-Fernique inequalities in Section 7.2 is based on the approach of J.-P. Kahane [180] and a smoothing argument of S. Chatterjee (see [11, Section 2.2]), and it follows [330, Section 6.1].

The relevance of comparison inequalities to random matrix theory was noticed by S. Szarek. The applications in Section 7.3 can derived from the work of Y. Gordon [140]. Our presentation there follows the argument in [94, Section II.c], reproduced also in [340, Section 5.3.1].

Sudakov inequality (Theorem 7.4.1) was originally proved by V. N. Sudakov. Our presentation follows [210, Theorem 3.18]; see [21, Section 4.2] for an alternative proof via duality.

Gaussian width, introduced in Section 7.5, originates in geometric functional analysis and asymptotic convex geometry [21, 246]. Starting from [290], the role of Gaussian width was recognized in applications to signal processing and high-dimensional statistics. Milman's "hyperbolic sketch" of a high-dimensional convex body discussed in Remark 7.5.10 is from [241]; for more on this, see the preface of [21] and [23].

The effective dimension introduced in Section 7.5.3 has some variants in the literature, including the statistical dimension for convex cones [228, 18, 267].

Theorem 7.6.1 on the diameters of random projections of sets is due to to V. Milman [245], see also [21, Proposition 5.7.1].

Talagrand contraction principle (Exercise 7.3) can be found in [210, Corollary 3.17], where one can find a more general result (with a convex and increasing function of the supremum).

Exercise 7.3 is adapted from [330, Exercise 7.4]. A more general version of Gaussian contraction inequality in Exercise 7.8 can be found in [210, Corollary 3.17].

Gordon inequality (Theorem 7.2.9) and its extensions can be found in [140, 141, 144, 180].

Exercise 7.11 on the norms of symmetric Gaussian matrices and Exercise 7.13(a) on the smallest singular values of Gaussian matrices are from [94, Section II.c].

Exercise 7.18 introduces the nuclear norm – one of the Schatten norms, which are the ℓ^p norms of the matrix's singular values. Special cases include the nuclear norm ($p = 1$), Frobenius norm ($p = 2$), and operator norm ($p = \infty$). The duality extends naturally to all Schatten norms, and can be deduced von Neumann trace inequality.

Exercises

7.1 ☕☕ (Covariance vs. increments) Consider a random process $(X_t)_{t\in T}$.

(a) Assuming that the zero random variable 0 belongs to the process, express the covariance function $\Sigma(t,s) = \mathbb{E}\, X_t X_s$ in terms of the metric $d(t,s) = \|X_t - X_s\|_{L^2}$.

(b) Do the same assuming that, the process contains the random variable $-X_t$ whenever it contains X_t.

(c) Show by example that without any assumption, it may be impossible to recover the covariance function from the metric.

7.2 ☕☕☕ (Symmetrization for random processes) Let $X_1(t), \dots, X_N(t)$ be N independent, mean-zero random processes indexed by points $t \in T$. Let $\varepsilon_1, \dots, \varepsilon_N$ be independent Rademacher random variables. Prove that

$$\frac{1}{2}\,\mathbb{E}\sup_{t\in T}\Big|\sum_{i=1}^N \varepsilon_i X_i(t)\Big| \le \mathbb{E}\sup_{t\in T}\Big|\sum_{i=1}^N X_i(t)\Big| \le 2\,\mathbb{E}\sup_{t\in T}\Big|\sum_{i=1}^N \varepsilon_i X_i(t)\Big|.$$

7.3 ☕☕☕ (Talagrand contraction principle) In this exercise, we get a first look at random processes, which we will dive into fully in the next chapter. Let $T \subset \mathbb{R}^n$ be a bounded set. Let $\phi_i : \mathbb{R} \to \mathbb{R}$ be contractions, i.e. Lipschitz functions with $\|\phi_i\|_{\mathrm{Lip}} \le 1$. Let $\varepsilon_1, \dots, \varepsilon_n$ be independent Rademacher random variables. Show that

$$\mathbb{E}\sup_{t\in T}\sum_{i=1}^n \varepsilon_i \phi_i(t_i) \le \mathbb{E}\sup_{t\in T}\sum_{i=1}^n \varepsilon_i t_i. \tag{7.25}$$

by going through the following steps:

(a) First let $n = 2$. Consider a subset $T \subset \mathbb{R}^2$ and contraction $\phi : \mathbb{R} \to \mathbb{R}$, and check that

$$\sup_{t\in T}(t_1 + \phi(t_2)) + \sup_{t\in T}(t_1 - \phi(t_2)) \le \sup_{t\in T}(t_1 + t_2) + \sup_{t\in T}(t_1 - t_2).$$

(b) Use induction on n to complete proof.

7.4 ☕ (General Talagrand contraction principle) Generalize Talagrand contraction principle (Exercise 7.3) for arbitrary Lipschitz functions $\phi_i : \mathbb{R} \to \mathbb{R}$ without restriction on their Lipschitz norms.

7.5 ☕☕ (Expressing a random walk as a canonical process) Express 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 some $T \subset \mathbb{R}^N$.

7.6 ☕☕☕ (Multivariate Gaussian integration by parts) Prove Lemma 7.2.4.

7.7 ☕☕☕ (Differentiating the expected log-partition function) In the proof of Sudakov-Fernique Theorem 7.2.8, one step left for you was to check that for $Z(u) = \sqrt{u}\,X + \sqrt{1-u}\,Y$, the log-partition function $f(x) = \frac{1}{\beta}\log\sum_{i=1}^n e^{\beta x_i}$ satisfies

$$\frac{d}{du}\,\mathbb{E}\, f(Z(u)) \le 0 \quad \text{for all } u \in [0,1].$$

Prove this by following these steps:

(a) Differentiate f and check that

$$\frac{\partial f}{\partial x_i} = \frac{e^{\beta x_i}}{\sum_k e^{\beta x_k}} =: p_i(x) \quad \text{and} \quad \frac{\partial^2 f}{\partial x_i \partial x_j} = \beta\left(\delta_{ij} p_i(x) - p_i(x) p_j(x)\right),$$

where δ_{ij} is the Kronecker delta, which equals 1 is $i = j$ and 0 otherwise.

(b) Use the Gaussian interpolation formula 7.2.5. Simplify the expression and deduce that

$$\frac{d}{du}\,\mathbb{E}\, f(Z(u)) = \frac{\beta}{4}\sum_{i \ne j}\Big[\mathbb{E}(X_i - X_j)^2 - \mathbb{E}(Y_i - Y_j)^2\Big]\,\mathbb{E}\, p_i(Z(u))\, p_j(Z(u)).$$

(c) Note that this expression is non-positive by assumption.

7.8 ☕☕ (Gaussian contraction inequality) The following is a Gaussian version of Talagrand contraction principle from Exercise 7.3. Consider a bounded subset $T \subset \mathbb{R}^n$, and let $g_1, \ldots, g_n$ be independent $N(0,1)$ random variables. Let $\phi_i : \mathbb{R} \to \mathbb{R}$ be contractions, i.e. Lipschitz functions with $\|\phi_i\|_{\mathrm{Lip}} \le 1$. Use Sudakov-Fernique inequality to prove that

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

7.9 ☕☕☕ (Gordon inequality) Prove Gordon inequality (Theorem 7.2.9) under the additional equal variances assumption:

$$\mathbb{E}\, X_{ut}^2 = \mathbb{E}\, Y_{ut}^2 \quad \text{for all } u, t.$$

Like with the Sudakov-Fernique inequality, the equal variances assumption can be dropped from Gordon inequality too. We wont prove this here.

7.10 ☕☕☕ (Frobenius distance between rank-one matrices) This bound was used in the proof of Theorem 7.3.1. For unit vectors $u, w \in \mathbb{R}^{n-1}$ and unit vectors $v, z \in \mathbb{R}^{m-1}$, show that

$$\|uv^{\mathsf{T}} - wz^{\mathsf{T}}\|_F^2 \le \|u - w\|_2^2 + \|v - z\|_2^2.$$

7.11 ☕☕☕ (Norms of GOE random matrices) Adapt the arguments in Section 7.3 to the *symmetric* case. Consider an $n \times n$ Gaussian random matrix A with independent normal entries on and above the diagonal, where the diagonal entries are $N(0,2)$ and the off-diagonal entries are $N(0,1)$. (This is the *Gaussian orthogonal ensemble* from Exercise 3.19.)

(a) Show the expectation bound:

$$\mathbb{E}\|A\| \le 2\sqrt{n}.$$

(b) Boost this to a high-probability bound: for any $t \ge 0$,

$$\mathbb{P}\{\|A\| \ge 2\sqrt{n}+t\} \le 2\exp(-ct^2).$$

7.12 ☕☕☕ (The norm of a Gaussian vector) For $g \sim N(0, I_n)$, show that the function

$$f(n) = \mathbb{E}\|g\|_2 - \sqrt{n}$$

is increasing on $[C,\infty)$, where C is a suitable absolute constant.

7.13 ☕☕ (Smallest singular values of Gaussian matrices) Let's use the Gordon inequality (Theorem 7.2.9) to get sharp bounds on the smallest singular value of an $m \times n$ random matrix A with independent $N(0,1)$ entries. For $m \ge n \ge C$ (where C is a suitable absolute constant), show that:

(a) (Expectation)

$$\mathbb{E}\, s_n(A) \ge \sqrt{m} - \sqrt{n};$$

(b) (High probability)

$$\mathbb{P}\big\{s_n(A) \le \sqrt{m} - \sqrt{n} - t\big\} \le 2\exp(-ct^2) \quad \text{for any } t \ge 0.$$

7.14 ☕☕ (Sudakov inequality for non-compact sets) Show that if (T,d) is not relatively compact, i.e. $\mathcal{N}(T,d,\varepsilon) = \infty$ for some $\varepsilon > 0$, then $\mathbb{E}\sup_{t\in T} X_t = \infty$.

7.15 ☕☕ (Properties of Gaussian width) Prove Properties (a)–(d) and (g) in Proposition 7.5.2.

7.16 ☕☕ (Gaussian width and diameter) Show by example that, for any dimension n, both bounds in Proposition 7.5.2(f) are attained up to absolute constant factors.

7.17 ☕ (Gaussian width of the ℓ^p ball) Let $1 \le p \le \infty$. Consider the unit ball of the ℓ^p norm in $\mathbb{R}^n$, i.e. $B_p^n = \{x \in \mathbb{R}^n : \|x\|_p \le 1\}$. Prove that

$$w(B_p^n) \asymp \begin{cases} \sqrt{p'}\, n^{1/p'}, & p' \le \log n \\ \sqrt{\log n}, & p' \ge \log n \end{cases}$$

where p' is the conjugate exponent for p, as in Hölder inequality (1.22).

7.18 ☕☕☕ (Nuclear norm) For an $n \times n$ matrix A, the operator norm $\|A\|$ is the ℓ^∞ norm of the singular values, and the Frobenius norm $\|A\|_F$ is the ℓ^2 norm (see Lemma 4.1.11). Define the *nuclear norm* as the ℓ^1 norm of the singular values:

$$\|A\|_* := \sum_{i=1}^n s_i(A).$$

(a) Just like ℓ^1 and ℓ^∞ norms are dual for vectors (see Exercise 1.19(b)), prove that the nuclear and operator norms are dual for matrices:[9]

$$\|A\|_* = \max\left\{\langle A, B\rangle :\ \|B\| \le 1\right\}.$$

(b) Conclude that the nuclear norm is indeed a norm.

7.19 ☕☕☕ (Gaussian width of matrices with bounded operator norm) Show that the unit ball of the operator norm for $n \times n$ matrices,

$$T := \{B :\ \|B\| \le 1\}, \quad \text{satisfies}^{10} \quad w(T) \asymp n^{3/2}.$$

7.20 ☕☕☕ (Gaussian width vs. Gaussian complexity) For any bounded set $T \subset \mathbb{R}^n$ and a point $y \in T$, show that

$$\gamma(T) \asymp w(T) + \|y\|_2$$

This is one of the statements in Proposition 7.5.11, which was left unproved.

7.21 ☕☕ (Effective dimension vs. algebraic dimension) We introduced the effective dimension of a bounded set $T \subset \mathbb{R}^n$ in Definition 7.5.12.

(a) Prove that the effective dimension is bounded by the linear-algebraic dimension: $d(T) \le \dim(T)$.

(b) Show that the equality holds when T is a Euclidean ball in a subspace of $\mathbb{R}^n$.

7.22 ☕ (Effective dimension of a finite set) For any finite set T, show that

$$d(T) \le C \log|T|.$$

7.23 ☕☕ (Ellipsoids) An ellipsoid is just a linear transformation of the unit ball B_2^n via some $m \times n$ matrix A, i.e. the set $A(B_2^n)$.

(a) (Gaussian width) Show that

$$w\left(A(B_2^n)\right) \asymp \|A\|_F$$

where the notation $\asymp$ hides positive absolute constant factors.

(b) (Effective dimension) Show that

$$d\big(A(B_2^n)\big) = r\big(A^\mathsf{T} A\big) = s(A)$$

where $d(\cdot)$ is the effective dimension, $r(\cdot)$ is the effective rank and $s(\cdot)$ is the stable rank (see Definition 7.5.12 and Remarks 5.6.3, 5.6.4).

7.24 ☕☕ (Modeling a random vector on the sphere) Let z be a fixed vector on the unit sphere S^{m-1}, and let B be an $n \times m$ matrix consisting of the first m columns of a random orthogonal matrix $U \sim \mathrm{Unif}(O(n))$. Check that

$$Bz \sim \mathrm{Unif}(S^{n-1}).$$

[9] Recall that $\langle A, B\rangle = \mathrm{tr}(A^\mathsf{T} B)$ by definition (4.7).

7.25 ☕☕ (Gaussian projections of sets) Prove a version of Theorem 7.6.1 for an $m \times n$ Gaussian random matrix G with independent $N(0,1)$ entries. Specifically, show that for any bounded set $T \subset \mathbb{R}^n$, we have

$$\mathbb{E}\,\mathrm{diam}(GT) \asymp w(T) + \sqrt{m}\,\mathrm{diam}(T).$$

Here $w(T)$ is the Gaussian width of T.

7.26 ☕☕☕ (Random projections of sets: a lower bound) Show the lower bound in Theorem 7.6.1: for any bounded set $T \subset \mathbb{R}^n$,

$$\mathbb{E}\,\mathrm{diam}(PT) \geq c\Big[w_s(T) + \sqrt{\frac{m}{n}}\,\mathrm{diam}(T)\Big].$$

7.27 ☕☕ (Matrix sketching) Suppose we have a large A be an $n \times k$ matrix. Here is how we can reduce it without sacrificing accuracy of the operator norm.

(a) Let P be a projection in $\mathbb{R}^n$ onto a random m-dimensional subspace (picked uniformly in the Grassmanian $G_{n,m}$). Prove that

$$\mathbb{E}\|PA\| \asymp \frac{1}{\sqrt{n}}\,\|A\|_F + \sqrt{\frac{m}{n}}\,\|A\|.$$

(b) Let G be an $m \times n$ random matrix with i.i.d. $N(0,1)$ entries. Prove that

$$\mathbb{E}\|GA\| \asymp \|A_F\| + \sqrt{m}\,\|A\|.$$

8

Chaining

This chapter covers some of the central methods for bounding random processes $(X_t)_{t\in T}$. In Section 8.1, we introduce the *chaining* technique and use it to bound $\mathbb{E}\sup_{t\in T} X_t$ in terms of covering numbers of T, a result known as Dudley inequality. We illustrate it in Section 8.2 with an application to Monte Carlo integration and *empirical processes*.

Then in Section 8.3 we introduce the *VC theory*, which gives combinatorial rather than metric insights into random processes. We apply it for *statistical learning theory* in Section 8.4.

In Section 8.5, we refine the chaining method to *generic chaining* and get optimal, two-sided bounds on random processes, in terms of Talagrand's $\gamma_2(T)$ functional (we only prove an the upper bound here). A useful consequence is Talagrand comparison inequality, which generalizes Sudakov-Fernique inequality for subgaussian processes.

Finally, in Section 8.6, we use Talagrand comparison inequality to derive *Chevet inequality*, a handy tool for random bilinear forms over general sets.

With everything you've learned, some exercises now feel like fun mini research results – Lipschitz law of large numbers in higher dimensions (Exercise 8.10), one-bit quantization (Exercise 8.26), the small ball method with applications to heavy-tailed random matrices (Exercise 8.27), and $p \to q$ norms of random matrices (Exercise 8.41), just to mention a few!

8.1 Dudley inequality

For a general Gaussian process $(X_t)_{t\in T}$, Sudakov inequality (Theorem 7.4.1) gives a *lower* bound on

$$\mathbb{E}\sup_{t\in T} X_t$$

in terms of the metric entropy of T. Now we will go for an *upper* bound. And instead of sticking just to Gaussian processes, we will take it further and handle more general, subgaussian ones.

[0] This material will be published by Cambridge University Press as *High-Dimensional Probability, 2nd Edition* by Roman Vershynin. This pre-publication version is free to view and download for personal use only. Not for re-distribution, re-sale or use in derivative works. ©Roman Vershynin 202X.
If you want to be notified when the printed version is available, let me know: rvershyn@uci.edu

Definition 8.1.1 (Subgaussian increments)**.** A random process $(X_t)_{t\in T}$ on a metric space (T,d) has *subgaussian increments* if there exists $K \ge 0$ such that

$$\|X_t - X_s\|_{\psi_2} \le K d(t,s) \quad \text{for all } t,s \in T. \tag{8.1}$$

Example 8.1.2 (Gaussian processes)**.** Let $(X_t)_{t\in T}$ be a Gaussian process on some set T. It naturally defines a *canonical metric* on T:

$$d(t,s) := \|X_t - X_s\|_{L^2}, \quad t,s \in T,$$

as we explained in (7.1). With respect to this metric, $(X_t)_{t\in T}$ clearly had subgaussian increments, with some absolute constant K.

Here is another (trivial) example: any random process can be made to have subgaussian increments by defining the metric as $d(t,s) := \|X_t - X_s\|_{\psi_2}$.

Now we give a bound on a general subgaussian random process $(X_t)_{t\in T}$ in terms of the metric entropy $\log \mathcal{N}(T,d,\varepsilon)$:

Theorem 8.1.3 (Dudley integral inequality)**.** *Let $(X_t)_{t\in T}$ be a mean-zero random process on a metric space (T,d) with subgaussian 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 inequality, let's compare it with Sudakov inequality (Theorem 7.4.1), which for Gaussian processes says:

$$\mathbb{E}\sup_{t\in T} X_t \ge c \sup_{\varepsilon>0} \varepsilon \sqrt{\log \mathcal{N}(T,d,\varepsilon)}.$$

Figure 8.1 illustrates both bounds. There is a clear gap between them, and it turns out that metric entropy alone cannot close it – we will explore this later.

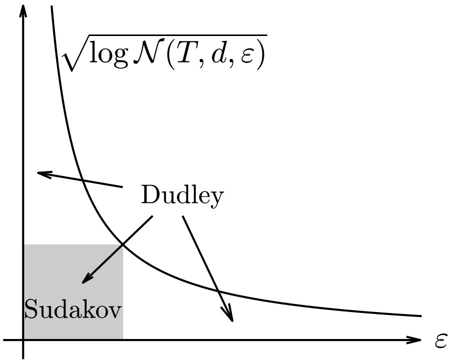


Figure 8.1 Dudley inequality bounds $\mathbb{E}\sup_{t\in T} X_t$ by the area under the curve. Sudakov inequality bounds it below by the largest area of a rectangle under the curve, up to constants.

The Dudley inequality hints that $\mathbb{E}\sup_{t\in T} X_t$ is a *multiscale* quantity – to bound it, we need to look at T across all scales ε. And that is exactly how the proof works. We will first prove a discrete version using dyadic scales $\varepsilon = 2^{-k}$ (like a Riemann sum), then move to the continuous version.

Theorem 8.1.4 (Discrete Dudley inequality)**.** *Let $(X_t)_{t \in T}$ be a mean-zero random process on a metric space (T, d) with subgaussian increments as in (8.1). Then*

$$\mathbb{E} \sup_{t \in T} X_t \leq CK \sum_{k \in \mathbb{Z}} 2^{-k} \sqrt{\log \mathcal{N}(T, d, 2^{-k})}. \tag{8.2}$$

The proof uses the important technique of *chaining*. It is basicaly a multiscale version of the ε-net argument we have used before, like in the proofs of Theorems 4.4.3 and 7.6.1.

In the ε-net argument, we approximate T by an ε-net $\mathcal{N}$ so every point $t \in T$ is close to some $\pi(t) \in \mathcal{N}$, with $d(t, \pi(t)) \leq \varepsilon$. Then the increment condition (8.1) gives

$$\|X_t - X_{\pi(t)}\|_{\psi_2} \leq K\varepsilon. \tag{8.3}$$

This leads to

$$\mathbb{E} \sup_{t \in T} X_t \leq \mathbb{E} \sup_{t \in T} X_{\pi(t)} + \mathbb{E} \sup_{t \in T} (X_t - X_{\pi(t)}).$$

We can handle the first term via a union bound over $|\mathcal{N}| = \mathcal{N}(T, d, \varepsilon)$ points $\pi(t)$. But the second term is tricky – it is not clear how to combine (8.3) with a union bound over all $t \in T$. To fix this, we don't stop at one net, but we choose smaller and smaller ε to get better approximations $\pi_1(t), \pi_2(t), \ldots$ to t with finer nets. That is the idea behind *chaining*, which we will now formalize.

Proof of Theorem 8.1.4. **Step 1: Chaining set-up.** Without loss of generality, we may assume that $K = 1$ (why?) and T is finite (see Remark 7.2.1). Define the dyadic scale

$$\varepsilon_k = 2^{-k}, \quad k \in \mathbb{Z} \tag{8.4}$$

and choose ε_k-nets T_k of T so that

$$|T_k| = \mathcal{N}(T, d, \varepsilon_k). \tag{8.5}$$

Only a part of the dyadic scale will be needed. Since T is finite, there exists a small enough number $\kappa \in \mathbb{Z}$ (defining the coarsest net) and a large enough number $K \in \mathbb{Z}$ (defining the finest net), such that

$$T_\kappa = \{t_0\} \text{ for some } t_0 \in T, \quad T_K = T. \tag{8.6}$$

For a point $t \in T$, let $\pi_k(t)$ denote a closest point in T_k, so we have

$$d(t, \pi_k(t)) \leq \varepsilon_k. \tag{8.7}$$

Since $\mathbb{E}\, X_{t_0} = 0$ by assumpiton, we have

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

Let's write $X_t - X_{t_0}$ as a telescoping sum, walking from t_0 to t along a chain of points $\pi_k(t)$ that mark progressively finer approximations to t:

$$X_t - X_{t_0} = (X_{\pi_\kappa(t)} - X_{t_0}) + (X_{\pi_{\kappa+1}(t)} - X_{\pi_\kappa(t)}) + \cdots + (X_t - X_{\pi_K(t)}), \tag{8.8}$$

see Figure 8.2. The first and last terms of this sum are zero by (8.6), so we have

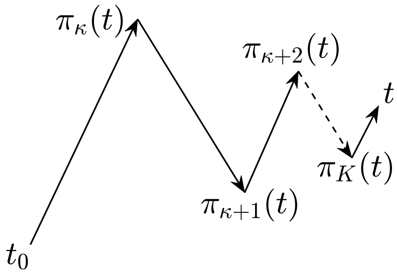


Figure 8.2 Chaining: a walk from a fixed point t_0 to an arbitrary point t in T along elements $\pi_k(T)$ of progressively finer nets of T

$$X_t - X_{t_0} = \sum_{k=\kappa+1}^{K} (X_{\pi_k(t)} - X_{\pi_{k-1}(t)}). \tag{8.9}$$

Since the supremum of the sum is bounded by the sum of suprema, we get

$$\mathbb{E} \sup_{t\in T} (X_t - X_{t_0}) \le \sum_{k=\kappa+1}^{K} \mathbb{E} \sup_{t\in T} (X_{\pi_k(t)} - X_{\pi_{k-1}(t)}). \tag{8.10}$$

Step 2: Controlling the increments. In (8.10), it looks like we are taking the supremum over all of T in each summand, but really it is over the smaller set of pairs $(\pi_k(t), \pi_{k-1}(t))$.The number of such pairs is

$$|T_k| \cdot |T_{k-1}| \le |T_k|^2,$$

a number that we can control through (8.5). Moreover, for a fixed t, we can bound the increments in (8.10) like this:

$$\begin{aligned}
\|X_{\pi_k(t)} - X_{\pi_{k-1}(t)}\|_{\psi_2} &\le d(\pi_k(t),\, \pi_{k-1}(t)) \quad \text{(by (8.1) and since } K = 1\text{)} \\
&\le d(\pi_k(t), t) + d(t,\, \pi_{k-1}(t)) \quad \text{(by triangle inequality)} \\
&\le \varepsilon_k + \varepsilon_{k-1} \quad \text{(by (8.7))} \\
&\le 2\varepsilon_{k-1}.
\end{aligned}$$

Recall from (2.22) that the expected maximum of N subgaussian random variables is at most $CL\sqrt{\log N}$, where L is the largest ψ_2 norm. We can use this to bound each term in (8.10):

$$\mathbb{E} \sup_{t\in T} (X_{\pi_k(t)} - X_{\pi_{k-1}(t)}) \le C\varepsilon_{k-1}\sqrt{\log |T_k|}. \tag{8.11}$$

Step 3: Summing up the increments. We have shown that

$$\mathbb{E}\sup_{t\in T}(X_t - X_{t_0}) \le C \sum_{k=\kappa+1}^{K} \varepsilon_{k-1}\sqrt{\log|T_k|}. \tag{8.12}$$

Now plug in the values $\varepsilon_k = 2^{-k}$ from (8.4) and the bounds (8.5) on $|T_k|$, and get

$$\mathbb{E}\sup_{t\in T}(X_t - X_{t_0}) \le C_1 \sum_{k=\kappa+1}^{K} 2^{-k}\sqrt{\log\mathcal{N}(T,d,2^{-k})}.$$

Theorem 8.1.4 is proved. □

Let us now deduce the integral form of Dudley inequality.

Proof of Dudley integral inequality, Theorem 8.1.3. 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

$$2\sum_{k\in\mathbb{Z}}\int_{2^{-k-1}}^{2^{-k}}\sqrt{\log\mathcal{N}(T,d,\varepsilon)}\,d\varepsilon = 2\int_0^\infty\sqrt{\log\mathcal{N}(T,d,\varepsilon)}\,d\varepsilon. \quad \square$$

To practice, show that the discrete and integral Dudley inequalities are actually equivalent (Exercise 8.3).

8.1.1 Variations and Examples

Remark 8.1.5 (Dudley inequality: supremum of increments)**.** A quick look at the proof shows that chaining actually gives

$$\mathbb{E}\sup_{t\in T}|X_t - X_{t_0}| \le CK\int_0^\infty\sqrt{\log\mathcal{N}(T,d,\varepsilon)}\,d\varepsilon \tag{8.13}$$

for any fixed $t_0 \in T$. We can combine with the same bound for $X_s - X_{t_0}$ and use triangle inequality to get:

$$\mathbb{E}\sup_{t,s\in T}|X_t - X_s| \le CK\int_0^\infty\sqrt{\log\mathcal{N}(T,d,\varepsilon)}\,d\varepsilon. \tag{8.14}$$

Remark 8.1.6 (Dudley inequality: a high-probability bound)**.** Dudley inequality gives only an expectation bound, but chaining actually gives a high-probability bound. Assuming T is finite,[1] for every $u \ge 0$, the bound

$$\sup_{t,s\in T}|X_t - X_s| \le CK\Big[\int_0^\infty\sqrt{\log\mathcal{N}(T,d,\varepsilon)}\,d\varepsilon + u\cdot\operatorname{diam}(T)\Big] \tag{8.15}$$

[1] As always, are are assuming T is finite to avoid measurability issues; the general case typically follows by approximation.

holds with probability at least $1 - 2\exp(-u^2)$ (Exercise 8.1). For Gaussian processes, this also follows directly from Gaussian concentration (Exercise 8.2).

Note that (8.14) and (8.15) do not need the mean zero assumption $\mathbb{E}\, X_t = 0$, but Dudley Theorem 8.1.3 does – otherwise it can fail. (Can you see why?)

Remark 8.1.7 (Limits of Dudley integral)**.** Even though the Dudley integral goes over $[0, \infty]$, we can cap it at the diameter of T, since for $\varepsilon > \operatorname{diam}(T)$ a single ε-ball covers T and so $\log \mathcal{N}(T, d, \varepsilon) = 0$. Thus:

$$\mathbb{E} \sup_{t \in T} X_t \le CK \int_0^{\operatorname{diam}(T)} \sqrt{\log \mathcal{N}(T, d, \varepsilon)}\, d\varepsilon. \tag{8.16}$$

If we apply Dudley inequality for the canonical Gaussian process $\langle g, t\rangle$, just like we did with Sudakov inequality in Corollary 7.4.2, we get:

Theorem 8.1.8 (Dudley inequality in $\mathbb{R}^n$)**.** *The Gaussian width of any bounded set $T \subset \mathbb{R}^n$ satisfies*

$$w(T) \le C \int_0^\infty \sqrt{\log \mathcal{N}(T, \varepsilon)}\, d\varepsilon. \tag{8.17}$$

Here $\mathcal{N}(T, \varepsilon)$ is the smallest number of Euclidean balls with radius ε and centers in T that cover T.

Example 8.1.9 (Dudley is sharp for the Euclidean ball)**.** Let's test Dudley inequality for the unit Euclidean ball $T = B_2^n$. Recall from (4.18) that $\mathcal{N}(B_2^n, \varepsilon) \le (3/\varepsilon)^n$ for $\varepsilon \in (0, 1]$, and of course $\mathcal{N}(B_2^n, \varepsilon) = 1$ for $\varepsilon > 1$. Then

$$w(B_2^n) \lesssim \int_0^1 \sqrt{n \log(3/\varepsilon)}\, d\varepsilon \lesssim \sqrt{n}.$$

This is optimal: as we know from (7.17), $w(B_2^n) \asymp \sqrt{n}$.

Remark 8.1.10 (Dudley can be loose – but not too loose)**.** In general, Dudley integral can overestimate the Gaussian width. Here is an example:

$$T = \Big\{ \frac{e_k}{\sqrt{1 + \log k}},\ k = 1, \dots, n \Big\}$$

with e_k the standard basis vectors in $\mathbb{R}^n$. Try Exercise 8.4 to show that

$$w(T) = O(1) \quad \text{while} \quad \int_0^\infty \sqrt{\log \mathcal{N}(T, d, \varepsilon)}\, d\varepsilon \to \infty$$

as $n \to \infty$. However, the good news:

(a) Dudley inequality is tight up to a logarithmic factor (Exercise 8.5);
(b) we will upgrade chaining to remove that logarithmic factor (Section 8.5).

For more practice, show the log-sharpness of Dudley and Sudakov (Exercise 8.5), refine the lower limit in Dudley's integral (Exercise 8.6), and prove subexponential (Exercise 8.7) and local (Exercise 8.8) versions of Dudley.

8.2 Application: empirical processes

Let's apply Dudley inequality to *empirical processes* – certain natural random processes indexed by functions. Here is a motivating example.

8.2.1 The Monte Carlo method

Suppose we want to compute an integral

$$\int_\Omega f \, d\mu$$

where $f : \Omega \to \mathbb{R}$ is a given function on some set Ω and μ is a probability measure on Ω. For instance, this could just be $\int_0^1 f(x)\,dx$ for a function $f : [0,1] \to \mathbb{R}$ (see Figure 8.3a).

We can do this *probabilistically*. Suppose X is a random point in Ω drawn according to μ, i.e. $\mathbb{P}\{X \in A\} = \mu(A)$ for any measurable set $A \subset \Omega$. (For instance, $X \sim \mathrm{Unif}[0,1]$ if you are integrating over $[0,1]$.) Then the integral becomes the expectation:

$$\int_\Omega f \, d\mu = \mathbb{E}\, f(X).$$

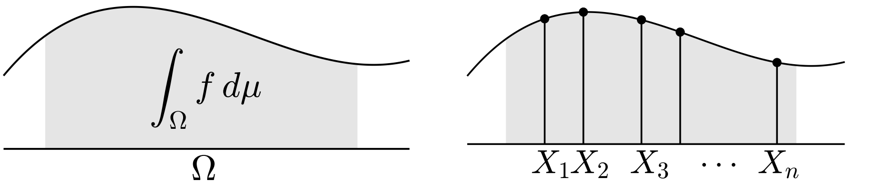
Now take i.i.d. samples $X_1, X_2, \ldots$ By the law of large numbers (Theorem 1.7.1),

$$\frac{1}{n} \sum_{i=1}^n f(X_i) \to \mathbb{E}\, f(X) \quad \text{almost surely} \tag{8.18}$$

as $n \to \infty$. So, we can approximate the integral with the arithmetic mean:

$$\int_\Omega f \, d\mu \approx \frac{1}{n} \sum_{i=1}^n f(X_i) \tag{8.19}$$

(see Figure 8.3b). This is the *Monte Carlo method* – compute the integral by averaging function values at random sample points.



(a) The problem is to compute the integral of f on a domain Ω.

(b) The integral is approximated by $\frac{1}{n}\sum_1^n f(X_i)$ with i.i.d. random points X_i.

Figure 8.3 Monte Carlo method for numerical integration.

Remark 8.2.1 (Error rate)**.** The expected error of the Monte Carlo estimate

(8.19) is $O(1/\sqrt{n})$. This comes from the rate of convergence in the law of large numbers (1.23):

$$\mathbb{E}\Big|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}\, f(X)\Big| \le \Big[\operatorname{Var}\Big(\frac{1}{n}\sum_{i=1}^{n} f(X_i)\Big)\Big]^{1/2} = O\Big(\frac{1}{\sqrt{n}}\Big). \tag{8.20}$$

Remark 8.2.2 (Monte Carlo is high-dimensional, agnostic)**.** Monte Carlo works well in high dimensions since the error (8.20) does not depend on dimension– unlike grid-based integration methods. You don't even need to know the measure μ; just being able to sample from it is enough. Same with f – you only need its values at a few random points.

8.2.2 Lipschitz law of large numbers

Can we use the same sample $X_1, \ldots, X_n$ to estimate the integral of *any* function $f : \Omega \to \mathbb{R}$? No. A badly chosen f could wiggle wildly between sample points (like in Figure 8.4), making the Monte Carlo estimate (8.19) totally off.

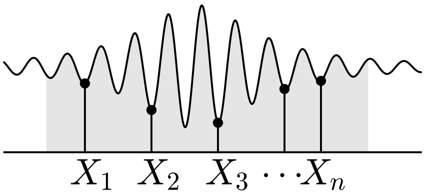


Figure 8.4 One sample $X_1, \ldots, X_n$ cannot be used to approximate the integral of *all* functions f.

But what if we stick to functions that do not wiggle too much – like Lipschitz ones? Then yes!

Theorem 8.2.3 (Lipschitz law of large numbers)**.** *Consider the class of functions*

$$\mathcal{F} := \{f : [0,1] \to \mathbb{R},\ \|f\|_{\text{Lip}} \le L\}\,, \tag{8.21}$$

where L is any number. Let $X, X_1, X_2, \ldots, X_n$ be i.i.d. random variables taking values in $[0,1]$. Then

$$\mathbb{E}\sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}\, f(X)\Big| \le \frac{CL}{\sqrt{n}}. \tag{8.22}$$

Remark 8.2.4 (One sample serves all Lipschitz functions)**.** Before the proof, let's reiterate the key point: the supremum over $f \in \mathcal{F}$ is *inside* the expectation. Thanks to Markov inequality, this means that a single sample $X_1, \ldots, X_n$ will, with high probability, work well simultaneously for all $f \in \mathcal{F}$. And "work well" means approximating each integral with error $O(1/\sqrt{n})$ – same rate as the usual law of large numbers for just one function. So, made the law of large numbers uniform without losing anything!

To make the proof of Theorem 8.2.3 more intuitive, it is helpful to view the left side of (8.22) as the maximal magnitude of a random process indexed by functions $f \in \mathcal{F}$. Such random processes are called *empirical processes*:

Definition 8.2.5 (Empirical process)**.** Let $\mathcal{F}$ be a class of real-valued functions $f : \Omega \to \mathbb{R}$ on some set Ω. Let X be a random point in Ω picked according to some probability distribution, 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

$$X_f := \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E}\, f(X) \tag{8.23}$$

is called an *empirical process* indexed by $\mathcal{F}$.

Proof of Theorem 8.2.3 Without loss of generality, it is enough to prove the theorem for the class

$$\mathcal{F} := \{f : [0,1] \to [0,1],\ \|f\|_{\mathrm{Lip}} \le 1\}. \tag{8.24}$$

(Why?) We would like to bound

$$\mathbb{E} \sup_{f \in \mathcal{F}} |X_f|$$

for the empirical process $(X_f)_{f \in \mathcal{F}}$ defined in (8.23).

Step 1: Checking subgaussian increments. Let's use Dudley inequality (Theorem 8.1.3). To apply it, we will check that the empirical process has subgaussian increments with respect to the L^∞ metric $d(f,g) = \|f - g\|_{L^\infty}$. So, fix a pair of functions $f, g \in \mathcal{F}$ and write

$$\|X_f - X_g\|_{\psi_2} = \frac{1}{n} \Big\| \sum_{i=1}^{n} Z_i \Big\|_{\psi_2} \quad \text{where} \quad Z_i := (f-g)(X_i) - \mathbb{E}(f-g)(X).$$

Since Z_i are independent, mean-zero random variables, Proposition 2.7.1 gives

$$\|X_f - X_g\|_{\psi_2} \lesssim \frac{1}{n} \Big(\sum_{i=1}^{n} \|Z_i\|_{\psi_2}^2 \Big)^{1/2}.$$

Now, using centering (Lemma 2.7.8) we have

$$\|Z_i\|_{\psi_2} \lesssim \|(f-g)(X_i)\|_{\psi_2} \lesssim \|f - g\|_{L^\infty}.$$

It follows that

$$\|X_f - X_g\|_{\psi_2} \lesssim \frac{1}{n} \cdot n^{1/2} \|f - g\|_{L^\infty} = \frac{1}{\sqrt{n}} \|f - g\|_{L^\infty}.$$

Step 2: Applying Dudley inequality. Now apply Dudley inequality (Theorem 8.1.3) in the form (8.13):

$$\mathbb{E} \sup_{f \in \mathcal{F}} |X_f| = \mathbb{E} \sup_{f \in \mathcal{F}} |X_f - X_0| \lesssim \frac{1}{\sqrt{n}} \int_0^1 \sqrt{\log \mathcal{N}(\mathcal{F}, \|\cdot\|_{L^\infty}, \varepsilon)}\, d\varepsilon. \tag{8.25}$$

(Here we used that the zero function belongs to $\mathcal{F}$, and the diameter of $\mathcal{F}$ in the

L^∞ metric is bounded by 1 by (8.24).) It is not difficult to bound the covering numbers of $\mathcal{F}$ like this:

$$\mathcal{N}(\mathcal{F}, \|\cdot\|_{L^\infty}, \varepsilon) \leq e^{C/\varepsilon},$$

as you will check in Exercise 8.9. Substitute this bound into the integral:

$$\mathbb{E} \sup_{f \in \mathcal{F}} |X_f| \lesssim \frac{1}{\sqrt{n}} \int_0^1 \sqrt{\frac{C}{\varepsilon}}\, d\varepsilon \lesssim \frac{1}{\sqrt{n}}.$$

Theorem 8.2.3 is proved. □

To get some practice, try Exercise 8.10, where you extend Theorem 8.2.3 to higher dimensions. It is marked with four coffee cups, but do not worry – it is more labor-intensive than tricky.

8.2.3 Empirical measure

For a broader perspective, take one more look at the Definition 8.2.5 of an empirical process. Given an i.i.d. sample $X_1, \ldots, X_n$ picked from Ω according to some probability measure μ, let's consider the *empirical measure* μ_n that assigns equal probabilities $1/n$ to each point, counting multiplicities:

$$\mu_n = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}. \tag{8.26}$$

Here δ_x is the Dirac probability measure at x, i.e. $\delta_x(A) = 1$ if $x \in A$ and 0 otherwise, for any set A. So, $\mu_n(A)$ is the fraction of sample points that fall in a set $A \subset \Omega$.

The integral of f with respect to the original measure μ is $\mathbb{E} f(X)$ (the "population" average of f), and the integral of f with respect to the empirical measure μ_n is $\frac{1}{n}\sum_{i=1}^n f(X_i)$ (the "sample", or empirical, average of f). The empirical process X_f in (8.23) tracks the deviation of the population expectation from the empirical expectation.

This deviation, which we bound in Theorem 8.2.3, can be thought as a distance between measures μ and μ_n, called the *Wasserstein distance* $W_1(\mu, \mu_n)$. It has an equivalent interpretation as the *transportation cost* of turning one measure into the other. The equivalence is provided by the Kantorovich-Rubinstein duality theorem. For this reason Theorem 8.2.3 is often called the *Wasserstein law of large numbers*.

Many tools you have learned also work for empirical processes. Try proving an empirical version of symmetrization in Exercise 8.11.

8.3 VC dimension

Now we introduce the notion of the VC (Vapnik–Chervonenkis) dimension, which plays a major role in statistical learning theory. We connect it to covering numbers, and then, through Dudley inequality, to random processes and uniform law of large numbers. Applications to statistical learning theory will be given in next section.

8.3.1 Definition and examples

The VC dimension measures how complex a class of Boolean functions is. Here, a Boolean function is a map $f : \Omega \to \{0,1\}$ on some set Ω, and we are looking at some collection $\mathcal{F}$ of these.

Definition 8.3.1 (VC dimension)**.** A subset $\Lambda \subseteq \Omega$ is *shattered* by a class of Boolean functions $\mathcal{F}$ if, for any possible binary labeling $g : \Lambda \to \{0,1\}$, there is some function $f \in \mathcal{F}$ that matches[2] it on Λ. The *VC dimension* of $\mathcal{F}$, denoted $\mathrm{vc}(\mathcal{F})$, is the largest cardinality of a subset $\Lambda \subseteq \Omega$ that is shattered. (And if there is no largest one, we say $\mathrm{vc}(\mathcal{F}) = \infty$.)

VC dimension can take some time to sink in, so let's walk through a few examples to make it clearer.

Example 8.3.2 (Intervals)**.** Let $\mathcal{F}$ consist of the indicators of all closed intervals in $\mathbb{R}$:

$$\mathcal{F} := \Big\{ \mathbf{1}_{[a,b]} :\ a, b \in \mathbb{R},\ a \le b \Big\}.$$

We claim that

$$\mathrm{vc}(\mathcal{F}) = 2.$$

To prove $\mathrm{vc}(\mathcal{F}) \ge 2$, we need to find a two-point set $\Lambda \subset \mathbb{R}$ that is shattered by $\mathcal{F}$. Take, for exampe, $\Lambda = 3, 5$. There are four possible binary labelings $g : \Lambda \to \{0,1\}$ on this set, and each one can be obtained by restricting some interval indicator $f = \mathbf{1}_{[a,b]}$ onto Λ. For instance, $g(3) = 1$, $g(5) = 0$ comes from $f = \mathbf{1}_{[2,4]}$. The other three cases are similar (see Figure 8.5), so Λ is indeed shattered by $\mathcal{F}$.

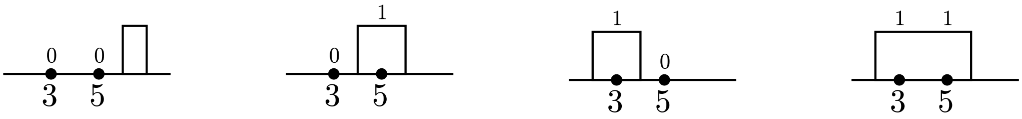


Figure 8.5 The binary function $g(3) = g(5) = 0$ is the restriction of $\mathbf{1}_{[6,7]}$ onto $\Lambda = \{3,5\}$ (left). The function $g(3) = 0$, $g(5) = 1$ is the restriction of $\mathbf{1}_{[4,6]}$ onto Λ (middle left). The function $g(3) = 1$, $g(5) = 0$ is the restriction of $\mathbf{1}_{[2,4]}$ onto Λ (middle right). The function $g(3) = g(5) = 1$ is the restriction of $\mathbf{1}_{[2,6]}$ onto Λ (right).

To prove $\mathrm{vc}(\mathcal{F}) < 3$, we need to show that no three-point set $\Lambda = \{p, q, r\}$ can be shattered by $\mathcal{F}$. To see this, assume $p < q < r$ and consider the labeling $g(p) = 1$, $g(q) = 0$, $g(r) = 1$. Then g can not be a restriction of any indicator $\mathbf{1}_{[a,b]}$ onto Λ, for otherwise $[a, b]$ must contain two points p and r but not the point q that lies between them, which is impossible.

[2] Formally, this means that the restriction of f onto Λ is g, i.e. $f(x) = g(x)$ for all $x \in \Lambda$.

Example 8.3.3 (Half-planes)**.** Let $\mathcal{F}$ consist of the indicators of all closed half-planes in $\mathbb{R}^2$. We claim that

$$\text{vc}(\mathcal{F}) = 3.$$

To prove $\text{vc}(\mathcal{F}) \geq 3$, we need to find a three-point set $\Lambda \subset \mathbb{R}^2$ that is shattered by $\mathcal{F}$. Let Λ consist of three points in general position like in Figure 8.6. Each of the $2^3 = 8$ binary labelings $g : \Lambda \to \{0, 1\}$ is a restriction of the indicator function of some half-plane. To see this, arrange the half-plane to contain exactly those points of Λ where g takes value 1. Thus, Λ is shattered.

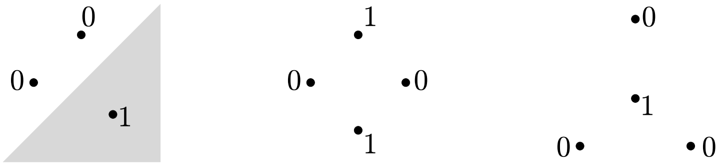


Figure 8.6 The proof that VC(half-planes)= 3 in Example 8.3.3 consists of two steps. To show VC ≥ 3, we find a three-point set Λ on which every binary labeling g is linearly separable (left). To show VC < 4, we demonstrate that on any four-point set Λ there exists a binary labeling g that is not linearly separable (middle and right).

To prove $\text{vc}(\mathcal{F}) < 4$, we need to show that no four-point set can be shattered by $\mathcal{F}$. There are two possible configurations of four-point sets Λ in general position, shown in Figure 8.6. (What if Λ is not in general position? Analyze this case.) In each of the two cases, there exists a binary labeling such that no half-plane can contain exactly the points labeled 1; see Figure 8.6. This means that always exists a binary labeling $g : \Lambda \to \{0, 1\}$ that is not a restriction of any indicator of a half-plane, and thus Λ is not shattered.

Example 8.3.4. Let $\Omega = \{1, 2, 3\}$. We can conveniently represent Boolean functions on Ω as binary strings of length three. Consider the class

$$\mathcal{F} := \{001, 010, 100, 111\}.$$

The set $\Lambda = \{1, 3\}$ is shattered by $\mathcal{F}$. Indeed, restricting the functions in $\mathcal{F}$ onto Λ amounts to dropping the second digit, thus producing strings $00, 01, 10, 11$. Thus, the restriction produces all possible binary strings of length two, or equivalently, all possible binary labelings $g : \Lambda \to \{0, 1\}$. Hence Λ is shattered by $\mathcal{F}$, and thus $\text{vc}(\mathcal{F}) \geq |\Lambda| = 2$. On the other hand, the (only) three-point set $\{1, 2, 3\}$ is not shattered by $\mathcal{F}$, as this would require all eight binary digits of length three to appear in $\mathcal{F}$, which is not true.

For more practice, find the VC dimension of pairs of intervals, circles, rectangles, squares, and polygons (Exercises 8.12–8.16), and also try Exercise 8.17 to extend Example 8.3.3 to higher dimensions like this:

Example 8.3.5 (Half-spaces)**.** A half-space in $\mathbb{R}^n$ is a set of the form

$$\{x : \langle a, x\rangle \le b\} \quad \text{where } a \in \mathbb{R}^n \text{ and } b \in \mathbb{R}.$$

Let $\mathcal{F}$ be the class of indicators of all half-spaces in $\mathbb{R}^n$. Then

$$\mathrm{vc}(\mathcal{F}) = n + 1.$$

Remark 8.3.6 (VC dimension vs. parameter count)**.** VC dimension of a function class often roughly matches the number of parameters – for instance, half-spaces in $\mathbb{R}^n$ are defined with $n+1$ parameters, which matches their VC dimension (see Exercise 8.3.5). This is not a hard rule but rather a useful heuristic.

8.3.2 Pajor Lemma

Suppose the domain Ω is finite and consists of n points. Then any class of Boolean functions $\mathcal{F}$ on Ω is also finite, and

$$2^{\mathrm{vc}(\mathcal{F})} \le |\mathcal{F}| \le 2^n. \tag{8.27}$$

(Why?) The upper bound is usually loose – most function classes are closer in size to the lower bound. This is not so obvious. To prepare for this result, let's first show that there are as many shattered subsets of Ω as the functions in $\mathcal{F}$.

Lemma 8.3.7 (Pajor lemma)**.** *Let $\mathcal{F}$ be a class of Boolean functions on a finite set Ω. Then*

$$|\mathcal{F}| \le \big|\{\Lambda \subseteq \Omega : \Lambda \textit{ is shattered by } \mathcal{F}\}\big|.$$

We include the empty set $\Lambda = \emptyset$ in the count on the right side.

Before the proof, let's illustrate this result using Example 8.3.4. Here, $|\mathcal{F}| = 4$ and there are six subsets Λ that are shattered by $\mathcal{F}$, namely $\{1\}$, $\{2\}$, $\{3\}$, $\{1,2\}$, $\{1,3\}$ and $\{2,3\}$. (Check!) Thus the inequality in Pajor lemma reads $4 \le 6$.

Proof of Lemma 8.3.7. We proceed by induction on the cardinality of Ω. The case $|\Omega| = 1$ is trivial, since we include the empty set in the counting. Assume the lemma holds for any n-point set Ω, and let us prove it for Ω with $|\Omega| = n+1$.

Chopping out one (arbitrary) point from the set Ω, we can express it as

$$\Omega = \Omega_0 \cup \{x_0\}, \quad \text{where} \quad |\Omega_0| = n.$$

The class $\mathcal{F}$ then naturally breaks into two sub-classes

$$\mathcal{F}_0 := \{f \in \mathcal{F} : f(x_0) = 0\} \quad \text{and} \quad \mathcal{F}_1 := \{f \in \mathcal{F} : f(x_0) = 1\}.$$

By the induction hypothesis, the counting function

$$S(\mathcal{F}) = \big|\{\Lambda \subseteq \Omega : \Lambda \text{ is shattered by } \mathcal{F}\}\big|$$

satisfies[3]

$$S(\mathcal{F}_0) \ge |\mathcal{F}_0| \quad \text{and} \quad S(\mathcal{F}_1) \ge |\mathcal{F}_1|. \tag{8.28}$$

[3] To properly use the induction hypothesis here, restrict the functions in $\mathcal{F}_0$ and $\mathcal{F}_1$ onto the n-point set Ω_0.

To complete the proof, all we need to check is

$$S(\mathcal{F}) \geq S(\mathcal{F}_0) + S(\mathcal{F}_1), \tag{8.29}$$

for then (8.28) would give $S(\mathcal{F}) \geq |\mathcal{F}_0| + |\mathcal{F}_1| = |\mathcal{F}|$, as needed.

Inequality (8.29) may seem trivial. Any set Λ that is shattered by $\mathcal{F}_0$ or $\mathcal{F}_1$ is automatically shattered by the larger class $\mathcal{F}$, and thus each set Λ counted by $S(\mathcal{F}_0)$ or $S(\mathcal{F}_1)$ is automatically counted by $S(\mathcal{F})$. The problem, however, lies in the double counting. Assume the same set Λ is shattered by *both* $\mathcal{F}_0$ and $\mathcal{F}_1$. The counting function $S(\mathcal{F})$ will not count Λ twice. However, a different set will be counted by $S(\mathcal{F})$, which was not counted by either $S(\mathcal{F}_0)$ or $S(\mathcal{F}_1)$ – namely, $\Lambda \cup \{x_0\}$. A moment thought reveals that this set is indeed shattered by $\mathcal{F}$. (Check!) This establishes inequality (8.29) and completes the proof of Pajor Lemma. □

Let's illustrate the key point in the proof of Pajor lemma:

Example 8.3.8. Let us again go back to Example 8.3.4. Following the proof of Pajor lemma, we chop out $x_0 = 3$ from $\Omega = \{1, 2, 3\}$, making $\Omega_0 = \{1, 2\}$. The class $\mathcal{F} = \{001, 010, 100, 111\}$ then breaks into two sub-classes

$$\mathcal{F}_0 = \{010, 100\} \quad \text{and} \quad \mathcal{F}_1 = \{001, 111\}.$$

There are exactly two subsets Λ shattered by $\mathcal{F}_0$, namely $\{1\}$ and $\{2\}$, and *the same* subsets are shattered by $\mathcal{F}_1$, making $S(\mathcal{F}_0) = S(\mathcal{F}_1) = 2$. Of course, the same two subsets are also shattered by $\mathcal{F}$, but we need two more shattered subsets to make $S(\mathcal{F}) \geq 4$ for the key inequality (8.29). Here is how we construct them: append $x_0 = 3$ to the already counted subsets Λ. The resulting sets $\{1, 3\}$ and $\{2, 3\}$ are also shattered by $\mathcal{F}$, and we have not counted them yet. Now have at least *four* subsets shattered by $\mathcal{F}$, making the key inequality (8.29) true.

8.3.3 Sauer-Shelah Lemma

We now deduce a remarkable upper bound on the cardinality of a function class in terms of VC dimension:

Lemma 8.3.9 (Sauer-Shelah Lemma)**.** *Let $\mathcal{F}$ be a class of Boolean functions on an n-point set Ω. Then*

$$|\mathcal{F}| \leq \sum_{k=0}^{d} \binom{n}{k} \leq \left(\frac{en}{d}\right)^d \quad \textit{where} \quad d = \mathrm{vc}(\mathcal{F}).$$

Proof Pajor Lemma states that $|\mathcal{F}|$ is bounded by the number of subsets $\Lambda \subseteq \Omega$ shattered by $\mathcal{F}$. The cardinality of each such set Λ is bounded by $d = \mathrm{vc}(\mathcal{F})$, according to the definition of VC dimension. Thus

$$|\mathcal{F}| \leq \big|\{\Lambda \subseteq \Omega : \ |\Lambda| \leq d\}\big| = \sum_{k=0}^{d} \binom{n}{k}$$

since the sum in right hand side gives the total number of subsets of an n-element set with cardinalities at most k. This proves the first inequality of Sauer-Shelah lemma. The second inequality follows from the bound on the binomial sum we proved in Exercise 0.6. □

Both Pajor and Sauer-Shelah lemma are generally sharp (try Exercise 8.19).

8.3.4 Growth function

Sauer-Shelah lemma assumes that the domain Ω is finite. What about function classes $\mathcal{F}$ on infinite domains like $\Omega = \mathbb{R}^n$? It is often convenient to measure the complexity of $\mathcal{F}$ by the growth function:

Definition 8.3.10 (Growth function)**.** Let $\mathcal{F}$ be a class of Boolean functions on a domain Ω. The growth function of $\mathcal{F}$ is defined as the maximum number of functions that can be obtained by restricting all functions in $\mathcal{F}$ to a subset of n elements:

$$\Pi_{\mathcal{F}}(n) = \sup \left\{ \left| \mathcal{F}|_{\Lambda} \right| : \; \Lambda \subset \Omega, \, |\Lambda| = n \right\}.$$

In this light, the VC dimension of $\mathcal{F}$ can be seen as the largest d for which $\Pi_{\mathcal{F}}(d) = 2^d$. Immediate bounds on the growth function are

$$2^d \le \Pi_{\mathcal{F}}(n) \le \left(\frac{en}{d} \right)^d \quad \text{for all } n \ge d \tag{8.30}$$

if $d = \text{vc}(\mathcal{F}) < \infty$. The lower bound is a restatement of (8.27) and the upper bound follows from Sauer-Shelah lemma (Lemma 8.3.9).

To see how the growth function can be useful, let's deduce from (8.30) the stability of VC dimension with respect to natural operations.

Proposition 8.3.11 (VC stability)**.** *Let $\mathcal{F}$ and $\mathcal{G}$ be two classes of Boolean functions on the same domain. Let*

$$\mathcal{F} \wedge \mathcal{G} = \{ f \wedge g : \; f \in \mathcal{F}, \, g \in \mathcal{G} \}$$

where $f \wedge g$ denotes the pointwise minimum of the functions f and g. Then

$$\text{vc}(\mathcal{F} \wedge \mathcal{G}) \le 10 \max \left(\text{vc}(\mathcal{F}), \, \text{vc}(\mathcal{G}) \right).$$

The same holds for the pointwise maximum $\vee$.

Proof Assume for contradiction that $n := \text{vc}(\mathcal{F} \wedge \mathcal{G}) > 10d$. Then

$$2^n \le \Pi_{\mathcal{F} \wedge \mathcal{G}}(n) \le \Pi_{\mathcal{F}}(n) \cdot \Pi_{\mathcal{G}}(n) \le \left(\frac{en}{d} \right)^{2d}.$$

(The first and last bounds follow from (8.30), and the middle bound should be clear by definition.) However, a simple calculation shows that $2^n > (en/d)^{2d}$ whenever $n > 10d$. Contradiction. □

Proposition 8.3.11 can be extended to any particular way of combining classes of functions – try Exercise 8.21. It can be helpful when we want to bound the VC dimension without computing it directly, such as in this example:

Example 8.3.12 (Strips)**.** A strip in $\mathbb{R}^n$ is a set of the form

$$\{x : |\langle a, x\rangle - b| \le c\} \quad \text{where } a \in \mathbb{R}^n \text{ and } b, c \in \mathbb{R},$$

see Figure 8.7. Let $\mathcal{F}$ be the class of indicators of strips. Example 8.3.5 gives

$$\operatorname{vc}(\mathcal{F}) \le 20(n+1) \le 40n.$$

Indeed, each strip can be represented as the intersection of two half-spaces $\{x : \langle a, x\rangle - b \le c\}$ and $\{x : \langle a, x\rangle - b \ge -c\}$. Thus the indicator of each strip is the pointwise minimum of the indicators of two half-spaces. Now apply the VC stability property (Proposition 8.3.11) and the result in Example 8.3.5.

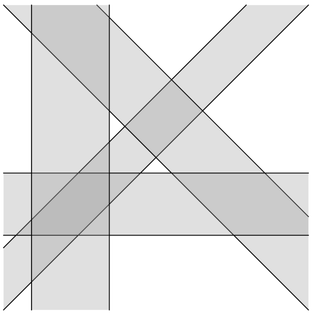


Figure 8.7 Four different strips in $\mathbb{R}^2$

For more practice with growth function, show a mind-blowing dichotomy – this function can grow either polynomially or exponentially (Exercise 8.20) and compute the VC dimension of the union (Exercise 8.22).

8.3.5 Covering numbers via VC dimension

Covering numbers usually grow exponentially with dimension (see (4.17) for example). Now, let's refine this heuristic by replacing the algebraic dimension with VC dimension – which can save us a lot.

Let $\mathcal{F}$ be a class of Boolean functions on some domain Ω, and let μ be any probability measure on Ω. Define distance between functions as

$$d(f, g) = \|f - g\|_{L^2(\mu)} = \left(\mathbb{E}(f-g)(X)^2\right)^{1/2}, \tag{8.31}$$

where X is a random variable with distribution μ. (If you have not done measure theory, just take an arbitrary random variable X taking values in Ω, and think of μ as the name for the distribution of X.)

Now let's bound the covering numbers $\mathcal{N}(\mathcal{F}, L^2(\mu), \varepsilon)$ of the class $\mathcal{F}$ with respect to the metric (8.31):

Theorem 8.3.13 (Covering numbers via VC dimension)**.** *Let $\mathcal{F}$ be a class of Boolean functions on a domain Ω with a probability measure μ on it. Then, for every $\varepsilon \in (0,1)$,*

$$\mathcal{N}(\mathcal{F}, L^2(\mu), \varepsilon) \le \Big(\frac{2}{\varepsilon}\Big)^{Cd} \quad \textit{where} \quad d = \operatorname{vc}(\mathcal{F}).$$

For a first attempt at proving Theorem 8.3.13, let's assume for a moment that Ω is finite, say $|\Omega| = n$. Then Sauer-Shelah Lemma 8.3.9 gives

$$\mathcal{N}(\mathcal{F}, L^2(\mu), \varepsilon) \le |\mathcal{F}| \le \Big(\frac{en}{d}\Big)^d.$$

This is not quite Theorem 8.3.13, but it comes close. To tighten the bound, we need to get rid of n, and we will do this by shrinking Ω. This lemma will help:

Lemma 8.3.14 (Dimension reduction)**.** *Let $\mathcal{F}$ be a finite class of Boolean functions on a domain Ω with a probability measure μ on it. Assume that all functions in $\mathcal{F}$ are ε-separated, that is*

$$\|f-g\|_{L^2(\mu)} > \varepsilon \quad \textit{for all distinct } f, g \in \mathcal{F}.$$

If $n \ge C\varepsilon^{-4}\log|F|$, then the empirical measure μ_n satisfies the following with probability at least 0.99*:*

$$\|f-g\|_{L^2(\mu_n)} > \varepsilon/2 \quad \textit{for all distinct } f, g \in \mathcal{F}.$$

By definition of empirical measure (see Section 8.2.3), $\|f-g\|_{L^2(\mu_n)}$ is the same as (8.31) but with the population average replaced by the sample average:

$$\|f-g\|_{L^2(\mu_n)} = \Big(\frac{1}{n}\sum_{i=1}^n (f-g)(X_i)^2\Big)^{1/2}, \tag{8.32}$$

where X_i are i.i.d. copies of X.

Proof of Lemma 8.3.14 Notice some similarity to another dimension reduction result – the Johnson–Lindenstrauss lemma (Theorem 5.3.1)? The proof is along the same lines – concentration plus a union bound.

Fix a pair of distinct functions $f, g \in \mathcal{F}$, and consider

$$\|f-g\|^2_{L^2(\mu_n)} - \|f-g\|^2_{L^2(\mu)} = \frac{1}{n}\sum_{i=1}^n h(X_i) - \mathbb{E}\, h(X),$$

where $h = (f-g)^2$. On the right, we have a sum of independent bounded (and thus subgaussian) random variables, so Hoeffding inequality (Theorem 2.7.3) gives

$$\mathbb{P}\Big\{ \Big| \|f-g\|^2_{L^2(\mu_n)} - \|f-g\|^2_{L^2(\mu)} \Big| > \frac{\varepsilon^2}{4} \Big\} \le 2\exp(-cn\varepsilon^4).$$

(Check!) Therefore, with probability at least $1 - 2\exp(-cn\varepsilon^4)$, we have

$$\|f-g\|^2_{L^2(\mu_n)} \ge \|f-g\|^2_{L^2(\mu)} - \frac{\varepsilon^2}{4} > \varepsilon^2 - \frac{\varepsilon^2}{4} > \frac{\varepsilon^2}{4}, \tag{8.33}$$

by lemma's assumption. Now take a union bound over all pairs of distinct functions $f, g \in \mathcal{F}$. There are at most $|\mathcal{F}|^2$ of them, so with probability at least

$$1 - |\mathcal{F}|^2 \cdot 2\exp(-cn\varepsilon^4), \tag{8.34}$$

the bound (8.33) holds simultaneously for all distinct $f, g \in \mathcal{F}$. Since $n \geq C\varepsilon^{-4}\log|\mathcal{F}|$ by assumption, by choosing the absolute constant C sufficiently big we can make the quantity in (8.34) at least 0.99. □

Proof of Theorem 8.3.13 By the packing-covering equiavelence (Lemma 4.2.8), we can find

$$N = \mathcal{N}(\mathcal{F}, L^2(\mu), \varepsilon)$$

functions in $\mathcal{F}$ that are ε-separated in the $L^2(\mu)$ metric. Set $n = \lceil C\varepsilon^{-4}\log N \rceil$ and apply Lemma 8.3.14 to the set of those functions. With positive probability, those functions stay $(\varepsilon/2)$-separated in the metric $L^2(\mu_n)$ defined in (8.32), so their restrictions onto $\Omega_n = \{X_1, \dots, X_n\}$ are all different.

Fix a realization of random variables $X_1, \dots, X_n$ for which this event holds. So there exists a subset $\Omega_n \subset \Omega$ with $|\Omega_n| \leq n \leq 2C\varepsilon^{-4}\log N$, such that the class $\mathcal{F}_n = \mathcal{F}|_{\Omega_n}$ obtained by restricting all functions onto Ω_n satisfies $|\mathcal{F}_n| \geq N$. Now apply Sauer-Shelah Lemma 8.3.9 for $\mathcal{F}_n$ and Ω_n to get

$$N \leq \Big(\frac{en}{d_n}\Big)^{d_n} \leq \Big(\frac{2C\varepsilon^{-4}\log N}{d_n}\Big)^{d_n}$$

where $d_n = \mathrm{vc}(\mathcal{F}_n)$. Simplifying,[4] we get

$$N \leq (2C\varepsilon^{-4})^{2d_n}.$$

Finally, replace $d_n = \mathrm{vc}(\mathcal{F}_n)$ by the larger quantity $d = \mathrm{vc}(\mathcal{F})$. □

8.3.6 VC law of large numbers

Back in Section 8.2.2, we bounded a general empirical process over all Lipschitz functions. Now, let's replace Lipschitz by any class of Boolean functions with finite VC dimension:

Theorem 8.3.15 (VC law of large numbers). *Let $\mathcal{F}$ be a class of Boolean functions with finite VC dimension on some domain Ω, and let $X, X_1, X_2, \dots, X_n$ be independent random points in Ω with common distribution. Then*

$$\mathbb{E}\sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}\, f(X)\Big| \leq C\sqrt{\frac{\mathrm{vc}(\mathcal{F})}{n}}.$$

[4] To do this, note that $\frac{\log N}{2d_n} = \log(N^{1/2d_n}) \leq N^{1/2d_n}$.

Proof We will combine Dudley inequality with the bound on the covering numbers from Theorem 8.3.13. But first, let's symmetrize the process using the empirical version of symmetrization (Exercise 8.11):

$$\mathbb{E} \sup_{f \in \mathcal{F}} \Big| \frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbb{E} f(X) \Big| \le \frac{2}{\sqrt{n}} \mathbb{E} \sup_{f \in \mathcal{F}} \underbrace{\Big| \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i f(X_i) \Big|}_{Z_f}.$$

Condition on (X_i), leaving all randomness in the random signs (ε_i). To use Dudley inequality for the process $(Z_f)_{f \in \mathcal{F}}$, we need to check that the increments are subgaussian. Triangle inequality gives

$$|Z_f - Z_g| \le \frac{1}{\sqrt{n}} \Big| \sum_{i=1}^n \varepsilon_i (f-g)(X_i) \Big|,$$

so using Proposition 2.7.1 and the obvious fact that $\|\varepsilon_i\|_{\psi_2} \lesssim 1$, we get:[5]

$$\|Z_f - Z_g\|_{\psi_2} \le \frac{1}{\sqrt{n}} \Big\| \sum_{i=1}^n \varepsilon_i (f-g)(X_i) \Big\|_{\psi_2} \lesssim \Big(\frac{1}{n} \sum_{i=1}^n (f-g)(X_i)^2 \Big)^{1/2} = \|f-g\|_{L^2(\mu_n)}$$

where μ_n is the empirical measure – recall (8.32).

Now use Dudley inequality[6] (Theorem 8.1.3) conditionally on (X_i), then remove the conditioning by taking expectation with respect to (X_i). We get

$$\frac{2}{\sqrt{n}} \mathbb{E} \sup_{f \in \mathcal{F}} Z_f \lesssim \frac{1}{\sqrt{n}} \mathbb{E} \int_0^1 \sqrt{\log \mathcal{N}(\mathcal{F}, L^2(\mu_n), \varepsilon)}\, d\varepsilon. \tag{8.35}$$

Finally, we use Theorem 8.3.13 to bound the covering numbers:

$$\log \mathcal{N}(\mathcal{F}, L^2(\mu_n), \varepsilon) \lesssim \mathrm{vc}(\mathcal{F}) \log(2/\varepsilon).$$

Substituting this into (8.35), we get the integral of $\sqrt{\log(2/\varepsilon)}$, which is bounded by an absolute constant, leading to

$$\frac{2}{\sqrt{n}} \mathbb{E} \sup_{f \in \mathcal{F}} Z_f \lesssim \sqrt{\frac{\mathrm{vc}(\mathcal{F})}{n}},$$

as claimed. □

Remark 8.3.16 (Rademacher complexity)**.** If $\mathcal{F}$ is a class of Boolean functions with finite VC dimension, then the expression

$$\mathbb{E} \sup_{f \in \mathcal{F}} \Big| \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(x_i) \Big|$$

is called the *Rademacher complexity* of $\mathcal{F}$ on a given set of points $x_1, x_2, \dots, x_n \in \Omega$. Rademacher complexity reflects how rich $\mathcal{F}$ is. In proving Theorem 8.3.15, a key step was relating it to another measure of richness – the VC dimension: we

[5] Keep in mind that here X_i and thus $(f-g)(X_i)$ are fixed numbers due to conditioning.

[6] The upper limit of the integral is the diameter of $\mathcal{F}$ thanks to (8.16); check that $\mathrm{diam}(\mathcal{F}) \le 1$.

showed that Rademacher complexity of $\mathcal{F}$ is bounded by $C\sqrt{\mathrm{vc}(\mathcal{F})/n}$ for any n-point set.

Let's apply Theorem 8.3.15 to a classical statistics problem: estimate the distribution of a random variable X from a sample. To estimate the CDF of X,

$$F(x) = \mathbb{P}\{X \le x\},$$

from an i.i.d. sample $X_1, \ldots, X_n$, a natural guess is to use the *empirical CDF* – the fraction of the sample points satisfying $X_i \le x$:

$$F_n(x) \coloneqq \frac{1}{n}\big|\{i : \; X_i \le x\}\big|.$$

Amazingly, F_n approximates F *uniformly* over all $x \in \mathbb{R}$:

Theorem 8.3.17 (Glivenko-Cantelli Theorem)**.** *Let $X_1, \ldots, X_n$ be independent random variables with common cumulative distribution function F. Then*

$$\mathbb{E}\|F_n - F\|_{L^\infty} = \mathbb{E}\sup_{x\in\mathbb{R}}|F_n(x) - F(x)| \le \frac{C}{\sqrt{n}}.$$

Proof This is just a restatement of Theorem 8.3.15 for $\Omega = \mathbb{R}$ and the class of indicators of half-infinite intervals

$$\mathcal{F} \coloneqq \big\{\mathbf{1}_{(-\infty,x]} : \; x \in \mathbb{R}\big\},$$

whose VC dimension is bounded by 2 as we noted in Example 8.3.2. □

Example 8.3.18 (Discrepancy)**.** Take an i.i.d. sample of n points from the uniform distribution on the unit square $[0,1]^2$, as in Figure 8.8, and apply Theorem 8.3.15 for the class $\mathcal{F}$ of all indicator functions of circles in that square, which has VC dimension at most 3 (see Exercise 8.13). Then, with high probability, the sample satisfies:

$$\text{fraction of points in } \mathcal{C} = \mathrm{Area}(\mathcal{C}) + O(1/\sqrt{n})$$

simultaneously for all circles $\mathcal{C}$ in the square. This is a classic result in *geometric discrepancy*, which also holds for half-planes, rectangles, polygons with few vertices, etc. – anything with finite VC dimension.

Remark 8.3.19 (Uniform Glivenko-Cantelli classes)**.** A class of real-valued functions $\mathcal{F}$ on a set Ω is called a *uniform Glivenko-Cantelli class* if, for any $\varepsilon > 0$,

$$\lim_{n\to\infty} \sup_{\mu} \mathbb{P}\Big\{\sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E} f(X)\Big| > \varepsilon\Big\} = 0,$$

where the supremum is taken over all probability measures μ on Ω, and where $X, X_1, \ldots, X_n$ are i.i.d. points in Ω with distribution μ. Theorem 8.3.15 followed by Markov inequality implies that any Boolean class with finite VC dimension is uniform Glivenko-Cantelli. The converse is also true (see Exercise 8.28) – so the two are equivalent.

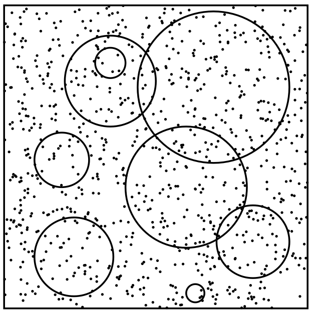


Figure 8.8 The VC law of large numbers implies that the number of points in each circle is proportional to its area with $O(\sqrt{n})$ error.

To practice, try Exercise 8.24 for a weaker but simpler version of Theorem 8.3.15. Also, work through remarkable applications – learning 1D marginals of high-dimensional distributions (Exercise 8.25), one-bit quanitzation (Exercise 8.26) and random matrices with no moment assumptions (Exercise 8.27 – sounds almost too good to be true!).

8.4 Application: statistical learning theory

Statistical learning (or machine learning) is about making predictions from data. Suppose there is an unknown function $T : \Omega \to \mathbb{R}$ on some set Ω (the *target function*), and we get to see its values at a few sample points $X_1, \ldots, X_n$ drawn independently from some distribution on Ω. So, our *training data* is

$$(X_i, T(X_i)), \quad i = 1, \ldots, n. \tag{8.36}$$

The goal is to use this sample to predict $T(X)$ for a new point X drawn from the same distribution (see Figure 8.9).

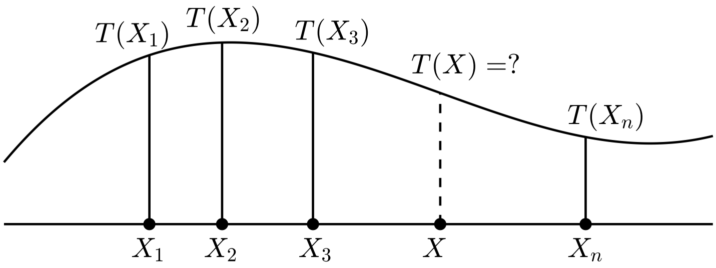


Figure 8.9 We want to learn a function $T : \Omega \to \mathbb{R}$ (a "target function") from its values on the i.i.d. training data $X_1, \ldots, X_n$, so we can predict $T(X)$ for a new random point X.

Example 8.4.1 (Classification)**.** An important type of learning problems is classification, where the function T is Boolean (takes values 0 and 1), classifying points in Ω into two classes. For instance, imagine a health study with n patients. For each patient, we record d health parameters like blood pressure or temperature – that is our vector $X_i \in \mathbb{R}^d$. Suppose we also know if they have diabetes: $T(X_i) = 0$ (healthy) or 1 (sick). The goal is to learn how to predict diabetes from data (8.36) – that is, to learn the function $T : \mathbb{R}^d \to \{0, 1\}$ so we can diagnose new patients based on their health parameters.

8.4.1 Risk, fit and complexity

Given the training data (8.36), we want to find a function $f : \Omega \to \mathbb{R}$ that approximates T, so we can later diagnose a new patient X by checking $f(X)$. We aim to minimize the *risk* of misdiagnosing a new patient, defined as:

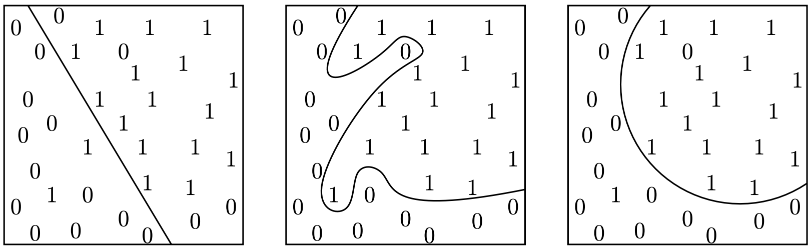
$$R(f) = \mathbb{E}\left(f(X) - T(X)\right)^2. \tag{8.37}$$

Example 8.4.2. In classification problems where T and f are Boolean functions, the risk is simply the probability of misclassification (check this!):

$$R(f) = \mathbb{P}\{f(X) \neq T(X)\}. \tag{8.38}$$

How much training data do we need? That depends on the complexity of the problem. If we believe the target function $T(X)$ behaves in a complicated way, we need more data. Since we usually do not know this up front, we limit our guesses f to some class of functions $\mathcal{F}$, called the *hypothesis class*.

But how do we pick $\mathcal{F}$? There is no universal rule, but it should balance fit and complexity. If $\mathcal{F}$ is too simplistic – say, only linear functions – we might *underfit* (see Figure 8.10a) and missing real patterns. Too complex, and we might *overfit*, just memorizing the training data rather than generalizing from it (Figure 8.10b), and we also need way more data. The sweet spot is a hypothesis class that is just complex enough to capture the real patterns, without chasing random noise (Figure 8.10c).



(a) Underfitting

(b) Overfitting

(c) Right fit

Figure 8.10 Trade-off between fit and complexity

8.4.2 Empirical risk minimization

Once we picked a hypothesis space $\mathcal{F}$, we might just choose the best function f^* in it – one that minimizes the risk (8.37):

$$f^* := \arg\min_{f \in \mathcal{F}} R(f).$$

The catch is that we cannot actually compute the risk $R(f)$, since this involves taking expectation over the whole population Ω. Solution? Take expectation over the training data instead:

Definition 8.4.3 (Empirical risk minimization)**.** For a function $f : \Omega \to \mathbb{R}$, define the *empirical risk* and the empirical minimizer as

$$R_n(f) := \frac{1}{n} \sum_{i=1}^{n} \left(f(X_i) - T(X_i)\right)^2, \quad f_n^* := \arg\min_{f \in \mathcal{F}} R_n(f). \tag{8.39}$$

Now we can compute $R_n(f)$ and f_n^* from the training data. The outcome of learning is f_n^*, and its quality is measured by the *generalization error* $R(f_n^*)$.

Example 8.4.4 (Classification)**.** In classification, where f and T take values 0 or 1, the empirical risk $R_n(f)$ is just the fraction of training points where f gets it wrong: $f(X_i) \neq T(X_i)$. So empirical risk minimization picks the $f \in \mathcal{F}$ that makes the fewest mistakes on the training data.

8.4.3 VC generalization bound

Let's use the VC theory to bound the generalization error in any classification problem (where the target T is Boolean).

Theorem 8.4.5 (VC generalization bound)**.** *Assume that the target T is a Boolean function, and the hypothesis space $\mathcal{F}$ is a class of Boolean functions with finite VC dimension. Then*

$$\mathbb{E}\, R(f_n^*) \le R(f^*) + C\sqrt{\frac{\mathrm{vc}(\mathcal{F})}{n}}.$$

Proof **Step 1: Excess risk.** The following bound holds pointwise:

$$R(f_n^*) - R(f^*) \le 2 \sup_{f \in \mathcal{F}} |R_n(f) - R(f)|. \tag{8.40}$$

To check it, denote $\varepsilon := \sup_{f \in \mathcal{F}} |R_n(f) - R(f)|$ and write

$$\begin{aligned} R(f_n^*) &\le R_n(f_n^*) + \varepsilon && \text{(since } f_n^* \in \mathcal{F} \text{ by construction)} \\ &\le R_n(f^*) + \varepsilon && \text{(since } f_n^* \text{ minimizes } R_n \text{ in the class } \mathcal{F}) \\ &\le R(f^*) + 2\varepsilon && \text{(since } f^* \in \mathcal{F} \text{ by construction).} \end{aligned}$$

Subtracting $R(f^*)$ from both sides gives (8.40).

Step 2: Applying VC law of large numbers. Thanks to (8.40), it is enough to show that

$$\mathbb{E}\sup_{f\in\mathcal{F}}|R_n(f)-R(f)| \lesssim \sqrt{\frac{\mathrm{vc}(\mathcal{F})}{n}}.$$

Recalling the definitions (8.39) and (8.37) of the empirical and true (population) risk, we can rewrite this as

$$\mathbb{E}\sup_{\ell\in\mathcal{L}}\Big|\frac{1}{n}\sum_{i=1}^{n}\ell(X_i)-\mathbb{E}\,\ell(X)\Big| \lesssim \sqrt{\frac{\mathrm{vc}(\mathcal{F})}{n}}, \tag{8.41}$$

where $\mathcal{L}=\{(f-T)^2:\ f\in\mathcal{F}\}$. A moment's thought reveals that $\mathrm{vc}(\mathcal{L})=\mathrm{vc}(\mathcal{F})$ (Exercise 8.29), so an application of Theorem 8.3.15 completes the proof. □

Example 8.4.6 (Classification)**.** Say we have n training data points $X_1,\ldots,X_n$ sampled uniformly from the unit square (as in Example 8.3.18), each labeled "sick" (1) if it lies in some fixed circle $\mathcal{C}$, and "healthy" (0) otherwise. Our goal is to learn that "sickness" circle $\mathcal{C}$ from the data. Let's do empirical risk minimization – pick a circle that best matches the labels, i.e. minimizes misclassifications (the total number of healthy points in the circle and sick points outside).

How well do we do? Since the true circle $\mathcal{C}$ gives zero error ($R(f^*)=0$), and the VC dimension of circles is at most 3 (Exercise 8.13), Theorem 8.4.5 tells us the risk for our learned circle is at most $O(1/\sqrt{n})$. So, new points can be classified just by checking if they are inside our learned circle – with misdiagnosis probability $O(1/\sqrt{n})$, which is dropping as we get more data.

Remark 8.4.7 (Bias-variance tradeoff)**.** The VC generalization bound (Theorem 8.4.5) identifies two main sources of error in learning. The *bias* term $R(f^*)$ comes from an imperfect choice of the hypothesis class (underitting). We can shrink the bias by including more functions in $\mathcal{F}$ – ideally enough to capture the true target function T, making the bias equal zero. But then the *variance* term $O(\sqrt{\mathrm{vc}(\mathcal{F})/n})$ grows. To keep it in check, we may use more training data (increase n) to avoid overfitting.

For practice, extend the theory to the more realistic setup where the labels are random but correlated with inputs (Exercise 8.30), get away from Boolean classes and show how to learn a Lipschitz function (Exercise 8.31), and demonstrate that learning fails if the VC dimension is infinite (Exercise 8.32).

8.5 Generic chaining

Although Dudley inequality is a simple and versatile tool, it can be loose (see Exercise 8.4). The covering numbers $\mathcal{N}(T,d,\varepsilon)$ just do not contain enough information to capture the magnitude of a random process $(X_t)_{t\in T}$.

Fortunately, there is a way to obtain accurate, two-sided bounds on $\mathbb{E}\sup_{t\in T}X_t$ 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 inequality (Theorem 8.1.4).

8.5.1 A makeover of Dudley inequality

Recall the bound (8.12) we obtained by chaining:

$$\mathbb{E} \sup_{t \in T} X_t \lesssim \sum_{k=\kappa+1}^{\infty} \varepsilon_{k-1} \sqrt{\log |T_k|}, \tag{8.42}$$

where $\varepsilon_k = 2^{-k}$, T_k are smallest ε_k-nets of T, so $|T_k| = \mathcal{N}(T, d, \varepsilon_k)$, and κ is chosen so that $|T_\kappa| = 1$.

Now, let's flip the approach: instead of fixing ε_k and minimizing $|T_k|$, fix $|T_k|$ and minimize ε_k. Specifically, pick subsets $T_k \subset T$ such that

$$|T_0| = 1, \quad |T_k| \le 2^{2^k}, \quad k = 1, 2, \ldots \tag{8.43}$$

and define

$$\varepsilon_k = \sup_{t \in T} d(t, T_k),$$

where $d(t, T_k)$ denotes the distance[7] from t to the set T_k. Each T_k is then an ε_k-net, and the chaining bound (8.42) becomes

$$\mathbb{E} \sup_{t \in T} X_t \lesssim \sum_{k=1}^{\infty} 2^{k/2} \sup_{t \in T} d(t, T_{k-1}),$$

of, after re-indexing,

$$\mathbb{E} \sup_{t \in T} X_t \lesssim \sum_{k=0}^{\infty} 2^{k/2} \sup_{t \in T} d(t, T_k). \tag{8.44}$$

8.5.2 The γ_2 functional and generic chaining

So far, we have just restated Dudley's inequality in a new form – nothing major yet. The important step will come now. The generic chaining will allow us to pull the supremum *outside* the sum in (8.44). The resulting quantity has a name:

Definition 8.5.1 (γ_2 functional)**.** Let (T, d) be a metric space. A sequence of subsets $(T_k)_{k=0}^{\infty}$ of T satisfying (8.43) is called an *admissible sequence.* The γ_2 *functional* of T is defined as

$$\gamma_2(T, d) = \inf_{(T_k)} \sup_{t \in T} \sum_{k=0}^{\infty} 2^{k/2} d(t, T_k) \tag{8.45}$$

where the infimum is over all admissible sequences.

[7] The distance from a point t to a set A in a metric space is $d(t, A) = \inf\{d(t, a) :\ a \in A\}$.

Since the supremum in the γ_2 functional is outside the sum, it is smaller than the Dudley sum in (8.44). That might seem like a small change, but it can make a real difference in some cases (see Exercise 8.34).

The good news is that we can improve Dudley inequality (Theorem 8.1.4) by replacing the Dudley sum (or integral) by a tighter quantity – the γ_2-functional:

Theorem 8.5.2 (Generic chaining bound)**.** *Let $(X_t)_{t\in T}$ be a mean-zero random process on a metric space (T,d) with subgaussian increments as in* (8.1)*. Then*

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

Proof We use the chaining method from the proof of Dudley inequality (Theorem 8.1.4), but do it more carefully this time.

Step 1: Chaining set-up. As before, we may assume that $K=1$ and that T is finite, which makes $\gamma_2(T,d)$ finite (why?). Let (T_k) be an admissible sequence of subsets of T which almost attains the supremum in (8.45):

$$\sup_{t\in T}\sum_{k=0}^{\infty} 2^{k/2} d(t,T_k) \le 2\gamma_2(T,d) < \infty. \tag{8.46}$$

Denote $T_0=\{t_0\}$. There must be some K for which $T_K=T$; otherwise some $t\in T$ would keep getting left out of infinitely many sets T_k, so $d(t,T_k)>\varepsilon$ for all those k and some fixed $\varepsilon>0$, making the series in (8.46) diverge.

We walk from t_0 to a general point $t\in T$ along the (finite) chain

$$t_0=\pi_0(t)\to\pi_1(t)\to\pi_2(t)\to\cdots\to t$$

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):

$$X_t-X_{t_0}=\sum_{k=1}^{\infty}(X_{\pi_k(t)}-X_{\pi_{k-1}(t)}). \tag{8.47}$$

Step 2: Controlling the increments. This is where we need to be more careful. We would like say that, with high probability, the following event holds:

$$\left|X_{\pi_k(t)}-X_{\pi_{k-1}(t)}\right| \lesssim 2^{k/2} d(t,T_k) \quad \forall k\in\mathbb{N}, \quad \forall t\in T. \tag{8.48}$$

Summing over all k would lead to a desired bound in terms of $\gamma_2(T,d)$.

To prove (8.48), let us fix k and t first. The subgaussian assumption gives

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

So, for every $u\ge 0$, the event

$$\left|X_{\pi_k(t)}-X_{\pi_{k-1}(t)}\right| \le Cu2^{k/2} d(\pi_k(t),\pi_{k-1}(t)) \tag{8.49}$$

holds with probability at least

$$1-2\exp(-8u^2 2^k).$$

(To get the constant 8, choose the absolute constant C large enough.)

Now unfix $t \in T$ by taking a union bound over

$$|T_k| \cdot |T_{k-1}| \le |T_k|^2 \le 2^{2^{k+1}}$$

pairs $(\pi_k(t), \pi_{k-1}(t))$. Also, unfix k by taking a union bound over all $k \in \mathbb{N}$. Then (8.49) holds simultaneously for all $t \in T$ and $k \in \mathbb{N}$ with probability at least

$$1 - \sum_{k=1}^{\infty} 2^{2^{k+1}} \cdot 2\exp(-8u^2 2^k) \ge 1 - 2\exp(-u^2).$$

if $u > c$. (Check the last inequality!)

Step 3: Summing up the increments. In the event that the bound (8.49) does holds for all $t \in T$ and $k \in \mathbb{N}$, we can sum up the inequalities over $k \in \mathbb{N}$ and plug the result into the chaining sum (8.47). We get

$$|X_t - X_{t_0}| \lesssim u \sum_{k=1}^{\infty} 2^{k/2} d(\pi_k(t), \pi_{k-1}(t)), \tag{8.50}$$

where the notation $\lesssim$ hides an absolute constant factor. By triangle inequality,

$$d(\pi_k(t), \pi_{k-1}(t)) \le d(t, \pi_k(t)) + d(t, \pi_{k-1}(t)).$$

Using that bound, reindexing, and plugging in (8.46), we get that the right-hand side of (8.50) is at most $Cu\gamma_2(T,d)$, that is

$$|X_t - X_{t_0}| \lesssim u\gamma_2(T,d).$$

(Check!) Taking the supremum over T yields

$$\sup_{t\in T} |X_t - X_{t_0}| \lesssim u\gamma_2(T,d).$$

Since this holds with probability at least $1 - 2\exp(-u^2)$ for any $u > c$, we get

$$\Big\| \sup_{t\in T} |X_t - X_{t_0}| \Big\|_{\psi_2} \lesssim \gamma_2(T,d).$$

This quickly implies the conclusion of Theorem 8.5.2 (check!) □

Remark 8.5.3 (Generic chaining: supremum of increments)**.** Similarly to Dudley inequality (Remark 8.1.5), the generic chaining actually gives

$$\mathbb{E} \sup_{t,s\in T} |X_t - X_s| \le CK\gamma_2(T,d).$$

which is valid even without the mean zero assumption $\mathbb{E} X_t = 0$.

Remark 8.5.4 (Generic chaining: a high-probability bound)**.** Theorem 8.5.2 gives only an expectation bound, but generic chaining actually gives a high-probability bound – we have seen this logic in Dudley inequality (Remark 8.1.6).

Assuming T is finite,[8] for every $u \geq 0$, the event

$$\sup_{t,s\in T} |X_t - X_s| \leq CK\Big[\gamma_2(T,d) + u \cdot \operatorname{diam}(T)\Big]$$

holds with probability at least $1 - 2\exp(-u^2)$ Try Exercise 8.35 to prove this! For Gaussian processes, you can deduce it directly from Gaussian concentration.

For practice, try applying generic chaining to empirical processes (Exercise 8.36).

8.5.3 Majorizing measure and comparison theorems

The γ_2 functional (Definition 8.5.1) is usually harder to compute than covering numbers in Dudley inequality (Theorem 8.1.3). But it is often worth the effort – unlike Dudley, generic chaining is sharp up to constants:

Theorem 8.5.5 (Talagrand majorizing measure theorem)**.** *Let $(X_t)_{t\in T}$ be a mean-zero Gaussian process on a set T, equipped with the canonical metric $d(t,s) = \|X_t - X_s\|_{L^2}$ as in (7.1). Then*

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

The upper bound in Theorem 8.5.5 comes straight from generic chaining (Theorem 8.5.2). The lower bound is trickier; its proof can be seen as a refined, multiscale extension of Sudakov inequality (Theorem 7.4.1), which we will not include here.

The upper bound, as we know from Theorem 8.5.2, holds not just for Gaussian but also for all *subgaussian* processes. Therefore, by combining the upper and lower bounds, we can bound any subgaussian process by the Gaussian one:

Corollary 8.5.6 (Talagrand comparison inequality)**.** *Let $(X_t)_{t\in T}$ be a mean-zero random process on a set T and let $(Y_t)_{t\in T}$ be a mean-zero Gaussian process. Assume*

$$\|X_t - X_s\|_{\psi_2} \leq K\|Y_t - Y_s\|_{L^2} \quad \textit{for all } t,s \in T.$$

Then

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

Proof Consider the canonical metric $d(t,s) = \|Y_t - Y_s\|_{L^2}$ on T. Now just use the generic chaining bound (Theorem 8.5.2) followed by the lower bound in the majorizing measure Theorem 8.5.5:

$$\mathbb{E}\sup_{t\in T} X_t \lesssim K\gamma_2(T,d) \lesssim K\,\mathbb{E}\sup_{t\in T} Y_t. \quad \square$$

Remark 8.5.7 (Sudakov-Fernique)**.** Corollary 8.5.6 extends the Sudakov-Fernique inequality (Theorem 7.2.8) to subgaussian processes – with only an absolute constant factor as the price for this generalization!

[8] As always, are are assuming T is finite to avoid measurability issues; the general case typically follows by approximation.

Let's apply Corollary 8.5.6 for the canonical Gaussian process $Y_x = \langle g, x\rangle$ on a set $T \subset \mathbb{R}^n$, where $g \sim N(0, I_n)$. Recall from Section 7.5 that

$$w(T) = \mathbb{E} \sup_{x\in T} \langle g, x\rangle$$

is the *Gaussian width* of the set T. We immediately get the following:

Corollary 8.5.8 (Talagrand comparison inequality: geometric form)**.** *Let $(X_x)_{x\in T}$ be a mean-zero random process on a subset $T \subset \mathbb{R}^n$. Assume that*

$$\|X_x - X_y\|_{\psi_2} \le K\|x - y\|_2 \quad \textit{for all } x, y \in T.$$

Then

$$\mathbb{E} \sup_{x\in T} X_x \le CKw(T).$$

Remark 8.5.9 (Subgaussian width $\lesssim$ Gaussian width)**.** A nice consequence: if X is a subgaussian random vector in $\mathbb{R}^n$, then

$$\mathbb{E} \sup_{t\in T} \langle X, t\rangle \le CKw(T) \quad \text{for any bounded set } T \subset \mathbb{R}^n,$$

where $K = \|X\|_{\psi_2}$. Just apply Corollary 8.5.8 to the process $(\langle X, x\rangle)_{x\in T}$, whose increments satisfy $\|\langle X, x\rangle - \langle X, y\rangle\|_{\psi_2} = \|\langle X, x - y\rangle\|_{\psi_2} \le K\|x - y\|_2$ by definition of a subgaussian random vector.

For practice, try proving a few versions of Talagrand comparison (Exercise 8.37) and use it to get sharp bounds on the ℓ^p norms of subgsaussian vectors (Exercise 8.38).

8.6 Chevet inequality

Talagrand comparison inequality (and more generally generic chaining) is a powerful tool that works in a wide range of settings. Let's use it to get a very general uniform bound for a random quadratic form:

$$\sup_{x\in T,\, y\in S} \langle Ax, y\rangle \le ?$$

where A is a random matrix and T, S are any bounded sets.

A special case where T and S are Euclidean balls leads to the operator norm of A, which we already studied in Theorems 4.4.3 and 7.3.1. Now let's go general and consider arbitrary sets, with the goal of bounding things using just two geometric quantities: the *Gaussian width* and the *radius*, defined as

$$\operatorname{rad}(T) \coloneqq \sup_{x\in T} \|x\|_2. \tag{8.51}$$

Theorem 8.6.1 (Subgaussian Chevet inequality)**.** *Let A be an $m \times n$ random matrix with independent, mean-zero, subgaussian rows A_i. Let $T \subset \mathbb{R}^n$ and $S \subset \mathbb{R}^m$ be arbitrary bounded sets. Then*

$$\mathbb{E} \sup_{x\in T,\, y\in S} \langle Ax, y\rangle \le CK\,[w(T)\operatorname{rad}(S) + w(S)\operatorname{rad}(T)]$$

where $K = \max_i \|A_i\|_{\psi_2}$*. The same holds if "rows" is replaced by "columns".*

Proof of Theorem 8.6.1 We will follow the proof of Theorem 7.3.1 but with Talagrand comparison inequality instead of Sudakov-Fernique's.

Without loss of generality, $K = 1$. We need to bound the random process

$$X_{uv} \langle Au, v \rangle, \quad u \in T,\, v \in S.$$

To check that the increments are subgaussian, fix$(u, v), (w, z) \in T \times S$ and write

$$X_{uv} - X_{wz} = X_{uv} - X_{wv} + X_{wv} - X_{wz} = \langle A(u-w), v\rangle + \langle Aw, v - z\rangle.$$

Using triangle inequality and the subgaussian assumption (see Exercise 3.34 – do it now if you skipped it), we get

$$\begin{aligned} \|X_{uv} - X_{wz}\|_{\psi_2} &\le \big\|\langle A(u-w), v\rangle\big\|_{\psi_2} + \big\|\langle Aw, v-z\rangle\big\|_{\psi_2} \\ &\lesssim \|u-w\|_2\|v\|_2 + \|v-z\|_2\|w\|_2 \\ &\le \|u-w\|_2 \operatorname{rad}(S) + \|v-z\|_2 \operatorname{rad}(T). \end{aligned} \tag{8.52}$$

Let's pick a simpler Gaussian process (Y_{uv}) for Talagrand comparison inequality (Corollary 8.5.6). The increment bound (8.52) points us to a good choice:

$$Y_{uv} := \langle g, u\rangle \operatorname{rad}(S) + \langle h, v\rangle \operatorname{rad}(T),$$

where $g \sim N(0, I_n)$ and $h \sim N(0, I_m)$ are independent. The increments of this process are

$$\|Y_{uv} - Y_{wz}\|_{L^2}^2 = \|u-w\|_2^2 \operatorname{rad}(S)^2 + \|v-z\|_2^2 \operatorname{rad}(T)^2.$$

(Check this like in the proof of Theorem 7.3.1!) Comparing to (8.52), we find that

$$\|X_{uv} - X_{wz}\|_{\psi_2} \lesssim \|Y_{uv} - Y_{wz}\|_{L^2},$$

where we used the inequality $a + b \le \sqrt{2(a^2+b^2)}$. Applying Talagrand comparison inequality (Corollary 8.5.6), we finish the proof:

$$\begin{aligned} \mathbb{E} \sup_{u\in T,\, v\in S} X_{uv} &\lesssim \mathbb{E} \sup_{u\in T,\, v\in S} Y_{uv} \\ &= \mathbb{E}\sup_{u\in T}\langle g, u\rangle \operatorname{rad}(S) + \mathbb{E}\sup_{v\in S}\langle h, v\rangle \operatorname{rad}(T) \\ &= w(T)\operatorname{rad}(S) + w(S)\operatorname{rad}(T). \end{aligned}$$ □

Remark 8.6.2 (Operator norms of random matrices)**.** For the special case $T = S^{n-1}$ and $S = S^{m-1}$, Chevet inequality gives us the familiar sharp bound on the operator norm:

$$\mathbb{E}\|A\| \le CK(\sqrt{n} + \sqrt{m}),$$

which we earlier proved in Section 4.4.2 using ε-nets. But this new approach is more flexible. For example, picking T and S as ℓ^p balls gives the $\|A\|_{p\to q}$ norm of a random matrix – try Exercise 8.41!

Remark 8.6.3 (Gaussian Chevet inequality)**.** For Gaussian matrices A with i.i.d. $N(0,1)$ entries, we can even prove Chevet inequality with sharp constant 1:

$$\mathbb{E} \sup_{x \in T,\, y \in S} \langle Ax, y \rangle \le w(T)\,\mathrm{rad}(S) + w(S)\,\mathrm{rad}(T),$$

and a reverse inequality up to a constant (Exercise 8.39). Later, we will further improve Gaussian Chevet inequality in Section 9.7.1.

For more practice, prove a high-probability version of Chevet (Exercise 8.40).

8.7 Notes

The idea of chaining already appears in Kolmogorov's proof of his continuity theorem for Brownian motion, see e.g. [253, Chapter 1]. Dudley integral inequality (Theorem 8.1.3) can be traced to the work of R. Dudley. Our exposition in Section 8.1 mostly follows [210, Chapter 11], [315, Section 1.2] and [330, Section 5.3].

Monte Carlo method (Section 8.2.1) is extremely popular in scientific computing, especially when combined with the power of Markov chains, see e.g. [63]. The rich theory of *empirical processes* (see Definitiondef: empirical process) has applications to statistics and machine learning, see [329, 328, 279, 232]. The Lipschitz law of large numbers (Theorem 8.2.3), also known Wasserstein law of large numbers, is loosely based on [330, Example 5.15], see [115, Chapter 11]. For an introduction to transportation of measure (Section 8.2.3), see [342, 85].

The concept of VC dimension introduced in Section 8.3 goes back to the foundational work of V. Vapnik and A. Chervonenkis [335]; modern treatments can be found e.g. in [329, Section 2.6.1], [210, Section 14.3], [330, Section 7.2], [225, Sections 10.2–10.3], [232, Section 2.2], [329, Section 2.6]. Pajor Lemma 8.3.7 is originally due to A. Pajor [268]; see [128], [210, Proposition], [330, Theorem 7.19], [329, Lemma 2.6.2].

What we now call Sauer-Shelah Lemma (Lemma 8.3.9) was proved independently by V. Vapnik and A. Chervonenkis [335], N. Sauer [294] and M. Perles and S. Shelah [300]. Various proofs of Sauer-Shelah lemma can be found in literature, e.g. [46, Chapter 17], [225, Sections 10.2–10.3], [210, Section 14.3]. A number of variants of Sauer-Shelah Lemma is known, see e.g. [158, 312, 313, 15, 337]. The growth function (Section 8.3.4) was originally introduced by Vapnik-Chervonenkis, who studied its various properties including the exponential-polynomial dichotomy (Exercise 8.20).

Theorem 8.3.13 is due to R. Dudley [114]; see [210, Section 14.3], [329, Theorem 2.6.4]. The dimension reduction Lemma 8.3.14 is implicit in Dudley proof; it was stated explicitly in [237] and reproduced in [330, Lemma 7.17]. For generalization of VC theory from $\{0,1\}$ to general real-valued function classes, see [237, 288], [330, Sections 7.3–7.4].

Since the foundational work of V. Vapnik and A. Chervonenkis [335], bounds on empirical processes via VC dimension like Theorem 8.3.15 were in the spotlight of the statistical learning theory, see e.g. [232, 32, 329, 288], [330, Chapter 7]. Our presentation of Theorem 8.3.15 is based on [330, Corollary 7.18]. Although explicit statements of this result are difficult to find in earlier literature, it can be derived from [32, Theorem 6], [56, Section 5].

Glivenko-Cantelli theorem (Theorem 8.3.17) is a result from 1933 [138, 75] which predated and partly motivated the later development of VC theory; see [210, Section 14.2], [329, 115] for more on Glivenko-Cantelli theorems and other uniform results in probability theory. Example 8.3.18 discusses a basic problem in discrepancy theory; see [224] for a comprehensive treatment of discrepancy theory.

In Section 8.4 we scratch the surface of statistical learning theory, which is a big area on the intersection of probability, statistics, and theoretical computer science. For deeper introduction to this subject, see e.g. the tutorials [52, 232] and books [171, 156, 196].

Generic chaining, presented in Section 8.5, was put forward by M. Talagrand since 1985 (after an earlier work of X. Fernique [124]) as a sharp method to obtain bounds on Gaussian processes. Our presentation is based on Talagrand's books [315, 316, 317], which discuss ramifications, applications and history of generic chaining in great detail. The upper bound on subgaussian processes (Theorem 8.5.2) can be found in [315, Theorem 2.2.22]; the lower bound (the majorizing measure Theorem 8.5.5) can be found in [315, Theorem 2.4.1]. Talagrand comparison inequality (Corollary 8.5.6) is borrowed from [315, Theorem 2.4.12]. Another presentation of generic chaining can be found in [330, Chapter 6]. A different proof of the majorizing measure theorem was recently given by R. van Handel in [332, 333]. A high-probability version of generic chaining bound (Remark 8.5.4) is from[316, Theorem 2.2.27] and [317, Theorem 2.7.13]; it was also proved by a different method by S. Dirksen [103].

Section 8.6 covers Chevet inequality for subgaussian processes. For Gaussian processes, it goes back to S. Chevet [84]; the constants were then improved by Y. Gordon [140], leading to the result we stated in Exercise 8.39. A exposition of this result can be found in [21, Section 9.4]. For variants and applications of Chevet inequality, see [321, 7].

The logarthmic gap between Sudakov and Dudley inequalities (Exercise 8.5) is optimal; the cross-polytope $T = B_1^n$ serves as an example (see [317, Exercise 2.5.11]).

Subexponential Dudley inequality (Exercise 8.7) extends easily to almost any kind of tail decay [210, Section 11.1].

Local Dudley inequality (Exercise 8.8) is a tool that has applications in statistical learning theory and for proving continuity of random processes, see [330, Section 5.4].

Exercise 8.10 extends the Lipschitz law of large numbers to higher dimensions. To learn about its history, extensions, improvements and (almost) optimality, see [85, Chapter 2].

The bound $O(dk \log k)$ we obtain in Exercise 8.21 is optimal [119].

The polynomial-exponential dichotomy of growth function (Exercise 8.20), as well as the characterization of uniform convergence in LLN (Exercises 8.28) and learnability (Exerfcise 8.32) as finite VC dimension essentially go back to the original work of Vapnik and Chervonenkis [335].

One-bit quantization (Exercise 8.26) has been extensively studied in signal processing and machine learning, see e.g. [89, 53, 355, 198, 199, 275, 274, 276, 13, 341, 92, 265, 189, 278, 168, 28, 351, 105, 129, 106, 104, 226, 81, 107]. The result in Exercise 8.26 follows from [265, Theorem 2.2].

Exercise 8.27 presents a simplified version of the *small ball method*, which was introduced in [191] as a way to weaken the assumptions on the distributions in random matrix theory and machine learning [234, 235, 208, 207].

The $p \to q$ norms of random matrices (Exercise 8.41) was first studied in [37] for $p = 2$ and $q \in [2, \infty)$ for bounded entires; more general recent results can be found in [9, 100, 200, 202, 201].

Exercises

8.1 ☕☕☕ (Dudley inequality: a high-probability bound) Modify the chaining argument in Section 8.1 to get the high-probability version of Dudley inequality stated in Remark 8.1.6. To do this, upgrade (8.11) to a high-probability version:

$$\sup_{t \in T} (X_{\pi_k(t)} - X_{\pi_{k-1}(t)}) \le C \varepsilon_{k-1} \left[\sqrt{\log |T_k|} + z_k \right] \tag{8.53}$$

with probability at least $1 - 2\exp(-z_k^2)$. Choose the values for z_k to be able to make a union bound over all terms in (8.12).

8.2 ☕☕ (Dudley for Gaussian processes: a high-probability bound) Give an alternative proof of the result in Remark 8.1.6 for Gaussian processes by combining Dudley inequality with Gaussian concentration (Theorem 7.1.11).

8.3 ☕☕ (Dudley integral and sum are equivalent) In the proof of Theorem 8.1.3, we bounded the Dudley sum by the Dudley integral. Show they are actually equivalent up to constants:

$$\sum_{k \in \mathbb{Z}} 2^{-k} \sqrt{\log \mathcal{N}(T, d, 2^{-k})} \asymp \int_0^\infty \sqrt{\log \mathcal{N}(T, d, \varepsilon)}\, d\varepsilon.$$

8.4 ☕☕☕ (Dudley inequality can be loose) Let e_k the standard basis vectors in $\mathbb{R}^n$, and

$$T := \Big\{ \frac{e_k}{\sqrt{1+\log k}},\ k = 1, \ldots, n \Big\}.$$

(a) Show that $w(T) \le C$.
(b) Show that $\int_0^\infty \sqrt{\log \mathcal{N}(T, d, \varepsilon)}\, d\varepsilon \to \infty$ as $n \to \infty$.

8.5 ☕☕☕ (Dudley and Sudakov inequalities are sharp up to log factors) Sudakov inequality (Corollary 7.4.2) and Dudley inequality (Theorem 8.1.8) give these bounds on the Gaussian width $w(T)$ for any bounded set $T \subset \mathbb{R}^n$:

$$s(T) \lesssim w(T) \lesssim d(T),$$

where

$$s(T) = \sup_{\varepsilon > 0} \varepsilon \sqrt{\log \mathcal{N}(T, \varepsilon)} \quad \text{and} \quad d(T) = \int_0^\infty \sqrt{\log \mathcal{N}(T, \varepsilon)}\, d\varepsilon.$$

Prove that these bounds are equivalent up to a logarithmic factor:

$$s(T) \le d(T) \le C \log(n)\, s(T).$$

In particular, both Sudakov inequality (Corollary 7.4.2) and Dudley inequality (Theorem 8.1.8) are tight up to a logarithmic factor.

8.6 ☕☕☕☕ (Dudley inequality with refined limits) We know from Remark 8.1.7 that the upper limit in the Dudley integral (8.17) can be set to the diameter of T. It would be useful to improve the lower limit, too, since the integral might blow up near 0. Show that the Gaussian width of any bounded set $T \subset \mathbb{R}^n$ satisfies

$$w(T) \le C \int_a^b \sqrt{\log \mathcal{N}(T, \varepsilon)}\, d\varepsilon \quad \text{where} \quad a = \frac{c\,w(T)}{\sqrt{n}}, \quad b = \operatorname{diam}(T).$$

8.7 ☕☕ (Subexponential Dudley inequality) Suppose that the increments of a mean-zero random process $(X_t)_{t \in T}$ are not subgaussian as in (8.1) but subexponential:

$$\|X_t - X_s\|_{\psi_1} \le K d(t, s) \quad \text{for all } t, s \in T.$$

Prove the following version of Dudley inequality (Theorem 8.1.3):

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

8.8 ☕☕☕ (Local Dudley inequality) Let $(X_t)_{t \in T}$ be a random process on a metric space (T, d) with subgaussian increments as in (8.1). Let $\delta > 0$. Prove that

$$\mathbb{E} \sup_{\substack{s,t \in T \\ d(s,t) \le \delta}} |X_s - X_t| \le CK \int_0^\delta \sqrt{\log \mathcal{N}(T, d, \varepsilon)}\, d\varepsilon.$$

8.9 ☕☕☕ (Covering numbers of Lipschitz functions) For the class of functions

$$\mathcal{F} := \big\{ f : [0,1] \to [0,1],\ \|f\|_{\mathrm{Lip}} \le 1 \big\},$$

show that

$$\mathcal{N}(\mathcal{F}, \|\cdot\|_{L^\infty}, \varepsilon) \le e^{C/\varepsilon} \quad \text{for any } \varepsilon \in (0,1).$$

8.10 ☕☕☕☕ (Lipschitz law of large numbers in high dimensions) Let's generalize Theorem 8.2.3 to high dimensions. Consider functions on the unit cube that are L-Lipschitz with respect to the $\|\cdot\|_\infty$ metric in $\mathbb{R}^n$:

$$\mathcal{F} := \left\{ f : [0,1]^d \to \mathbb{R},\ \|f\|_{\mathrm{Lip}} \le L \right\}.$$

Let $X, X_1, X_2, \ldots, X_n$ be i.i.d. random variables taking values in $[0,1]$. Show that

$$\mathbb{E} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbb{E} f(X) \right| \le CL \begin{cases} \log(n) n^{-1/2}, & d = 2; \\ n^{-1/d}, & d \ge 3. \end{cases}$$

To prove this, assume $L = 1$ and follow these steps:

(a) Extend the bound on the covering numbers (8.9) to higher dimensions:

$$\mathcal{N}(\mathcal{F}, \|\cdot\|_{L^\infty}, \varepsilon) \le e^{C/\varepsilon^d}.$$

(b) If you just plug this into (8.25), the integral will diverge because of the singularity at zero. To fix this, refine (8.25) to

$$\mathbb{E} \sup_{f \in \mathcal{F}} |X_f| \lesssim \delta + \frac{1}{\sqrt{n}} \int_\delta^1 \sqrt{\log \mathcal{N}(\mathcal{F}, \|\cdot\|_{L^\infty}, \varepsilon)}\, d\varepsilon \quad \text{for any } \delta \in [0,1]. \tag{8.54}$$

8.11 ☕☕ (Symmetrization for empirical processes) Modify the proof of Symmetrization Lemma 6.3.2 to obtain the following version of symmetrization. Let $\mathcal{F}$ be a class of Boolean functions on some domain Ω, and let $X, X_1, X_2, \ldots, X_n$ be independent random points in Ω with common distribution. Prove that

$$\mathbb{E} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbb{E} f(X) \right| \le 2\, \mathbb{E} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(X_i) \right|$$

where $\varepsilon_1, \varepsilon_2, \ldots$ are independent Rademacher random variables (which are also independent of $X_1, X_2, \ldots$).

8.12 ☕☕ (VC dimension of pairs of intervals) Let $\mathcal{F}$ be the class of indicators of sets of the form $[a,b] \cup [c,d]$ in $\mathbb{R}$. Show that

$$\mathrm{vc}(\mathcal{F}) = 4.$$

8.13 ☕☕☕ (VC dimension of circles) Let $\mathcal{F}$ be the class of indicators of all circles in $\mathbb{R}^2$. Show that

$$\mathrm{vc}(\mathcal{F}) = 3.$$

8.14 ☕☕☕ (VC dimension of rectangles) Let $\mathcal{F}$ be the class of indicators of all closed axis-aligned rectangles, i.e. product sets $[a,b] \times [c,d]$, in $\mathbb{R}^2$. Show that

$$\mathrm{vc}(\mathcal{F}) = 4.$$

8.15 ☕☕☕ (VC dimension of squares) Let $\mathcal{F}$ be the class of indicators of all closed axis-aligned squares – sets of the form $[a, a+d] \times [b, b+d]$ in $\mathbb{R}^2$. Show that

$$\mathrm{vc}(\mathcal{F}) = 3.$$

8.16 ☕☕ (VC dimension of polygons) Let $\mathcal{F}$ be the class of indicators of all convex polygons in $\mathbb{R}^2$, without any restriction on the number of vertices. Show that

$$\mathrm{vc}(\mathcal{F}) = \infty.$$

8.17 ☕☕☕ (VC dimension of half-spaces) Prove the result in Example 8.3.5: if $\mathcal{F}$ is the class of indicators of all half-planes in $\mathbb{R}^n$, then

$$\mathrm{vc}(\mathcal{F}) = n + 1.$$

8.18 ☕☕ (VC dimension vs. algebraic dimension) Show that for any finite-dimensional class of Boolean functions $\mathcal{F}$ on a domain Ω, we have

$$\mathrm{vc}(\mathcal{F}) \le \dim(\mathcal{F}),$$

where $\dim(\mathcal{F})$ denotes the linear algebraic dimension of $\mathcal{F}$, i.e. the maximal number of linearly independent functions in $\mathcal{F}$.

8.19 ☕☕ (Sharpness of Pajor and Sauer-Shelah lemmas) Show that both Pajor Lemma 8.3.7 and Sauer-Shelah Lemma 8.3.9 are sharp by considering the set $\mathcal{F}$ of binary strings of length n with at most d ones (this set is often called a *Hamming ball*).

8.20 ☕ (VC dichotomy) Let $\mathcal{F}$ be a class of Boolean functions on some domain. Prove that the growth function $\Pi_{\mathcal{F}}(n)$ can only grow either polynomially or exponentially, and nothing in between. More rigorously, prove that either $\Pi_{\mathcal{F}}(n) = 2^n$ for all n, or there exists a polynomial $p(n)$ such that $\Pi_{\mathcal{F}}(n) \le p(n)$ for all n.

8.21 ☕☕ (VC stability) In Proposition 8.3.11, we observed how the VC dimension is stable when we combined two function classes by taking pointwise minima or maxima. Let us generalize this observation for any number of classes, and for any way of combining them. Let $\mathcal{F}_1, \ldots, \mathcal{F}_k$ be classes of Boolean functions on the same domain. Fix any function $\phi : \{0,1\}^k \to \{0,1\}$, which we can think of a Boolean formula. Consider the class

$$\mathcal{F} = \{\phi(f_1, \ldots, f_k) : \; f_1 \in \mathcal{F}_1, \ldots, f_k \in \mathcal{F}_k\}.$$

Prove that if $\mathrm{vc}(F_i) \le d$ for every i, then

$$\mathrm{vc}(\mathcal{F}) \le Cdk \log k$$

where C is an absolute constant.

8.22 ☕☕☕☕ (VC dimension of the union)

(a) Let $\mathcal{F}$ and $\mathcal{G}$ classes of Boolean functions on the same domain. Prove that

$$\mathrm{vc}(\mathcal{F} \cup \mathcal{G}) \le \mathrm{vc}(\mathcal{F}) + \mathrm{vc}(\mathcal{G}) + 1.$$

(b) Give an example where $\mathcal{F}$ and $\mathcal{G}$ have positive VC dimensions, and

$$\mathrm{vc}(\mathcal{F} \cup \mathcal{G}) = \mathrm{vc}(\mathcal{F}) + \mathrm{vc}(\mathcal{G}) + 1.$$

8.23 ☕☕ Theorem 8.3.13 is stated for $\varepsilon \in (0, 1)$. What bound holds for larger ε?

8.24 ☕☕☕ (A simpler, weaker VC law of large numbers) Use Sauer-Shelah Lemma directly, instead of Pajor Lemma, to prove a weaker version Theorem 8.3.15, with $C\sqrt{\frac{d}{n}\log\frac{en}{d}}$ in the right hand side, where $d = \mathrm{vc}(\mathcal{F})$.

8.25 ☕☕ (Learning 1-dimensional marginals) Learning high-dimensional distributions from data is hard. But we can learn all 1-dimensional marginals from just $O(n)$ samples! Let $X, X_1, \dots, X_m$ be i.i.d. random vectors in $\mathbb{R}^n$. Show that the CDFs of all 1-dimensional marginals $\langle X, u\rangle$ can be estimated uniformly from the data, using the empirical CDFs:

$$\mathbb{E} \sup_{u \in S^{n-1},\, t \in \mathbb{R}} \left| \frac{1}{m} \sum_{i=1}^{m} \mathbf{1}_{\{\langle X_i, u\rangle \le t\}} - \mathbb{P}\{\langle X, u\rangle \le t\} \right| \lesssim \sqrt{\frac{n}{m}}.$$

8.26 ☕☕☕ (One-bit quantization) The simplest way to quantize a unit vector $u \in \mathbb{R}^n$ is to take the sign of each coordinate: set $\Psi(u) = \mathrm{sign}(u)$, where the sign[9] function is applied to each coordinate of u. But this can be quite inaccurate – very close vectors u and v can be sent to very different outputs. (Can you see why?) Let's try something smarter.

(a) Let A be an $m \times n$ random matrix with independent standard normal entries. Consider the map $\Phi : S^{n-1} \to \{-1, 1\}^m$ given by

$$\Phi(u) = \mathrm{sign}(Au)$$

where the sign function is applied to each coordinate of Au. Check that

$$\mathbb{E}\, \frac{1}{m} d\left(\Phi(u), \Phi(v)\right) = \frac{1}{\pi}\rho(u, v) \quad \text{for all } u, v \in S^{n-1}.$$

Here $d(\cdot, \cdot)$ denotes the Hamming distance on the binary cube $\{-1, 1\}^m$ and $\rho(\cdot, \cdot)$ denotes the geodesic distance[10] on the sphere S^{n-1}.

(b) Use the VC law of large numbers (Theorem 8.3.15) to show that the following event occurs with probability at least 0.99:

$$\left| \frac{1}{m} d\left(\Phi(u), \Phi(v)\right) - \frac{1}{\pi}\rho(u, v) \right| \le C\sqrt{\frac{n}{m}} \quad \text{for all } u, v \in S^{n-1}.$$

8.27 ☕☕☕☕ (Random matrices with no moment assumptions) Most random matrix results assume something nice about the distribution – like subgaussian tails (Theorem 4.6.1) or at least a finite second moment (see Exercise 5.31). But now, you will use the impressive power of VC theory (specifically, Theorem 8.3.15) to prove that a tall random matrix is nicely invertible, even with no moment assumptions at all.

[9] Let's disallow zero values, redefining the sign as $\mathrm{sign}(t) = 1$ if $t \ge 0$ and $\mathrm{sign}(t) = -1$ if $t < 0$.

[10] For Hamming distance, recall Definition 4.2.14. The geodesic distance on the sphere is the length of the smallest arc that connects the two points.

(a) Let $\varepsilon, \delta > 0$ and let X be a random vector in $\mathbb{R}^n$ that satisfies the following *anti-concentration* assumption:

$$\mathbb{P}\{|\langle X, u\rangle| \geq \varepsilon\} \geq \delta \quad \forall u \in S^{n-1}. \tag{8.55}$$

Let A be an $m \times n$ random matrix whose rows are i.i.d. copies of the random vector X. Prove that if $m \geq C\delta^{-2}n$, then with probability at least 0.99, the smallest singular value of A is bounded away from zero:

$$s_n(A) \geq 0.99\varepsilon\sqrt{\delta m}.$$

(b) Demonstrate that the bound in part ((a)) is optimal. For every positive integers $m \geq n$ and positive reals ε and $\delta < 1/3$, find a random matrix A whose rows are i.i.d. copies of a random vector X that satisfies (8.55), and such that

$$\mathbb{E}\, s_n(A) \leq 1.01\varepsilon\sqrt{\delta m}.$$

8.28 ☕☕☕ (Glivenko-Cantelli = finite VC dimension) In Remark 8.3.19, we noted that if a Boolean class has finite VC dimension, it is uniform Glivenko-Cantelli. Show the converse: if a Boolean class is uniform Glivenko-Cantelli, then its VC dimension must be finite.

8.29 ☕☕ (VC dimension of $(f-h)^2$) Let $\mathcal{F}$ be a class of Boolean functions on some set Ω, and let h be a Boolean function on Ω. Show that the class

$$\{(f-h)^2 : f \in \mathcal{F}\}$$

has the same VC dimension as $\mathcal{F}$.

8.30 ☕☕☕ (Learning with random labels) In our setup of a learning problem (Section 8.4), we assumed that labels $T(X)$ were fully determined by X, which rarely holds in real life. For instance, it is unrealistic to think a diagnosis $T(X) \in 0, 1$ is entirely determined by health data X. More often, a label Y is a random variable that is just correlated with X. Try to extend the learning theory (up to Theorem 8.4.5) to training data of the form

$$(X_i, Y_i), \quad i = 1, \ldots, n,$$

where (X_i, Y_i) are independent copies of some random pair (X, Y) with $X \in \Omega$ and $Y \in \mathbb{R}$.

8.31 ☕☕☕ (Learning a Lipschitz function) We proved the generalization bound (Theorem 8.4.5) only for Boolean classes, but we often have data with real labels. Say we want to learn a Lipschitz function $T : [0,1] \to [0,1]$ from its values on a random sample $X_1, \ldots, X_n \sim \mathrm{Unif}[0,1]$. So, consider the hypothesis class

$$\mathcal{F} := \big\{f : [0,1] \to [0,1],\ \|f\|_{\mathrm{Lip}} \leq 1\big\}$$

and assume that $T \in \mathcal{F}$. Show that the empirical risk minimization algorithm from Section 8.4.2 has a generalization bound similar to Theorem 8.4.5:

$$\mathbb{E}\, R(f_n^*) \leq R(f^*) + \frac{C}{\sqrt{n}}.$$

8.32 ☕☕☕ (No learning with infinite VC dimension) Theorem 8.4.5 hints that to learn well, we need at least as many training points n as the VC dimension of the hypothesis class. Show that if $n < \frac{1}{2}\,\mathrm{vc}(\mathcal{F})$, then there exists a distribution on Ω where no algorithm can reliably learn the target function in $\mathcal{F}$. (Make this precise!)

8.33 ☕☕☕ (γ_2 functional is at least as good as Dudley) Show that γ_2 functional is bounded by Dudley integral:

$$\gamma_2(T,d) \le C \int_0^\infty \sqrt{\log \mathcal{N}(T,d,\varepsilon)}\, d\varepsilon.$$

8.34 ☕☕☕ (γ_2 functional can outperform Dudley) Sometimes the γ_2 functional (8.45) can be significantly smaller than the Dudley sum in (8.44). To see this, let's revisit example in Exercise 8.4:

$$T := \{0\} \cup \Big\{ \frac{e_k}{\sqrt{1+\log k}},\ k = 1,\ldots,n \Big\}.$$

(a) Construct an admissible sequence (T_k) to check that the γ_2 functional of T (with respect to the Euclidean metric) is bounded:

$$\gamma_2(T,d) = \inf_{(T_k)} \sup_{t \in T} \sum_{k=0}^\infty 2^{k/2} d(t, T_k) \le C.$$

(b) Show that Dudley sum is unbounded:

$$\inf_{(T_k)} \sum_{k=0}^\infty 2^{k/2} \sup_{t\in T} d(t,T_k) \to \infty \quad \text{as } n \to \infty.$$

8.35 ☕☕☕☕ (Generic chaining: a high-probability bound) Prove the result in Remark 8.5.4 by chaining with the first big leap:

$$t_0 \to \pi_\kappa(t) \to \pi_{\kappa+1}(t) \to \pi_{\kappa+2}(t) \to \cdots \to t$$

where κ is chosen so that $u \asymp 2^{\kappa/2}$.

8.36 ☕☕☕ (Empirical processes: a generic chaining bound) So far, we have bounded empirical processes for two kinds of function classes: Lipschitz (Theorem 8.2.3) and Boolean with finite VC dimension (Theorem 8.3.15). Now let's go general. Take any class $\mathcal{F}$ of real-valued functions on a domain Ω, and let $X, X_1, X_2, \ldots, X_n$ be i.i.d. random points in Ω. Let d be a metric on $\mathcal{F}$ satisfying

$$\|f(X) - g(X)\|_{\psi_2} \le d(f,g) \quad \text{for all } f, g \in \mathcal{F}.$$

(For instance, $d(f,g) = C\|f-g\|_{L^\infty}$ works.) Then show:

$$\mathbb{E} \sup_{f\in\mathcal{F}} \Big| \frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbb{E}\, f(X) \Big| \le \frac{C\gamma_2(F,d)}{\sqrt{n}}.$$

8.37 ☕☕ (Talagrand comparison inequality: a geometric form) Let's explore some useful variants of Corollary 8.5.8. Let $(X_x)_{x\in T}$ be a random process (not necessarily mean-zero) on a subset $T \subset \mathbb{R}^n$ such that[11] $X_0 = 0$. Assume that

$$\|X_x - X_y\|_{\psi_2} \le K\|x-y\|_2 \quad \text{for all } x, y \in T \cup \{0\}.$$

(a) (Expectation) Show that

$$\mathbb{E} \sup_{x\in T} |X_x| \le CK\gamma(T)$$

where $\gamma(T) = \mathbb{E}\sup_{x\in T} |\langle g, x\rangle|$ is the Gaussian complexity of T (see Section 7.5.3).

(b) (High-probability bound) Show that for every $u \ge 0$,

$$\sup_{x\in T} |X_x| \le CK\big(w(T) + u \cdot \mathrm{rad}(T)\big)$$

with probability at least $1 - 2e^{-u^2}$, where $\mathrm{rad}(T) = \sup_{x\in T} \|x\|_2$ is the radius of T.

(c) (Moments) Conclude that

$$\big(\mathbb{E} \sup_{x\in T} |X_x|^p\big)^{1/p} \le C\sqrt{p}\, K\gamma(T).$$

8.38 ☕☕ (Expected ℓ^p norm of a random vector) In Exercise 3.5, we gave optimal bounds on the expected ℓ^p norm of a random vector with independent subgaussian coordinates. Now, let's drop the independence assumption. For a subgaussian random vector X in $\mathbb{R}^N$ and $p \in [1, \infty]$, show that

$$\mathbb{E}\|X\|_p \le \begin{cases} CK\sqrt{p}N^{1/p}, & p \le \log N \\ CK\sqrt{\log N}, & p \ge \log N \end{cases}$$

where $K = \|X\|_{\psi_2}$.

8.39 ☕☕☕ (Gaussian Chevet inequality) Let A be an $m \times n$ random matrix with i.i.d. $N(0,1)$ entires. Let $T \subset \mathbb{R}^n$ and $S \subset \mathbb{R}^m$ be arbitrary bounded sets.

(a) Show that Theorem 8.6.1 holds with sharp constant 1:

$$\mathbb{E} \sup_{x\in T,\, y\in S} \langle Ax, y\rangle \le w(T)\,\mathrm{rad}(S) + w(S)\,\mathrm{rad}(T).$$

(b) Prove that the converse inequality:

$$\mathbb{E} \sup_{x\in T,\, y\in S} \langle Ax, y\rangle \ge c\,[w(T)\,\mathrm{rad}(S) + w(S)\,\mathrm{rad}(T)]\,.$$

8.40 ☕☕ (Chevet inequality: a high-probability version) Under the assumptions of Theorem 8.6.1, prove a tail bound for $\sup_{x\in T,\, y\in S}\langle Ax, y\rangle$.

8.41 ☕☕ (The $p \to q$ norm of a random matrix) Now we've got tools to handle all p, q – not just a few special cases like before (see Exercise 4.44)!

[11] If the set T does not contain the origin, simply define $X_0 := 0$.

(a) Let A be an $m \times n$ random matrix with independent, mean-zero, subgaussian rows A_i, and take any $p, q \in [1, \infty]$. Prove that

$$\mathbb{E}\|A\|_{p \to q} \leq CK \left[r(n,p)w(m,q) + r(m,q')w(n,p')\right],$$

where $K = \max_i \|A_i\|_{\psi_2}$ and

$$r(n,p) = \begin{cases} 1, & p \in [1,2] \\ n^{\frac{1}{2}-\frac{1}{p}}, & p \in [2,\infty], \end{cases} \qquad w(n,p) = \begin{cases} \sqrt{p}\, n^{1/p}, & p \in [1, \log n] \\ \sqrt{\log n}, & p \in [\log n, \infty] \end{cases}.$$

Here p' and q' are the usual Hölder conjugates of p and q as in (1.22).

(b) Show this bound is tight for Gaussian matrices: if A has i.i.d. $N(0,1)$ entires, then

$$\mathbb{E}\|A\|_{p \to q} \geq c \left[r(n,p)w(m,q) + r(m,q')w(n,p')\right].$$

9

Deviations of Random Matrices on Sets

How does an $m \times n$ random matrix act on a general set $T \subset \mathbb{R}^n$? In Section 9.1, we prove a core result – the matrix deviation inequality (Theorem 9.1.1), which we then apply it to a variety of high-dimensional problems, both familiar and new.

In Sections 9.2 we quickly deduce two-sided bounds for random matrices, sharper bounds for random projections of sets, covariance estimation for low-dimensional data, and the Johnson-Lindenstrauss lemma (even for infinite sets).

Section 9.3 features two elegant results: how random slicing shrinks high-dimensional sets (the M^* bound), and how it can totally miss them (the escape theorem).

Then in Sections 9.4–9.5, we apply these ideas to a basic data science problem: learning structured high-dimensional linear models.

In Section 9.6, we extend the matrix deviation inequality to any norms, and use it in Section 9.7 to sharpen Chevet inequality and deduce the Dvoretzky-Milman theorem, which says that random low-dimensional projections of high-dimensional sets look nearly round.

In the exercises, you will explore matrix and process-level deviation bounds (like in Exercise 9.5), high-dimensional estimation methods including *Lasso* for sparse regression (Exercises 9.20–9.21), the Garnaev-Gluskin theorem on random slicing the cross-polytope (Exercise 9.28) and other ℓ^p balls (Exercise 9.14), extensions of Johnson-Lindenstrauss lemma to general norms (Exercises 9.37–9.39), and more.

9.1 Matrix deviation inequality

Take an $m \times n$ random matrix A with independent, isotropic subgaussian rows. The concentration of norm (Theorem 3.1.1) tells us that, for any fixed vector $x \in \mathbb{R}^n$, the approximation

$$\|Ax\|_2 \approx \sqrt{m}\|x\|_2 \tag{9.1}$$

holds with high probability.

Now let's ask something bigger: is it true that with high probability, (9.1)

[0] This material will be published by Cambridge University Press as *High-Dimensional Probability, 2nd Edition* by Roman Vershynin. This pre-publication version is free to view and download for personal use only. Not for re-distribution, re-sale or use in derivative works. ©Roman Vershynin 202X.
If you want to be notified when the printed version is available, let me know: rvershyn@uci.edu

holds *simultaneously for many vectors* $x \in \mathbb{R}^n$? To quantify how many, pick any bounded set $T \subset \mathbb{R}^n$ and ask if the approximation holds simultaneously for all $x \in T$. It turns out that the maximal error is about $\gamma(T)$, the Gaussian complexity of T – a close cousin of Gaussian width from Section 7.5.3:

Theorem 9.1.1 (Matrix deviation inequality). *Let A be an $m \times n$ random matrix with independent, isotropic and subgaussian rows A_i. Then for any subset $T \subset \mathbb{R}^n$,*

$$\mathbb{E} \sup_{x \in T} \Big| \|Ax\|_2 - \sqrt{m}\|x\|_2 \Big| \le CK^2 \gamma(T),$$

where $\gamma(T)$ is the Gaussian complexity (see Section 7.5.3) and $K = \max_i \|A_i\|_{\psi_2}$.

Our plan is to deduce this result from Talagrand comparison inequality (Corollary 8.5.8). To do that, we just need to check that the random process

$$Z_x := \|Ax\|_2 - \sqrt{m}\|x\|_2 \tag{9.2}$$

indexed by vectors $x \in \mathbb{R}^n$ has subgaussian increments. Here is the claim:

Theorem 9.1.2 (Subgaussian increments). *Let A be an $m \times n$ random matrix with independent, isotropic and subgaussian rows A_i. Then the random process (9.2) has subgaussian increments:*

$$\|Z_x - Z_y\|_{\psi_2} \le CK^2 \|x - y\|_2 \quad \textit{for all } x, y \in \mathbb{R}^n. \tag{9.3}$$

Here $K = \max_i \|A_i\|_{\psi_2}$.

Once we have proved this theorem, we plug it into Talagrand comparison inequality (Corollary 8.5.8, or more precisely Exercise 8.37(a)) and get

$$\mathbb{E} \sup_{x \in T} |Z_x| \le CK^2 \gamma(T)$$

which gives us Theorem 9.1.1. So, all that's left is to prove Theorem 9.1.2 – and that is easier since it is only about *fixed* x and y.

Proof of Theorem 9.1.2 Although this argument is a bit longer than usual, we will make it easier by starting with simpler cases and building up from there.

Step 1: Unit vector x and zero vector y. If

$$\|x\|_2 = 1 \quad \text{and} \quad y = 0,$$

the inequality in (9.3) becomes

$$\Big\| \|Ax\|_2 - \sqrt{m} \Big\|_{\psi_2} \le CK^2. \tag{9.4}$$

The random vector $Ax \in \mathbb{R}^m$ has independent, subgaussian coordinates $\langle A_i, x \rangle$, which satisfy $\mathbb{E}\langle A_i, x \rangle^2 = 1$ by isotropy. So, (9.4) follows from the concentration of norm (Theorem 3.1.1).

Step 2: Unit vectors x, y and the squared process. Assume now that

$$\|x\|_2 = \|y\|_2 = 1.$$

In this case, the inequality in (9.3) becomes

$$\big\| \|Ax\|_2 - \|Ay\|_2 \big\|_{\psi_2} \le CK^2 \|x-y\|_2. \tag{9.5}$$

Since the *squared* ℓ^2 norm would be simpler to work with (no square root), let's first prove a version of (9.5) with squared norms. Here's a good guess for what it should look like: with high probability,

$$\begin{aligned} \|Ax\|_2^2 - \|Ay\|_2^2 &= \big(\|Ax\|_2 + \|Ay\|_2\big) \cdot \big(\|Ax\|_2 - \|Ay\|_2\big) \\ &\lesssim \sqrt{m} \cdot \|x-y\|_2. \end{aligned} \tag{9.6}$$

This seems reasonable because $\|Ax\|_2$ and $\|Ay\|_2$ are roughly $\sqrt{m}$ by (9.4), and we expect (9.5) to hold.

Let's go ahead and prove this. Expand the matrix-vector product:

$$\|Ax\|_2^2 - \|Ay\|_2^2 = \sum_{i=1}^m \big(\langle A_i, x\rangle^2 - \langle A_i, y\rangle^2\big) = \sum_{i=1}^m \langle A_i, x+y\rangle \langle A_i, x-y\rangle,$$

and divide both sides by $\|x-y\|_2$, getting

$$\Delta := \frac{\|Ax\|_2^2 - \|Ay\|_2^2}{\|x-y\|_2} = \sum_{i=1}^m \langle A_i, u\rangle\langle A_i, v\rangle, \tag{9.7}$$

where

$$u := x+y \quad \text{and} \quad v := \frac{x-y}{\|x-y\|_2}.$$

Our goal is to show that $|\Delta| \lesssim \sqrt{m}$ with high probability.

What do we see in (9.7)? A sum of *independent* random variables $\langle A_i, u\rangle\langle A_i, v\rangle$. They are *mean-zero*, because by construction we have

$$\langle A_i, u\rangle\langle A_i, v\rangle = \frac{\langle A_i, x\rangle^2 - \langle A_i, y\rangle^2}{\|x-y\|_2},$$

and by isotropy, $\mathbb{E}\,[\langle A_i, x\rangle^2 - \langle A_i, y\rangle^2] = 1 - 1 = 0$. And they are *subexponential*: Lemma 2.8.6 and the subgaussian assumption on A_i give

$$\|\langle A_i, u\rangle\langle A_i, v\rangle\|_{\psi_1} \le \|\langle A_i, u\rangle\|_{\psi_2} \cdot \|\langle A_i, v\rangle\|_{\psi_2} \le K\|u\|_2 \cdot K\|v\|_2 \le 2K^2$$

where in the last step we used that $\|u\|_2 \le \|x\|_2 + \|y\|_2 \le 2$ and $\|v\|_2 = 1$. So we can apply Bernstein inequality (Theorem 2.9.1) and get

$$\mathbb{P}\{|\Delta| \ge t\sqrt{m}\} \le 2\exp\left[-c\min\left(\frac{t^2}{K^4}, \frac{t\sqrt{m}}{K^2}\right)\right] \le 2\exp\left(-\frac{c_1 t^2}{K^4}\right) \tag{9.8}$$

for any[1] $0 \le t \le \sqrt{m}$.

Step 3: Unit vectors x, y and the original process. Now let's get rid of the squares and prove the original inequality (9.5) for all unit vectors x and y.

[1] To get the last inequality in (9.8), recall that K is bounded below by a positive absolute constant (why?) and choose the constant $c > 0$ small enough.

Using the definition of subgaussian norm (recall Proposition 2.6.6(i) and Remark 2.6.3), inequality (9.5) becomes

$$p(s) := \mathbb{P}\Big\{\frac{|\,\|Ax\|_2 - \|Ay\|_2|}{\|x-y\|_2} \ge s\Big\} \le 4\exp\Big(-\frac{cs^2}{K^4}\Big) \quad \forall s > 0. \tag{9.9}$$

(The constant 4 instead of 2 will give us a little more room for maneuver.)

Case 1: $s \le 2\sqrt{m}$. Let's use the result of Step 2. To apply it, multiply both sides of the inequality that defines $p(s)$ by $\|Ax\|_2 + \|Ay\|_2$, and recall the definition (9.7) of Δ to get

$$p(s) = \mathbb{P}\{|\Delta| \ge s\,(\|Ax\|_2 + \|Ay\|_2)\} \le \mathbb{P}\{|\Delta| \ge s\|Ax\|_2\}.$$

We know from (9.4) that $\|Ax\|_2 \approx \sqrt{m}$ with high probability. So it makes sense to consider two cases: the likely case where $\|Ax\|_2 \ge \sqrt{m}/2$ and thus $|\Delta| \ge s\sqrt{m}/2$, and the unlikely case where $\|Ax\|_2 < \sqrt{m}/2$ (and where we drop the clause about Δ, only increasing the probability). This leads to

$$p(s) \le \mathbb{P}\Big\{|\Delta| \ge \frac{s\sqrt{m}}{2}\Big\} + \mathbb{P}\{\|Ax\|_2 < \frac{\sqrt{m}}{2}\} =: p_1(s) + p_2(s).$$

The result of Step 2 handles the likely case:

$$p_1(s) \le 2\exp\left(-\frac{cs^2}{K^4}\right),$$

and the result (9.4) of Step 1 together with the triangle inequality handle the unlikely case:

$$p_2(s) \le \mathbb{P}\Big\{|\,\|Ax\|_2 - \sqrt{m}| > \frac{\sqrt{m}}{2}\Big\} \le 2\exp\Big(-\frac{cs^2}{K^4}\Big).$$

Adding up the two, we get the desired bound

$$p(s) \le 4\exp\Big(-\frac{cs^2}{K^4}\Big).$$

Case 2: $s > 2\sqrt{m}$. By triangle inequality, $|\,\|Ax\|_2 - \|Ay\|_2| \le \|A(x-y)\|_2$, so

$$\begin{aligned} p(s) &\le \mathbb{P}\{\|Au\|_2 \ge s\} \quad \Big(\text{where } u = \frac{x-y}{\|x-y\|_2} \text{ as before}\Big) \\ &\le \mathbb{P}\{\|Au\|_2 - \sqrt{m} \ge s/2\} \quad (\text{since } s > 2\sqrt{m}) \\ &\le 2\exp\Big(-\frac{cs^2}{K^4}\Big) \quad (\text{by the result (9.4) of Step 1 for } u \text{ instead of } x). \end{aligned}$$

Therefore, in either case we obtain the desired bound (9.9).

Step 4: Full generality. Finally, let's show (9.3) for arbitrary $x, y \in \mathbb{R}^n$. By scaling, we can assume without loss of generality that

$$\|x\|_2 = 1 \quad \text{and} \quad \|y\|_2 \ge 1.$$

Project y onto the unit sphere, i.e. consider $\bar{y} := y/\|y\|_2$ (see Figure 9.1), then use triangle inequality:

$$\|Z_x - Z_y\|_{\psi_2} \le \|Z_x - Z_{\bar{y}}\|_{\psi_2} + \|Z_{\bar{y}} - Z_y\|_{\psi_2}.$$

Since both x and $\bar{y}$ are unit vectors, the result of Step 3 handles the first term:

$$\|Z_x - Z_{\bar{y}}\|_{\psi_2} \le CK^2 \|x - \bar{y}\|_2.$$

To handle the second term, note that $\bar{y}$ and y are collinear vectors. So, by homogeneity,

$$\|Z_{\bar{y}} - Z_y\|_{\psi_2} = \|\bar{y} - y\|_2 \cdot \|Z_{\bar{y}}\|_{\psi_2}.$$

(Check!) Now, since $\bar{y}$ is a unit vector, the result of Step 1 gives $\|Z_{\bar{y}}\|_{\psi_2} \le CK^2$. Combining the two terms, we conclude that

$$\|Z_x - Z_y\|_{\psi_2} \le CK^2 \left(\|x - \bar{y}\|_2 + \|\bar{y} - y\|_2 \right). \tag{9.10}$$

At first this looks bad – we want to bound the right-hand side by $\|x - y\|_2$, but triangle inequality usually works the other way! Luckily, in our case (where $\bar{y}$ is y projected onto the unit sphere, see Figure 9.1), the triangle inequality can be approximately reversed:

$$\|x - \bar{y}\|_2 + \|\bar{y} - y\|_2 \le \sqrt{2}\|x - y\|_2$$

(check this in Exercise 9.1). Plugging this into (9.10), we get the desired bound:

$$\|Z_x - Z_y\|_{\psi_2} \le \sqrt{2} CK^2 \|x - y\|_2,$$

which proves Theorem 9.1.2. □

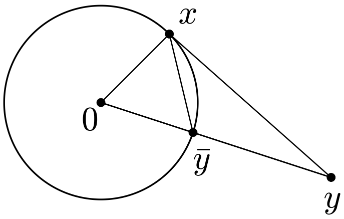


Figure 9.1 The triangle inequality can be approximately reversed for these three vectors: $\|x - \bar{y}\|_2 + \|\bar{y} - y\|_2 \le \sqrt{2}\|x - y\|_2$.

Remark 9.1.3 (Matrix deviations from the mean)**.** A quick centering trick turns Theorem 9.1.1 into a deviation inequality around the mean $\mathbb{E}\|Ax\|_2$; see Exercise 9.2 to check it.

Remark 9.1.4 (Matrix deviations: a high-probability bound)**.** We only stated Theorem 9.1.1 as an expectation bound, but thanks to the high-probability version of Talagrand inequality (see Exercise 8.37(b)), it automatically upgrades to a tail bound. For any $u \geq 0$, the event

$$\sup_{x \in T} \Big| \|Ax\|_2 - \sqrt{m}\|x\|_2 \Big| \leq CK^2 \Big[w(T) + u \cdot \mathrm{rad}(T) \Big] \tag{9.11}$$

holds with probability at least $1 - 2\exp(-u^2)$. Here $\mathrm{rad}(T) = \sup_{x \in T} \|x\|_2$ is the radius of T. Can you see why (9.11) implies the expectation bound?

Remark 9.1.5 (Matrix deviations of squares)**.** If you are interested in deviations of the quadratic process $\|Ax\|_2^2$, one can easily deduce this from Theorem 9.1.1:

$$\mathbb{E} \sup_{x \in T} \Big| \|Ax\|_2^2 - m\|x\|_2^2 \Big| \leq CK^4 \gamma(T)^2 + CK^2 \sqrt{m}\, \mathrm{rad}(T) \gamma(T).$$

Check this in Exercise 9.3.

To practice, extend matrix deviation inequality to empirical processes (Exercise 9.5) and prove a version for random projections (Exercise 9.6 – this one is challenging!).

9.2 Random matrices, covariance estimation, and Johnson-Lindenstrauss

Matrix deviation inequality has lots of useful consequences. We will go over a few of them through the rest of the chapter.

9.2.1 Singular values of random matrices

Applying the matrix deviation inequality for the unit Euclidean sphere $T = S^{n-1}$ gives us the singular value bounds from Section 4.6.

Here is the quick check: since for the sphere we have

$$\mathrm{rad}(T) = 1 \quad \text{and} \quad w(T) \leq \sqrt{n},$$

the matrix deviation inequality (9.11) shows that the event

$$\sqrt{m} - CK^2(\sqrt{n} + u) \leq \|Ax\|_2 \leq \sqrt{m} + CK^2(\sqrt{n} + u) \quad \forall x \in S^{n-1}$$

holds with probability at least $1 - 2\exp(-u^2)$. Using (4.14), this translates to

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

recovering Theorem 4.6.1 we proved earlier in a different way.

9.2.2 Random projections of sets

Another immediate application of matrix deviation inequality is a (sharper!) version of the random projection bound from Section 7.6:

Proposition 9.2.1 (Sizes of random projections of sets)**.** *Let $T \subset \mathbb{R}^n$ be a bounded set, and let A be an $m \times n$ matrix with independent, isotropic and subgaussian rows A_i. Then the scaled matrix $P = \frac{1}{\sqrt{n}} A$ (a "subgaussian projection") satisfies*

$$\mathbb{E}\,\mathrm{diam}(PT) \le \sqrt{\frac{m}{n}}\,\mathrm{diam}(T) + CK^2 w_s(T).$$

Here $K = \max_i \|A_i\|_{\psi_2}$ and $w_s(T)$ is the spherical width of T (recall Section 7.5.1).

This is a bit sharper that our older bounds (Theorem 7.6.1 and Exercise 7.25) – there is no constant factor in front of $\sqrt{m/n}$.

Proof Theorem 9.1.1 implies via triangle inequality:

$$\mathbb{E} \sup_{x \in T} \|Ax\|_2 \le \sqrt{m} \sup_{x \in T} \|x\|_2 + CK^2 \gamma(T),$$

which we can rewrite in terms of the radii of AT and T:

$$\mathbb{E}\,\mathrm{rad}(AT) \le \sqrt{m}\,\mathrm{rad}(T) + CK^2 \gamma(T).$$

Apply this bound for the difference set $T - T$ instead of T to get

$$\mathbb{E}\,\mathrm{diam}(AT) \le \sqrt{m}\,\mathrm{diam}(T) + 2CK^2 w(T),$$

where we used Lemma 7.5.11(a) to pass from Gaussian complexity to Gaussian width. Divide both sides by $\sqrt{n}$ to complete the proof. □

Now try this: upgrade the expectation bound in Proposition 9.2.1 to a high-probability one (Exercise 9.7) and analyze the actual random projections (rather than "Gaussian" ones) in Exercise 9.8.

9.2.3 Covariance estimation for low-dimensional distributions

Let's revisit the covariance estimation problem from Section 4.7, where we want to estimate the covariance matrix $\Sigma = \mathbb{E}\, XX^\mathsf{T}$ of an n-dimensional distribution from m i.i.d. samples using the sample covariance matrix $\Sigma_m = \frac{1}{m} \sum_{i=1}^m X_i X_i^\mathsf{T}$.

In general, $m = O(n \log n)$ samples are enough (Section 5.6), but for subgaussian distributions, $m = O(n)$ is enough (Section 4.7).

It gets even better for approximately low-dimensional distributions. If a distribution concentrates near an r-dimensional subspace, $m = O(r \log n)$ samples suffice (Remark 5.6.3). Now we will show that for subgaussian distributions, $m = O(r)$ samples suffices:

Theorem 9.2.2 (Covariance estimation for low-dimensional distributions)**.** *Let X be a subgaussian random vector in $\mathbb{R}^n$. More specifically, assume that there exists $K \ge 1$ such that*

$$\|\langle X, x\rangle\|_{\psi_2} \le K \|\langle X, x\rangle\|_{L^2} \quad \textit{for any } x \in \mathbb{R}^n.$$

Then, for every positive integer m, we have

$$\mathbb{E}\|\Sigma_m - \Sigma\| \le CK^4\Big(\sqrt{\frac{r}{m}} + \frac{r}{m}\Big)\,\|\Sigma\|,$$

where $r = \operatorname{tr}(\Sigma)/\|\Sigma\|$ is the effective rank of Σ – a measure of the effective dimension of the data (see Remark 5.6.3).

Proof We start as in Theorem 4.7.1 by bringing the distribution to the isotropic position: $X = \Sigma^{1/2}Z$ and $X_i = \Sigma^{1/2}Z_i$ where Z and Z_i are isotropic, and

$$\begin{aligned}
\|\Sigma_m - \Sigma\| &= \|\Sigma^{1/2}R_m\Sigma^{1/2}\| \quad (\text{where } R_m = \tfrac{1}{m}\textstyle\sum_{i=1}^m Z_iZ_i^\mathsf{T} - I_n)\\
&= \max_{x\in S^{n-1}}|x^\mathsf{T}\Sigma^{1/2}R_m\Sigma^{1/2}x| \quad (\text{see Remark 4.1.12})\\
&= \max_{x\in T}|x^\mathsf{T}R_mx| \quad (\text{if we define the ellipsoid } T := \Sigma^{1/2}S^{n-1})\\
&= \max_{x\in T}\Big|\frac{1}{m}\sum_{i=1}^m\langle Z_i,x\rangle^2 - \|x\|_2^2\Big| \quad (\text{by definition of } R_m)\\
&= \frac{1}{m}\max_{x\in T}\Big|\|Ax\|_2^2 - m\|x\|_2^2\Big|,
\end{aligned}$$

where A is the $m\times n$ matrix with rows Z_i. As in the proof of Theorem 4.7.1, Z_i are isotropic, and satisfy $\|Z_i\|_{\psi_2} \lesssim 1$. (For simplicity, let's hide the dependence on K in this argument.) This allows us to apply matrix deviation inequality for A (in the form given in Exercise 9.3), which gives

$$\mathbb{E}\|\Sigma_m - \Sigma\| \lesssim \frac{1}{m}\big(\gamma(T)^2 + \sqrt{m}\,\operatorname{rad}(T)\gamma(T)\big).$$

The radius and Gaussian complexity of the ellipsoid $T = \Sigma^{1/2}S^{n-1}$ satisfy

$$\operatorname{rad}(T) = \|\Sigma\|^{1/2} \quad\text{and}\quad \gamma(T) \le (\operatorname{tr}\Sigma)^{1/2}$$

(check!). So,

$$\mathbb{E}\|\Sigma_m - \Sigma\| \lesssim \frac{1}{m}\Big(\operatorname{tr}\Sigma + \sqrt{m\|\Sigma\|\operatorname{tr}\Sigma}\Big).$$

Substitute $\operatorname{tr}(\Sigma) = r\|\Sigma\|$ and simplify the bound to complete the proof. □

Remark 9.2.3 (Covariance estimation: a high-probability guarantee)**.** Just like before (see Remarks 4.7.3 and 5.6.5), we can upgrade the expectation bound in Theorem 9.2.2 to a high-probability one. For any $u \ge 0$, we have

$$\|\Sigma_m - \Sigma\| \le CK^4\Big(\sqrt{\frac{r+u}{m}} + \frac{r+u}{m}\Big)\,\|\Sigma\|$$

with probability at least $1 - 2e^{-u}$. Try proving this in Exercise 9.9!

9.2.4 Johnson-Lindenstrauss lemma for infinite sets

The matrix deviation inequality quickly recovers Johnson-Lindenstrauss lemma from Section 5.3 – and extends it to general, possibly infinite, sets.

To get a version of Johnson-Lindenstrauss (Theorem 5.3.1) from matrix deviation, fix any N-point set $\mathcal{X} \in \mathbb{R}^n$ and consider the normalized differences:

$$T := \Big\{ \frac{x-y}{\|x-y\|_2} : \; x, y \in \mathcal{X} \text{ distinct} \Big\}.$$

The Gaussian complexity of T satisfies

$$\gamma(T) \le C\sqrt{\log N} \tag{9.12}$$

(check this as in Example 7.5.9). Matrix deviation inequality (Theorem 9.1.1) shows that with high probability,[2]

$$\sup_{x,y\in\mathcal{X}} \left| \frac{\|Ax-Ay\|_2}{\|x-y\|_2} - \sqrt{m} \right| \lesssim \sqrt{\log N}.$$

Rearranging the terms, rewrite this as follows: the random matrix $Q := \frac{1}{\sqrt{m}}A$ is an approximate isometry on $\mathcal{X}$, i.e.

$$(1-\varepsilon)\|x-y\|_2 \le \|Qx-Qy\|_2 \le (1+\varepsilon)\|x-y\|_2 \quad \text{for all } x, y \in \mathcal{X},$$

for some $\varepsilon \asymp \sqrt{\log(N)/m}$. Equivalently, if we fix $\varepsilon > 0$ and choose

$$m \gtrsim \varepsilon^{-2} \log N,$$

then with high probability Q is an ε-isometry on $\mathcal{X}$, which recovers a version of the classical Johnson-Lindenstrauss Lemma (Theorem 5.3.1).

The argument we just gave does not care if $\mathcal{X}$ is finite or not – all that matters is the Gaussian width. So we can extend Johnson-Lindenstrauss to any set:

Lemma 9.2.4 (Additive Johnson-Lindenstrauss lemma)**.** *Let $\mathcal{X} \subset \mathbb{R}^n$ be a bounded set, and let A be an $m \times n$ random matrix with independent, isotropic and subgaussian rows A_i. Then, with high probability (say,* 0.99*), the scaled matrix $Q = \frac{1}{\sqrt{m}}A$ satisfies*

$$\big| \|Qx-Qy\|_2 - \|x-y\|_2 \big| \le \delta \quad \textit{for all } x, y \in \mathcal{X}$$

where $\delta = CK^2 w(\mathcal{X})/\sqrt{m}$ and $K = \max_i \|A_i\|_{\psi_2}$.

Proof Apply matrix deviation inequality (Theorem 9.1.1) for the set of differences $T = \mathcal{X} - \mathcal{X} = \{x - y : \; x \in X, \; y \in Y\}$. Then, with high probability,

$$\sup_{x,y\in\mathcal{X}} \Big| \|Ax-Ay\|_2 - \sqrt{m}\|x-y\|_2 \Big| \le CK^2\gamma(\mathcal{X}-\mathcal{X}) = 2CK^2 w(\mathcal{X}),$$

thanks to Lemma 7.5.11(a). Divide both sides by $\sqrt{m}$ to complete the proof. □

[2] To keep things simple, we will settle for 99% success probability (via Markov's inequality) and ignore the dependence on the subgaussian norm K. You will make this precise in Exercise 9.10.

Unlike the classical Johnson-Lindenstrauss Lemma for finite sets (Theorem 5.3.1), which gives a relative error, here we get an absolute error δ. It is a small difference – but in general, a necessary one (see Exercise 9.11).

Remark 9.2.5 (Effective dimension)**.** To better understand the additive Johnson-Lindenstrauss lemma, let's restate it using the effective dimension of the data $d(\mathcal{X}) \asymp w(\mathcal{X})^2 / \operatorname{diam}(\mathcal{X})^2$ (see Definition 7.5.12). If we choose

$$m \gtrsim \varepsilon^{-2} d(T)$$

(ignoring the dependence on K for simplicity), then we can make $\delta = \varepsilon \operatorname{diam}(\mathcal{X})$ in 9.2.4, so Q preserves distances up to a small fraction of diameter – in other words, it reduces the dimension of the data down to its effective dimension.

9.3 Random sections: the M^* bound and escape theorem

Here is a surprising high-dimensional fact: if you slice a convex set $T \subset \mathbb{R}^n$ with a random subspace E of codimension m, the slice $T \subset E$ is often tiny – even when $m \ll n$ and E is nearly full-dimensional! Let's see how this follows from the matrix deviation inequality.

9.3.1 The M^* bound

It is handy to model a random subspace E as the kernel of an $m \times n$ random matrix: $E = \ker A$. We always have

$$\dim(E) \geq n - m,$$

and if A has a continuous distribution, $\dim(E) = n - m$ almost surely.

A great example is a Gaussian matrix A with i.i.d. $N(0,1)$ entries – by rotation invariance, $E = \ker(A)$ is uniformly distributed in the Grassmanian:

$$E \sim \operatorname{Unif}(G_{n,n-m}).$$

Theorem 9.3.1 (M^* bound)**.** *Let $T \subset \mathbb{R}^n$ be a bounded set, and let A be an $m \times n$ random matrix with independent, isotropic and subgaussian rows A_i. Then the random subspace $E = \ker A$ satisfies*

$$\mathbb{E} \operatorname{diam}(T \cap E) \leq \frac{CK^2 w(T)}{\sqrt{m}},$$

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

Proof Apply Theorem 9.1.1 for $T - T$:

$$\mathbb{E} \sup_{x,y \in T} \Big| \|Ax - Ay\|_2 - \sqrt{m} \|x - y\|_2 \Big| \leq CK^2 \gamma(T - T) = 2CK^2 w(T),$$

by Lemma 7.5.11(a). Considering only the points x, y in the kernel of A makes

$\|Ax - Ay\|_2$ disappear since $Ax = Ay = 0$. Divide on both sides by $\sqrt{m}$ to get

$$\mathbb{E} \sup_{x,y \in T \cap \ker A} \|x - y\|_2 \leq \frac{CK^2 w(T)}{\sqrt{m}},$$

which is exactly what we claimed. □

Example 9.3.2 (The cross-polytope)**.** Let's apply the M^* bound to the cross-polytope B_1^n – the unit ball of the ℓ^1 norm. Since its Gaussian width is roughly $\sqrt{\log n}$ due to (7.19), we get

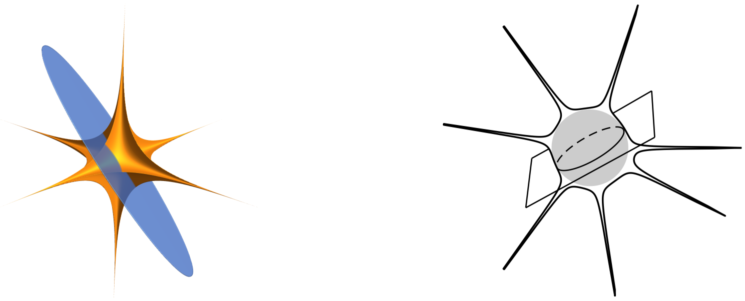
$$\mathbb{E} \operatorname{diam}(B_1^n \cap E) \lesssim \sqrt{\frac{\log n}{m}}.$$

For example, if $m = 0.01n$, then

$$\mathbb{E} \operatorname{diam}(T \cap E) \lesssim \sqrt{\frac{\log n}{n}}. \tag{9.13}$$

So, a random $0.99n$-dimensional slice of a cross-polytope is tiny!

How can this be? For intuition, recall the *hyperbolic sketch* of the cross-polytope (Figure 7.4a): the "bulk" of B_1^n is concentrated near the inscribed ball of radius $1/\sqrt{n}$, while the rest stretches out into long, thin "spikes" along the coordinate axes. A random subspace E probably misses those spikes and cuts through the bulk (Figure 9.2a), so the slice ends up with diameter about $O(1/\sqrt{n})$, maybe with a log factor[3] as in (9.13). This intuition extends to general convex sets too (Figure 9.2b).



(a) The octahedron B_1^n

(b) General convex set

Figure 9.2 Slicing a convex set with a random subspace.

Remark 9.3.3 (Effective dimension)**.** To get more intuition, write the M^* bound using the effective dimension $d(T) \asymp w(T)^2/\operatorname{diam}(T)^2$ (see Definition 7.5.12). The M^* bound shows that slicing shrinks the diameter:

$$\mathbb{E} \operatorname{diam}(T \cap E) \leq 0.01 \cdot \operatorname{diam}(T)$$

[3] You will remove the logarithmic factor from (9.13) later in Exercise 9.28.

as long as[4] $m \gtrsim d(T)$. Since $\dim(E) = n - m$, this condition is equivalent to

$$\dim(E) + cd(T) \le n.$$

That lines up with linear algebra intuition: if T is a centered Euclidean ball in some subspace $F \subset \mathbb{R}^n$, slicing can shrink the diameter of T only when $\dim E + \dim F \le n$ (why?).

To practice, extend the M^* bound to affine sections (Exercise 9.12), prove its high-probability version (Exercise 9.13), and compute the diameter of a random slice of the ℓ^p ball (Exercise 9.14).

9.3.2 The escape theorem

When does a random subspace E miss a given set T entirely with high probability? Not if T contains the origin – but if T lies on the unit sphere (see Figure 9.3), then it does as long as the codimension of E is not too small:

Theorem 9.3.4 (Escape theorem)**.** *Let $T \subset S^{n-1}$ be any set, and let A be an $m \times n$ matrix with independent, isotropic and subgaussian rows A_i. If*

$$m \ge CK^4 w(T)^2, \tag{9.14}$$

then the random subspace $E = \ker A$ satisfies

$$T \cap E = \varnothing$$

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 (see Remark 9.1.4): with probability at least $1 - 2\exp(-u^2)$,

$$\sup_{x \in T} \Big| \|Ax\|_2 - \sqrt{m} \Big| \le C_1 K^2 (w(T) + u). \tag{9.15}$$

Suppose (9.15) occurs. If $T \cap E \ne \varnothing$, then for any $x \in T \cap E$ we have $Ax = 0$, so

$$\sqrt{m} \le C_1 K^2 (w(T) + u).$$

Set $u = \sqrt{m}/(2C_1K^2)$ and simplify this bound to get

$$\sqrt{m} \le 2C_1 K^2 w(T),$$

which contradicts the assumption if the absolute constant C is large enough. So, with that choice of u, the event (9.15) implies $T \cap E = \varnothing$. Done! □

Now show the tightness of the escape theorem (Exercise 9.15 and prove its cousin for a point set instead of a subspace (Exercise 9.16).

[4] We suppress the dependence on K for simplicity.

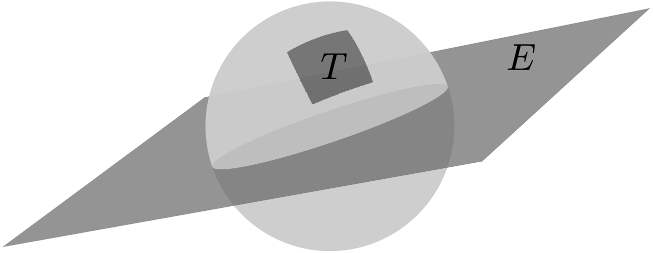


Figure 9.3 The escape theorem quantifies when a random subspace E misses a given subset T of the sphere.

9.4 Application: high-dimensional linear models

Let's use our tools on a classic data science problem: learning a linear model in high dimensions. Imagine there is an unknown vector $x \in \mathbb{R}^n$ that we want to learn from m linear and possibly noisy observations, or measurements:

$$y_i = \langle A_i, x \rangle + w_i, \quad i = 1, \ldots, m.$$

Here $A_i \in \mathbb{R}^n$ are known, and w_i are unknown numbers representing noise. In matrix form, it is

$$y = Ax + w, \tag{9.16}$$

where A is an known $m \times n$ matrix and $w \in \mathbb{R}^m$ is the unknown noise (see Figure 9.4). The goal is to recover x from y and A as accurately as possible.

We assume the rows A_i of A are random and independent – this is reasonable in many statistical settings (think about i.i.d. observations), and perfect for applying tools from high-dimensional probability.

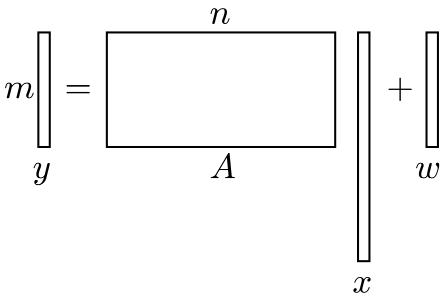


Figure 9.4 A high-dimensional linear model: recover x from $y = Ax + w$.

Example 9.4.1 (Audio sampling)**.** In signal processing, x could be a digitized audio signal, and y the result of sampling it at m random time points (see Figure 9.5).

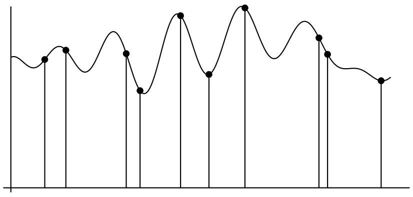


Figure 9.5 Signal recovery problem in audio sampling: recover an audio signal x from its values at m random time points.

Example 9.4.2 (Linear regression)**.** A core problem in statistics is linear regression, where we want to learn a linear relationship between n predictor variables and a response variable from m samples. It is written as

$$Y = X\theta + w$$

where X is an $m \times n$ matrix of predictors, $Y \in \mathbb{R}^m$ is the vector of responses, $\theta \in \mathbb{R}^n$ is the parameter vector we are trying to learn, and w is noise. For example, in genetics, you might want to predict a disease from gene expression data. You collect data from m patients, where X_{ij} is how active gene j is in patient i, and Y_i indicates whether the patient has the disease. The goal is to learn the parameter vector θ, which tells you how each gene influences the disease.

Remark 9.4.3 (The high-dimensional regime)**.** In modern problems, we often have less data than parameters:

$$m \ll n,$$

For example, in a typical genetic study (see Example 9.4.2), there might be ~ 100 patients but $\sim 10,000$ genes. In this *high-dimensional* setting, even solving the noiseless problem $y = Ax$ is impossible – there are too many possible solutions since they live in a large subspace of dimension at least $n - m$.

Still, we might be able to recover x if we have some prior information about its structure – something we know or believe about x. We can write this as:

$$x \in T \tag{9.17}$$

for some known set $T \subset \mathbb{R}^n$. For instance, if x is sparse (mostly zeros), we pick T as the set of all sparse vectors, or x is a band-limited signal, we pick T as a set of band-limited signals, and so on. Let's see how this idea helps.

9.4.1 Constrained recovery

Consider the noiseless case first:

$$y = Ax, \quad x \in T.$$

How do we solve this high-dimensional constrained linear problem?

A simple idea is just to pick any vector $x' \in T$ that matches the observations:

$$\text{find } x' : \; y = Ax', \quad x' \in T. \tag{9.18}$$

If T is convex, this is a convex program, and many algorithms exists to numerically solve it. Now let's check how accurate this solution is:

Theorem 9.4.4 (Constrained recovery)**.** *Suppose the rows A_i of A are independent, isotropic and subgaussian random vectors. Then any solution $\widehat{x}$ of the program* (9.18) *satisfies*

$$\mathbb{E}\|\widehat{x} - x\|_2 \le \frac{CK^2 w(T)}{\sqrt{m}}, \tag{9.19}$$

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

Proof Since $x, \widehat{x} \in T$ and $Ax = A\widehat{x} = y$, we have

$$x, \widehat{x} \in T \cap E_x, \quad \text{where} \quad E_x = x + \ker A$$

(see Figure 9.6). Then the M^* bound (in the form of Exercise 9.12) gives

$$\mathbb{E}\|\widehat{x} - x\|_2 \le \mathbb{E}\,\text{diam}(T \cap E_x) \le \frac{CK^2 w(T)}{\sqrt{m}}. \quad \square$$

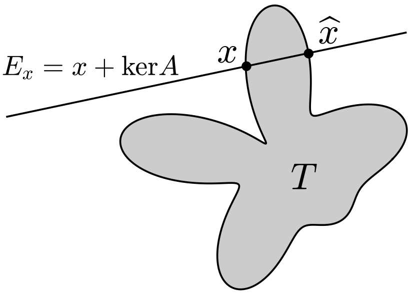


Figure 9.6 The geometry of a constrained high-dimensional linear problem.

Remark 9.4.5 (Effective dimension)**.** To get some intuition, rewrite the accuracy guarantee (9.19) using the effective dimension $d(T) \asymp w(T)^2/\,\text{diam}(T)^2$ (see Definition 7.5.12). We get a non-trivial error bound

$$\mathbb{E}\|\widehat{x} - x\|_2 \le 0.01 \cdot \text{diam}(T)$$

as long as the number of observations satisfies[5]

$$m \gtrsim d(T).$$

Since $d(T)$ can be much smaller than the ambient dimension n, recovery is often possible even in the high-dimensional regime where $m \ll n$.

[5] We suppress the dependence on K for simplicity.

Remark 9.4.6 (Convex relaxation)**.** If T is not convex, we can just replace it with its convex hull $\operatorname{conv}(T)$. This turns (9.18) into a convex (and thus tractable) problem. The recovery guarantees from Theorem 9.4.4 do not change, since $w(\operatorname{conv}(T)) = w(T)$ by Proposition 7.5.2(c).

The approach we covered is pretty flexible. Try it out yourself: analyze the mean squared error (Exercise 9.17), recovery via optimization (Exercise 9.18) and the noisy case (9.16) (Exercise 9.19).

Remark 9.4.7 (Unconstrained optimization)**.** Forcing strict rules on the solution like $y = Ax'$ and $x' \in T$ in (9.18) can be too rigid – noise or a bad choice of T might mean no solution exists. A safer move is to relax the rules and just *penalize* how much they are broken.

So, given noisy measurements

$$y = Ax + w,$$

a good way to recover x is by solving the *unconstrained* convex problem:

$$\text{minimize } \|y - Ax'\|_2^2 + \lambda \|x'\|_T \quad \text{over all } x' \in \mathbb{R}^n, \tag{9.20}$$

where $\|\cdot\|_T$ is any norm you like and $\lambda > 0$ tunes the tradeoff between fitting the observations y and keeping the solution x' structured (small $\|x\|_T$). Adjusting λ lets you decide which to prioritize.

Now, try Exercise 9.20 to prove the recovery guarantee: if A is an $m \times n$ random matrix, and λ is chosen well, then the solution $\widehat{x}$ satisfies

$$\mathbb{E}\|\widehat{x} - x\|_2 \lesssim \frac{w(T)\|x\|_T + \|w\|_2}{\sqrt{m}},$$

where T is the unit ball of $\|\cdot\|_T$. In short: if x is well structured (small $\|x\|_T$) and the noise w is small ($o(1)$ per observation), then you can recover x accurately from $m \asymp d(T)$ observations, where $d(T)$ is the effective dimension of T.

If this feels a bit abstract, two examples will make it clear.

9.4.2 Example: sparse recovery

Sometimes we believe that x is *sparse* – most of its entries are zero or nearly zero. For instance, in the genetic study (Example 9.4.2), maybe only ~ 10 genes really affect a disease, and we want to find those. In other applications, x may be sparse in some basis: the audio signal from Example 9.4.1 is not sparse in time (see Figure 9.5), but its Fourier transform could be sparse – band-limited to a small frequency range.

We can quantify the sparsity of $x \in \mathbb{R}^n$ by the number of nonzero entries:

$$\|x\|_0 \coloneqq |\operatorname{supp}(x)| = |\{i : x_i \neq 0\}|, \tag{9.21}$$

and we say that x is s-sparse if $\|x\|_0 \le s$. The "ℓ^0 norm" $\|x\|_0$ is not really a norm, but is a limit of ℓ^p norms as $p \to 0$ (see Exercise 9.23).

A quick dimension count shows that we can recover x from $y = Ax$ if A is in general position and we have enough observations: $m \ge 2\|x\|_0$ (try Exercise 9.22). Sounds great – we can recover a sparse vector from only a few observations! The catch? It is computationally hard unless we already know the support of x. Without that, searching over all possible supports is too expensive – there are $\binom{n}{s} \ge 2^s$ subsets to check. Fortunately, *random* observations can help. Here is how.

Let's try to use the tools from Section 9.4.1. If we believe x is s-sparse, it is tempting to pick as our prior

$$T = \{x \in \mathbb{R}^n : \ \|x\|_0 \le s\},$$

but T is highly non-convex, so solving (9.18) would be computationally hard.

Here is a simple fix: *replace the "ℓ^0 norm" by the ℓ^1 norm* – the closest ℓ^p that is actually a norm (see Exercise 9.23). Since s-sparse vectors with $\|x\|_2 \le 1$ satisfy $\|x\|_1 \le \sqrt{s}$ (check!), it makes sense to pick the convex set

$$T := \sqrt{s} B_1^n \tag{9.22}$$

as our prior. The recovery program (9.18) becomes

$$\text{find } x' : \ y = Ax', \quad \|x'\|_1 \le \sqrt{s}, \tag{9.23}$$

which is convex and computationally tractable. And Theorem 9.4.4 gives:

Corollary 9.4.8 (Sparse recovery)**.** *Suppose the rows A_i of A are independent, isotropic and subgaussian random vectors. Assume an unknown s-sparse vector $x \in \mathbb{R}^n$ satisfies $\|x\|_2 \le 1$. Then any solution $\widehat{x}$ of the program* (9.23) *satisfies*

$$\mathbb{E}\|\widehat{x} - x\|_2 \le CK^2 \sqrt{\frac{s \log n}{m}},$$

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

Proof Set $T = \sqrt{s} B_1^n$. Then the result follows from Theorem 9.4.4 and the bound (7.19) on the Gaussian width of the ℓ^1 ball: $w(T) = \sqrt{s}w(B_1^n) \le C\sqrt{s \log n}$. □

Remark 9.4.9 (Observations scale almost linearly with sparsity)**.** Corollary 9.4.8 gives a small error as long as

$$m \gtrsim s \log n \tag{9.24}$$

(if the hidden constant is appropriately large). That's great news – we can efficiently recover a sparse vector (with $s \ll n$) from way fewer observations m than the full dimension n.

Now try Exercise 9.24 to extend Corollary 9.4.8 for approximately sparse vectors.

Remark 9.4.10 (A logarithmic improvement)**.** The set $S_{n,s}$ of unit s-sparse vectors in $\mathbb{R}^n$ can be convexified a little tighter. Instead of using the ℓ^1 ball

$T = \sqrt{s}B_1^n$ as in (9.22), use the *truncated* ℓ^1 ball $T_{n,s} = \sqrt{s}B_1^n \cap B_2^n$. This relaxation is pretty tight – try Exercise 9.25 to show

$$\operatorname{conv}(S_{n,s}) \subset T_{n,s} \subset 2\operatorname{conv}(S_{n,s}).$$

This tightening gives a logarithmic improvement to (9.24), showing that

$$m \gtrsim s\log(en/s)$$

observations suffice for sparse recovery (see Exercise 9.26).

For more practice, find the Gaussian width of $T_{n,s}$ and $S_{n,s}$ (Exercises 9.26–9.27), prove the neat Garnaev-Gluskin theorem about slicing the ℓ^1 ball (Exercise 9.28), and discover *Lasso* – a popular method for sparse regression that is just a special case of (9.20) with the ℓ^1 norm (Exercise 9.21).

9.4.3 Example: low-rank recovery

Here is one more example of a high-dimensional linear problem: recover an $d \times d$ matrix X (instead of a vector) from m linear observations:

$$y_i = \langle A_i, X\rangle, \quad i = 1, \dots, m, \tag{9.25}$$

where A_i are known, independent random matrices, and the inner product is $\langle A, B\rangle = \operatorname{tr}(A^\mathsf{T} B)$ as in (4.7).

Normally, you would need d^2 observations – one per entry. To get away with fewer, we need some structure in X. A common one is *low rank*. Just like sparsity counts nonzero entires, rank counts nonzero singular values.

So let's try to argue like in Section 9.4.2. The rank – the ℓ^0 norm of the singular values – is a highly nonconvex matrix function. To fix this, we make a convex relaxation by replacing it with the ℓ^1 norm, or the sum of the singular values, known as the *nuclear norm*:

$$\|X\|_* := \sum_{i=1}^d s_i(X).$$

We looked at the nuclear norm back in Exercise 7.18 – if you skipped it, now is a good time to go back and try it.

Since every vector x with at most s nonzero entries and $\|x\|_2 \le 1$ satisfies $\|x\| \le \sqrt{s}$, every matrix with rank at most r and $\|X\|_F \le 1$ satisfies $\|X\|_* \le \sqrt{r}$ (consider the vector of singular values to make the connection). So it makes sense to consider $T = \sqrt{r}B_*$ as our prior, where B_* is the unit ball of the nuclear norm:

$$B_* := \left\{X \in \mathbb{R}^{d\times d} : \ \|X\|_* \le 1\right\}.$$

The recovery program (9.18) becomes

$$\text{find } X' : \ y_i = \langle A_i, X'\rangle \ \forall i = 1, \dots, m; \quad \|X'\|_* \le \sqrt{r}, \tag{9.26}$$

which is convex and computationally tractable. And Theorem 9.4.4 gives:

Corollary 9.4.11 (Low-rank matrix recovery)**.** *Suppose A_i are independent Gaussian random matrices with all i.i.d. $N(0,1)$ entires.*[6] *Assume an unknown $d \times d$ matrix X has rank at most r and $\|X\|_F \le 1$. Then any solution $\widehat{X}$ of the program* (9.26) *satisfies*

$$\mathbb{E}\|\widehat{X} - X\|_F \le C\sqrt{\frac{rd}{m}}.$$

Proof Using the duality between the nuclear and operator norms (Exercise 7.18(a)), we get for a $d \times d$ Gaussian matrix G with i.i.d. $N(0,1)$ entries:

$$w(B_*) = \mathbb{E} \sup_{\|X\|_* \le 1} \langle G, X\rangle = \mathbb{E}\|G\| \le 2\sqrt{d}$$

by Theorem 7.3.1. Now just apply Theorem 9.4.4 for $T = \sqrt{r}B_*$. □

Remark 9.4.12 (Recovering a low-rank matrix from few observations)**.** Corollary 9.4.8 gives a small error as long as

$$m \gtrsim rd,$$

allowing us to recover a low-rank matrix (with $r \ll d$) from way fewer observations m than the number of entires d^2. This is similar to matrix completion (which we studied in Section 6.5), where we can recover a low-rank matrix from about $m \asymp rd \log d$ randomly chosen entries.

For practice, extend low-rank recovery to rectangular and approximately low-rank matrices (Exercise 9.29).

9.5 Application: exact sparse recovery

In the noiseless case, we can do even better – we can recover a sparse vector x from $y = Ax$ *exactly* (and algorithmically effective)! We will look at two ways to get this surprising result:

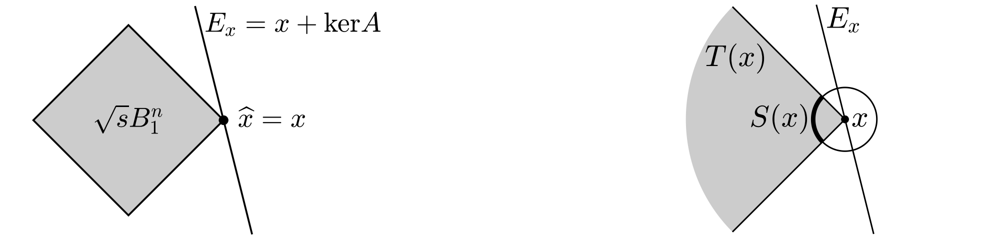
1. Use the escape theorem (Theorem 9.3.4).
2. Identify a deterministic condition on the matrix A that guarantees exact recovery (the "restricted isometry property") – then show random matrices satisfy it with high probability.

9.5.1 Exact recovery based on the escape theorem

To see how exact recovery is possible, let's get some geometric intuition (Figure 9.6). Suppose we are trying to recover an unknown s-sparse unit vector x from $y = Ax$ by solving the convex program (9.23). A solution $\widehat{x}$ lies in the intersection of the prior set $T = \sqrt{s}B_1^n$ (an ℓ^1 ball) and the affine subspace $E_x = x + \ker A$.

T is a cross-polytope, and x sits on one of its $(s-1)$-dimensional edges (see

[6] This assumption can be relaxed – how?



(a) Exact sparse recovery happens when the random subspace E_x is tangent to the ℓ^1 ball at the point x.

(b) 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 9.7 Exact sparse recovery

Figure 9.7a). With some probability, the random subspace E_x is *tangent* to the polytope at x. If so, x is the only point where T and E_x intersect, so the solution $\widehat{x}$ must be exact:

$$\widehat{x} = x.$$

To justify this argument, we just need to show that a random subspace E_x is tangent to the ℓ^1 ball with high probability. That's where the escape theorem (Theorem 9.3.4) is helpful. Zoom in near x (Figure 9.7b): E_x is tangent if and only if the *tangent cone* $T(x)$ (all rays coming from x into the ℓ^1 ball) intersects E_x only at x. This happens 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 – and that is exactly what the escape theorem can guarantee!

Let's now formalize this. We want to recover x from

$$y = Ax$$

by solving the optimization problem

$$\text{minimize } \|x'\|_1 \text{ subject to } y = Ax'. \tag{9.27}$$

Theorem 9.5.1 (Exact sparse recovery)**.** *An $m \times n$ random matrix A with independent, isotropic, subgaussian rows A_i satisfies the following with probability at least $1 - 2\exp(-cm/K^4)$, where $K = \max_i \|A_i\|_{\psi_2}$. If the number of observations m satisfies*

$$m \geq CK^4 s \log n,$$

then, for any s-sparse vector $x \in \mathbb{R}^n$, a solution $\widehat{x}$ of the program (9.27) is exact:

$$\widehat{x} = x.$$

To prove this, we need to show that the recovery error is zero:

$$h := \widehat{x} - x = 0.$$

First, let's show a weaker claim: h has more mass on the support of x than off it.

Lemma 9.5.2 (The error is heavier on x's support)**.** *Set $S := \operatorname{supp}(x)$, and let $h_S \in \mathbb{R}^S$ denote the restriction of h onto S (and similarly for S^c). Then*

$$\|h_{S^c}\|_1 \le \|h_S\|_1.$$

Proof Since $\widehat{x}$ is the minimizer in the program (9.27), we have

$$\|\widehat{x}\|_1 \le \|x\|_1. \tag{9.28}$$

But there is also a lower bound

$$\|\widehat{x}\|_1 = \|x + h\|_1 = \|x_S + h_S\|_1 + \|x_{S^c} + h_{S^c}\|_1 \ge \|x\|_1 - \|h_S\|_1 + \|h_{S^c}\|_1,$$

where the last line follows by triangle inequality and using $x_S = x$ and $x_{S^c} = 0$. Substitute this into (9.28) and simplify to complete the proof. □

Lemma 9.5.3 (The error is approximately sparse)**.** *The error vector satisfies*

$$\|h\|_1 \le 2\sqrt{s}\|h\|_2.$$

Proof Using Lemma 9.5.2 and then the Cauchy-Schwarz inequality, we get

$$\|h\|_1 = \|h_S\|_1 + \|h_{S^c}\|_1 \le 2\|h_S\|_1 \le 2\sqrt{s}\|h_S\|_2 \le 2\sqrt{s}\|h\|_2. \quad \square$$

Proof of Theorem 9.5.1 Assume $h = \widehat{x} - x \neq 0$. Lemma 9.5.3 gives

$$\frac{h}{\|h\|_2} \in T_s := \left\{z \in S^{n-1} : \ \|z\|_1 \le 2\sqrt{s}\right\},$$

and since also $Ah = A\widehat{x} - Ax = y - y = 0$, we have

$$\frac{h}{\|h\|_2} \in T_s \cap \ker A. \tag{9.29}$$

The escape theorem (Theorem 9.3.4) shows that this intersection is empty with high probability as long as $m \ge C_1 K^4 w(T_s)^2$. Now, since $T_s \subset 2\sqrt{s}B_1^n$, we get

$$w(T_s) \le 2\sqrt{s}w(B_1^n) \le C_2\sqrt{s\log n}, \tag{9.30}$$

due to (7.19). Thus, if $m \ge CK^4 s\log n$, the intersection in (9.29) is empty with high probability, which means that the inclusion in (9.29) cannot hold. So, our assumption that $h \neq 0$ is false with high probability. The proof is complete. □

Remark 9.5.4 (Improving the logarithmic factor)**.** By slightly tightening (9.30), we can improve the sufficient number of observations in Theorem 9.5.1 to

$$m \ge CK^4 s\log(en/s).$$

This follows from Exercise 9.26 (how?).

For practice, try interpreting the exact recovery proof geometrically (Exercise 9.30), extend it to the noisy case (Exercise 9.31), and discover a useful nullspace property for exact recovery (Exercise 9.32).

9.5.2 Restricted isometries

Let's find a *deterministic* condition that ensures a matrix A works for sparse recovery, and prove that random matrices satisfy this condition. It is called the restricted isometry property (RIP):

Definition 9.5.5 (RIP)**.** An $m \times n$ matrix A satisfies the *restricted isometry property* (RIP) with parameters α, β and s if the inequality

$$\alpha \|v\|_2 \leq \|Av\|_2 \leq \beta \|v\|_2$$

holds for all vectors $v \in \mathbb{R}^n$ with at most s nonzero entries.

RIP just says that the singular values of all $m \times s$ submatrices[7] A_I of A satisfy:

$$\alpha \leq s_s(A_I) \leq s_1(A_I) \leq \beta. \tag{9.31}$$

(Why?) And if $\alpha \approx \beta \approx 1$, RIP tells us that all those submatrices are approximate isometries (recall Section 4.1.7).

Theorem 9.5.6 (RIP implies exact recovery)**.** *Suppose an $m \times n$ matrix A satisfies RIP with some parameters α, β and $(1+\lambda)s$, where $\lambda > (\beta/\alpha)^2$. Then every s-sparse vector $x \in \mathbb{R}^n$ can be exactly recovered from $y = Ax$ by solving* (9.27)*, i.e. the solution satisfies*

$$\widehat{x} = x.$$

Proof As in the proof of Theorem 9.5.1, we need to show that the error

$$h = \widehat{x} - x$$

is zero. To do this, we decompose h in a way similar to Exercise 9.25. (If you've solved that exercise, great, but it is not necessary to understand this proof.)

Step 1: Decomposing the support. Let I_0 be the support of x. Let I_1 index the λs largest entries of $h_{I_0^c}$ in magnitude, let I_2 index the next λs largest entries of $h_{I_0^c}$ in magnitude, and so on. Finally, set $I_{01} = I_0 \cup I_1$. Since

$$Ah = A\widehat{x} - Ax = y - y = 0,$$

the triangle inequality gives

$$0 = \|Ah\|_2 \geq \|A_{I_{01}} h_{I_{01}}\|_2 - \|A_{I_{01}^c} h_{I_{01}^c}\|_2. \tag{9.32}$$

Next, let's look at the two terms on the right-hand side.

Step 2: Applying RIP. Since $|I_{0,1}| \leq s + \lambda s$, RIP gives

$$\|A_{I_{01}} h_{I_{01}}\|_2 \geq \alpha \|h_{I_{01}}\|_2$$

and triangle inequality followed by RIP gives

$$\|A_{I_{01}^c} h_{I_{01}^c}\|_2 \leq \sum_{i \geq 2} \|A_{I_i} h_{I_i}\|_2 \leq \beta \sum_{i \geq 2} \|h_{I_i}\|_2.$$

[7] Formally, A_I is the $m \times s$ submatrix made by picking the columns indexed by some s-element subset $I \subset \{1, \ldots, n\}$.

Plugging into (9.32), we get

$$\beta \sum_{i \ge 2} \|h_{I_i}\|_2 \ge \alpha \|h_{I_{0,1}}\|_2. \tag{9.33}$$

Step 3: Summing up. Next, we bound the sum in the left like we did in Exercise 9.25. By definition of I_i, each entry of h_{I_i} is bounded in magnitude by the average of the entries of $h_{I_{i-1}}$, i.e. by $\frac{1}{\lambda s}\|h_{I_{i-1}}\|_1$ for $i \ge 2$. Thus

$$\|h_{I_i}\|_2 \le \frac{1}{\sqrt{\lambda s}} \|h_{I_{i-1}}\|_1.$$

Summing up, we get

$$\begin{aligned} \sum_{i \ge 2} \|h_{I_i}\|_2 &\le \frac{1}{\sqrt{\lambda s}} \sum_{i \ge 1} \|h_{I_i}\|_1 = \frac{1}{\sqrt{\lambda s}} \|h_{I_0^c}\|_1 \le \frac{1}{\sqrt{\lambda s}} \|h_{I_0}\|_1 \quad \text{(by Lemma 9.5.2)} \\ &\le \frac{1}{\sqrt{\lambda}} \|h_{I_0}\|_2 \le \frac{1}{\sqrt{\lambda}} \|h_{I_{0,1}}\|_2. \end{aligned}$$

Putting this into (9.33) we conclude that

$$\frac{\beta}{\sqrt{\lambda}} \|h_{I_{0,1}}\|_2 \ge \alpha \|h_{I_{0,1}}\|_2.$$

But this implies that $h_{I_{0,1}} = 0$ since $\beta/\sqrt{\lambda} < \alpha$ by assumption. And since $I_{0,1}$ contains the largest entries of h, it must be that $h = 0$. □

While we do not know how to construct deterministic matrices A that satisfy RIP with good parameters, we can show that random matrices satisfy it with high probability:

Theorem 9.5.7 (Random matrices satisfy RIP)**.** *Consider an $m \times n$ matrix A with independent, isotropic, subgaussian rows A_i. Assume that*

$$m \ge CK^4 s \log(en/s).$$

where $K = \max_i \|A_i\|_{\psi_2}$. Then, with probability at least $1 - 2\exp(-cm/K^4)$, the random matrix A satisfies RIP with parameters $\alpha = 0.9\sqrt{m}$, $\beta = 1.1\sqrt{m}$ and s.

Proof We need to check (9.31) for all $m \times s$ sub-matrices A_I. First, fix I. By Theorem 4.6.1, we get

$$0.9\sqrt{m} \le s_s(A_I) \le s_1(A_I) \le 1.1\sqrt{m} \tag{9.34}$$

with probability at least $1 - 2\exp(-2cm/K^4)$ (set $t = \sqrt{2cm}/K^2$ and use the assumption on m, choosing constants c and C appropriately).

Now take a union bound over all $\binom{n}{s}$ possible s-element subsets $I \subset \{1, \ldots, n\}$. Then (9.34) holds with probability at least

$$1 - 2\exp(-2cm/K^4) \cdot \binom{n}{s} > 1 - 2\exp(-cm/K^4),$$

using the bound $\binom{n}{s} \le \exp(s\log(en/s))$ from (0.6) and the assumption on m. The proof is complete. □

We just learned an alternative approach to exact recovery:

Second proof of Theorem 9.5.1 By Theorem 9.5.7, A satisfies RIP with $\alpha = 0.9\sqrt{m}$, $\beta = 1.1\sqrt{m}$ and $3s$. Thus, Theorem 9.5.6 for $\lambda = 2$ guarantees exact recovery. So Theorem 9.5.1 holds – and we even get the logarithmic improvement from Exercise 9.5.4! □

To practice, show that random projections satisfy RIP (Exercise 9.33).

9.6 Deviations of random matrices for general norms

Let's generalize the matrix deviation inequality (Theorem 9.1.1) to work for any norm – not just the Euclidean one. Actually, we don't even need the norm to be nonnegative – just homogeneity and the triangle inequality will do.

Definition 9.6.1 (Allowing a norm to take negative values)**.** A real-valued function f on a linear vector space V is called:

- *Positive-homogeneous* if $f(\alpha x) = \alpha f(x)$ for all $\alpha \ge 0$ and $x \in V$;
- *Subadditive* if $f(x+y) \le f(x) + f(y)$ for all $x, y \in V$.

Example 9.6.2. These functions are positive-homogeneous and subadditive:

(a) any *norm*;
(b) any real-valued linear function (called a *linear functional*);
(c) an particular, the function $f(x) = \langle x, y\rangle$ for any fixed vector $y \in \mathbb{R}^m$;
(d) the *support function* of any bounded set $S \subset \mathbb{R}^n$, defined by

$$f(x) := \sup_{y\in S}\langle x, y\rangle, \quad x \in \mathbb{R}^m. \tag{9.35}$$

(check this!).

Here is a version of Theorem 9.1.1 that works for all norms (and even positive-homogeneous, subadditive functions), but with a tradeoff – it applies only to Gaussian matrices:

Theorem 9.6.3 (General matrix deviation inequality)**.** *Let A be an $m \times n$ random matrix with i.i.d. $N(0,1)$ entries. Let $f : \mathbb{R}^m \to \mathbb{R}$ be a bounded, positive-homogeneous and subadditive function, and let $b \in \mathbb{R}$ be such that*

$$f(x) \le b\|x\|_2 \quad \textit{for all } x \in \mathbb{R}^n. \tag{9.36}$$

Then for any subset $T \subset \mathbb{R}^n$,

$$\mathbb{E}\sup_{x\in T}\Big|f(Ax) - \mathbb{E} f(Ax)\Big| \le Cb\gamma(T),$$

where $\gamma(T)$ is the Gaussian complexity (see Section 7.5.3).

Exactly as in Section 9.1, Theorem 9.6.3 would immediately follow from Talagrand comparison inequality once we show that the random process

$$Z_x := f(Ax) - \mathbb{E} f(Ax) \tag{9.37}$$

has subgaussian increments. Let us do this now:

Theorem 9.6.4 (Subgaussian 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* (9.36)*. Then the random process* (9.2) *has subgaussian increments:*

$$\|Z_x - Z_y\|_{\psi_2} \le Cb\|x-y\|_2 \quad \textit{for all } x, y \in \mathbb{R}^n. \tag{9.38}$$

Proof of Theorem 9.6.4 Without loss of generality we may assume that $b = 1$. (Why?) Just like in the proof of Theorem 9.1.2, first assume that

$$\|x\|_2 = \|y\|_2 = 1.$$

In this case, the inequality in (9.38)becomes

$$\big\|f(Ax) - f(Ay)\big\|_{\psi_2} \le C\|x-y\|_2. \tag{9.39}$$

Step 1: Creating independence. Consider the vectors

$$u := \frac{x+y}{2}, \quad v := \frac{x-y}{2}. \tag{9.40}$$

Then $x = u + v$ and $y = u - v$, and thus

$$Ax = Au + Av, \quad Ay = Au - Av$$

(see Figure 9.8). Since u and v are orthogonal (check!), the Gaussian random vectors Au and Av are independent (recall Exercise 3.20).

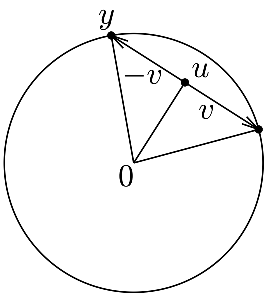


Figure 9.8 Creating a pair of orthogonal vectors u, v out of x, y.

Step 2: Using Gaussian concentration. Let's condition on $a := Au$ and study the conditional distribution of

$$f(Ax) = f(a + Av).$$

By independence, $a + Av$ is a Gaussian random vector that we can write as

$$a + Av = a + \|v\|_2 \, g, \quad \text{where} \quad g \sim N(0, I_m)$$

(again by Exercise 3.20.) We claim that the function

$$z \mapsto f(a + \|v\|_2 \, z)$$

is Lipschitz with respect to the Euclidean norm on $\mathbb{R}^m$, with Lipschitz norm bounded by $\|v\|_2$. To check this, fix any $t, s \in \mathbb{R}^m$ and use subadditivity of f (in the form of Exercise 9.34) to get

$$\begin{aligned} f(a + \|v\|_2 \, t) - f(a + \|v\|_2 \, s) &\le f(\|v\|_2 \, t - \|v\|_2 \, s) \\ &= \|v\|_2 \, f(t - s) \quad \text{(by positive homogeneity)} \\ &\le \|v\|_2 \, \|t - s\|_2 \quad \text{(using (9.36) with } b = 1\text{)}, \end{aligned}$$

proving our claim.

Concentration in the Gauss space (Theorem 5.2.3) then yields

$$\big\| f(a + Av) - \mathbb{E}_a \, f(a + Av) \big\|_{\psi_2(a)} \le C \|v\|_2, \tag{9.41}$$

where the index "a" reminds us that these bounds are valid for the conditional distribution, with $a = Au$ fixed.

Step 3: Removing the conditioning. Since the random vector $a - Av$ has the same distribution as $a + Av$ (why?), it satisfies the same bound:

$$\big\| f(a - Av) - \mathbb{E}_a \, f(a - Av) \big\|_{\psi_2(a)} \le C \|v\|_2, \tag{9.42}$$

Subtract (9.42) from (9.41), 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 holds conditionally for any fixed $a = Au$. Therefore, it holds for the original distribution, too:

$$\big\| f(a + Av) - f(a - Av) \big\|_{\psi_2} \le 2C \|v\|_2.$$

(Why?) Passing back to the x, y notation by (9.40), we obtain the desired inequality (9.39).

We proved the theorem for any unit vectors x, y. To extend it to the general case, argue exactly as in Step 4 in the proof of Theorem 9.1.2 (check!). □

Remark 9.6.5. It is an open question if Theorem 9.6.3 holds for general subgaussian matrices A.

To practice, try Exercises 9.35 and 9.36 to get an anisotropic and a high-probability versions of Theorem 9.6.3.

9.7 Two-sided Chevet inequality and Dvoretzky-Milman theorem

Just like the original matrix deviation inequality from Chapter 9, the more general Theorem 9.1.1 has many applications. For instance, you can now get a *Johnson-Lindenstrauss-type lemma* in any norm, not just the Euclidean one – try it out in Exercises 9.37–9.39!

9.7.1 Two-sided Chevet inequality

Another consequence of general matrix deviation is a sharper version of Chevet inequality, which we which we first looked at in Section 8.6.

Theorem 9.7.1 (Two-sided Chevet 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*

$$\mathbb{E} \sup_{x \in T} \Big| \sup_{y \in S} \langle Ax, y \rangle - w(S) \|x\|_2 \Big| \le C\gamma(T) \operatorname{rad}(S),$$

where $\gamma(T)$ is the Gaussian complexity (see Section 7.5.3) and $\operatorname{rad}(T) = \sup_{x \in T} \|x\|_2$ *is the radius.*

Applying triangle inequality, you can see that Theorem 9.7.1 gives a stronger, two-sided version of Chevet inequality (Theorem 8.6.1).

Proof Let's apply Theorem 9.6.3 for the support function of S from (9.35):

$$f(x) = \sup_{y \in S} \langle x, y \rangle.$$

This is a bounded function, since Cauchy-Schwarz inequality gives

$$f(x) \le \sup_{y \in S} \|x\|_2 \|y\|_2 = \operatorname{rad}(S) \|x\|_2 \quad \text{for all } x \in \mathbb{R}^n. \tag{9.43}$$

Since Ax has the same distribution as $g\|x\|_2$ where $g \in N(0, I_m)$ (see Exercise 3.20), we have

$$\begin{aligned} \mathbb{E} f(Ax) &= \|x\|_2 \, \mathbb{E} f(g) \quad \text{(by positive homogeneity)} \\ &= \|x\|_2 \, \mathbb{E} \sup_{y \in S} \langle g, y \rangle \quad \text{(by definition of } f\text{)} \\ &= \|x\|_2 w(S) \quad \text{(by definition of the Gaussian width).} \end{aligned} \tag{9.44}$$

Substitute (9.43) and (9.44) into Theorem 9.6.3 to complete the proof. □

9.7.2 Dvoretzky-Milman Theorem

We will now prove an amazing result: if you randomly project any bounded set in $\mathbb{R}^n$ to a low-dimensional subspace, it will look *approximately round* with high probability (see Figure 9.9).

It's easier to work with Gaussian projections,[8] where the result says:

[8] By a Gaussian projection here we mean an $m \times n$ matrix with i.i.d. $N(0,1)$ entires. Heuristically, if $m \ll n$, it behaves almost like a random projection scaled up by $\sqrt{n}$ – can you see why?

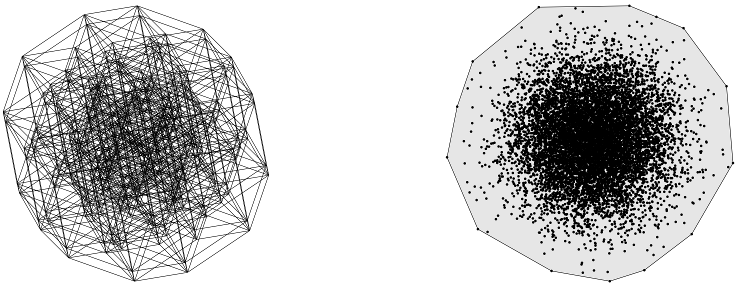


Figure 9.9 A random projection of a 8-dimensional cube (left) and 10^4 Gaussian points (right) onto the plane

Theorem 9.7.2 (Dvoretzky-Milman theorem)**.** *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:*

$$r_- B_2^m \subset \operatorname{conv}(AT) \subset r_+ B_2^m \tag{9.45}$$

where B_2^m denotes the unit Euclidean ball in $\mathbb{R}^m$, and

$$r_\pm = w(T) \pm C\sqrt{m}\,\operatorname{rad}(T).$$

The left inclusion holds only if r_- is nonnegative; the right inclusion, always.

Proof Let's express two-sided Chevet inequality (Theorem 9.7.1) in the following form:

$$\mathbb{E}\sup_{y\in S}\Big|\sup_{x\in T}\langle Ax, y\rangle - w(T)\|y\|_2\Big| \le C\gamma(S)\operatorname{rad}(T),$$

where $T \subset \mathbb{R}^n$ and $S \subset \mathbb{R}^m$. (To get this, just apply the theorem to $A^\top$ with T and S swapped.)

Let S be the sphere S^{m-1}; its Gaussian complexity satisfies $\gamma(S) \le \sqrt{m}$. Then, by Markov inequality, the following holds with probability at least 0.99:

$$\Big|\sup_{x\in T}\langle Ax, y\rangle - w(T)\Big| \le C\sqrt{m}\operatorname{rad}(T) \quad \text{for every } y \in S^{m-1}.$$

By triangle inequality and definition of $r_\pm$, this implies

$$r_- \le \sup_{x\in T}\langle Ax, y\rangle \le r_+ \quad \text{for every } y \in S^{m-1}.$$

Rewriting $\sup_{x\in T}\langle Ax, y\rangle$ as $\sup_{x\in AT}\langle x, y\rangle$ and using homogeneity, we get

$$r_-\|y\|_2 \le \sup_{x\in AT}\langle x, y\rangle \le r_+\|y\|_2 \quad \text{for every } y \in \mathbb{R}^m.$$

By duality, this is the same as (9.45) (try Exercise 9.40 to work out the details of the duality argument). □

Remark 9.7.3 (The effective dimension)**.** Assume that T is bounded, convex and contains the origin, and let

$$m \le c d(T)$$

where $d(T) \asymp w(T)^2/\operatorname{rad}(T)^2$ is the effective dimension (see Definition 7.5.12). If we pick the absolute constant c small enough, we can make $C\sqrt{m}\operatorname{rad}(T) \le 0.01 w(T)$, so the that Dvoretzky-Milman Theorem 9.7.2 gives

$$0.99B \subset AT \subset 1.01B$$

with $B = w(T)B_2^n$ is the Euclidean ball of radius $w(T)$. In short: *projecting any bounded convex set T onto a random subspace of dimension about $d(T)$ makes it look almost like a round ball!*

Example 9.7.4 (Almost round projections of the cube)**.** Consider the cube $T = [-1,1]^n$. By (7.18), $w(T) = \sqrt{2/\pi} \cdot n$ and $\operatorname{diam}(T) = 2\sqrt{n}$, so the effective dimension is $d(T) \asymp n$. So, if $m \le cn$, then with high probability we have

$$0.99B \subset A[-1,1]^n \subset 1.01B$$

where B is the Euclidean ball with radius $\sqrt{2/\pi} \cdot n$. In short: *projecting an n-dimensional cube onto a subspace of dimension $m = cn$ makes it look almost like a round ball!* Figure 9.9 illustrates this remarkable fact.

Remark 9.7.5 (Summary of random projections)**.** In Sections 7.6 and 9.2.2, we found that a random projection P of a set T onto an m-dimensional subspace in $\mathbb{R}^n$ undergoes a phase transition. In the high-dimensional regime ($m \gtrsim d(T)$), the projection shrinks the diameter of T by the factor of order $\sqrt{m/n}$:

$$\operatorname{diam}(PT) \asymp \sqrt{\frac{m}{n}} \operatorname{diam}(T)$$

Moreover, the additive Johnson-Lindenstrauss Lemma 9.2.4 shows that in this regime, the random projection P approximately preserves the geometry of T (the distances between all points in T shrink roughly by the same scaling factor).

In the low-dimensional regime ($m \lesssim d(T)$), shrinking stops:

$$\operatorname{diam}(PT) \asymp w_s(T) \asymp \frac{w(T)}{\sqrt{n}}$$

regardless of how small m is. Dvoretzky-Milman theorem explains why: PT is now an *approximately the round ball* of radius of order $w_s(T)$ (see Exercise 9.43), which obviously does not shrink under any projection.

To practice, get a high-probability version of Dvoretzky-Milman theorem (Exercise 9.41), show that a Gaussian cloud is nearly round in low dimensions (Exercise 9.42), and get at version of Dvoretzky-Milman theorem for actual (rather than Gaussian) projections (Exercise 9.43).

9.8 Notes

The matrix deviation inequality (Theorem 9.1.1) and its proof come from [213], though many related results existed earlier. For a Gaussian matrix A and a set T on the unit sphere, it can be derived from Gaussian comparison inequalities: the upper bound from Sudakov-Fernique (Theorem 7.2.8), and the lower from Gordon (Theorem 7.2.9). Schechtman [295] proved a version of matrix deviation inequality for Gaussian A and general norms; we will cover that in Section 9.6. For subgaussian A, versions appear in [187, 236, 103]; see [213, Section3] for comparisons. A particularly clean result for quadratic processes is in [79, Theorem 3.2.1]. Other extensions include versions for sparse A [54], for ℓ^p norms [301], and for independent columns [277].

The quadratic dependence on K in Theorem 9.1.1 was improved to the optimal $K\sqrt{\log K}$ in [176]. This automatically improves the dependence on subgaussian norms in all results that follow from Theorem 9.1.1, like Theorems 3.1.1, 4.6.1, 5.3.1, 9.3.1, 9.3.4, 9.4.4, Proposition 9.2.1, Corollary 9.4.8, and others.

A version of the bound on random projections of sets (Proposition 9.2.1) goes back to V. Milman [245]; see [21, Proposition 5.7.1].

Theorem 9.2.2 on covariance estimation for lower-dimensional distributions is due to V. Koltchinskii and K. Lounici [190]; they used a different approach that is also based on the majorizing measure theorem. R. van Handel [331] gave an alternative proof for Gaussian distributions using decoupling, conditioning, and Slepian Lemma. For Gaussian matrices, the bound in Theorem 9.2.2 is tight, and many extensions now exist – see the end of Chapter 4 for references.

A version of additive Johnson-Lindenstrauss lemma (Lemma 9.2.4) is from [213].

The M^* bound (Theorem 9.3.1) has been around for a while in geometric functional analysis. Early versions came from V. Milman [243, 244]; a version with the right dependence on m was proved by Pajor and Tomczak [269], and Gordon later gave an even sharper form with exact constants [143]. For more on this, including proofs and variants, see [21, Sections 7.3–7.4, 9.3], [143, 233, 341]. The version we gave in Theorem 9.3.1 comes from [213].

The escape theorem (Theorem 9.3.4), also known as "escape from the mesh", was first proved by Y.Gordon [143] for Gaussian matrices, with a sharp constant in (9.14), using his comparison inequality (Theorem 7.2.9). Matching lower bounds are known for spherically convex sets [308, 18], and in that case, the exact hitting probability can be calculated using tools from integral geometry [18]. Oymak and Tropp [267] showed how this result can be extended beyond the Gaussian case. The version we gave in Theorem 9.3.4 comes from [213].

The applications in Sections 9.4–9.5 come from high-dimensional statistics and signal processing (specifically, compressed sensing). The tutorial [341] offers a unified treatment of these two kinds of problems, which we followed in this chapter. The books [67, 157, 344, 137] discuss statistical aspects, while the survey [91] and book [127] focus on signal processing side.

Recovery based on M^* bound discussed in Section 9.4.1 is based on [341], which has various versions of Theorem 9.4.4 and Corollary 9.4.8.

The survey [93] offers a comprehensive overview of the low-rank matrix recovery problem discussed in Section 9.4.3. Our presentation is based on [341, Section 10].

Exact sparse recovery discussed in Section 9.5 was discovered in the area of compressed sensing; see [91] and book [127]. The approach via escape theorem (Section 9.5.1) was first discovered in [290]; here we loosely follow [341, Section 9]. See also [80, 307] and especially [322] for applications of the escape theorem to sparse recovery. Sharp bounds on how many observations are needed for exact recovery were first proved in [112], and later extended in [111, 108, 109, 110]. Phase transitions for more general sets T and matrices A were studied in [18, 266, 267].

The RIP-based approach to sparse recovery (Section 9.5.2) was pioneered by E. Candes and T. Tao [73]; see [127, Chapter 6] for a comprehensive introduction. A version of Theorem 9.5.6 appears in their work, and our proof is based on an argument from Y. Plan, similar to [70]. The fact that random matrices satisfy RIP (Theorem 9.5.7) is a cornerstone of compressed sensing – see [127, 340].

For more on deviations for quadratic processes (Exercise 9.5), see [79, Theorem 3.2.1] – it is a pretty similar result.

Exercises 9.19–9.21 discuss a popular tool for sparse linear regression, known in statistics literature as Lasso (least absolute shrinkage and selection operator). It was pioneered by R. Tibshirani [319]. The books [157, 67, 344] offer a comprehensive introduction into statistical problems with sparsity constraints; these books discuss Lasso and its many variants.

Sparse recovery based on the nullspace property (Exercise 9.32) goes back to Cohen, Dahmen and DeVore [88], see [281, 127, 344].

Garnaev-Gluskin bound (Exercise 9.28) was first proved in [132], see also [222] and [127, Chapter 10].

General matrix deviation inequality (Theorem 9.6.3) and its proof is due to G. Schechtman [295].

The original version Chevet inequality was proved by S. Chevet [84] and the constant factors there were improved by Y. Gordon [140]; see also [21, Section 9.4], [210, Theorem 3.20] and [321, 7]. The version of Chevet inequality that we stated in Theorem 9.7.1) can be reconstructed from the work of Y. Gordon [140, 142], see [210, Corollary 3.21].

Dvoretzky-Milman theorem has a long history in functional analysis. Proving a conjecture of A. Grothendieck, A. Dvoretzky [117, 118] showed that any n-dimensional normed space has an m-dimensional almost Euclidean subspace, where $m = m(n)$ grows to infinity with n. V. Milman gave a probabilistic proof of this theorem [242] and pioneered the study of the best possible dependence $m(n)$. Theorem 9.7.2 is due to V. Milman [242]; it is optimal [247], see [21, Theorem 5.3.3]. The tutorial [23] contains a a light introduction into Dvoretzky-Milman theorem. For a full exposition of Dvoretzky-Milman theorem and many of its ramifications, see e.g. [21, Chapter 5 and Section 9.2], [210, Section 9.1] and the references there. A "distributional" version of Dvoretzky-Milman theorem, where one asks if a random m-dimensional marginal of an n-dimensional distribution is approximately normal, has been pioneered for log-concave distributions by B. Klartag [186] (the central limit therem for convex bodies) and for discrete distributions by E. Meckes [230].

The phase transition phenomenon noted in Remark 9.7.5 was put forth by V. Milman [245]; see [21, Proposition 5.7.1].

Exercises

9.1 ☕☕ (Reverse triangle inequality) Let's check the geometric observation used in the proof of matrix deviation (Theorem 9.1.2). Consider any vectors $x, y \in \mathbb{R}^n$ satisfying $1 = \|x\|_2 \le \|y\|_2$, and let $\bar{y} := y/\|y\|$ (see Figure 9.1). Show that

$$\|x - y\|_2 \le \|x - \bar{y}\|_2 + \|\bar{y} - y\|_2 \le \sqrt{2}\|x - y\|_2.$$

9.2 ☕☕ (Matrix deviation inequality: deviations from the mean) Deduce from Theorem 9.1.1 a centered version of matrix deviation:

$$\mathbb{E}\sup_{x \in T}\Big|\|Ax\|_2 - \mathbb{E}\|Ax\|_2\Big| \le CK^2\gamma(T).$$

9.3 ☕☕ (Quadratic matrix deviation) Let's prove the claim in Remark 9.1.5: under the conditions of Theorem 9.1.1, we have

$$\mathbb{E}\sup_{x \in T}\Big|\|Ax\|_2^2 - m\|x\|_2^2\Big| \le CK^4\gamma(T)^2 + CK^2\sqrt{m}\,\mathrm{rad}(T)\gamma(T).$$

9.4 ☕☕ (Anisotropic matrix deviation) In Theorem 9.1.1, we assumed that the rows of the random matrix are isotropic. Let's remove this assumption. Let B be an $m \times n$ random matrix with independent rows B_i, which satisfy

$$\mathbb{E}\,B_i B_i^{\mathsf{T}} = \Sigma, \quad \|\langle B_i, x\rangle\|_{\psi_2} \le K\|\langle B_i, x\rangle\|_{L^2} \quad \text{for any } x \in \mathbb{R}^n,$$

for some matrix Σ and number K. Show that for any subset $T \subset \mathbb{R}^n$,

$$\mathbb{E} \sup_{x \in T} \left| \|Bx\|_2 - \sqrt{m}\|\Sigma^{1/2}x\|_2 \right| \le CK^2 \gamma(\Sigma^{1/2}T).$$

9.5 ☕☕☕ (Quadratic empirical process) In Exercise 8.36, we bounded the deviation of the empirical mean from the sample mean for any subgaussian process. Now, let's do the same for the L^2 mean by extending Theorem 9.1.1. Take any star-shaped[9] class $\mathcal{F}$ of real-valued functions on a domain Ω, and let $X, X_1, X_2, \dots, X_n$ be i.i.d. random points in Ω. Consider the L^2 metric on $\mathcal{F}$

$$d(f,g) := \left(\mathbb{E}(f(X) - g(X))^2 \right)^{1/2}$$

and assume that

$$\|f(X) - g(X)\|_{\psi_2} \le K d(f,g) \quad \text{for all } f, g \in \mathcal{F}.$$

Then show:

$$\mathbb{E} \sup_{f \in \mathcal{F}} \left| \left(\frac{1}{m} \sum_{i=1}^m f(X_i)^2 \right)^{1/2} - \left(\mathbb{E} f(X)^2 \right)^{1/2} \right| \le \frac{CK^2 \gamma_2(F,d)}{\sqrt{m}},$$

where in the right hand side we have the γ_2 functional from Section 8.5. Can you see why this result contains Theorem 9.1.1 if $T \subset S^{n-1}$?

9.6 ☕☕☕☕ (Deviation of random projections) Prove a version of matrix deviation inequality (Theorem 9.1.1) for random projections. Let P be the orthogonal projection in $\mathbb{R}^n$ on an m-dimensional subspace uniformly distributed in the Grassmanian $G_{n,m}$. Show that for any subset $T \subset \mathbb{R}^n$, we have

$$\mathbb{E} \sup_{x \in T} \left| \|Px\|_2 - \sqrt{\frac{m}{n}}\|x\|_2 \right| \le \frac{C\gamma(T)}{\sqrt{n}}.$$

9.7 ☕☕ (Sizes of projections: a high-probability bound) Let's get a high-probability version of Proposition 9.2.1. Let $T \subset \mathbb{R}^n$ be a bounded set, and let A be an $m \times n$ matrix with independent, isotropic and subgaussian rows A_i, and consider the "subgaussian projection" $P = \frac{1}{\sqrt{n}}A$. Show that for $\varepsilon > 0$, the bound

$$\operatorname{diam}(PT) \le (1+\varepsilon)\sqrt{\frac{m}{n}} \operatorname{diam}(T) + CK^2 w_s(T)$$

holds with probability at least $1 - \exp(-c\varepsilon^2 m/K^4)$. Here $K = \max_i \|A_i\|_{\psi_2}$ and $w_s(T)$ is the spherical width of T.

9.8 ☕☕☕ Prove a version of Proposition 9.2.1 for the original model of P considered in Section 7.6, i.e. for the projection P onto a random m-dimensional subspace $E \sim \mathrm{Unif}(G_{n,m})$.

9.9 ☕ (Covariance estimation with high probability) Check the high-probability guarantee on the covariance estimation mentioned in Remark 9.2.3.

[9] Star-shaped means that if f is in $\mathcal{F}$, then so is af for any $a \in [0,1]$. If $\mathcal{F}$ is not shar-shaped, no problem – just add to it all af for $a \in [0,1]$; the Gaussian complexity stays the same.

9.10 ☕☕ (From matrix deviation to Johnson-Lindenstrauss) Use matrix deviation inequality to give an alternative solution of Exercise 5.14. (Quantify the success probability and dependence on the subgaussian norm.)

9.11 ☕ (Additive Johnson-Lindenstrauss lemma) Argue that the error in the conclusion of Lemma 9.2.4 must be absolute as opposed to relative.

9.12 ☕☕ (M^* bound for random affine sections) We proved the M^* bound (Theorem 9.3.1) for random sections through the origin. Extend it for all affine sections:

$$\mathbb{E} \max_{z \in \mathbb{R}^n} \operatorname{diam}\big(T \cap (z + \ker A)\big) \le \frac{CK^2 w(T)}{\sqrt{m}}.$$

9.13 ☕☕ (M^* bound with high probability) Prove a high-probability version of the M^* bound (Theorem 9.3.1).

9.14 ☕☕☕ (Slicing the ℓ^p ball) Let B_p^n be the unit ℓ^p ball in $\mathbb{R}^n$ (see Figure 1.2 for some examples). Let E be a random k-dimensional subspace, with $1 \le k \le 0.99n$.

(a) Show that

$$\mathbb{E} \operatorname{diam}(B_p^n \cap E) \asymp_p n^{\frac{1}{2} - \frac{1}{p}} \quad \text{for all } p \in (1, \infty),$$

where the notation $\asymp_p$ hides positive constants that may depend only on p.

(b) Check that this is equivalent to the radius of the *inscribed* Euclidean ball in B_p^n for $p \le 2$, and *circumscribed* for $p \ge 2$. Can you explain this intuitively?

9.15 ☕ (Tightness of the escape theorem) Show that the escape theorem (Theorem 9.3.4) is generally optimal for all $m \le n$ by taking T to be the unit sphere in a subspace of $\mathbb{R}^n$.

9.16 ☕☕ (Putting a sticker on the soccer ball) Here is another version of the escape theorem. Let $T \subset S^{n-1}$ be any set (think of a sticker on a soccer ball) and let $\mathcal{X} \subset S^{n-1}$ be an N-point set (like ink marks). Assume that

$$\sigma_{n-1}(T) < \frac{1}{N}$$

where σ_{n-1} is the normalized surface area on the sphere. Show that there exists a rotation $U \in O(n)$ such that

$$UT \cap \mathcal{X} = \varnothing$$

(so we can place the sticker without covering any marks).

9.17 ☕☕ (Constrained recovery: mean squared error) Extend the error bound Theorem 9.4.4 for the bigger *mean squared error*:

$$\mathbb{E}\|\widehat{x} - x\|_2^2.$$

9.18 ☕☕☕ (Recovery by optimization) Let T be the unit ball of some norm $\|\cdot\|_T$ in $\mathbb{R}^n$. Show that the conclusion of Theorem 9.4.4 holds also for the following optimization program:

$$\text{minimize } \|x'\|_T \text{ subject to } y = Ax'.$$

9.19 ☕☕☕ (Constrained optimization) Let's extend the constrained recovery result (Theorem 9.4.4) to the noisy model considered in (9.16):

$$y = Ax + w, \quad x \in T,$$

where w is some unknown noise vector (maybe even dependent on A). To recover x, we find a vector that lies in T and fits the measurements y as close as possible:

$$\text{minimize } \|y - Ax'\|_2 \text{ subject to } x' \in T.$$

Show that the solution $\widehat{x}$ satisfies the following accuracy guarantee:

$$\mathbb{E}\|\widehat{x} - x\|_2 \lesssim \frac{K^2 w(T) + \|w\|_2}{\sqrt{m}}.$$

9.20 ☕☕☕ (Unconstrained optimization) Let's make Remark 9.4.7 a bit more rigorous. Let $x \in \mathbb{R}^n$ be any vector. Let A be a $m \times n$ random matrix with independent, isotropic and subgaussian rows A_i. Consider the noisy linear model

$$y = Ax + w$$

where $w \in \mathbb{R}^m$ be any noise vector (possibly even dependent on A). To recover x from y and A, consider the unconstrained convex problem

$$\text{minimize } \|y - Ax'\|_2^2 + \lambda \|x'\|_T \quad \text{over all } x' \in \mathbb{R}^n,$$

where $\|\cdot\|_T$ is any norm and $\lambda > 0$ is a tuning parameter. Show that if we choose $\lambda \asymp \|w\|_2^2 / \|x\|_T$, then the solution $\widehat{x}$ satisfies

$$\mathbb{E}\|\widehat{x} - x\|_2 \lesssim \frac{K^2 w(T) \|x\|_T + \|w\|_2}{\sqrt{m}},$$

where T is the unit ball of the norm $\|\cdot\|_T$ and $K = \max_i \|A_i\|_{\psi_2}$.

9.21 ☕ (Lasso) Let's specialize Exercise 9.20 to the ℓ^1 norm and analyze Lasso – a popular method for sparse regression. Suppose $x \in \mathbb{R}^n$ is s-sparse, and A is an $m \times n$ random matrix with independent, isotropic, subgaussian rows A_i. Consider the noisy linear model

$$y = Ax + w.$$

To recover x from y and A, consider the unconstrained convex problem

$$\text{minimize } \|y - Ax'\|_2^2 + \lambda \|x'\|_1 \quad \text{over all } x' \in \mathbb{R}^n.$$

Show that if we choose $\lambda \asymp \|w\|_2^2 / \|x\|_1$, then the solution $\widehat{x}$ satisfies

$$\mathbb{E}\|\widehat{x} - x\|_2 \lesssim \frac{K^2 \sqrt{s \log n} + \|w\|_2}{\sqrt{m}},$$

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

9.22 ☕☕ (The sparse recovery problem is well posed) Let A be an $m \times n$ matrix in general position (choose a convenient definition of general position yourself).

(a) Show that if $m \geq 2\|x\|_0$, then the equation $y = Ax$ has a unique solution (if it exists).

(b) In this case, how can you find x algorithmically efficiently if you know the support of x (i.e. which entries are nonzero)?

9.23 ☕☕☕ (The "ℓ^p norms" for $0 \leq p < 1$) In (9.21), we defined $\|x\|_0$ as the number of nonzero entries of the vector x.

(a) Check that $\|\cdot\|_0$ is not a norm on $\mathbb{R}^n$.

(b) Check that $\|\cdot\|_p$ is not a norm on $\mathbb{R}^n$ if $0 < p < 1$ (Figure 9.10 shows that the unit balls are not convex).

(c) Show that, for every $x \in \mathbb{R}^n$,

$$\|x\|_0 = \lim_{p \to 0_+} \|x\|_p^p.$$

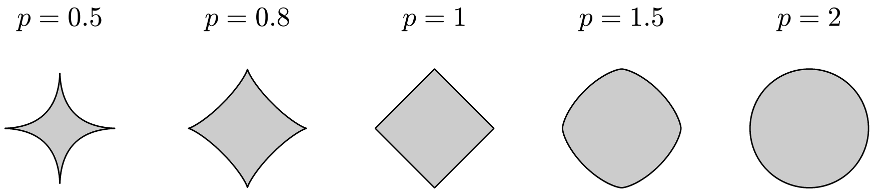


Figure 9.10 The unit ℓ^p balls for various values of p in $\mathbb{R}^2$.

9.24 ☕☕☕ (Approximate sparse recovery) Let's extend sparse recovery (Corollary 9.4.8) to accommodate approximately sparse signals.

(a) Show that any s-sparse signal x can be recovered from measurements $y = Ax$ by solving the optimization problem

$$\text{minimize } \|x'\|_1 \text{ subject to } y = Ax',$$

whose solution $\widehat{x}$ gives the recovery error

$$\mathbb{E}\|\widehat{x} - x\|_2 \leq CK^2 \sqrt{\frac{s \log n}{m}} \|x\|_2.$$

(b) Argue that a similar result holds for approximately sparse signals. State and prove such a guarantee.

9.25 ☕☕☕ (Convexifying the set of sparse vectors) Consider the set of unit s-sparse vectors:

$$S_{n,s} := \left\{x \in \mathbb{R}^n : \|x\|_0 \leq s, \|x\|_2 \leq 1\right\} \tag{9.46}$$

and the truncated ℓ^1 ball:

$$T_{n,s} := \sqrt{s} B_1^n \cap B_2^n = \left\{x \in \mathbb{R}^n : \|x\|_1 \leq \sqrt{s}, \|x\|_2 \leq 1\right\}. \tag{9.47}$$

Show that

$$\operatorname{conv}(S_{n,s}) \subset T_{n,s} \subset 2\operatorname{conv}(S_{n,s}).$$

To prove the second inclusion, fix $x \in T_{n,s}$ and partition the support of x into disjoint subsets $I_1, I_2, \ldots$ so that I_1 indexes the s largest coefficients of x in magnitude, I_2 indexes the next s largest coefficients, and so on. Show that $\sum_{i \geq 1} \|x_{I_i}\|_2 \leq 2$, where $x_I \in \mathbb{R}^T$ denotes the restriction of x onto a set I.

9.26 ☕☕ (A logarithmic improvement in sparse recovery) Use Exercise 9.25 to show that

$$w(T_{n,s}) \leq 2w(S_{n,s}) \leq C\sqrt{s\log(en/s)}.$$

Improve the logarithmic factor in the sparse recovery guarantee (Corollary 9.4.8) to

$$\mathbb{E}\|\widehat{x} - x\|_2 \leq CK^2\sqrt{\frac{s\log(en/s)}{m}},$$

showing that $m \gtrsim s\log(en/s)$ measurements suffice.

9.27 ☕☕☕☕ (The Gaussian width of sparse vectors) Show that the set of unit s-sparse vectors (9.46) and its (approximate) convex hull (9.47) have this Gaussian width:

$$w(T_{n,s}) \asymp w(S_{n,s}) \asymp \sqrt{s\log(en/s)}$$

where the notation $\asymp$ hides positive absolute constants.

9.28 ☕☕☕ (Garnaev-Gluskin theorem) Improve the logarithmic factor in the bound (9.3.2) on random slices of the ℓ^1 ball:

$$\mathbb{E}\operatorname{diam}(B_1^n \cap E) \lesssim \sqrt{\frac{\log(en/m)}{m}}.$$

In particular, the logarithmic factor can be removed from (9.13).

9.29 ☕☕ (Extensions of low-rank matrix recovery) Let's get some versions of low rank recovery from Section 9.4.3.

(a) Show that the conclusion of Corollary 9.4.11 holds also for the following optimization program:

$$\text{minimize } \|X'\|_* \text{ subject to } y_i = \langle A_i, X'\rangle \; \forall i = 1, \ldots, m.$$

(b) Extend the matrix recovery result for *approximately* low-rank matrices.

(c) Extend the matrix recovery result to $d_1 \times d_2$ matrices.

9.30 ☕☕ (Geometry of exact sparse recovery) Give a geometric interpretation of the proof of Theorem 9.5.1 (see Figure 9.7b). What does the proof say about the tangent cone $T(x)$? Its spherical part $S(x)$?

9.31 ☕☕☕ (Noisy measurements) Extend the exact sparse recovery result (Theorem 9.5.1) to noisy measurements $y = Ax + w$. (Modify the recovery program (9.27) accordingly.)

9.32 ☕☕☕ (Nullspace property) Here is a handy condition ensuring exact recovery. Let's say that an $m \times n$ matrix A satisfies the *nullspace property of order s* if the inequality

$$\|h_S\|_1 < \|h_{S^c}\|_1$$

holds for any nonzero vector $h \in \ker(A)$ and any s-element subset $S \subset \{1, \ldots, n\}$.

(a) Show that A satisfies the nullspace property if and only if every s-sparse vector $x \in \mathbb{R}^n$ is the unique solution to (9.27) with $y = Ax$.

(b) Show that a random matrix A with $m \gtrsim s \log(en/s)$ satisfy the nullspace property with high probability (make this statement precise as in Theorem 9.5.7).

9.33 ☕☕☕ (Random projections satisfy RIP) Let P be the orthogonal projection in $\mathbb{R}^n$ onto a random m-dimensional subspace (uniformly distributed in the Grassmanian).

(a) Prove that P satisfies RIP (similar to Theorem 9.5.7, up to a normalization).

(b) Conclude a version of Theorem 9.5.1 for exact recovery from random projections.

9.34 ☕ (Subadditivity) Let $f : V \to \mathbb{R}$ be a subadditive function on a vector space V. Show that

$$f(x) - f(y) \le f(x - y) \quad \text{for all } x, y \in V.$$

9.35 ☕☕ (General matrix deviation for anisotropic distributions) Extend Theorem 9.6.3 to $m \times n$ matrices A whose rows are independent $N(0, \Sigma)$ random vectors, showing that

$$\mathbb{E} \sup_{x \in T} \Big| f(Ax) - \mathbb{E} f(Ax) \Big| \le C b \gamma(\Sigma^{1/2} T).$$

9.36 ☕☕ (General matrix deviation: a high-probability bound) Prove a high-probability version of Theorem 9.6.3.

9.37 ☕☕ (Johnson-Lindenstrauss lemma for general norms) Use the general matrix deviation inequality (Theorem 9.6.3) to get a version of Johnson-Lindenstrauss lemma for any norm on $\mathbb{R}^m$, not just the Euclidean one.

9.38 ☕☕ (Johnson-Lindenstrauss lemma for the ℓ^1 norm) Specialize Exercise 9.37 to the ℓ^1 norm. So, take any set $\mathcal{X}$ 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 pick any $\varepsilon \in (0, 1)$. Assume that

$$m \ge C(\varepsilon) \log N.$$

Show that with high probability, the matrix $Q = \sqrt{\frac{\pi}{2}} \frac{1}{m} A$ satisfies

$$(1 - \varepsilon)\|x - y\|_2 \le \|Qx - Qy\|_1 \le (1 + \varepsilon)\|x - y\|_2 \quad \text{for all } x, y \in \mathcal{X}.$$

9.39 ☕☕ (Johnson-Lindenstrauss embedding into ℓ^∞) Specialize Exercise 9.37 to the ℓ^∞ norm. So, take any set $\mathcal{X}$ 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 pick any $\varepsilon \in (0, 1)$. Assume that

$$m \ge N^{C(\varepsilon)}.$$

Show that with high probability, the matrix $Q = C(\log m)^{-1/2} A$ satisfies

$$(1-\varepsilon)\|x-y\|_2 \le \|Qx - Qy\|_\infty \le (1+\varepsilon)\|x-y\|_2 \quad \text{for all } x, y \in \mathcal{X},$$

for an appropriate choice of the absolute constant C. Note that in this case $m \ge N$, so Q is not a dimension reduction map but rather an *embedding*.

9.40 ☕☕☕ (Duality) Show that for a closed, bounded set $V \subset \mathbb{R}^m$, its support function is close to the Euclidean norm if and only if its convex hull is close to the Euclidean ball. Specifically, for $r_-, r_+ \ge 0$, prove that

$$r_- B_2^m \subset \operatorname{conv}(V) \subset r_+ B_2^m$$

holds if and only if

$$r_- \|y\|_2 \le \sup_{x \in V} \langle x, y \rangle \le r_+ \|y\|_2 \quad \text{for all } y \in \mathbb{R}^m.$$

9.41 ☕☕ State and prove a high-probability version of Dvoretzky-Milman theorem.

9.42 ☕☕ (Gaussian cloud is nearly round) Consider i.i.d. random vectors $g_1, \ldots, g_n \sim N(0, I_m)$. Suppose that

$$m \le c \log n.$$

Show that with high probability, the convex hull of these points is approximately a Euclidean ball with radius $\asymp \sqrt{\log n}$ (see Figure 9.9).

9.43 ☕☕☕ (Actual random projections) We stated Dvoretzky-Milman theorem for "Gaussian projections". Prove a version of it for a (true) projection P onto a random m-dimensional subspace in $\mathbb{R}^n$. Under the same assumptions, the conclusion should be that

$$(1-\varepsilon)B \subset \operatorname{conv}(PT) \subset (1+\varepsilon)B$$

where B is a Euclidean ball with radius $w_s(T)$, where $w_s(T)$ is the spherical width of T (recall Definition 7.5.4).

Hints for the exercises

0.1 Recall that $\|x-y\|_2^2 = \|x\|_2^2 - 2\langle x,y\rangle + \|y\|_2^2$. (This follows by expanding $\|x-y\|_2^2 = \langle x-y, x-y\rangle$.) Use this formula for $\|Z - \mathbb{E}\,Z\|_2^2$.

0.2 Check the identity $\mathbb{E}\|Z-a\|_2^2 - \mathbb{E}\|Z-\mu\|_2^2 = \|a-\mu\|_2^2$ where $\mu = \mathbb{E}\,Z$.

0.4 (a) Select the signs independently at random. Calculate the expected squared norm of the random vector $\pm x_1 \pm x_2 \pm \cdots \pm x_n$ using Example 0.3.

0.5 Choose $T = \{e_1, \ldots, e_n\}$ where e_i are the standard basis vectors. Then $\operatorname{conv}(T)$ is an $(n-1)$-dimensional simplex; draw a picture for $n=3$. Let x be the center of the simplex. All that remains is to calculate the distance from x to each $(k-1)$-dimensional face of the simplex.

0.6 To prove the upper bound, multiply the sum of binomial coefficients by the quantity $(k/n)^k$, replace this quantity by $(k/n)^j$ in the left side, and use the binomial theorem. To prove the lower bound, use the definition of the binomial coefficient to express it as a product of k fractions; check that each fraction is lower bounded by n/k.

0.7 Recall the scaling property of the volume in $\mathbb{R}^n$ used in the beginning of the proof of Theorem 0.0.4: the ball of radius r has volume r^n times the volume of the unit ball.

0.8 Compute the CDF of $\|X\|_2$, deduce the probability density function by differentiation, and then compute the expectation.

0.9 First, improve the bound on the number of balls in Corollary 0.0.3 using the following fact from elementary combinatorics: the number of ways to choose an unordered subset of k elements from an N-element set, with possible repetitions, is $\binom{N+k-1}{k}$. Substitute $k = k_0 = n/\log(eN/n)$ and use Exercise 0.6 to bound the binomial coefficient by C^n. Then follow the proof of Theorem 0.0.4.

1.3 (a) Use induction on m. At the induction step, represent $\sum_{i=1}^m \lambda_i x_i$ as a convex combination of two vectors, one of which is x_m and the other is some convex combination of $x_1, \ldots, x_{m-1}$.

1.4 To prove the upper bound, express a point $x \in \operatorname{conv}(T)$ as a convex combination of some points in T and use Jensen inequality from Exercise 1.3.

1.7 Condition on the value of n, but otherwise follow the proof of the result in Example 1.4.2.

1.8 What is the probability that a given subset of k students is independent? How many subsets consisting of k students are there? Answer these questions and use the union bound.

1.9 Following the proof of the result in Example 1.4.2, the problem reduces to checking that $n(1-p_n)^{n-1} \to 0$.

1.10 (b) Expanding yields $\mathbb{E}\,S_n^2 = \sum_{i=1}^n \mathbb{E}\,X_i^2 + \sum_{i\neq j} \mathbb{E}\,X_i X_j$. Interpret each term $\mathbb{E}\,X_i X_j$ as the probability that both students i and j are friendless. Compute this probability.

1.11 Use Jensen inequality.

1.12 Write $\mathbb{E}|X|^p$ as $\mathbb{E}\left[|X|\,|X|^{p-1}\right]$ and bound the second factor by its supremum.

1.13 (a) To prove the first inequality, use Jensen inequality (1.19) for the random vector $X = (X_1, \ldots, X_n)$. Guess which norm you should use here. To prove the second inequality, bound the maximum of n nonnegative numbers by the sum.
(c) Consider independent Bernoulli random variables $\mathrm{Ber}(p_n)$; find the value of p_n that make the argument work.

1.14 Both bounds follow from Jensen inequality. For the first bound, use (1.19) for the for the random vector $X = (X_1, \ldots, X_n)$ and the ℓ^p norm. For the second bound, consider the convex function $\phi(x) = x^p$.

1.15 (b) Use the integrated tail formula for $f(X)$ and make a change of variable $t = f(s)$.

1.16 Consider the event $E = \{X > \varepsilon \mathbb{E}\, X\}$ and decompose $\mathbb{E}\, X$ into $\mathbb{E}\, X\mathbf{1}_E$ and $\mathbb{E}\, X\mathbf{1}_{E^c}$.

1.17 (a) To prove the lower bound for $q < \infty$, assume first that all coefficients of x satisfy $|x_i| \le 1$, deduce that $|x_i|^q \le |x_i|^p$ and sum these inequalities. To prove the upper bound for $q < \infty$, use the Hölder inequality with exponent p/q.

1.18 Use the result of Exercise 1.17 for $q = \infty$.

1.19 (a) Assume first that $x_i \ge 0$ for all i. In the case where $p = 1, p' = \infty$, set $y = (1, \ldots, 1)$. In the case where $p = \infty, p' = 1$, set $y = (0, \ldots, 0, 1, 0, \ldots, 0)$ where the value 1 is at the coordinate i_0 for which $|x_{i_0}| = \|x\|_\infty$. In the case where $p, p' \in (1, \infty)$, set $y_i := |x_i|^{p/p'}$ for all i.

2.1 To prove the upper bound, take square root on both sides of the inequality $Y_n \ge \mathbb{E}\, Y_n$ and apply Markov inequality. To prove the lower bound, consider the event where all $X_i \ge 1/2$.

2.3 (b) Using the formula from part (a), express the Gaussian tail as $-\frac{1}{\sqrt{2\pi}} \int_t^\infty \frac{f'(x)}{x}\, dx$ and integrate by parts with $u = 1/x$ and $v = f(x)$. Repeat.

2.4 (b) Integrate by parts and then use Proposition 2.1.2.

2.5 Compare Taylor expansions of both sides term by term.

2.7 Use the exponential moment method to bound the probability $\mathbb{P}\{\sum_{i=1}^N (-X_i/\varepsilon) \ge -N\}$.

2.8 By convexity, the graph of the function $f(x) = e^{\lambda x}$ lies below the linear segment that joins the points $(a, f(a))$ and $(b, f(b))$. Write down this observation as an inequality, substitute $x = X$ and take expectation on both sides.

2.9 (a) Use translation, dilation and the comparison inequality (Exercise 2.8).
(b) You should get $K(\lambda) = \lambda a + \log(b - ae^\lambda)$. Check that $K''(\lambda) = -abe^\lambda/(-ae^\lambda + b)^2$ and use the AM-GM inequality $\sqrt{xy} \le \frac{x+y}{2}$ for $x = -ae^\lambda$ and $y = b$. Write down the linear approximation of $K(\lambda)$ using Taylor theorem with a remainder in Lagrange form.

2.10 Follow the proof of Theorem 2.2.1. Use Hoeffding lemma to bound the MGF of each term.

2.11 Note that $\mathbb{P}\{S_N \le t\} = \mathbb{P}\{-S_N \ge -t\}$ and proceed as in the proof of Chernoff inequality.

2.12 Express the probability in terms of binomial coefficients. To lower bound the binomial coefficient, use the result of Exercise 0.6. To handle one of the remaining terms $(1 - \mu/N)^{N-t}$, prove that the smaller quantity $(1 - \mu/N)^{N-\mu}$ is bounded below by $e^{-\mu}$.

2.14 Check and use the numeric inequality $\ln(1 + x) \ge x/(1 + x/2)$ for all $x \ge 0$.

2.15 Use the MGF comparison inequality (Exercise 2.8).

2.16 Argue that we can assume that $t \ge 10$ without loss of generality. Choose $B = \lfloor t^2/4 \rfloor$. If B does not divide N, make one of the blocks larger than B.

2.17 (a) Using triangle inequality, check that the ratio of the densities of $X_i \sim \mathrm{Lap}(0, 1)$ and $Y_i \sim \mathrm{Lap}(\mu, 1)$ is uniformly bounded by $e^{-\mu}$. Write down the densities of X and Y. Express the probabilities $\mathbb{P}\{X \in B\}$ and $\mathbb{P}\{Y \in B\}$ in terms of the densities of X and Y.

2.18 The number of "bad" vertices has binomial distribution. Compute its mean and use Markov inequality.

2.19 The upper bound follows from Chernoff inequality (Theorem 2.3.1) as in Proposition 2.5.1. The lower bound is trickier because the degrees are not independent. (Why?) You can either try the second moment method (Exercise 1.10) or make independent proxies for the degrees as follows. Divide the set of vertices into two subsets V' and V'' of roughly the same size. For each vertex $i \in V''$, consider the number of vertices in V' connected to i. These *degrees into* V', denoted d_i', are independent and provide a lower bound for the true degrees d_i. Use the reverse Chernoff inequality (Exercise 2.12).

2.20 For a given pair of subsets S, T, the number of edges $e(S,T)$ is a binomial random variable. Use Chernoff inequality followed by a union bound over all pair of subsets S, T.

2.21 Apply Hoeffding inequality for the indicators of the wrong answers.

2.22 (a) Express $\mathbb{E}|g|^p$ as an integral; change variables to express it as the gamma function (1.30).
(b) Use Stirling approximation (1.31).

2.23 Use Jensen inequality for the exponential function.

2.26 In the forward direction, use the definition and apply Jensen inequality.

2.27 (a) Use Gaussian tail bound (Proposition 2.1.2).

2.28 Use the integrated tail formula (Lemma 1.6.1) and the result of Exercise 2.27.

2.30 Combine Paley-Zygmund inequality (Exercise 1.16) with the subgaussian Khinchine inequality (Theorem 2.7.5).

2.31 Restate the bound (2.14) in terms of the subgaussian norm. Use it for $a_i = 1/N$ and apply centering (Lemma 2.7.8).

2.32 To prove that $\|X+Y\|_{\psi_2} \ge c\|X\|_{\psi_2}$, compute the MGF of $X+Y$. Prove and use the fact that the MGF of Y is bounded below by 1.

2.33 (a) Express the MGF of the sum in terms the MGF of X.

2.34 (a) Let X_i follow a scaled, symmetrized Bernoulli distribution, meaning that X_i takes values $\pm q_i$ with probability $p_i/2$ each and 0 with probability $1-p_i$. Let the parameters p_i decay rapidly (at a doubly exponential rate) and choose the scaling q_i to make $\|X_i\|_{\psi_2} \asymp 1$.

2.35 (a) Assume $b=1$ and bound the L^p norm of X by $C\sqrt{p/\log(2/a)}$.
(b) Consider a scaled Bernoulli random variable and use Exercise 2.24(e).

2.36 (a) Represent $Z^2 = Z^{1/2}Z^{3/2}$ and use the Cauchy-Schwarz inequality.
(b) Combine the extrapolation trick with Khintchine inequality (Theorem 2.7.5) for $p=3$.

2.37 Use the union bound along with the subgaussian tail bound (Proposition 2.6.6(i)).

2.38 (a) Use Jensen inequality for $\exp(\lambda \cdot \mathbb{E}\max_{i \le N} g_i)$ and replace the maximum of exponentials by their sum. Optimize in λ.
(b) Using Proposition 2.1.2, show that $\max_{i \le N}|g_i| \ge \sqrt{(1-\varepsilon)2\ln N}$ with probability $\to 1$.

2.39 On the one hand, the probability that $\max_i|X_i| < 2\sqrt{\log N}$ can be bounded below by Markov inequality. On the other hand, it can be expressed in terms of the tail probability of $|X|$.

2.40 (a) To prove the triangle inequality $\|X+Y\|_G \le \|X\|_G + \|Y\|_G$, first assume that X and Y are mean-zero. To bound the MGF of $X+Y$, use the Hölder inequality with conjugate exponents that you optimize in the end.
(c) is also convenient to prove for mean-zero random variables first.

2.42 (a) To check that $\|X\|_\psi = 0$ implies $X=0$ a.s., use Jensen inequality. To prove the triangle inequality, write $\frac{|X+Y|}{K+L} \le \frac{|X|+|Y|}{K+L} = \frac{K}{K+L}\frac{|X|}{K} + \frac{L}{K+L}\frac{|Y|}{L}$.

2.43 Comparing Propositions 2.6.1 (corresponding to $\alpha = 2$) and 2.8.1 (corresponding to $\alpha = 1$), you should be able guess the result for general α.

2.46 Express the bound in Corollary 2.9.2 as a sum of two exponentials, and use the integrated tail formula as in the proof of Proposition 2.6.1(i)⇒(ii).

2.47 Check the numeric inequalty $e^z \le 1 + z + \frac{z^2/2}{1-|z|/3}$ for z satisfying $|z| < 3$, apply it for $z = \lambda X$, and take expectations on both sides.

3.1 First use Exercise 0.2 for $a = \sqrt{n}$, and then use Theorem 3.1.1.

3.2 Begin as in the solution of Exercise 3.1, and then use the identity $x - y = (x^2 - y^2)/(x + y)$ for $x = \|X\|_2$ and $y = \sqrt{n}$.

3.3 To prove the bound on $\mathbb{E}\|X\|_2$, consider the Taylor approximation of $\sqrt{z}$ around $z = 1$ up to a cubic term, and use it for $z = \frac{1}{n}\|X\|_2^2$.

3.4 (a) Write the spectral projection as $P_k = \sum_{i=1}^k v_i v_i^\mathsf{T}$, compute $\|P_k X\|_2^2$ and use (3.6).
(b) Write the projection as $P = \sum_{i=1}^k u_i u_i^\mathsf{T}$ for some orthonormal vectors u_i. Proceed as in part (a). Arguing like in the proof of Proposition 3.2.2, express $\mathbb{E}\|PX\|_2^2$ as $\sum_{j=1}^n \lambda_j a_j$ for some a_j satisfying $a_j \le 1$ and $\sum_{j=1}^n \lambda_j a_j \le \sum_{j=1}^n \lambda_k$. Conclude that $\sum_{j=1}^n \lambda_j a_j \le \sum_{j=1}^k \lambda_j$.

3.5 For $p \le \log n$, use Jensen inequality as in Exercise 1.14 and then recall how the absolute moments of subgaussian random variables grow.

3.6 For $p \ge \log n$, use Exercise 2.38. For $p \le \log n$, partition the set $\{1, \dots, n\}$ into approximately n/e^p disjoint subsets of cardinality approximately e^p each. When computing the ℓ^p norm of X, replace the sum of $|X_i|^p$ on each subset by the maximum; use Exercise 1.14 to push the expected value inside, and then use Exercise 2.38 to lower bound each expected maximum.

3.7 (a) While this inequality does not follow directly from the result of Exercise 2.7 (why?), you can prove it by a similar argument. Assume that $a = 0$, square both sides of the inequality $\|X\|_2 \le \varepsilon\sqrt{n}$, choose a parameter $\lambda > 0$, multiply both sides by $-\lambda^2/\varepsilon^2$, exponentiate, and apply Markov inequality. At the end, optimize the bound in λ.
(b) Rewrite the conclusion of part (a) as $\mathbb{P}\{Z \le \delta\} \le \delta^n$ where $Z = CKX/\sqrt{n}$, and compute $\mathbb{E}[1/Z]$ by the integrated tail formula. To avoid a possible singularity, remember that you can always bound a probability by 1.

3.8 (a) With $\mu := \mathbb{E}\, X$, start by expressing $\|\mu\|_2^2 = \langle \mathbb{E}\, X, \mu \rangle$.

3.10 (b) Assume $\mu = 0$ for simplicity. If Σ is invertible, $Z = \Sigma^{-1/2} X$ does the job. Otherwise, arrange the eigenvalues λ_i in (3.7) so that they are are nonzero for $i \le r$ and zero for $i > r$. Invert $\Sigma^{1/2}$ where it can be inverted, and complete it by isotropy on the rest of the space. For example, pick any mean-zero, isotropic random vector Y, and set $Z = \big(\sum_{i \le r} \lambda_i^{-1/2} v_i v_i^\mathsf{T}\big) X + \big(\sum_{i > r} v_i v_i^\mathsf{T}\big) Y$. Verify that $\Sigma^{1/2} Z = X$. (Since $\mathbb{E}\langle X, v_i \rangle^2 = 0$ for $i > r$, we have $\langle X, v_i \rangle = 0$ a.s.)

3.11 (a) For any fixed i, note that $\sigma(i)$ is uniformly distributed over all n indices. Similarly, for any fixed pair of distinct indices (i, j), the pair $(\sigma(i), \sigma(j))$ is uniformly distributed over all $n^2 - n$ (ordered) pairs of distinct indices.

3.13 (a) Combine Theorem 3.1.1 with Proposition 2.7.6.
(b) Prove separately that $\mathbb{E}\|X_1\|_2 \gtrsim \sqrt{n}$ and $\mathbb{E} \max_{i \le N} \|X_i\|_2 \gtrsim \sqrt{\log N}$ using Exercise 2.38.

3.14 Recall (3.6).

3.16 (a) Use Cramer-Wold device from the proof of Proposition 3.3.5. First, verify that X has finite mean μ and covariance matrix Σ. Then, compute the means and variances of all 1D marginals of X. Using the assumption, show that these match the means and variances of the 1D marginals of $Y \sim N(\mu, \Sigma)$.

3.17 Randomly flip the sign of $X \sim N(0,1)$.

3.18 Think of G and UG as vectors in $\mathbb{R}^{n\times n}$. Check that the mapping $G \mapsto UG$ is a linear orthogonal transformation, then use the rotation invariance of the normal distribution.

3.19 Represent A using part (a) and use the invariance for G established in Exercise 3.18.

3.20 Check that the random vector $(Gu, Gv) \in \mathbb{R}^{2m}$ obtained by concatenation is isotropic.

3.23 (a) Choose x to be the unit vector in the direction of g_1. Check that, with high probability, $\langle g_1, x\rangle > \sqrt{n}/2$ while $\langle g_1, x\rangle < \sqrt{n}/2$ for all $j = 2, \dots, N$.

3.24 Let $r = \|X\|_2$. Compute the CDF of r and differentiate to deduce the density.

3.25 Use Exercise 3.24 to check that $\mathbb{E}\|Y\|_2^2 = n$. Then argue like in the proof of Proposition 3.3.8.

3.26 Argue that we can write $\|X\|_\infty = \|g\|_\infty / \|g\|_2$ where $g \sim N(0, I_n)$. To prove the upper bound, use the Cauchy-Schwarz inequality and a small ball probability bound (Exercise 3.7). For the lower bound, use concentration of the norm to control $\|g\|_2$ and a direct computation using independence to control $\|g\|_\infty$.

3.27 First, consider $n = 2$ (the unit circle in the plane), and compute the density of the first coordinate of X. Make a plot and note that the density at x is proportional to the arclength that projects onto the interval $[x, x+\epsilon]$. Check that this is proportional to $g(x) = \varepsilon/\sqrt{1-x^2} + o(\varepsilon)$. Now try $n = 3$ and then a general n.

3.28 (b) You should be able to express the squared distance to the cube as $\sum_{i=1}^n (|g_i| - a)_+^2$, where $x_+ = \max(x, 0)$. Show that the expected distance is small if a is large enough. On the other hand, $\|g\|_2$ is unlikely to be very small due to the concentration of the norm.

3.30 For example, use the criterion in Exercise 3.29.

3.33 Note that $\langle UX, v\rangle = \langle X, U^\mathsf{T} v\rangle$ and that U^T is an orthogonal matrix.

3.34 Note that Av is a mean-zero random vector with independent coordinates. Use Lemma 3.4.2 to bound its subgaussian norm.

3.35 Consider a one-dimensional marginal of $\sum_i X_i$, and use Proposition 2.7.1 to bound its subgaussian norm.

3.36 (a) Argue as in Lemma 3.4.2, but use triangle inequality instead of Proposition 2.7.1.

3.38 Exercise 3.14 describes the distribution of 1D marginals of X. Exercise 2.24(c) gives their subgaussian norm. Proposition 3.2.2 allows you to maximize that expression.

3.39 (a) Apply Lemma 2.8.5 and the triangle inequality.

3.40 (a) If Y is an independent copy of X, you can use Bernstein's inequality to bound $|\langle X, Y\rangle|$ above, and concentration of norm to bound $\|X\|_2$ and $\|Y\|_2$ below, all with high probability. (b) An idea is to slightly modify an isotropic vector near the origin. For example, flip a biased coin. If it is heads (likely), pick X as the first basis vector, appropriately scaled. If it is tails (unlikely), sample X from a normal distribution on the orthogonal hyperplane $\mathbb{R}^{n-1}$.

3.41 (a) Use a conditioning trick: condition on Y and apply Theorem 3.4.5.

3.42 Use the decomposition from Exercise 3.24.

3.43 To bound the subgaussian norm of a 1D marginal $\langle X, v\rangle$, first get basic bounds on the L^2 and L^∞ norms, and then interpolate them using Exercise 2.35. For the lower bound, pick v as a standard basis vector and use Exercise 2.24(e).

3.44 (a) Argue that the density of X_1 at $u \in [-r, r]$ is proportional to $(r - |u|)^{n-1}$.
(b) Use symmetry to check that the coordinates of X are uncorrelated.
(c) Compute $\mathbb{P}\{X_1 > r/2\}$.

3.45 Note that $\|\langle X, x\rangle\|_{\psi_2} \le K\|x\|_2$ and write down what it means by definition of the subgaussian norm.

3.46 (b) Combine Exercise 3.45 with the identity $\mathbb{E}\|X\|_2^2 = n$, which follows from isotropy.

3.47 (a)⇒(b): for any fixed x, the function $y \mapsto x^\mathsf{T} Ay$ is linear and thus convex. Now recall (1.5) and Exercise 1.4.

3.49 (b) First, check that the function $f(z) = z^\mathsf{T} Az$ is convex, and so it must attain its maximum in $[-1, 1]^n$ on the vertices of the cube. Now fix any $x, y \in \{-1, 1\}^n$, use polarization identity to bound the bilinear form $|x^\mathsf{T} Ay|$ by two quadratic forms, and bound these two quadratic forms using the fact above. Finally, apply the classical Grothendieck inequality (Theorem 3.5.1).

3.50 (b) Check that $f(z) = |z^\mathsf{T} Az|$ is separately convex. Then proceed as in Exercise 3.49.

3.52 (a) Express the objective function as $\frac{1}{2}\operatorname{tr}(\tilde{A}ZZ^\mathsf{T})$, where $\tilde{A} = \left[\begin{smallmatrix} 0 & A \\ A^\mathsf{T} & 0 \end{smallmatrix}\right]$ is called the *Hermitian dilation* of the matrix A, and $Z = \left[\begin{smallmatrix} X \\ Y \end{smallmatrix}\right]$ where X and Y are the matrices with rows X_i^T and Y_j^T, respectively. Note that ZZ^T is the Gram matrix of the unit vectors $X_1, \dots, X_m, Y_1, \dots, Y_n$. Proceed as in the proof of Proposition 3.5.6. (b) Use Grothendieck inequality.

3.53 Use the rotation invariance of the standard normal distribution to reduce the problem to $\mathbb{R}^2$. Once in the plane, apply rotation invariance again to show that the probability of $\langle g, u\rangle$ and $\langle g, v\rangle$ having opposite signs is α/π, where $\alpha \in [0, \pi]$ is the angle between u and v. A picture might help!

3.54 Consider cutting G repeatedly. To find a lower bound on the probability of a success (finding a large cut), use Paley-Zygmund inequality (Exercise 1.16). Express the the expected number of experiments until success in terms of this probability.

3.56 (a) Argue that we can assume that $v = (1, 0, \dots, 0)$ and $u = (u_1, u_2, 0, \dots, 0)$.
(b) Expand $\mathbb{E}\, Z_u Z_v$ into four terms, and use Exercise 3.9 and part (a).

3.57 Sum up the identities from Exercise 3.56(b) for $u = u_i$ and $v = u_j$ with weights a_{ij}.

3.58 Use randomized rounding (3.34).

4.4 (a) Use the spectral representation of both norms (Lemma 4.1.11). (b) Use Proposition 3.2.1(b). (c) Use (b) to write $\|BA\|_F^2 = \mathbb{E}\|BAg\|_2^2$ where $g \sim N(0, I_n)$.

4.5 In fact, any nonnegative, weakly decreasing numbers s_k whose squares sum to a^2 satisfy $s_k \le a/\sqrt{k}$.

4.6 Express $\|A^k x\|_2^2$ in terms of the SVD of A. Show that the first term of the sum dominates.

4.7 (a) Write $Ax = \sum_{i=1}^n x_i A_i$. (b) For a contradiction, assume that $\|A\| = \|A_1\|_2$ but A_1 is not orthogonal to A_2. Take $x = (1, \varepsilon, 0, 0, \dots, 0)$ and show that the function $f(\varepsilon) = \|Ax\|_2^2 - \|A\|^2\|x\|_2^2$ can take positive values, contradicting the definition of the operator norm.

4.8 To bound $|x^\mathsf{T} Ay|$ for unit vectors x and y, use the triangle inequality: $|x^\mathsf{T} Ay| = |\sum_{ij} A_{ij} x_i y_j| \le \sum_{ij} |A_{ij} x_i y_j|$. Now apply the Cauchy-Schwarz inequality, splitting $|A_{ij}|$ between the two terms.

4.10 To prove achievability, consider a matrix whose all entries equal 1, and on the other hand Walsh matrix (Exercise 4.9).

4.11 For a unit vector x, find $\|Px - x/2\|_2^2$ and $\|Qx - x/2\|_2^2$, then apply the triangle inequality.

4.12 (a) A neat trick: block-diagonalize $P - Q$. Take an orthogonal matrix $\bar{U} = \begin{bmatrix} U & U_\perp \end{bmatrix}$ whose first columns form U and span $\operatorname{Im} P$. Similarly, take an orthogonal matrix $\bar{V} = \begin{bmatrix} V & V_\perp \end{bmatrix}$ whose first columns form V and span $\operatorname{Im} Q$. Then compute $\bar{U}^\mathsf{T}(P - Q)\bar{V}^\mathsf{T}$ and realize that it is a block matrix whose norm is the maximal norm of the two blocks.

4.13 Use Exercise 4.12 and Lemma 4.1.16; follow the Davis-Kahan proof (Theorem 4.1.15).

4.14 For "the only" part, compute H^2. It is a block-diagonal matrix, so its eigenvalues should be easy to compute.

4.15 Apply Davis-Kahan inequality (Exercise 4.13 and Theorem 4.1.15) for the Hermitian dilation of A (Exercise 4.14).

4.17 In one direction, assume that property (c) in Lemma 4.1.17 holds, and consider the spectral projection of P onto the top n eigenvalues of AA^T (what are they?). For the opposite direction, use Weyl inequality (Lemma 4.1.14).

4.19 Use convexity: recall from (1.3) that B_1^n, the unit ball of $\mathbb{R}^n$ in the ℓ^1 norm, is the convex hull of ± canonical basis vectors; then use the maximal principle (Exercise 1.4). Deduce the result for the $2 \to \infty$ norm by duality (Exercise 4.18(c)).

4.20 (a) Write $\|A\|_{\infty\to 1}$ as the maximum of $|x^\mathsf{T}Ay|$ over the unit cubes, and use convexity as in Exercise 4.19. (b) One direction is trivial by setting $Z = xy^\mathsf{T}$. In the other direction, express a rank-one matrix Z as $Z = uv^\mathsf{T}$, argue that $\|u\|_\infty\|v\|_\infty \le 1$, and rescale u and v to rewrite Z as $Z = xy^\mathsf{T}$ with $\|x\|_\infty = 1$ and $\|y\|_\infty \le 1$.

4.21 To bound $\|A\|_{\infty\to 1}$, express it as in Exercise 4.20(a). Decompose $x^\mathsf{T}Ay = \sum_{i,j} A_{ij}x_iy_j$ into four sums according to the value of $(x_i, y_j) \in \{-1,1\}^2$, and bound each sum by the cut norm of A. To bound the cut norm of A, fix subsets I and J and write $\sum_{i\in I\, j\in J} A_{ij}$ as $\sum_{i,j} A_{ij}(\frac{1+x_i}{2})(\frac{1+y_i}{2})$ for some vectors x and y with ±1 entries. Expand into four sums and bound each by the $\infty \to 1$ norm of A.

4.22 For the $\|A\|_{\infty\to 1}$ norm, use Exercise 3.52. For the $\|A\|_{\infty\to 2}$ norm, express it as the maximum of the quadratic form $\|Ax\|_2^2 = x^\mathsf{T}(A^\mathsf{T}A)x$ and use Theorem 3.5.7.

4.24 (a) Consider a disconnected metric space, such as $T = \{-1, 1\}$ with with the usual Euclidean metric. (b) Argue by contradiction: if two balls are not ε-separated, the arithmetic mean of their centers lies in both $\varepsilon/2$-balls.

4.25 If an $\varepsilon/2$-ball centered at $x \notin K$ covers some point in K, say $y \in K$, the ε-ball centered at y works even better in terms of covering K.

4.26 (a) Removing a point from K makes it impossible to place a ball with that center. Make an example where K is a 3-point set. (b) can be proved similarly to Exercise 4.25.

4.27 (a) Try induction on n.

4.28 Use Exercise 3.7 for i.i.d. random variables $X_i \sim \mathrm{Unif}[-\frac{1}{2}, \frac{1}{2}]$; it gives a bound on the volume of the intersection of a unit cube with the Euclidean ball of radius $\varepsilon\sqrt{n}$. Take a very small ε.

4.29 (a) This is a version of the integrated tail formula (Exercise 1.15(b)) for volume instead of probability. To deduce it, modify the proof of Lemma 1.6.1 by writing $f(\|x\|) = -\int_{\|x\|}^\infty f'(t)dt$.

4.30 (a) Exercise 1.17(a)) will help with the sandwiching.

4.31 To bound $|\mathcal{N}|$, note that the cubes $y + \frac{\varepsilon}{2\sqrt{n}}(-1,1)^n$ where $y \in \mathcal{N}$ are disjoint and contained in $(1 + \frac{\varepsilon}{2})B_2^n$. Run the volumetric argument like in the proof of Corollary 4.2.11.

4.33 Deduce from the error correction assumption that the closed balls of radius r centered at points $E(x)$ are disjoint. Then run a version of a volumetric argument.

4.34 (a) Construct the expansion iteratively. Approximate x by $x_1 \in \mathcal{N}$, normalize the residual $x - x_1$ and approximate it by $x_2 \in \mathcal{N}$, etc.

4.36 Use Lemma 4.4.1 to approximate $\|A\|$ with $\|Ax\|_2$ for some $x \in \mathcal{N}$; then use Exercise 4.35 to approximate $\|Ax\|_2$ with $\langle Ax, y\rangle$ for some $y \in \mathcal{M}$. For symmetric matrices, modify the proof of Lemma 4.4.1 using the identity $x^\mathsf{T}Ax - x_0^\mathsf{T}Ax_0 = (x - x_0)^\mathsf{T}A(x + x_0)$.

4.37 Assume without loss of generality that $\mu = 1$. Represent $\|Ax\|_2^2 - 1$ as a quadratic form $\langle Rx, x\rangle$ where $R = A^\mathsf{T} A - I_n$. Use Exercise 4.36 to compute the maximum of this quadratic form on a net.

4.38 Use the ε-net expansion (Exercise 4.34).

4.39 Fix an $\varepsilon/2$-net $\mathcal{M}$ of RB_2^n with good size. It is enough to show that each point $y \in \mathcal{M}$ is within distance $\varepsilon/2$ from some point $g_i/\sqrt{n} \in RB_2^n$. For each $y \in \mathcal{M}$, fix a ball B_y with radius $r = \frac{1}{4}\min(\varepsilon, R)$ satisfying $y \in B_y \subset RB_2^n$. Find a lower bound on the probability that $g_1/\sqrt{n}$ lands in a given ball B_y. Use independence to show that at least one of the N vectors $g_i/\sqrt{n}$ does so with high probability. Finally, apply a union bound over $\mathcal{M}$.

4.40 Use Exercise 4.39 to show that the scaled random vectors $g_i/\sqrt{n}$, $i > 1$, form an good net of ball of radius 10. Use Exercise 4.38 to show that the convex hull of these points contains a ball of radius 5. Meanwhile, with high probability, the first vector $g_1/\sqrt{n}$ falls in the radius 5 ball, and thus in the convex hull of the other vectors.

4.41 Use the integrated tail formula (Lemma 1.6.1).

4.42 Bound the operator norm of A below by the Euclidean norm of the first column and first row. Use concentration of the norm (3.2) to complete the proof.

4.43 The proof of Theorem 4.4.3 only relies on one property of A: that $\langle Ax, y\rangle = \langle A, xy^\mathsf{T}\rangle$ is subgaussian for any unit vectors x, y.

4.44 (a) Combine Exercise 4.19(a) with (2.22). (b) Combine Exercise 4.19(b) with Exercise 3.13.

4.45 The challenge is that, if q is small, the second eigenvalue of the expected adjacency matrix $D = \mathbb{E}\,A$ might not be well separated from the first, making it hard to use Davis-Kahan inequality. To fix this, modify the algorithm to consider *both* top eigenvectors of D. The spectral projections onto the top two eigenvectors of D and A will still be close, thanks to a higher-dimensional version of Davis-Kahan (Lemma 4.1.16, Exercise 4.13).

4.47 Using the min-max theorem (Corollary 4.1.7), reduce the problem to the smallest singular value of a $m \times k$ random matrix. (Pick E to be a coordinate subspace.)

4.48 Rewrite the quantity in question as $\|\bar{\Sigma}_m - I_n\|$ where $\bar{\Sigma}_m$ is the sample covariance matrix of the standard score of X.

4.50 (b) The idea is to use SVD to approximate a matrix $A = U\Sigma V^\mathsf{T} \approx U_0\Sigma_0 V_0^\mathsf{T}$ by replacing one factor at a time. Construct an $(\varepsilon/3)$-net to approximate U using (b), an $(\varepsilon/3)$-net to approximate V (similarly) and $(\varepsilon/3)$-net to approximate Σ. You will find Exercise 4.4 helpful. (c) Consider $m \times n$ matrices with Frobenius norm ≤ 1 and with only r first nonzero columns.

4.51 (b) Make a *quantitative* conclusion for 99.5% points: not just that the signs of $\langle X_i, u\rangle$ and θ_i agree, but that $\langle X_i, u\rangle$ is well separated from zero, e.g. $|\langle X_i, u\rangle - \theta_i\beta| \le \frac{\beta}{2}$ where $\beta = \|\mu\|_2$. (e) Now that you know from (d) that $v \approx u$, conclude that 99.5% of the points X_i satisfy $|\langle X_i, v\rangle - \langle X_i, u\rangle| \le 0.1\beta$. (To do this, write the sum of the squares of these quantities, and express it in terms of Σ_m and $u - v$.) Now combine with (b).

5.3 If the conclusion of the first part fails, the complement $B \coloneqq (A_s)^c$ satisfies $\sigma(B) \ge 1/2$. Apply the blow-up Lemma 5.1.6 for B.

5.6 For the upper bound, assume that $\|Z - \mathbb{E}\,Z\|_{\psi_2} \le K$ and use the definition of the median to show that $|M - \mathbb{E}\,Z| \le CK$.

5.7 First replace the expectation by the median using Exercise 5.6. Then apply the assumption for the function $f(x) \coloneqq \operatorname{dist}(x, A) = \inf\{d(x,y) :\ y \in A\}$ whose median is zero.

5.8 The ε-neighborhood of a half-space is still a half-space, and its Gaussian measure should be easy to compute.

5.9 (a) Use Gaussian concentration for $f(x) = \max_i x_i$. (b) Argue that we can express $X_i = \mu_i + \langle g, v_i\rangle$ for some non-random μ_i, v_i and $g \in N(0, I_n)$. Then use Gaussian concentration.

5.10 Bound $|\mu_p - \mu_1|$ where $\mu_p := \|Z\|_{L^p}$.

5.14 Apply norm concentration (Theorem 3.1.1) for the random vector Az.

5.15 Consider an orthogonal basis in $\mathbb{R}^N$.

5.17 (a) Symmetric commuting matrices are simultaneously diagonalizable.
(b) Find 2×2 matrices such that $0 \preceq X \preceq Y$ but $X^2 \npreceq Y^2$.

5.18 (a) First consider the case where $Y = I_n$. Next, multiply the inequality $0 \preceq X \preceq Y$ by $Y^{-1/2}$ (justify why you can do that), and use the identity case proved before.
(c) Argue that the matrix version of the formula in (b) is $\ln X = \int_0^\infty \left[(1+t)^{-1} I_n - (X + tI_n)^{-1}\right] dt$, and use part (a).

5.20 Check that matrix Bernstein inequality implies that $\left\|\sum_{i=1}^N X_i\right\| \lesssim \left\|\sum_{i=1}^N \mathbb{E}\, X_i^2\right\|^{1/2} \sqrt{\log n + u} + K(\log n + u)$ with probability at least $1 - 2e^{-u}$. Then use the integrated tail formula from Lemma 1.6.1.

5.21 Proceed like in the proof of Theorem 5.4.1. Instead of Lemma 5.4.10, check that $\mathbb{E}\exp(\lambda \varepsilon_i A_i) \preceq \exp(\lambda^2 A_i^2/2)$ just like in the proof of Hoeffding inequality (Theorem 2.2.1).

5.22 Use the integrated tail formula from Exercise 1.15(c).

5.23 Use Hermitian dilation (Exercise 4.14).

5.24 Use Hermitian dilation (Exercise 4.14).

5.25 Loops contribute only to the diagonal of the adjacency matrix. Isolate this diagonal matrix and bound its norm.

5.28 (a) Let $X \sim \mathrm{Unif}(\sqrt{n}e_1, \ldots, \sqrt{n}e_n)$ where e_i are the standard basis vectors in $\mathbb{R}^n$. Check that X is isotropic. Argue that, if $m \ll n \log n$, the sample $X_1, \ldots, X_n$ with high probability misses at least one basis vector (this is related to a coupon collector problem), making the sample covariance matrix Σ_m have at least one zero on the diagonal.

5.30 Use covariance estimation for the isotropic random vector $X \sim \mathrm{Unif}(\sqrt{m}u_1, \ldots, \sqrt{m}u_m)$ (recall Proposition 3.3.11(iv)).

5.31 Just like in the proof of Theorem 4.6.1, derive the conclusion from a bound on $\frac{1}{m}A^\mathsf{T} A - I_n = \frac{1}{m}\sum_{i=1}^m A_i A_i^\mathsf{T} - I_n$. Use the high-probability version of covariance estimation (5.28).

5.32 Consider the random vector $X \in \mathbb{R}^n$ that picks the random row of A, i.e. $X = A_i$ with probability $1/N$, and use the general covariance estimation result (Theorem 5.6.1) followed by Weyl inequality (Lemma 4.1.14).

6.1 Modify the proof of Theorem 6.1.1. Remove the diagonal from A and do steps 1 and 2. Then, in step 3, include the diagonal into Z.

6.2 (a) Use Theorem 6.1.1 for $F(x) = x^p$. (b) Which F is naturally suited for this?

6.4 We want to decouple $\mathbb{E}\|\sum_{i,j} A_{ij} X_i X_j\|$ where X_i are i.i.d. $\mathrm{Ber}(p)$ and A_{ij} is the $n \times n$ matrix with only the (i, j) entry of A kept. Follow the proof of Theorem 6.1.1 using the operator norm as F. Step 3 is trivial thanks to Exercise 4.2(d).

6.5 Let X have a rotation-invariant distribution, and let $\|X\|_2$ take two values: n with probability $1/n$ and $\sqrt{n}$ with probability $1 - 1/n$.

6.6 (a) Use Proposition 6.2.1.

6.7 Use the singular value decomposition for A and rotation invariance of $X \sim N(0, I_n)$ to simplify and control the quadratic form $X^\mathsf{T} AX$.

6.8 The quadratic form is now $X^\mathsf{T} AX$ where X is a $d \times n$ random matrix with columns X_i. Apply Gaussian replacement (Lemma 6.2.3) and redo the Gaussian MGF computation (Lemma 6.2.4).

6.9 First swap X for $g \sim N(0, I_m)$ like in Proposition 6.2.1. Then, compute using Gaussian properties as in Lemma 6.2.4.

6.10 Start as in the proof of Proposition 6.2.1 and use the MGF bound from Exercise 6.9.

6.11 Find B such that $A = B^\mathsf{T} B$ and use Exercise 6.10.

6.12 First, use Exercise 6.10 to show that if a random vector Y has mean-zero, covariance matrix Σ and satisfies $\|\langle Y, u\rangle\|_{\psi_2} \le K\|\langle Y, u\rangle\|_{L^2}$ for all $u \in \mathbb{R}^n$, then $\|Y\|_2 \le CK\sqrt{\operatorname{tr}\Sigma} + CK\sqrt{\|\Sigma\| \log(1/\alpha)}$ with probability at least $1-\alpha$. Then apply this for $Y = \frac{1}{\sqrt{N}}\sum_{i=1}^N (X_i - \mu)$.

6.13 Apply Hanson-Wright inequality (Theorem 6.2.2) for $A = B^\mathsf{T} B$ and $t = \varepsilon\|B\|_F^2$, and replace $\|B^\mathsf{T} B\|_F$ by $\|B^\mathsf{T}\|\|B\|_F$ using Exercise 4.4(c). This will give a deviation inequality for $\|BX\|_2^2$. Convert it to a deviation inequality for $\|BX\|_2$ like in the proof of Theorem 3.1.1.

6.15 Express the cut as a quadratic form like in (3.31) and apply Hanson-Wright inequality (Theorem 6.2.2). You can compute the Frobenius norm of A exactly, and the operator norm can only be smaller.

6.18 To prove $\mathbb{E}\|X + v\| \ge \mathbb{E}\|X\|$, argue that $\mathbb{E}\|X + v\| = \mathbb{E}\|X + \varepsilon v\|$ where ε is an independent Rademacher, and push the expectation over ε inside the norm.

6.21 Use the result of Exercise 6.20 with $F(x) = \exp(\lambda x)$ to bound the moment generating function, or with $F(x) = \exp(cx^2)$.

6.22 Argue that the probability equals $\mathbb{P}\{|\sum_{i=1}^N \varepsilon_i X_i| \ge t(\sum_{i=1}^N X_i^2)^{1/2}\}$ where ε_i are independent Rademacher random variables. Condition on (X_i) and use Hoeffding inmequality to bound the conditional probability. Then use the law of total expectation.

6.23 (a) Use the ℓ^p norm definition to express it as a sum, swap p for 2 by monotonicity (Exercise 1.11), apply the variance sum formula, and use monotonicity again.
(b) Use symmetrization (Exercise 6.20), condition on (X_i) and use part (a).

6.24 (a) Swap exponent 2 for p by monotonicity (Exercise 1.11), use the ℓ^p norm definition to express it as a sum, apply Khintchine inequality (Theorem 2.7.5) and then the triangle inequality triangle inequality in $(\mathbb{R}^n, \|\cdot\|_{p/2})$.
(b) Use symmetrization as in Exercise 6.23(b).

6.25 Repeat the proof of Theorem 0.0.2, using Exercises 6.23 and 6.24 instead of the variance-of-sum identity.

6.26 Apply symmetrization, then use Khintchine inequality conditionally on (X_i), and finally the triangle inequality in $L^{p/2}$.

6.27 Using the equality $a^2 - b^2 = (a - b)(a + b)$, reduce the problem to bounding the L^p norm of $\|X\|_2^2 - n$. Expand $\|X\|_2^2$ and use Marcinkiewicz-Zygmund inequality (Exercise 6.26).

6.28 Apply Theorem 6.4.1 for the Hermitial dilation of A (see Exercise 4.14).

6.29 Let A be a block-diagonal matrix with n/k independent blocks, each a $k \times k$ symmetric random matrix with independent Rademacher entries on and above diagonal. Condition on a block being all ones, then pick the value of k at the end.

6.30 Fix i and use Bernstein inequality to get a tail bound for $\sum_{j=1}^n (\delta_{ij} - p)^2$. Conclude by taking a union bound over $i = 1, \ldots, n$.

6.33 Use symmetrization and matrix Khintchine inequality (Theorem 5.4.14) tp bound $\mathbb{E}\|S - \mathbb{E}\, S\|$ in terms of $\mathbb{E}\|\sum_i Z_i^2\|^{1/2}$, then bound the latter quantity by $(\mathbb{E}\max_i \|Z_i\| \cdot \mathbb{E}\|S\|)^{1/2}$. This should give you $\mathbb{E}\|S - \mathbb{E}\, S\| \lesssim \sqrt{\mathbb{E}\|S\| \cdot L}$ where $L = \log(n)\, \mathbb{E}\max_i \|Z_i\|$. Finally, replace $\mathbb{E}\|S\|$ with the larger quantity $\mathbb{E}\|S - \mathbb{E}\, S\| + \|\mathbb{E}\, S\|$ and solve the resulting inequality.

6.34 Apply Theorem 6.33 for $Z_i = \frac{1}{m} X_i X_i^\mathsf{T}$. To bound L, use the assumption and Proposition 3.2.1(b).

6.36 Use symmetrization, contraction principle (Theorem 6.6.1) conditioned on (X_i), and finish by applying symmetrization again.

7.1 (b) Combine $\|X_t - X_s\|_{L^2}$ and $\|X_t + X_s\|_{L^2}$.

7.2 Argue like in the proof of Lemma 6.3.2.

7.3 (a) For any $t, s \in T$, we need to find $t', s' \in T$ so that $t_1 + \phi(t_2) + s_1 - \phi(s_2) \le t'_1 + t'_2 + s'_1 - s'_2$. Set (t', s') as either (t, s) or (s, t). Rewrite the inequality for both cases to see how to choose.
(b) Condition on $\varepsilon_1, \dots, \varepsilon_{n-1}$ and apply part (a).

7.4 Theorem 6.6.1 may help.

7.5 It might be simpler to think about increments $\|X_t - X_s\|_{L^2}$ instead of the covariance matrix.

7.6 Write $X = \Sigma^{1/2} Z$ for $Z \sim N(0, I_n)$ and note that $X_i = \sum_{k=1}^n (\Sigma^{1/2})_{ik} Z_k$ and $\mathbb{E}\, X_i f(X) = \sum_{k=1}^n (\Sigma^{1/2})_{ik}\, \mathbb{E}\, Z_k f(\Sigma^{1/2} Z)$. Apply the univariate Gaussian integration by parts (Lemma 7.2.3) for $\mathbb{E}\, Z_k f(\Sigma^{1/2} Z)$ conditionally on all random variables except $Z_k \sim N(0, 1)$, and simplify.

7.7 (b) The following identity can help simplify the expression: if $\sum_{i=1}^n p_i = 1$, then $\sum_{i,j=1}^n \sigma_{ij}(\delta_{ij} p_i - p_i p_j) = \frac{1}{2} \sum_{i \ne j} (\sigma_{ii} + \sigma_{jj} - 2\sigma_{ij}) p_i p_j$. Check it and use it for $\sigma_{ij} = \Sigma^X_{ij} - \Sigma^Y_{ij}$ and $p_i = p_i(Z(u))$.

7.9 Use Gaussian Interpolation Lemma 7.2.5 for $f(x) = \prod_i \left[1 - \prod_j h(x_{ij})\right]$ where $h(x)$ is an approximation to the indicator function $\mathbf{1}_{\{x \le \tau\}}$, as in the proof of Slepian inequality.

7.10 Add and subtract the cross-term wv^T, factor, and expand the Frobenius norm squared like this: $\|A - B\|_F^2 = \|A\|_F^2 + \|B\|_F^2 - 2\operatorname{tr}(A^\mathsf{T} B)$.

7.12 Consider Taylor expansion of $\sqrt{y}$ about $y = 1$ with three leading terms, and substitute $y = \|g\|_2^2 / n$ to approximate $\mathbb{E}\|g\|_2 / \sqrt{n}$.

7.13 (a) 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 inequality (without the requirement of equal variances) to show that $\mathbb{E}\, s_n(A) \ge \mathbb{E}\|h\|_2 - \mathbb{E}\|g\|_2$ where $h \sim N(0, I_m)$ and $g \sim N(0, I_n)$. Then use Exercise 7.12.
(b) Proceed like in the proof of Corollary 7.3.2, using Weyl inequality (Lemma 4.1.14).

7.15 (b) Use rotation invariance of Gaussian distribution.
(e) Use property (d), replace T with $-T$, and use property (d) again.
(g) Use Sudakov-Fernique comparison inequality.

7.16 The proof of Proposition 7.5.2(f) suggests the two extreme examples: an interval and a ball.

7.18 Argue that we can assume that A is diagonal. To show $\langle A, B \rangle \le \|A\|_* \|B\|$, write the trace as a sum and use that $\max_i |B_{ii}| \le \|B\|$. For the reverse inequality, pick B to be diagonal with entries $\operatorname{sign}(A_{ii})$.

7.17 Use duality (1.6) together with Exercises 3.5 and 3.6.

7.19 Use Exercise 7.18 to write the Gaussian width as the expected nuclear norm of a Gaussian random matrix. Then use the operator norm bound (see Remark 4.4.4) to get an upper bound, and the intermediate singular value bound (Exercise 4.47 with $k = cn$) for the lower bound.

7.20 For the upper bound, write $x = (x - y) + y$, then use triangle inequality and Proposition 7.5.11(a).

7.21 (a) By rotation invariance, assume that E is a coordinate subspace.

7.23 (a) It is easier to compute the squared version of the Gaussian width, $h(A(B_2^n))$, which is equivalent to the original one (Lemma 7.5.11).

7.24 It is enough to check the rotation invariance of the distribution of Bz.

7.26 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.

7.27 Express the operator norm of PA to the diameter of the ellipsoid $P(AB_2^k)$ and use Theorem 7.6.1 in part (a) and Exercise 7.25 in part (b).

8.1 Use Gaussian concentration (Theorem 7.1.11) to get (8.53). Choosing, for example, $z_k = u + \sqrt{k - \kappa}$ should make the union bound go through.

8.4 (a) should be straightforward from Exercise 2.37.
(b) The first m vectors in T form a $(1/\sqrt{\log m})$-separated set.

8.5 Assume $\mathrm{diam}(T) = 1$, replace the integral by the sum using Exercise 8.3, and split the sum into two parts: where $2^{-k} \leq n^{-2}$ (say) and $n^{-2} < 2^{-k} \leq 1$. In the first sum, use the volumetric bound (4.17) to bound the covering numbers. The second sum has $O(\log n)$ terms.

8.6 Start chaining (8.9) at the coarsest scale κ where a single ball covers the entire T, i.e. where $2^{-\kappa} \approx \mathrm{diam}(T)$, but stop early – at scale K where roughly $2^{-K} \approx w(T)/8\sqrt{n}$ (say). The last term in (8.8) may not be zero as before, but rather $\mathbb{E}\sup_{t\in T}(X_t - X_{\pi_K(t)})$. Bound this term by $\frac{1}{2}w(T)$, using that $\|t - \pi_K(t)\|_2 \leq 2^{-K}$.

8.7 Redo the chaining argument in Section 8.1.

8.8 Consider the set $V = \{(s,t) \in T \times T :\ d(s,t) \leq \delta\}$, define the random process $Y_u = X_s - X_t$ indexed by $u = (s,t) \in V$, and define the metric on V by $\rho(u,u') = \frac{1}{2K}\|Y_u - Y_{u'}\|_{\psi_2}$. Check that $\mathrm{diam}(V,\rho) \leq \delta$, bound the covering numbers of (V,ρ) by those of (T,d), and apply Dudley inequality (8.16).

8.9 Lay a grid on the square $[0,1]^2$ with step size ε. Given $f \in \mathcal{F}$, show that $\|f - f_0\|_{L^\infty} \leq \varepsilon$ for some "staircase" function f_0 that follows the grid by stepping up/down by ε or staying flat (see Figure H.1). Bound the number of all staircase functions by $e^{C/\varepsilon}$. Next, use Exercise 4.25.

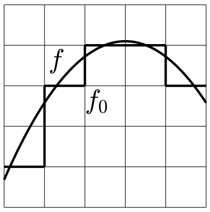


Figure H.1 Approximating a Lipschitz function f by a mesh-following function f_0 (Exercise 8.9).

8.10 (b) Instead of applying Dudley inequality directly, redo at the chaining argument (proof of Theorem 8.1.4) and make an *early stopping* at scale δ, like in the solution of Exercise 8.6. Once you proved (8.54), plug in the covering number bound from (a) and optimize δ.

8.18 The restriction of $\mathcal{F}$ onto a shattered subset $\Lambda \subset \Omega$ gives consists of all Boolean functions on Λ, so its linear algebraic dimension must be at least $|\Lambda|$ (check!).

8.22 (a) Argue as in the proof of Proposition 8.3.11 and use Sauer-Shelah lemma (Lemma 8.3.9).
(b) Try to find an example where $\mathrm{vc}(\mathcal{F}) = \mathrm{vc}(\mathcal{G}) = 1$ while $\mathrm{vc}(\mathcal{F} \cup \mathcal{G}) = 3$.

8.24 Proceed similarly to the proof of Theorem 8.3.15. Combine a concentration inequality with a union bound over the entire class $\mathcal{F}$ restricted onto $\{X_1, \ldots, X_n\}$. Control the cardinality of $\mathcal{F}$ using Sauer-Shelah Lemma.

8.25 Apply the VC law of large numbers (Theorem 8.3.15) to the class of indicators of half-spaces.

8.26 (a) Let $X_1, \dots, X_m$ denote the rows of A. Then $\Phi(u) = \operatorname{sign}(Au) = (\operatorname{sign}\langle X_i, u\rangle)_{i=1}^m$.
(b) Use the uniform law of large numbers (Theorem 8.3.15) for the function class $\mathcal{F}$ consisting of the indicators of the wedges $\{\langle x, u\rangle \neq \langle x, v\rangle\}$, where $u, v \in S^{n-1}$. Use the stability property of VC dimension (Proposition 8.3.11) to bound the VC dimension of the wedges in terms of the VC dimension of half-spaces, which was computed in Example 8.3.5.

8.27 (a) Let $X_1, \dots, X_m$ denote the rows of A. Use the uniform law of large numbers (Theorem 8.3.15) for the function class $\mathcal{F}$ consisting of the indicators of the strips $\{x : \ |\langle x, u\rangle| \geq \varepsilon\}$, where $u \in S^{n-1}$. Note that $|\langle X_i, u\rangle| \geq \varepsilon$ implies $\langle X_i, u\rangle^2 \geq \varepsilon$.
(b) Consider a random vector X that with probability δ equals some appropriately chosen multiple of the first basis vector e_1, and with probability $1 - \delta$ takes values in the subspace orthogonal to e_1.

8.28 Pick an arbitrarily large subset $\Omega_0 \subset \Omega$ shattered by $\mathcal{F}$, and let $X \sim \operatorname{Unif}(\Omega_0)$.

8.29 Note that $(f - h)^2$ is obtained from f by flipping the bits $f(x)$ where $h(x) = 1$.

8.31 Show that the random process $X_f = R_n(f) - R(f)$ has subgaussian increments: $\|X_f - X_g\|_{\psi_2} \lesssim \frac{1}{\sqrt{n}}\|f - g\|_\infty$ for all $f, g \in \mathcal{F}$. Use Dudley inequality to deduce that $\mathbb{E}\sup_{f \in \mathcal{F}}|R_n(f) - R(f)| \lesssim 1/\sqrt{n}$ (see the proof of Theorem 8.2.3). Then argue like in the proof of Theorem 8.4.5.

8.32 Take a uniform distribution over a shattered set $\Omega_0 \subset \Omega$ of size $\operatorname{vc}(\mathcal{F})$. The sample only sees the target function T on half of Ω_0, but T could do anything on the rest – so it can't be learn reliably.

8.34 (a) Set $T_0 = \{0\}$. For each $k \in \mathbb{N}$, let T_k contain 0 and the largest $2^{2^k} - 1$ elements of T.

8.35 Write $|X_t - X_{t_0}| \leq |X_t - X_{\pi_\kappa(t)}| + |X_{\pi_\kappa(t)} - X_{t_0}|$ and bound the two terms separately, each with probability $1 - \exp(u^2)$. The first term can be bounded by $C\gamma_2(T, d)$ via chaining $\pi_\kappa(t) \to \pi_{\kappa+1}(t) \to \pi_{\kappa+2}(t) \to \cdots \to t$ (by the choice of κ, the failure probabilities in this finer steps are much smaller than $\exp(-u^2)$). The second term, corresponding to the first big leap in chaining, can be bounded by $Cu\operatorname{diam}(T)$ by taking a union bound over T_κ.

8.36 Following the proof of Theorem 8.2.3, check that the process $Z_f = \frac{1}{\sqrt{n}}\sum_{i=1}^n f(X_i) - \sqrt{n}\,\mathbb{E} f(X)$ has subgaussian increments: $\|Z_f - Z_g\|_{\psi_2} \lesssim d(f, g)$. Then apply the generic chaining bound (Theorem 8.5.2).

8.37 (a) Use Remark 8.5.3 and the majorizing measure theorem to get a bound in terms of the Gaussian width $w(T \cup \{0\})$, then pass to Gaussian complexity using Exercise 7.20.
(b) Use Remark 8.5.4 and Exercise 7.20.

8.38 Use duality (1.6) and Talagrand comparison inequality (Remark 8.5.9).

8.39 (a) Follow the proof of Theorem 7.3.1 more closely instead of applying triangle inequality. Use Sudakov-Fernique inequality (Theorem 7.2.8) instead of Talagrand comparison inequality.
(b) Note that $\mathbb{E}\sup_{x \in T,\, y \in S}\langle Ax, y\rangle \geq \sup_{x \in T}\mathbb{E}\sup_{y \in S}\langle Ax, y\rangle$.

8.40 Use the result of Exercise 8.37(b).

8.41 Use Chevet inequality. Exercise 1.17(a) and Exercise 7.17 should help to compute the radius and Gaussian width.

9.2 Bound the difference between $\mathbb{E}\|Ax\|_2$ and $\sqrt{m}\|x\|_2$ using the concentration of norm (Theorem 3.1.1). Use Lemma 7.5.11(b) to show that this difference can be absorbed by the main error.

9.3 Use the identity $a^2 - b^2 = (a - b)(a + b)$.

9.4 Reduce to the isotropic case: write $B_i = \Sigma^{1/2}A_i$ for some isotropic A_i and apply Theorem 9.1.1 for $\Sigma^{1/2}T$.

9.5 To see why this result generalizes Theorem 9.1.1, express $\|Ax\|_2$ as $\left(\sum_{i=1}^m \langle A_i, x\rangle^2\right)^{1/2}$ where A_i are the rows of the matrix A, and consider the class of linear functions on $\mathbb{R}^n$. To prove the generalization, follow the proof of Theorem 9.1.1.

9.6 To make step 2 in Theorem 9.1.2), use rotation invariance to assume $x = (1, 0, 0, \dots, 0)$ and $y = (\sqrt{1 - \varepsilon^2}, \varepsilon, 0, 0, \dots, 0)$ where $\varepsilon \asymp \|x - y\|_2$. Expand $\|Px\|_2^2 - \|Py\|_2^2$ and reduce the problem to controlling one entry of P, namely P_{12}.

9.7 Use the high-probability version of matrix deviation inequality, given in (9.11).

9.8 If $m \ll n$, the random matrix A in the matrix deviation inequality is an approximate projection: this follows from Section 4.6.

9.9 Use the high-probability version (9.11) of matrix deviation inequality to get a high-probability version of the quadratic deviation (Exercise 9.3); then use it in the proof of Theorem 9.2.2.

9.11 If $\mathcal{X}$ has nonempty interior, dimension reduction with relative error is impossible.

9.13 Use the high-probability version of matrix deviation inequality (see Remark 9.1.4).

9.14 (a) For the upper bound, combine the M^* bound with Exercise 7.17. For the lower bound, try to show a stronger statement – that it holds even if $\dim(E) = 1$. Check that $\mathrm{diam}(B_p^n \cap E) = \|g\|_2 / \|g\|_p$ where $g \sim N(0, I_n)$.

9.16 Consider a random rotation $U \in \mathrm{Unif}(SO(n))$ as in Section 5.2.5, and use a union bound to show that the probability that there exists $x \in \mathcal{X}$ such that $U^{-1}x \in T$ is smaller than 1.

9.17 Modify the M^* bound accordingly.

9.19 Check that $\|A(x - \widehat{x})\|_2 \lesssim \|w\|_2$. Plug this into the matrix deviation inequality (Theorem 9.1.1) for $T - T$, noting that $x, \widehat{x} \in T$.

9.20 Bound both $\|y - Ax\|_2^2 + \lambda\|x\|_T$ and $\|y - A\widehat{x}\|_2^2 + \lambda\|\widehat{x}\|_T$ by $\|w\|_2^2 + \lambda\|x\|_T$. Using triangle inequality, deduce a bound on $\|A(\widehat{x} - x)\|_2$ and show that $\widehat{x} - x \in C\|x\|_T(T - T)$. Now use the matrix deviation inequality (Theorem 9.1.1) as in Exercise 9.19.

9.25 Note that $\|x_{I_1}\|_2 \le 1$. Next, for $i \ge 2$, note that each coordinate of x_{I_i} is smaller in magnitude than the average coordinate of $x_{I_{i-1}}$; conclude that $\|x_{I_i}\|_2 \le (1/\sqrt{s})\|x_{I_{i-1}}\|_1$. Then sum up the bounds.

9.27 To prove a lower bound on $w(S_{n,s})$, construct a large ε-separated subset of $S_{n,s}$ and thus deduce a lower bound on the covering numbers of $S_{n,s}$. Then use Sudakov inequality (Theorem 7.4.1).

9.28 Fix $\rho > 0$ and apply the M^* bound for the truncated cross-polytope $T_\rho := B_1^n \cap \rho B_2^n$. Use Exercise 9.26 to bound the Gaussian width of T_ρ. 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.

9.32 (a) Assume the contrary and follow the proof of Lemma 9.5.2.
(b) Assume the contrary and follow the proof of Theorem 9.5.1.

9.36 Follow Remark 9.1.4.

9.37 Argue as in Section 9.2.4.

9.40 Use the hyperplane separation theorem.

9.42 Let T be the canonical basis $\{e_1, \dots, e_n\}$ in $\mathbb{R}^n$. Express the points as $g_i = Ae_i$, and apply Theorem 9.7.2.

References

[1] E. Abbe. Community detection and stochastic block models: Recent developments. *Journal of Machine Learning Research*, 18(177):1–86, 2018.

[2] E. Abbe, A. S. Bandeira, and G. Hall. Exact recovery in the stochastic block model. *IEEE Transactions on information theory*, 62(1):471–487, 2015.

[3] P. Abdalla and S. Mendelson. Covariance estimation with direction dependence accuracy. *Probability Theory and Related Fields*, 2025.

[4] P. Abdalla and N. Zhivotovskiy. Covariance estimation: Optimal dimension-free guarantees for adversarial corruption and heavy tails. *Journal of the European Mathematical Society*, 2024.

[5] D. Achlioptas and F. McSherry. Fast computation of low-rank matrix approximations. *Journal of the ACM*, 54(2):9–es, 2007.

[6] R. Adamczak. A note on the Hanson-Wright inequality for random vectors with dependencies. *Electronic Communications in Probability*, 20(72):1–13, 2015.

[7] R. Adamczak, R. Latala, A. E. Litvak, A. Pajor, and N. Tomczak-Jaegermann. Chevet type inequality and norms of submatrices. *Studia Mathematica*, 210(1):35–56, 2012.

[8] R. Adamczak, R. Latala, and R. Meller. Hanson–Wright inequality in Banach spaces. *Ann. Inst. Henri Poincaré Probab. Stat.*, 56:2356–2376, 2020.

[9] R. Adamczak, J. Prochno, M. Strzelecka, and M. Strzelecki. Norms of structured random matrices. *Mathematische Annalen*, 388(4):3463–3527, 2024.

[10] K. Adiprasito, I. Bárány, N. Mustafa, and T. Terpai. Theorems of Carathéodory, Helly, and Tverberg without dimension. *Discrete & Computational Geometry*, 64(2):233–258, 2020.

[11] R. Adler and E. Taylor. *Random fields and geometry.* Springer, 2009.

[12] R. Ahlswede and A. Winter. Strong converse for identification via quantum channels. *IEEE Transactions on Information Theory*, 48(3):569–579, 2002.

[13] A. Ai, A. Lapanowski, Y. Plan, and R. Vershynin. One-bit compressed sensing with non-gaussian measurements. *Linear Algebra and its Applications*, 441:222–239, 2014.

[14] F. Albiac and N. J. Kalton. *Topics in Banach space theory.* Springer, 2006.

[15] S. Alesker. A remark on the Szarek-Talagrand theorem. *Combinatorics, Probability and Computing*, 6(2):139–144, 1997.

[16] N. Alon and A. Naor. Approximating the cut-norm via Grothendieck's inequality. In *Proceedings of the thirty-sixth annual ACM symposium on theory of computing*, pages 72–80, 2004.

[17] N. Alon and J. H. Spencer. *The probabilistic method.* John Wiley & Sons, 2015.

[18] D. Amelunxen, M. Lotz, M. B. McCoy, and J. A. Tropp. Living on the edge: Phase transitions in convex programs with random data. *Information and Inference: A Journal of the IMA*, 3(3):224–294, 2014.

[19] A. Anandkumar, R. Ge, D. Hsu, S. Kakade, and M. Telgarsky. Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, 15(1):2773–2832, 2014.

[20] G. W. Anderson, A. Guionnet, and O. Zeitouni. *An introduction to random matrices.* Cambridge University Press, 2010.

[21] S. Artstein-Avidan, A. Giannopoulos, and V. D. Milman. *Asymptotic geometric analysis, Part II.* American Mathematical Society, 2021.

[22] D. Bakry and M. Ledoux. Lévy–Gromov's isoperimetric inequality for an infinite dimensional diffusion generator. *Inventiones Mathematicae*, 123(2):259–281, 1996.

[23] K. Ball. An elementary introduction to modern convex geometry. In *Flavors of Geometry*, volume 31 of *MSRI Publications*, pages 1–58. Cambridge University Press, 1997.

[24] A. Bandeira. Ten lectures and forty-two open problems in the mathematics of data science. *Lecture notes*, 2015.

[25] A. Bandeira, M. Boedihardjo, and R. van Handel. Matrix concentration inequalities and free probability. *Inventiones Mathematicae*, 234(1):419–487, 2023.

[26] A. Bandeira, G. Cipolloni, D. Schröder, and R. van Handel. Matrix concentration inequalities and free probability II. two-sided bounds and applications. *Preprint arXiv:2406.11453*, 2024.

[27] A. Bandeira and R. van Handel. Sharp nonasymptotic bounds on the norm of random matrices with independent entries. *The Annals of Probability*, 44(4):2479–2506, 2016.

[28] R. Baraniuk, S. Foucart, D. Needell, Y. Plan, and M. Wootters. One-bit compressive sensing of dictionary-sparse signals. *Information and Inference: A Journal of the IMA*, 7(1):83–104, 2018.

[29] I. Bárány. Sylvester's question: The probability that n points are in convex position. *The Annals of Probability*, 27(4):2020–2034, 1999.

[30] F. Barthe and B. Maurey. Some remarks on isoperimetry of Gaussian type. *Annales de l'Institut Henri Poincare (B) – Probability and Statistics*, 36:419–434, 2000.

[31] F. Barthe and E. Milman. Transference principles for log-Sobolev and spectral-gap with applications to conservative spin systems. *Communications in Mathematical Physics*, 323(2):575–625, 2013.

[32] P. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

[33] M. Belkin and K. Sinha. Polynomial learning of distribution families. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 103–112. IEEE, 2010.

[34] P. Bellec. Concentration of quadratic forms under a Bernstein moment assumption. *Preprint arXiv:1901.08736*, 2019.

[35] G. Bennett. Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association*, 57(297):33–45, 1962.

[36] G. Bennett. Upper bounds on the moments and probability inequalities for the sum of independent, bounded random variables. *Biometrika*, 52(3/4):559–569, 1965.

[37] G. Bennett, V. Goodman, and C. Newman. Norms of random matrices. *Pacific Journal of Mathematics*, 59(2):359–365, 1975.

[38] S. N. Bernstein. On a modification of Chebyshev's inequality and of the error formula of Laplace. *Ann. Sci. Inst. Sav. Ukraine, Sect. Math.*, 1(5), 1924. Reprinted in: Math. Sbornik, 1933, 34(1): 1-17.

[39] S. N. Bernstein. *Theory of Probability.* Gosizdat, 1927.

[40] S. N. Bernstein. On certain modifications of Chebyshev's inequality. *Doklady Akademii Nauk SSSR*, 17(6):275–277, 1937.

[41] R. Bhatia. *Matrix analysis.* Springer, 2013.

[42] P. Billingsley. *Probability and Measure.* Wiley, 2012.

[43] A. Blum, S. Har-Peled, and B. Raichel. Sparse approximation via generating point sets. *ACM Transactions on Algorithms*, 15(3):1–16, 2019.

[44] C. R. Blyth and P. K. Pathak. A note on easy proofs of Stirling's theorem. *The American Mathematical Monthly*, 93(5):376–379, 1986.

[45] S. G. Bobkov. An isoperimetric inequality on the discrete cube, and an elementary proof of the isoperimetric inequality in Gauss space. *The Annals of Probability*, 25(1):206–214, 1997.

[46] B. Bollobás. *Combinatorics: set systems, hypergraphs, families of vectors, and combinatorial probability.* Cambridge University Press, 1986.

[47] B. Bollobás. *Random Graphs.* Cambridge University Press, 2001.

[48] C. Bordenave, M. Lelarge, and L. Massoulié. Non-backtracking spectrum of random graphs: community detection and non-regular Ramanujan graphs. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, pages 1347–1357. IEEE, 2015.

[49] E. Borel. *Introduction geometrique a quelques theories physiques.* Gauthier-Villars, 1914.

[50] C. Borell. The Brunn-Minkowski inequality in gauss space. *Inventiones mathematicae*, 30(2):207–216, 1975.

[51] J. M. Borwein and A. S. Lewis. *Convex Analysis and Nonlinear Optimization: Theory and Examples.* Springer, 2005.

[52] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence.* Oxford University Press, 2013.

[53] P. Boufounos and R. G. Baraniuk. 1-bit compressive sensing. In *2008 42nd Annual Conference on Information Sciences and Systems*, pages 16–21. IEEE, 2008.

[54] J. Bourgain, S. Dirksen, and J. Nelson. Toward a unified theory of sparse dimensionality reduction in euclidean space. In *Proceedings of the forty-seventh annual ACM symposium on Theory of Computing*, pages 499–508, 2015.

[55] J. Bourgain and L. Tzafriri. Invertibility of 'large' submatrices with applications to the geometry of Banach spaces and harmonic analysis. *Israel Journal of Mathematics*, 57:137–224, 1987.

[56] O. Bousquet, S. Boucheron, and G. Lugosi. Introduction to statistical learning theory. In *Summer School on Machine Learning*, pages 169–207. Springer, 2003.

[57] C. Boutsidis, P. Drineas, and M. Magdon-Ismail. Near-optimal column-based matrix reconstruction. *SIAM Journal on Computing*, 43(2):687–717, 2014.

[58] S. Boyd and L. Vandenberghe. *Convex Optimization.* Cambridge University Press, 2004.

[59] T. Brailovskaya and R. van Handel. Universality and sharp matrix concentration inequalities. *Geometric and Functional Analysis*, 34(6):1734–1838, 2024.

[60] M. Braverman, K. Makarychev, Y. Makarychev, and A. Naor. The Grothendieck constant is strictly smaller than Krivine's bound. In *Forum of Mathematics, Pi*, volume 1, page e4, 2013.

[61] S. Brazitikos, A. Giannopoulos, P. Valettas, and B.-H. Vritsiou. *Geometry of isotropic convex bodies.* American Mathematical Society, 2014.

[62] A. Brieden, P. Gritzmann, R. Kannan, V. Klee, L. Lovász, and M. Simonovits. Deterministic and randomized polynomial-time approximation of radii. *Mathematika*, 48(1-2):63–105, 2001.

[63] S. Brooks, A. Gelman, G. Jones, and X. L. Meng. *Handbook of Markov Chain Monte Carlo.* CRC Press, 2011.

[64] S. Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3-4):231–357, 2015.

[65] A. Buchholz. Operator Khintchine inequality in non-commutative probability. *Mathematische Annalen*, 319(1):1–16, 2001.

[66] A. Buchholz. Optimal constants in Khintchine type inequalities for fermions, Rademachers and q-Gaussian operators. *Bulletin of the Polish Academy of Sciences. Mathematics*, 53(3):315–321, 2005.

[67] P. Bühlmann and S. van De Geer. *Statistics for high-dimensional data: methods, theory and applications.* Springer, 2011.

[68] V. V. Buldygin and Yu. V. Kozachenko. Sub-Gaussian random variables. *Ukrainian Mathematical Journal*, 32:483–489, 1980.

[69] T. Cai, Z. Ren, and H. Zhou. Estimating structured high-dimensional covariance and precision matrices: Optimal rates and adaptive estimation. *Electronic Journal of Statistics*, 10(1):1–59, 2016.

[70] E. J. Candès. The restricted isometry property and its implications for compressed sensing. *Comptes rendus. Mathematique*, 346(9-10):589–592, 2008.

[71] E. J. Candès and Y. Plan. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Transactions on Information Theory*, 57(4):2342–2359, 2011.

[72] E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Communications of the ACM*, 55(6):111–119, 2012.

[73] E. J. Candès and T. Tao. Decoding by linear programming. *IEEE transactions on information theory*, 51(12):4203–4215, 2005.

[74] E. J. Candès and T. Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE transactions on information theory*, 56(5):2053–2080, 2010.

[75] F. P. Cantelli. Sulla determinazione empirica delle leggi di probabilita. *Giornale dell'Istituto Italiano degli Attuari*, 4(421-424), 1933.

[76] B. Carl. Inequalities of Bernstein-Jackson-type and the degree of compactness of operators in Banach spaces. *Annales de l'institut Fourier*, 35(3):79–118, 1985.

[77] B. Carl and A. Pajor. Gelfand numbers of operators with values in a Hilbert space. *Inventiones Mathematicae*, 94(3):479–504, 1988.

[78] P. G. Casazza, G. Kutyniok, and F. Philipp. Introduction to finite frame theory. *Finite frames: theory and applications*, pages 1–53, 2013.

[79] Djalil Chafaï, Olivier Guédon, Guillaume Lecué, and Alain Pajor. *Interactions Between Compressed Sensing Random Matrices and High Dimensional Geometry*, volume 37 of *Panoramas et Synthèses*. Société Mathématique de France, 2012.

[80] V. Chandrasekaran, B. Recht, P. Parrilo, and A. S. Willsky. The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, 12(6):805–849, 2012.

[81] J. Chen and M. Yuan. One-bit phase retrieval: Optimal rates and efficient algorithms. *Preprint arXiv:2405.04733*, 2024.

[82] R. Y. Chen, A. Gittens, and J. A. Tropp. The masked sample covariance estimator: an analysis using matrix concentration inequalities. *Information and Inference: A Journal of the IMA*, 1(1):2–20, 2012.

[83] H. Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Annals of Mathematical Statistics*, 23(4):493–507, 1952.

[84] S. Chevet. Séries de variables aléatoires gaussiennes à valeurs dans $E\hat{\otimes}_\varepsilon F$. Application aux produits d'espaces de Wiener abstraits. *Séminaire Maurey-Schwartz*, pages 1–15, 1978.

[85] S. Chewi, J. Niles-Weed, and P. Rigollet. Statistical optimal transport, 2024. Lecture Notes for École d'Été de Probabilités de Saint-Flour XLIX.

[86] P. Chin, A. Rao, and V. Vu. Stochastic block model and community detection in sparse graphs: A spectral algorithm with optimal rate of recovery. In *Conference on Learning Theory*, pages 391–423. PMLR, 2015.

[87] B. S. Cirel'son, L. A. Ibragimov, and V. N. Sudakov. Norm of Gaussian sample function. In *Proceedings of the 3rd Japan-USSR Symposium on Probability Theory. Lecture Notes in Math*, volume 550, pages 20–41, 1976.

[88] A. Cohen, W. Dahmen, and R. DeVore. Compressed sensing and best k-term approximation. *Journal of the American Mathematical Society*, 22(1):211–231, 2009.

[89] Z. Cvetkovic, I. Daubechies, and B. F. Logan. Single-bit oversampled A/D conversion with exponential accuracy in the bit rate. *IEEE Transactions on Information Theory*, 53(11):3979–3989, 2007.

[90] N. Dafnis, A. Giannopoulos, and A. Tsolomitis. Asymptotic shape of a random polytope in a convex body. *Journal of Functional Analysis*, 257(9):2820–2839, 2009.

[91] M. A. Davenport, M. F. Duarte, Y. C. Eldar, and G. Kutyniok. Introduction to compressed sensing. In *Compressed Sensing: Theory and Applications*, pages 1–68. Cambridge University Press, 2012.

[92] M. A. Davenport, Y. Plan, E. Van Den Berg, and M. Wootters. 1-bit matrix completion. *Information and Inference: A Journal of the IMA*, 3(3):189–223, 2014.

[93] M. A. Davenport and J. Romberg. An overview of low-rank matrix recovery from incomplete observations. *IEEE Journal of Selected Topics in Signal Processing*, 10(4):608–622, 2016.

[94] K. Davidson and S. J. Szarek. Local operator theory, random matrices and Banach spaces. In *Handbook of the geometry of Banach spaces*, volume 1, pages 317–366. Elsevier, 2001.

[95] C. Davis and W. M. Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

[96] V. de la Peña and W. Giné. *Decoupling: From Dependence to Independence.* Springer, 1999.

[97] V. de la Pena and S. J. Montgomery-Smith. Decoupling inequalities for the tail probabilities of multivariate U-statistics. *The Annals of Probability*, pages 806–816, 1995.

[98] A. Deshpande and S. Vempala. Adaptive sampling and fast low-rank matrix approximation. In *International Workshop on Approximation Algorithms for Combinatorial Optimization*, pages 292–303. Springer, 2006.

[99] L. Devroye, M. Lerasle, G. Lugosi, and R. I. Oliveira. Sub-gaussian mean estimators. *The Annals of Statistics*, 44(6):2695–2725, 2016.

[100] S. Dhara, D. Mukherjee, and K. Ramanan. On r-to-p norms of random matrices with nonnegative entries: Asymptotic normality and ℓ^∞-bounds for the maximizer. *The Annals of Applied Probability*, 34(6):5076–5115, 2024.

[101] P. Diaconis and D. Freedman. An elementary proof of Stirling's formula. *The American Mathematical Monthly*, 93(2):123–125, 1986.

[102] P. Diaconis and D. Freedman. A dozen de Finetti-style results in search of a theory. *Annales de l'I. H. P., section B*, 23(S2):397–423, 1987.

[103] S. Dirksen. Tail bounds via generic chaining. *Electronic Journal of Probability*, 20:1–29, 2015.

[104] S. Dirksen, J. Maly, and H. Rauhut. Covariance estimation under one-bit quantization. *The Annals of Statistics*, 50(6):3538–3562, 2022.

[105] S. Dirksen and S. Mendelson. Non-Gaussian hyperplane tessellations and robust one-bit compressed sensing. *Journal of the European Mathematical Society*, 23(9):2913–2947, 2021.

[106] S. Dirksen, S. Mendelson, and A. Stollenwerk. Sharp estimates on random hyperplane tessellations. *SIAM Journal on Mathematics of Data Science*, 4(4):1396–1419, 2022.

[107] S. Dirksen, S. Mendelson, and A. Stollenwerk. Fast metric embedding into the hamming cube. *SIAM Journal on Computing*, 53(2):315–345, 2024.

[108] D. Donoho, M. Gavish, and A. Montanari. The phase transition of matrix recovery from Gaussian measurements matches the minimax MSE of matrix denoising. *Proceedings of the National Academy of Sciences*, 110(21):8405–8410, 2013.

[109] D. Donoho, A. Javanmard, and A. Montanari. Information-theoretically optimal compressed sensing via spatial coupling and approximate message passing. *IEEE Transactions on Information Theory*, 59(11):7434–7464, 2013.

[110] D. Donoho, I. Johnstone, and A. Montanari. Accurate prediction of phase transitions in compressed sensing via a connection to minimax denoising. *IEEE transactions on information theory*, 59(6):3396–3433, 2013.

[111] D. Donoho, A. Maleki, and A. Montanari. The noise-sensitivity phase transition in compressed sensing. *IEEE Transactions on Information Theory*, 57(10):6920–6941, 2011.

[112] D. Donoho and J. Tanner. Counting faces of randomly projected polytopes when the projection radically lowers dimension. *Journal of the American Mathematical Society*, 22(1):1–53, 2009.

[113] P. Drineas and R. Kannan. Pass efficient algorithm for approximating large matrices. In *Proceedings of the 14th ACM–SIAM Symposium on Discrete Algorithms (SODA)*, pages 223–232, 2003.

[114] R. M. Dudley. Central limit theorems for empirical measures. *The Annals of Probability*, pages 899–929, 1978.

[115] R. M. Dudley. *Uniform central limit theorems*, volume 142. Cambridge University Press, 2014.

[116] R. Durrett. *Probability: Theory and Examples*, volume 49. Cambridge University Press, 2019.

[117] A. Dvoretzky. A theorem on convex bodies and applications to Banach spaces. *Proceedings of the National Academy of Sciences*, 45(2):223–226, 1959.

[118] A. Dvoretzky. Some results on convex bodies and Banach spaces. *Matematika*, 8(1):73–102, 1964.

[119] D. Eisenstat and D. Angluin. The VC dimension of k-fold union. *Information Processing Letters*, 101(5):181–184, 2007.

[120] J. Eldridge, M. Belkin, and Y. Wang. Unperturbed: spectral analysis beyond Davis-Kahan. In *Proceedings of the Algorithmic Learning Theory*, volume 83 of *Proceedings of Machine Learning Research*, pages 321–358. PMLR, 2018.

[121] P. Erdős and A. Rényi. On the strength of connectedness of a random graph. *Acta Mathematica Hungarica*, 12(1):261–267, 1961.

[122] Paul Erdös. On a lemma of littlewood and offord. *Bull. Amer. Math. Soc.*, 51:898–902, 1945.

[123] W. Feller. *An introduction to probability theory and its applications. Vol. I.* John Wiley & Sons, Inc., New York-London-Sydney, third edition, 1968.

[124] X. M. Fernique. Regularité des trajectoires des fonctions aléatoires gaussiennes. In *École d'Été de Probabilités de Saint-Flour. IV – 1974*, volume 480 of *Lecture Notes in Mathematics*, pages 1–96. Springer, Berlin, Heidelberg, 1975.

[125] G. B. Folland. *A course in Abstract Harmonic Analysis.* CRC press, 2016.

[126] S. Fortunato and D. Hric. Community detection in networks: A user guide. *Physics Reports*, 659:1–44, 2016.

[127] S. Foucart and H. Rauhut. *A mathematical introduction to compressive sensing.* Springer, 2013.

[128] P. Frankl. On the trace of finite sets. *Journal of Combinatorial Theory, Series A*, 34(1):41–45, 1983.

[129] M. P. Friedlander, H. Jeong, Y. Plan, and Ö. Yılmaz. NBIHT: An efficient algorithm for 1-bit compressed sensing with optimal error decay rate. *IEEE Transactions on Information Theory*, 68(2):1157–1177, 2021.

[130] A. Frieze and R. Kannan. Quick approximation to matrices and applications. *Combinatorica*, 19(2):175–220, 1999.

[131] A. Frieze and M. Karonski. *Introduction to Random Graphs.* Cambridge University Press, 2016.

[132] A. Yu. Garnaev and E. D. Gluskin. On diameters of the Euclidean sphere. *Dokl. Akad. Nauk SSSR*, 277(5):200–204, 1984.

[133] A. Gasull and F. Utzet. Approximating mills ratio. *Journal of Mathematical Analysis and Applications*, 420(2):1832–1853, 2014.

[134] A. A. Giannopoulos and V. D. Milman. Euclidean structure in finite dimensional normed spaces. In *Handbook of the geometry of Banach spaces*, volume 1, pages 707–779. Elsevier, 2001.

[135] E. Giné, F. Götze, and D. M. Mason. When is the Student t-statistic asymptotically standard normal? *The Annals of Probability*, 25(3):1514–1531, 1997.

[136] E. Giné and R. Nickl. *Mathematical Foundations of Infinite-dimensional Statistical Models.* Cambridge University Press, 2021.

[137] C. Giraud. *Introduction to High-Dimensional Statistics.* Chapman and Hall/CRC, 2021.

[138] V. Glivenko. Sulla determinazione empirica delle leggi di probabilita. *Giornale dell'Istituto Italiano degli Attuari*, 4:92–99, 1933.

[139] M. X. Goemans and D. P. Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of the ACM*, 42(6):1115–1145, 1995.

[140] Y. Gordon. Some inequalities for Gaussian processes and applications. *Israel Journal of Mathematics*, 50:265–289, 1985.

[141] Y. Gordon. Elliptically contoured distributions. *Probability Theory and Related Fields*, 76(4):429–438, 1987.

[142] Y. Gordon. Gaussian processes and almost spherical sections of convex bodies. *The Annals of Probability*, pages 180–188, 1988.

[143] Y. Gordon. On Milman's inequality and random subspaces which escape through a mesh in $\mathbb{R}^n$. In *Geometric Aspects of Functional Analysis: Israel Seminar (GAFA) 1986–87*, pages 84–106. Springer, 1988.

[144] Y. Gordon. Majorization of Gaussian processes and geometric applications. *Probability Theory and Related Fields*, 91(2):251–267, 1992.

[145] F. Götze, H. Sambale, and A. Sinulis. Concentration inequalities for polynomials in α-sub-exponential random variables. *Electronic Journal of Probability*, 26(95):1–22, 2021.

[146] N. Goyal, S. Vempala, and Y. Xiao. Fourier PCA and robust tensor decomposition. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 584–593, 2014.

[147] M. Gromov. Paul Levy's isoperimetric inequality. IHES preprint, 1980.

[148] D. Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Transactions on Information Theory*, 57(3):1548–1566, 2011.

[149] A. Grothendieck. Résumé de la théorie métrique des produits tensoriels topologiques. *Boletim da Sociedade Matemática de São Paulo*, 8(4):1–79, 1956.

[150] O. Guédon. Concentration phenomena in high dimensional geometry. In *ESAIM: Proceedings*, volume 44, pages 47–60. EDP Sciences, 2014.

[151] O. Guédon and R. Vershynin. Community detection in sparse networks via grothendieck's inequality. *Probability Theory and Related Fields*, 165(3):1025–1049, 2016.

[152] U. Haagerup. The best constants in the Khintchine inequality. *Studia Mathematica*, 70(3):231–283, 1981.

[153] B. Hajek, Y. Wu, and J. Xu. Achieving exact cluster recovery threshold via semidefinite programming. *IEEE Transactions on Information Theory*, 62(5):2788–2797, 2016.

[154] D. L. Hanson and F. T. Wright. A bound on tail probabilities for quadratic forms in independent random variables. *The Annals of Mathematical Statistics*, 42(3):1079–1083, 1971.

[155] L. H. Harper. Optimal numberings and isoperimetric problems on graphs. *Journal of Combinatorial Theory*, 1(3):385–393, 1966.

[156] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer, 2nd edition, 2017.

[157] T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations.* Chapman and Hall/CRC, 2015.

[158] D. Haussler and P. M. Long. A generalization of Sauer's lemma. *Journal of Combinatorial Theory, Series A*, 71(2):219–240, 1995.

[159] Y. He, K. Wang, and Y. Zhu. Sparse Hanson-Wright inequalities with applications. *Preprint arXiv:2410.15652*, 2024.

[160] P. Hitczenko and S. Kwapien. On the rademacher series. In *Probability in Banach Spaces, 9*, volume 35 of *Progress in Probability*, pages 31–36. Springer, 1994.

[161] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

[162] T. Hofmann, B. Schölkopf, and A. J. Smola. Kernel methods in machine learning. *The Annals of Statistics*, 36(3):1171–1220, 2008.

[163] P. W. Holland, K. B. Laskey, and S. Leinhardt. Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137, 1983.

[164] S. Hoory, N. Linial, and Avi Wigderson. Expander graphs and their applications. *Bulletin of the American Mathematical Society*, 43(4):439–561, 2006.

[165] D. Hsu and S. M. Kakade. Learning mixtures of spherical gaussians: moment methods and spectral decompositions. In *Proceedings of the 4th conference on Innovations in Theoretical Computer Science*, pages 11–20, 2013.

[166] D. Hsu, S. M. Kakade, and T. Zhang. A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*, 17(52):1–6, 2012.

[167] H. Huang and K. Tikhomirov. On dimension-dependent concentration for convex Lipschitz functions in product spaces. *Electronic Journal of Probability*, 28:1–23, 2023.

[168] J. Huang, Y. Jiao, X. Lu, and L. Zhu. Robust decoding from 1-bit compressive sampling with ordinary and regularized least squares. *SIAM Journal on Scientific Computing*, 40(4):A2062–A2086, 2018.

[169] F. W. Huffer. Slepian's inequality via the central limit theorem. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique*, pages 367–370, 1986.

[170] D. Hug. Random polytopes. In *Stochastic Geometry, Spatial Statistics and Random Fields: Asymptotic Methods*, pages 205–238. Springer, 2012.

[171] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning.* Springer, 2013.

[172] G. J. O. Jameson. A simple proof of Stirling's formula for the gamma function. *The Mathematical Gazette*, 99(544):68–74, 2015.

[173] S. Janson. Graphons, cut norm and distance, couplings and rearrangements. *New York Journal of Mathematics Monographs*, 4, 2013.

[174] S. Janson, T. Luczak, and A. Rucinski. *Random Graphs.* John Wiley & Sons, 2011.

[175] A. Javanmard, A. Montanari, and F. Ricci-Tersenghi. Phase transitions in semidefinite relaxations. *Proceedings of the National Academy of Sciences*, 113(16):E2218–E2223, 2016.

[176] H. Jeong, X. Li, Y. Plan, and O. Yilmaz. Sub-gaussian matrices on sets: optimal tail dependence and applications. *Communications on Pure and Applied Mathematics*, 75(8):1713–1754, 2022.

[177] M. Jerrum, L. Valiant, and V. Vazirani. Random generation of combinatorial structures from a uniform distribution. *Theoretical Computer Science*, 43:186–188, 1986.

[178] W. B. Johnson, J. Lindenstrauss, and G. Schechtman. Extensions of Lipschitz maps into Banach spaces. *Israel Journal of Mathematics*, 54(2):129–138, 1986.

[179] J.-P. Kahane. Propriétés locales des fonctions à séries de Fourier aléatoires. *Studia Mathematica*, 19(1):1–25, 1960.

[180] J.-P. Kahane. Une inégalité du type de Slepian et Gordon sur les processus gaussiens. *Israel Journal of Mathematics*, 55:109–110, 1986.

[181] A. T. Kalai, A. Moitra, and G. Valiant. Disentangling gaussians. *Communications of the ACM*, 55(2):113–120, 2012.

[182] R. H. Keshavan, A. Montanari, and S. Oh. Matrix completion from a few entries. *IEEE Transactions on Information Theory*, 56(6):2980–2998, 2010.

[183] A. Khintchine. Über dyadische brüche. *Mathematische Zeitschrift*, 18(1):109–116, 1923.

[184] S. Khot, G. Kindler, E. Mossel, and R. O'Donnell. Optimal inapproximability results for MAX-CUT and other 2-variable CSPs? *SIAM Journal on Computing*, 37(1):319–357, 2007.

[185] S. Khot and A. Naor. Grothendieck-type inequalities in combinatorial optimization. *Communications on Pure and Applied Mathematics*, 65:992–1035, 2012.

[186] B. Klartag. A central limit theorem for convex sets. *Inventiones Mathematicae*, 168(1):91–131, 2007.

[187] B. Klartag and S. Mendelson. Empirical processes and random projections. *Journal of Functional Analysis*, 225(1):229–245, 2005.

[188] Y. Klochkov and N. Zhivotovskiy. Uniform Hanson-Wright type concentration inequalities for unbounded entries via the entropy method. *Electronic Journal of Probability*, 25, 2020. Paper No. 22.

[189] K. Knudson, R. Saab, and R. Ward. One-bit compressive sensing with norm estimation. *IEEE Transactions on Information Theory*, 62(5):2748–2758, 2016.

[190] V. Koltchinskii and K. Lounici. Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli*, pages 110–133, 2017.

[191] V. Koltchinskii and S. Mendelson. Bounding the smallest singular value of a random matrix without concentration. *International Mathematics Research Notices*, 2015(23):12991–13008, 2015.

[192] H. König. On the best constants in the Khintchine inequality for Steinhaus variables. *Israel Journal of Mathematics*, 203:23–57, 2014.

[193] J. Kovačević and A. Chebira. An introduction to frames. *Foundations and Trends in Signal Processing*, 2(1):1–94, 2008.

[194] M. G. Krein, M. A. Krasnoselskii, and D. P. Milman. On the defect numbers of linear operators in a Banach space and on some geometric questions. *Sbornik Trudov Instisuta Matematiki Akademii Nauk USSR*, 11:97–112, 1948.

[195] J.-P. Krivine. Constantes de Grothendieck et fonctions de type positif sur les spheres. *Séminaire Maurey-Schwartz*, pages 1–17, 1978.

[196] S. Kulkarni and G. Harman. *An elementary introduction to statistical learning theory.* John Wiley & Sons, 2011.

[197] K. G. Larsen and J. Nelson. Optimality of the Johnson-Lindenstrauss lemma. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 633–638. IEEE, 2017.

[198] J. N. Laska, Z. Wen, W. Yin, and R. G. Baraniuk. Trust, but verify: Fast and accurate signal recovery from 1-bit compressive measurements. *IEEE Transactions on Signal Processing*, 59(11):5289–5301, 2011.

[199] L. Jacquesand J. N. Laska, P. Boufounos, and R. G. Baraniuk. Robust 1-bit compressive sensing via binary stable embeddings of sparse vectors. *IEEE transactions on Information Theory*, 59(4):2082–2102, 2013.

[200] R. Latala and M. Strzelecka. Chevet-type inequalities for subexponential Weibull variables and estimates for norms of random matrices. *Electronic Journal of Probability*, 29:1–19, 2024.

[201] R. Latala and M. Strzelecka. Operator $\ell^p \to \ell^q$ norms of Gaussian matrices. *arXiv preprint arXiv:2502.02186*, 2025.

[202] R. Latala and M. Strzelecka. Operator $\ell^p \to \ell^q$ norms of random matrices with iid entries. *Journal of Functional Analysis*, 288(3):110720, 2025.

[203] R. Latala, R. van Handel, and P. Youssef. The dimension-free structure of nonhomogeneous random matrices. *Inventiones Mathematicae*, 214:1031–1080, 2018.

[204] Rafal Latala. On the spectral norm of Rademacher matrices. *arXiv preprint arXiv:2405.13656*, 2024.

[205] M. Laurent and F. Vallentin. Semidefinite optimization. Lecture notes, April 2016.

[206] C. Le, E. Levina, and R. Vershynin. Concentration and regularization of random graphs. *Random Structures & Algorithms*, 51(3):538–561, 2017.

[207] G. Lecué and S. Mendelson. Regularization and the small-ball method II: complexity dependent error rates. *Journal of Machine Learning Research*, 18(146):1–48, 2017.

[208] G. Lecué and S. Mendelson. Regularization and the small-ball method I: sparse recovery. *Annals of Statistics*, 46(2):611–641, 2018.

[209] M. Ledoux. *The concentration of measure phenomenon.* American Mathematical Society, 2001.

[210] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: isoperimetry and processes.* Springer, 2013.

[211] E. Levina and R. Vershynin. Partial estimation of covariance matrices. *Probability Theory and Related Fields*, 153(3):405–419, 2012.

[212] S. Li and T. Schramm. Spectral clustering in the Gaussian mixture block model. *arXiv preprint arXiv:2305.00979*, 2023.

[213] C. Liaw, A. Mehrabian, Y. Plan, and R. Vershynin. A simple tool for bounding the deviation of random matrices on geometric sets. In *Geometric Aspects of Functional Analysis: Israel Seminar (GAFA) 2014–2016*, pages 277–299, 2017.

[214] E. Liberty. Simple and deterministic matrix sketching. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 581–588, 2013.

[215] J. Lindenstrauss and A. Pelczynski. Absolutely summing operators in l^p-spaces and their applications. *Studia Mathematica*, 29(3):275–326, 1968.

[216] J. E. Littlewood. On bounded bilinear forms in an infinite number of variables. *The Quarterly Journal of Mathematics*, (1):164–174, 1930.

[217] J. E. Littlewood and A. C. Offord. On the number of real roots of a random algebraic equation. III. *Rec. Math. [Mat. Sbornik] N.S.*, 12/54:277–286, 1943.

[218] K. Löwner. Über monotone matrixfunktionen. *Mathematische Zeitschrift*, 38(1):177–216, 1934.

[219] G. Lugosi and S. Mendelson. Mean estimation and regression under heavy-tailed distributions–a survey. *Foundations of Computational Mathematics*, 19(5):1145–1190, 2019.

[220] F. Lust-Piquard. Inégalites de Khintchine dans c_p $(1 < p < \infty)$. *Comptes Rendus de l'Académie des Sciences. Série I*, 01 1986.

[221] F. Lust-Piquard and G. Pisier. Non commutative Khintchine and Paley inequalities. *Arkiv för matematik*, 29:241–260, 1991.

[222] Yu. Makovoz. A simple proof of an inequality in the theory of n-widths. In *Proceedings of the Conference on Constructive Theory of Functions (Varna, 1987)*, pages 305–308, Sofia, 1988. Bulgarian Academy of Sciences.

[223] P.-G. Martinsson and J. A. Tropp. Randomized numerical linear algebra: Foundations and algorithms. *Acta Numerica*, 29:403–572, 2020.

[224] J. Matousek. *Geometric discrepancy: An illustrated guide*, volume 18. Springer, 2009.

[225] J. Matousek. *Lectures on discrete geometry*, volume 212. Springer Science, 2013.

[226] N. Matsumoto and A. Mazumdar. Binary iterative hard thresholding converges with optimal number of measurements for 1-bit compressed sensing. *Journal of the ACM*, 71(5):1–64, 2024.

[227] B. Maurey. Construction de suites symétriques. *Comptes Rendus de l'Académie des Sciences, Série A–B*, 288(14):A679–A681, 1979.

[228] M. B. McCoy and J. A. Tropp. From Steiner formulas for cones to concentration of intrinsic volumes. *Discrete & Computational Geometry*, 51:926–963, 2014.

[229] F. McSherry. Spectral partitioning of random graphs. In *Proceedings 42nd IEEE Symposium on Foundations of Computer Science*, pages 529–537. IEEE, 2001.

[230] E. Meckes. Projections of probability distributions: A measure-theoretic Dvoretzky theorem. In *Geometric Aspects of Functional Analysis: Israel Seminar (GAFA) 2006–2010*, pages 317–326. Springer, 2012.

[231] M. L. Mehta. *Random matrices.* Elsevier, 2004.

[232] S. Mendelson. A few notes on statistical learning theory. In *Advanced Lectures on Machine Learning*, volume 2600 of *Lecture Notes in Computer Science*, pages 1–40. Springer, 2003.

[233] S. Mendelson. A remark on the diameter of random sections of convex bodies. In *Geometric Aspects of Functional Analysis: Israel Seminar (GAFA) 2011-2013*, pages 395–404. Springer, 2014.

[234] S. Mendelson. Learning without concentration. *Journal of the ACM*, 62(3):1–25, 2015.

[235] S. Mendelson. Extending the scope of the small-ball method. *Studia Mathematica*, 256(2):147–167, 2021.

[236] S. Mendelson, A. Pajor, and N. Tomczak-Jaegermann. Reconstruction and subgaussian operators in asymptotic geometric analysis. *Geometric and Functional Analysis*, 17(4):1248–1282, 2007.

[237] S. Mendelson and R. Vershynin. Entropy and the combinatorial dimension. *Inventiones Mathematicae*, 152(1):37–55, 2003.

[238] S. Mendelson and N. Zhivotovskiy. Robust covariance estimation under $L_4 - L_2$ norm equivalence. *The Annals of Statistics*, 48(3):1648–1664, 2020.

[239] C. D. Meyer. *Matrix analysis and applied linear algebra.* SIAM, 2023.

[240] F. Mezzadri. How to generate random matrices from the classical compact groups. *Notices of the American Mathematical Society*, 54(5):592–604, 2007.

[241] V. Milman. Surprising geometric phenomena in high-dimensional convexity theory. In *European Congress of Mathematics: Budapest, July 22–26, 1996 Volume II*, pages 73–91. Springer, 1998.

[242] V. D. Milman. A new proof of A. Dvoretzky's theorem on cross-sections of convex bodies. *Functional Analysis and Its Applications*, 5(4):288–295, 1971.

[243] V. D. Milman. Geometrical inequalities and mixed volumes in the local theory of Banach spaces. *Astérisque*, 131:373–400, 1985.

[244] V. D. Milman. Random subspaces of proportional dimension of finite dimensional normed spaces: approach through the isoperimetric inequality. In *Banach Spaces: Proceedings of the Missouri Conference*, pages 106–115. Springer, 1985.

[245] V. D. Milman. A note on a low M^*-estimate. In *Geometry of Banach Spaces (Proc. Conf., Strobl, 1989)*, volume 158 of *Lecture Notes in Mathematics*, pages 219–229. Cambridge University Press, 1990.

[246] V. D. Milman and G. Schechtman. *Asymptotic theory of finite dimensional normed spaces*, volume 1200. Springer Science, 1986.

[247] V. D. Milman and G. Schechtman. Global vs. local asymptotic theories of finite dimensional normed spaces. *Duke Mathematical Journal*, 90:73–93, 1997.

[248] S. Minsker. On some extensions of Bernstein's inequality for self-adjoint operators. *Statistics & Probability Letters*, 127:111–119, 2017.

[249] M. Mitzenmacher and E. Upfal. *Probability and computing: Randomization and probabilistic techniques in algorithms and data analysis.* Cambridge University Press, 2017.

[250] A. Moitra. *Algorithmic aspects of machine learning.* Cambridge University Press, 2018.

[251] A. Moitra and G. Valiant. Settling the polynomial learnability of mixtures of gaussians. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 93–102. IEEE, 2010.

[252] S. J. Montgomery-Smith. The distribution of Rademacher sums. *Proceedings of the American Mathematical Society*, 109(2):517–522, 1990.

[253] P. Mörters and Y. Peres. *Brownian motion*, volume 30. Cambridge University Press, 2010.

[254] E. Mossel, J. Neeman, and A. Sly. Belief propagation, robust reconstruction and optimal recovery of block models. In *Conference on Learning Theory*, pages 356–370. PMLR, 2014.

[255] R. Murray, J. Demmel, and M. Mahoney et al. Randomized numerical linear algebra: A perspective on the field with an eye to software. *arXiv preprint arXiv:2302.11474*, 2023.

[256] A. Nemirovsky and D. Yudin. *Problem Complexity and Method Efficiency in Optimization.* Wiley Interscience, 1983.

[257] M. Newman. *Networks: An Introduction.* Oxford University Press, 2010.

[258] H. H. Nguyen and V. Vu. Small ball probability, inverse theorems, and applications. In *Erdös Centennial*, volume 25 of *Bolyai Society Mathematical Studies*, pages 409–463. Springer, 2013.

[259] R. I. Oliveira. Concentration of the adjacency matrix and of the Laplacian in random graphs with independent edges. *arXiv preprint arXiv:0911.0600*, 2009.

[260] R. I. Oliveira. Sums of random hermitian matrices and an inequality by Rudelson. *Electronic Communications in Probability*, 15:203–212, 2010.

[261] R. I. Oliveira and Z. F. Rico. Improved covariance estimation: optimal robustness and sub-gaussian guarantees under heavy tails. *The Annals of Statistics*, 52(5):1953–1977, 2024.

[262] S. O'Rourke, V. Vu, and K. Wang. Random perturbation of low rank matrices: Improving classical bounds. *Linear Algebra and its Applications*, 540:26–59, 2018.

[263] S. O'Rourke, V. Vu, and K. Wang. Optimal subspace perturbation bounds under Gaussian noise. In *2023 IEEE International Symposium on Information Theory (ISIT)*, pages 2601–2606. IEEE, 2023.

[264] M. I. Ostrovskii. Topologies on the set of all subspaces of a Banach space and related questions of banach space geometry. *Quaestiones Mathematicae*, 17(3):259–319, 1994.

[265] S. Oymak and B. Recht. Near-optimal bounds for binary embeddings of arbitrary sets. *arXiv preprint arXiv:1512.04433*, 2015.

[266] S. Oymak, C. Thrampoulidis, and B. Hassibi. The squared-error of generalized lasso: A precise analysis. In *2013 51st Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1002–1009. IEEE, 2013.

[267] S. Oymak and J. A. Tropp. Universality laws for randomized dimension reduction, with applications. *Information and Inference: A Journal of the IMA*, 7(3):337–446, 2018.

[268] A. Pajor. *Sous espaces ℓ_1^n des espaces de Banach.* Editions Hermann, 1985.

[269] A. Pajor and N. Tomczak-Jaegermann. Subspaces of small codimension of finite-dimensional Banach spaces. *Proceedings of the American Mathematical Society*, 97(4):637–642, 1986.

[270] D. Petz. A survey of certain trace inequalities. *Banach Center Publications*, 30(1):287–298, 1994.

[271] G. Pisier. Remarques sur un résultat non publié de B. Maurey. *Séminaire d'Analyse fonctionnelle (dit "Maurey-Schwartz")*, pages 1–12, 1981. Exposé no 5.

[272] G. Pisier. *The volume of convex bodies and Banach space geometry.* Cambridge University Press, 1999.

[273] G. Pisier. Grothendieck's theorem, past and present. *Bulletin of the American Mathematical Society*, 49(2):237–323, 2012.

[274] Y. Plan and R. Vershynin. Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach. *IEEE Transactions on Information Theory*, 59(1):482–494, 2012.

[275] Y. Plan and R. Vershynin. One-bit compressed sensing by linear programming. *Communications on Pure and Applied Mathematics*, 66(8):1275–1297, 2013.

[276] Y. Plan and R. Vershynin. Dimension reduction by random hyperplane tessellations. *Discrete & Computational Geometry*, 51(2):438–461, 2014.

[277] Y. Plan and R. Vershynin. Random matrices acting on sets: Independent columns. *arXiv preprint arXiv:2502.16827*, 2025.

[278] Y. Plan, R. Vershynin, and E. Yudovina. High-dimensional estimation with geometric constraints. *Information and Inference: A Journal of the IMA*, 6(1):1–40, 2017.

[279] D. Pollard. *Empirical Processes: Theory and Applications*, volume 2 of *NSF-CBMS Regional Conference Series in Probability and Statistics.* Institute of Mathematical Statistics, 1990.

[280] Y. Polyanskiy and Y. Wu. *Information Theory: From Coding to Learning.* Cambridge University Press, 2025.

[281] H. Rauhut. Compressive sensing and structured random matrices. In *Theoretical Foundations and Numerical Methods for Sparse Recovery*, volume 9 of *Radon Series on Computational and Applied Mathematics*, pages 1–92. De Gruyter, 2010.

[282] B. Recht. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12(12), 2011.

[283] R. E. Rietz. A proof of the Grothendieck inequality. *Israel Journal of Mathematics*, 19(3):271–276, 1974.

[284] P. Rigollet. High-dimensional statistics. *Lecture notes*, 2010.

[285] O. Rivasplata. Subgaussian random variables: An expository note. *Technical report, University of Alberta*, 2012.

[286] H. Robbins. A remark on Stirling's formula. *Amer. Math. Monthly*, 62:26–29, 1955.

[287] M. Rudelson. Random vectors in the isotropic position. *Journal of Functional Analysis*, 164(1):60–72, 1999.

[288] M. Rudelson and R. Vershynin. Combinatorics of random processes and sections of convex bodies. *Annals of Mathematics*, 164(2):603–648, 2006.

[289] M. Rudelson and R. Vershynin. Sampling from large matrices: An approach through geometric functional analysis. *Journal of the ACM*, 54(4):21–es, 2007.

[290] M. Rudelson and R. Vershynin. On sparse reconstruction from Fourier and Gaussian measurements. *Communications on Pure and Applied Mathematics*, 61(8):1025–1045, 2008.

[291] M. Rudelson and R. Vershynin. Non-asymptotic theory of random matrices: extreme singular values. In *Proceedings of the International Congress of Mathematicians 2010 (ICM 2010)*, pages 1576–1602. World Scientific, 2010.

[292] M. Rudelson and R. Vershynin. Hanson-Wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18:1–9, 2013.

[293] H. Sambale. Some notes on concentration for α-subexponential random variables. In *High Dimensional Probability IX: The Ethereal Volume*, pages 167–192. Springer, 2023.

[294] N. Sauer. On the density of families of sets. *Journal of Combinatorial Theory, Series A*, 13(1):145–147, 1972.

[295] G. Schechtman. Two observations regarding embedding subsets of Euclidean spaces in normed spaces. *Advances in Mathematics*, 200(1):125–135, 2006.

[296] G. Schechtman and J. Zinn. On the volume of the intersection of two l_p^n balls. *Proceedings of the American Mathematical Society*, 110(1):217–224, 1990.

[297] R. Schneider and W. Weil. *Stochastic and integral geometry.* Springer, 2008.

[298] I. Schur. Bemerkungen zur Theorie der beschränkten Bilinearformen mit unendlich vielen Veränderlichen. 1911.

[299] Y. Seginer. The expected norm of random matrices. *Combinatorics, Probability and Computing*, 9(2):149–166, 2000.

[300] S. Shelah. A combinatorial problem; stability and order for models and theories in infinitary languages. *Pacific Journal of Mathematics*, 41(1):247–261, 1972.

[301] Y.-C. Sheu and T.-C. Wang. Matrix deviation inequality for ℓ_p-norm. *Random Matrices: Theory and Applications*, 12(04):2350007, 2023.

[302] I. Shevtsova. On the absolute constants in the Berry-Esseen type inequalities for identically distributed summands. *arXiv preprint arXiv:1111.6554*, 2011.

[303] M. Simonovits. How to compute the volume in high dimension? *Mathematical Programming*, 97:337–374, 2003.

[304] D. Slepian. The one-sided barrier problem for Gaussian noise. *Bell System Technical Journal*, 41(2):463–501, 1962.

[305] David Slepian. On the zeros of Gaussian noise. In *Time Series Analysis*, pages 104–115. Wiley, 1963.

[306] G. W. Stewart and J.-G. Sun. *Matrix Perturbation Theory.* Academic Press, 1990.

[307] M. Stojnic. Various thresholds for ℓ_1-optimization in compressed sensing. *arXiv preprint arXiv:0907.3666*, 2009.

[308] M. Stojnic. Regularly random duality. *arXiv preprint arXiv:1303.7295*, 2013.

[309] V. Sudakov. Gaussian random processes and measures of solid angles in Hilbert space. *Doklady Akademii Nauk*, 197(1):43–45, 1971.

[310] V. Sudakov. *Geometric problems in the theory of infinite-dimensional probability distributions.* American Mathematical Society, 1979.

[311] S. Szarek. On the best constants in the Khinchin inequality. *Studia Mathematica*, 2(58):197–208, 1976.

[312] S. J. Szarek and M. Talagrand. An "isomorphic" version of the Sauer-Shelah lemma and the Banach-Mazur distance to the cube. In *Geometric Aspects of Functional Analysis: Israel Seminar (GAFA) 1987–88*, pages 105–112. Springer, 1989.

[313] S. J. Szarek and M. Talagrand. On the convexified Sauer–Shelah theorem. *Journal of Combinatorial Theory, Series B*, 69(2):183–192, 1997.

[314] M. Talagrand. A new look at independence. *The Annals of Probability*, pages 1–34, 1996.

[315] M. Talagrand. *The generic chaining: upper and lower bounds of stochastic processes.* Springer, 2005.

[316] M. Talagrand. *Upper and lower bounds for stochastic processes.* Springer, 2014.

[317] M. Talagrand. *Upper and lower bounds for stochastic processes: decomposition theorems.* Springer, 2022.

[318] T. Tao and V. Vu. From the Littlewood-Offord problem to the circular law: universality of the spectral distribution of random matrices. *Bulletin of the American Mathematical Society*, 46(3):377–396, 2009.

[319] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996.

[320] K. Tikhomirov. Sample covariance matrices of heavy-tailed distributions. *International Mathematics Research Notices*, 2018(20):6254–6289, 2018.

[321] N. Tomczak-Jaegermann. *Banach-Mazur distances and finite-dimensional operator ideals.* Pitman, 1989.

[322] J. Tropp. Convex recovery of a structured signal from independent random linear measurements. In *Sampling Theory, a Renaissance: Compressive Sensing and Other Developments*, pages 67–101. Birkhäuser, 2015.

[323] J. A. Tropp. Norms of random submatrices and sparse approximation. *Comptes Rendus. Mathématique*, 346(23-24):1271–1274, 2008.

[324] J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12:389–434, 2012.

[325] J. A. Tropp and R. J. Webber. Randomized algorithms for low-rank matrix approximation: Design, analysis, and applications. *arXiv preprint arXiv:2306.12418*, 2023.

[326] J. A. Tropp, A. Yurtsever, M. Udell, and V. Cevher. Practical sketching algorithms for low-rank matrix approximation. *SIAM Journal on Matrix Analysis and Applications*, 38(4):1454–1485, 2017.

[327] Joel A Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends in Machine Learning*, 8(1-2):1–230, 2015.

[328] S. van de Geer. *Applications of empirical process theory.* Cambridge University Press, 2000.

[329] A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics.* Springer, 2nd edition, 2023.

[330] R. van Handel. Probability in high dimension. *Lecture Notes (Princeton University)*, 2(3):2–3, 2014.

[331] R. van Handel. Structured random matrices. In *Convexity and Concentration*, volume 161 of *IMA Volumes in Mathematics and Applications*, pages 107–165. Springer, 2017.

[332] R. van Handel. Chaining, interpolation, and convexity. *Journal of the European Mathematical Society*, 20(10):2413–2435, 2018.

[333] Ramon van Handel. Chaining, interpolation and convexity II: The contraction principle. *The Annals of Probability*, 46(3):1764–1805, 2018.

[334] J. H. van Lint. *Introduction to Coding Theory.* Springer, 1998.

[335] V. N. Vapnik and A. Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications*, 16(2):264–280, 1971.

[336] S. Vempala. Geometric random walks: A survey. In *Combinatorial and Computational Geometry*, volume 52 of *Mathematical Sciences Research Institute Publications*, pages 573–612. Cambridge University Press, 2005.

[337] R. Vershynin. Integer cells in convex sets. *Advances in Mathematics*, 197(1):248–273, 2005.

[338] R. Vershynin. Golden-Thompson inequality. *Lecture notes*, 2009.

[339] R. Vershynin. A note on sums of independent random matrices after Ahlswede-Winter. *Lecture notes*, 2009.

[340] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing: Theory and Applications*, pages 210–268. Cambridge University Press, 2012.

[341] Roman Vershynin. Estimation in high dimensions: a geometric perspective. In *Sampling Theory, a Renaissance: Compressive Sensing and Other Developments*, pages 3–66. Springer, 2015.

[342] C. Villani. *Topics in Optimal Transportation.* American Mathematical Society, 2021.

[343] V. Vu. Singular vectors under random perturbation. *Random Structures & Algorithms*, 39(4):526–538, 2011.

[344] M. J. Wainwright. *High-dimensional Statistics: A non-Asymptotic Viewpoint*, volume 48. Cambridge University Press, 2019.

[345] K. Wang. Analysis of singular subspaces under random perturbations. *arXiv preprint arXiv:2403.09170*, 2024.

[346] P.-A. Wedin. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12:99–111, 1972.

[347] P.-A. Wedin. On angles between subspaces of a finite dimensional inner product space. In *Matrix Pencils: Proceedings of a Conference Held at Pite Havsbad, Sweden, March 22–24, 1982*, volume 973 of *Lecture Notes in Mathematics*, pages 263–285. Springer, 2006.

[348] J. G. Wendel. A problem in geometric probability. *Mathematica Scandinavica*, 11(1):109–111, 1962.

[349] A. Wigderson and D. Xiao. Derandomizing the Ahlswede-Winter matrix-valued Chernoff bound using pessimistic estimators, and applications. *Theory of Computing*, 4(1):53–76, 2008.

[350] F. T. Wright. A bound on tail probabilities for quadratic forms in independent random variables whose distributions are not necessarily symmetric. *The Annals of Probability*, 1(6):1068–1070, 1973.

[351] C. Xu and L. Jacques. Quantized compressive sensing with rip matrices: The benefit of dithering. *Information and Inference: A Journal of the IMA*, 9(3):543–586, 2020.

[352] Y. Yu, T. Wang, and R. Samworth. A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015.

[353] A. Y. Zhang and H. H. Zhou. Minimax rates of community detection in stochastic block models. *The Annals of Statistics*, 44(5):2252–2280, 2016.

[354] S. Zhou. Sparse Hanson–Wright inequalities for subgaussian quadratic forms. *Bernoulli*, 25(3):1603–1639, 2019.

[355] A. Zymnis, S. Boyd, and E. Candes. Compressed sensing with quantized measurements. *IEEE Signal Processing Letters*, 17(2):149–152, 2009.

Index