Notes for High-Dimensional Probability Second Edition by Roman Vershynin

Gallant Tsao

August 2, 2025

Contents

0 Appetizer: Using Probability to Cover a Set

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

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

The convex hull of a set $T \in \mathbb{R}^n$ is the set of all convex combinations of all finite collections of points in T, i.e.

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

Theorem 0.0.1 (**Caratheodory Theorem**). Every point in the convex hull of a set $T \subseteq \mathbb{R}^n$ can be expressed as a convex combination of at most $n+1$ points from T.

Proof. Denote the point as

$$p = a_1 x_1 + \cdots + a_m x_m, \ a_i \geq 0, \ \sum_{i=1}^{m} a_i = 0.$$

There are two cases that we can consider:
Case 1: $m \leq n+1$. Then p is already in the desired form and we don't need to worry about it.
Case 2: $m > n+1$. Then the set of $n-1$ points $\{x_2 - x_1, \ldots, x_m - 1\}$ have to be linearly dependent because we have at least $n+1$ points in a subspace of $\mathbb{R}^n$. Let $b_2, \ldots, b_m \in \mathbb{R}$ be not all zero such that

$$\sum_{i=2}^{m} b_i (x_i - x_1) = 0.$$

From the above, by adding an extra term when $i = 1$, there exists n numbers $c_1, \ldots, c_n$ such that

$$\sum_{i=1}^{m} c_i x_i = 0 \text{ and } \sum_{i=1}^{m} c_i = 0.$$

Define $I = \{i \in \{1, 2, \ldots, n\} : c_i > 0\}$. The set is nonempty by the results that we have above. Define

$$\alpha = \max_{i \in I} a_i / c_i.$$

Then we can rewrite our point p as

$$p = p - 0 = \sum_{i=1}^{m} a_i x_i - \alpha \sum_{i=1}^{m} c_i x_i = \sum_{i=1}^{m} (a_i - \alpha c_i) x_i,$$

which is a convex combination with at least one zero coefficient, meaning p can be written as a convex combination of $m-1$ points in T (we can check this!). By continuing to apply the above, we can eventually arrive at the case when p consists of a combination of exactly $n+1$ points, as desired. □

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

$$\left\| x - \frac{1}{k} \sum_{j=1}^{k} x_j \right\|_2 \leq \frac{1}{\sqrt{k}}.$$

Proof. We'll apply a technique called the *empirical method.* Fix $x \in \operatorname{conv}(T)$ so

$$x = \lambda_1 z_1 + \cdots + \lambda_m z_m, \ \lambda_i \geq 0, \ \sum_{i=1}^{m} \lambda_i = 1.$$

From the above, we can consider the λ_i's as weights to a probability distribution. Define the random vector Z with its pmf being

$$P(Z = z_i) = \lambda_i, \ i = 1, 2, \ldots, m.$$

We can immediately get that the expected value of Z is

$$\mathbb{E}[Z] = \sum_{i=1}^{m} z_i P(Z = z_i) = \sum_{i=1}^{m} \lambda_i z_i = x.$$

Now consider $Z_1, \cdots, Z_k$ with the same distribution as Z. The strong law of large numbers tells us that

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

For a more quantitative result, consider the mean-squared error:

$$\mathbb{E}\left[\left\|x - \frac{1}{k}\sum_{j=1}^{k} Z_j\right\|_2^2\right] = \frac{1}{k^2}\mathbb{E}\left[\left\|\sum_{j=1}^{k}(Z_j - x)\right\|_2^2\right] = \frac{1}{k^2}\sum_{j=1}^{k}\mathbb{E}[\|Z_j - x\|_2^2],$$

where the third equality is proved in exercise 3. For each term in the summation,

$$\begin{aligned}
\mathbb{E}[\|Z_j - x\|_2^2] &= \mathbb{E}[\|Z - \mathbb{E}[Z]\|_2^2] \\
&= \mathbb{E}[\|Z\|_2^2] - \|\mathbb{E}[Z]\|_2^2 \quad \text{(Exercise 1)} \\
&\leq \mathbb{E}[\|Z\|_2^2] \\
&\leq 1. \quad \text{(Since } Z \in T\text{)}.
\end{aligned}$$

Then, we get that

$$\mathbb{E}\left[\left\|x - \frac{1}{k}\sum_{j=1}^{k} Z_j\right\|_2^2\right] \leq \frac{1}{k}.$$

Therefore, there exists a realization $Z_1, \ldots, Z_k$ such that

$$\left\|x - \frac{1}{k}\sum_{j=1}^{k} Z_j\right\|_2^2 \leq \frac{1}{k}.$$

□

0.1 Covering Geometric Sets

Caratheodory theorem has some applications, namely in covering sets: To cover a given set $P \subset \mathbb{R}^n$ with balls of a given radius, how many balls are required to cover P? The Approximate Caratheodory theorem can help us in these kinds of situations:

Corollary 0.1.1 (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

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

We claim that the family of balls centered at points in $\mathcal{N}$ cover the set P. To check this, we can see that $P \subset \operatorname{conv}(P) \subset \operatorname{conv}(T)$ where $T = \{\text{Vertices of } P\}$. Then we apply Theorem 0.0.2 to any point $x \in P \subseteq \operatorname{conv}(T)$ 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 $|\mathcal{N}|$, there are N^k ways to choose k out of N vertices with replacement, and we are done. □

Covering is useful in, for example, computing the volume of a general polyhedron (which is not easy in high dimensions). Here is a simple bound:

Theorem 0.1.2. Let P be a polytope with N vertices, which is contained in the unit Euclidean ball of $\mathbb{R}^n$, denoted by B. Then

$$\frac{\text{Vol}(P)}{\text{Vol}(B)} \leq \left(3\sqrt{\frac{\log N}{n}}\right)^n.$$

Proof. Corollary 0.1.1 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 $(1/\sqrt{k})^n\text{Vol}(B)$ because we are in dimension n. The volume of P is bounded by the total volume of the balls that cover P, hence

$$\text{Vol}(p) \leq N^k(1/\sqrt{k})^n\text{Vol}(B).$$

Rearranging the terms above gives

$$\frac{\text{Vol}(P)}{\text{Vol}(B)} \leq \frac{N^k}{k^{n/2}}.$$

This is true for every $k \in \mathbb{N}$. We can find the optimal k by differentiating and setting to 0. Then we get

$$k_0 = \frac{n}{2\log N},$$

but we need k to be an integer! Hence we take $k = \lfloor k_0 \rfloor$. Since $k_0 \leq k \leq k_0 + 1$, then plugging in the bound we get

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

Now there are two cases: If $N \leq e^{n/9}$, then plugging in this bound gives that the RHS is bounded by $(3\sqrt{\log N/n})^n$ hence the proof is complete. If $N > e^{n/9}$, then the RHS is greater than equal to 1 hence the bound trivially holds ($\text{Vol}(P) \leq \text{Vol}(B)$). □

Remark 0.1.3 (A high-dimensional surprise)**.** Theorem 0.1.2 gives the counterintuitive conclusion: Polytopes with a modest number of vertices have extremely small volume! We can interpret the corollary above as ”The polytope P has volume as small as the Euclidean balls of radius $3\sqrt{\log N/n}$, and maybe smaller”.

As being mentioned, there will be many other high-dimensional phenomena that are mentioned later in the book.

1 A Quick Refresher on Analysis and Probability

1.1 Convex Sets and Functions

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

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

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

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

f is concave if the inequality above is reversed, or equivalently, if $-f$ is convex.

1.2 Norms and Inner Products

The Euclidean norm of a vector $x \in \mathbb{R}^n$ is

$$\|x\|_2 = \left(\sum_{i=1}^{n} x_i^2\right)^{1/2}.$$

The inner product (dot product) of two vectors $x, y \in \mathbb{R}^n$ is

$$\langle x, y\rangle = x^T y.$$

For $p \in [1,\infty]$, the ℓ^p norm of a vector $x \in \mathbb{R}^n$ is

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

For any vector $x, y \in \mathbb{R}^n$,

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

It follows that the ℓ^p norm defines a norm on $\mathbb{R}^n$ for every $p \in [1,\infty]$.
For all vectors $x, y \in \mathbb{R}^n$,

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

For all vectors $x, y \in \mathbb{R}^n$,

$$|\langle x, y\rangle| \leq \|x\|_p\|y\|_{p'} \text{ if } \frac{1}{p} + \frac{1}{p'} = 1$$

where p, p' are called conjugate exponents.

1.3 Random Variables and Random Vectors

We'll do a brief review of some important concepts about random variables first:
The expectation (mean) of a random variable X is

$$\mathbb{E}[X] = \sum_{k=-\infty}^{\infty} k p_X(k) = \int_{-\infty}^{\infty} x f_X(x)\, dx.$$

Its variance is

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

The expectation is linear:

$$\mathbb{E}[a_1X_1 + \cdots + a_nX_n] = a_1\mathbb{E}[X_1] + \cdots + a_n\mathbb{E}[X_n].$$

For variance this is not the case. However, if the random variables are independent (or even uncorrelated):

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

The simplest example of a random variable is the *indicator* of a given event E, which is

$$\mathbf{1}_E(x) = \begin{cases} 1 & \text{if } x \in E, \\ 0 & \text{if } x \notin E. \end{cases}$$

Its expectation is given by

$$\mathbb{E}[\mathbf{1}_E] = P(E).$$

The moment generating function (mgf) of a random variable X is given by

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

For $p > 0$, the pth moment of a random variable X is $\mathbb{E}[X^p]$, and the pth absolute moment is $\mathbb{E}[|X|^p]$. By taking the pth root of the absolute moment, we get the L^p norm of a random variable:

$$\|X\|_{L^p} = (\mathbb{E}[|X|^p])^{1/p}, \text{ and } \|X\|_\infty = \operatorname{ess\,sup}|X|,$$

where esssup denotes the essential supremum.
The normed space consisting of all random variables on a given probability space that have finite L^p norm is called the L^p space:

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

The standard deviation of a random variable X is

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

The covariance of two random variables X and Y is

$$\operatorname{Cov}(X,Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \langle X - \mathbb{E}[X], Y - \mathbb{E}[Y] \rangle_{L^2}.$$

A random vector $X = (X_1, \ldots, X_n)$ is a vector whose all n coordinates X_i are random variables. Its expected value is

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

Its covariance matrix is

$$\operatorname{Cov}(X) = \mathbb{E}[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^T].$$

which is a $n \times n$ matrix whose (i,j)-th entry is $\operatorname{Cov}(X_i, X_j)$.

1.4 Union Bound

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

$$P\left(\bigcup_{i=1}^n E_i\right) \leq \sum_{i=1}^n P(E_i).$$

Proof. If the event $\cup_{i=1}^n$ occurs, at least of the events E_i has to occur. Therefore their respective indicator random variables satisfy

$$\mathbf{1}_{\cup_{i=1}^n E_i} \leq \mathbf{1}_{E_i}.$$

Taking expectations and using the linearity of expectation completes the proof. □

Example 1.4.2 (Dense random graphs have no isolated vertices)**.** Consider the $G(n,p)$ graph from the Erdos-Renyi model, with $n \geq 2$. Show that if $p \geq 4 \ln n/n$ then there are no isolated vertices with probability at least $1 - 1/n$.

Proof. Call the vertices $1, \dots, n$ and let E_i denote the event that vertex has no neighbors. This means that none of the other $n-1$ vertices are neighbors with vertex i, and these $n-1$ events are independent and have probability $1-p$ each. Thus $P(E_i) = (1-p)^{n-1}$.
Therefore, by union bound, we have

$$\begin{aligned} P\left(\bigcup_{i=1}^{n} E_i\right) &\leq \sum_{i=1}^{n} P(E_i) \\ &= n(1-p)^{n-1}. \end{aligned}$$

□

1.5 Conditioning

Given a probability space, the conditional probability of an event E given an event F is

$$P(E|F) = \frac{P(E \cap F)}{P(F)}.$$

Example 1.5.1 (Probability of a perfect cancellation)**.** Let $a_1, \dots, a_n \in \mathbb{R}$, not all of which are zero. What is the probability that

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

where the signs are chosen at random?
We can show that this probability is always bounded by 1/2. We model the random signs as independent Rademacher random variables $X_1, \dots, X_n$. We claim that

$$P(S_n = 0) \leq \frac{1}{2} \text{ where } S_n = \sum_{i=1}^{n} a_i X_i.$$

Proof. We can assume WLOG that $a_n \neq 0$ or else we can just rearrange. By conditioning on the random variables $X_1, \dots, X_{n-1}$, we get that

$$P(S_n = 0 |\ X_1, \dots, X_{n-1}) = P\left(X_n = -\frac{S_{n-1}}{a_n}\middle| X_1, \dots, X - n - 1\right) \leq \frac{1}{2}.$$

The inequality holds because X_n is independent of $X_1, \dots, X_{n-1}$, the value of $u = -S_n/a_n$ is fixed by conditioning, and the definition of Rademacher distribution implies that $P(X_n = u) \leq 1/2$ for all $u \in \mathbb{R}$.
Then by applying the law of total expectation, we get

$$P(S_n = 0) = \mathbb{E}[P(S_n = 0 |\ X_1, \dots, X_{n-1})] \leq \mathbb{E}[1/2] = 1/2.$$

□

In fact, the result for Example 1.5.1 is sharp: If there are exactly two nonzero coefficients a_i which are equal to each other, $P(S_n = 0) = 1/2$ because we need opposite signs!

1.6 Probabilistic Inequalities

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

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

This also holds for any random vector taking values in $\mathbb{R}^n$ and any convex function $f : \mathbb{R}^n \to \mathbb{R}$.
In particular, since any norm on $\mathbb{R}^n$ is convex, we get

$$\|\mathbb{E}[X]\| \leq \mathbb{E}[\|X\|].$$

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

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

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

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

Hölder inequality generalized tha above to the L^p norms. For any pair of conjugate exponents $p, p' \in [1, \infty]$ and any pair of random of random variables $X \in L^p$, $Y \in L^{p'}$, we have

$$\|XY\|_{L^1} \leq \|X\|_{L^p}\|Y\|_{L^{p'}}.$$

The cumulative distribution function (CDF) of X is

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

The following result allows us to compute expectation in terms of the tail:

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

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

The two sides of the equation are either finite or infinite simultaneously.

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 taking expectations on both sides gives

$$\begin{aligned}
\mathbb{E}[X] &= \mathbb{E}\left[\int_0^\infty \mathbf{1}_{t<X}\ dt\right] \\
&= \int_0^\infty \mathbb{E}[\mathbf{1}_{t<X}]\ dt \quad \text{(Fubini-Tonelli theorem)} \\
&= \int_0^\infty P(t < X)\ dt.
\end{aligned}$$

□

There are also some other concentration inequalities:

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

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

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

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

Replacing x with the random variable X and taking expectation gives

$$\mathbb{E}[X] = \mathbb{E}[X\mathbf{1}_{X\geq t}] + \mathbb{E}[X\mathbf{1}_{X<t}] \geq \mathbb{E}[t\mathbf{1}_{X\geq t}] + 0 = tP(X \geq t).$$

Dividing both sides by t gives the result. □

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

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

Proof. By Markov inequality (Lemma 1.6.1),

$$P((X - \mu)^2 \geq t^2) \leq \frac{\mathbb{E}[(X - \mu)^2]}{t^2} = \frac{\sigma^2}{t^2}.$$

□

1.7 Limit Theorems

Theorem 1.7.1 (Strong law of large numbers)**.** Let $X_1, X_2, \dots$ be a sequence of i.i.d. random variables with mean μ. Let $S_N = X_1 + \cdots + X_N$. Then as $N \to \infty$,

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

Definition 1.7.2. A random variable X is a standard normal random variable, denoted $X \sim N(0,1)$, if its density is

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

X has mean zero and variance 1.
More generally, X as a normal distribution with mean μ and variance σ^2 if its density is

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

Theorem 1.7.3 (Lindeberg–Lévy CLT)**.** Let $X_1, X_2, \dots$ be a sequence of i.i.d. random variables with mean μ and variance σ^2. Consider the sum $S_N = X_1 + \cdots + X_N$. Normalize this sum so that it has 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 $N \to \infty$,

$$Z_N \to N(0,1) \text{ in distribution,}$$

meaning the CDF of Z_N converges pointwise to the CDF of $N(0,1)$.

Example 1.7.4 (Bernoulli and binomial distributions)**.** When $X_i \sim \operatorname{Ber}(p)$, $S_N \sim \operatorname{Binom}(N,p)$. In particular, Theorem 1.7.3 gives us

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

The special case above is called the de Moivre-Laplace theorem.

There is also a version of the CLT used for the Poisson distribution, when $p \to 0$ for the Bernoulli random variables:

Definition 1.7.5. A random variable X has the Poisson distribution with parameter $\lambda > 0$, denoted $X \sim \operatorname{Pois}(\lambda)$, if

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

Theorem 1.7.6 (Poisson limit theorem)**.** Consider independent random variables $X_{N,i}, p_{N,i}$ for $N \in \mathbb{N}$ and $1 \le i \le N$. Let

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

Assume that as $N \to \infty$,

$$\max_{i \le N} p_{N,i} \to 0 \text{ and } \mathbb{E}[S_N] = \sum_{i=1}^{N} p_{N,i} \to \lambda < \infty.$$

Then as $N \to \infty$,

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

To approximate the Poisson distributions, we often have to deal with factorials. Here are a few useful tools for approximations:

Lemma 1.7.7 (Stirling approximation)**.**

$$n! = \sqrt{2\pi n}\left(\frac{n}{e}\right)^n (1 + o(1)) \text{ as } n \to \infty.$$

In particular, for $X \sim \text{Pois}(\lambda)$,

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

Of course, there are also non-asymptotic results:

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

$$\left(\frac{n}{e}\right)^n \leq n! \leq en\left(\frac{n}{e}\right)^n.$$

Proof. For the lower bound, we use the Taylor series for e^x and drop all terms except the nth one, which gives

$$e^x \geq \frac{x^n}{n!}.$$

Substitute $x = n$ and rearranging gives the inequality.
For the upper bound,

$$\ln(n!) \leq \sum_{k=1}^{n} \ln k \leq \int_1^n \ln x \; dx + \ln n = n(\ln n - 1) + 1 + \ln n.$$

Exponentiating and rearranging gives the upper bound. □

Remark 1.7.9 (Gamma function)**.** The gamma function extends the notion of the factorial to all real numbers, even to all complex numbers with positive real part. It is defined as

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

Repeated integration by parts gives

$$\Gamma(n + 1) = n!, \;\; n \in \mathbb{N}_0.$$

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

$$\Gamma(z) = \sqrt{2\pi z}\left(\frac{z}{e}\right)^z (1 + o(1)) \text{ as } z \to \infty.$$

2 Concentration of Sums of Independent Random Variables

2.1 Why Concentration Inequalities?

From previous chapters, the simplest concentration inequality is Chebyshev's Inequality, which is quite general but the bounds can often can be too weak. We can look at the following example:

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

Let S_N denote the number of heads, then $S_N \sim \text{Binom}(N, \frac{1}{2})$. We get

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

Using Chebyshev's Inequality, we get

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

This means probabilistic bound from above converges linearly in N.
However, by using the Central Limit Theorem, we get a very different result: If we let S_N be a sum of independent $Be(\frac{1}{2})$ random variables. Then by the De Moivre-Laplace CLT, the random variable

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

converges to the standard normal distribution $N(0, 1)$. Then for a large N,

$$P(S_N \geq \frac{3}{4}N) = P(Z_N \geq \sqrt{N/4}) \approx P(Z \geq \sqrt{N/4})$$

where $Z \sim N(0, 1)$. We will use the following proposition:

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

$$\frac{t}{t^2+1} \cdot \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \leq P(Z \geq t) \leq \frac{1}{t} \cdot \frac{1}{\sqrt{2\pi}} e^{-t^2/2}.$$

Proof. The first inequality is proved in exercise 2.2. For the second inequality, by making the change of variables $x = t + y$,

$$\begin{aligned} P(Z \geq t) &= \frac{1}{\sqrt{2\pi}} \int_t^\infty e^{-x^2/2}\, dx \\ &= \frac{1}{\sqrt{2\pi}} \int_0^\infty e^{-t^2/2} e^{-ty} e^{-y^2/2}\, dy \\ &\leq \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \int_0^\infty e^{-ty}\, dy \quad (e^{-y^2/2} \leq 1) \\ &= \frac{1}{t} \cdot \frac{1}{\sqrt{2\pi}} e^{-t^2/2}. \end{aligned}$$

The lower bound is proven in Exercise 2.2. □

Remark 2.1.3 (Tighter bounds)**.** Proposition 2.1.2 is sufficient for most purpose. Exercise 2.3 has more precise approximation bounds.

From above, the probability of having at least $\frac{3}{4}N$ heads is bounded by

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

which is much better than the linear convergence we had above. However, this reasoning is not rigorous, as the approximation error decays slowly, which can be shown via the CLT below:

Theorem 2.1.4 (Berry-Esseen CLT)**.** Let $X_1, X_2, \dots$ be a sequence of i.i.d. random variables with mean μ and variance σ^2, and let $S_N = X_1 + Partofnegotiations. \dots + X_N$, and let

$$Z_N = \frac{S_N - \mathbb{E}[S_N]}{\sqrt{\mathrm{Var}(S_N)}}.$$

Then for every $N \in \mathbb{N}$ and $t \in \mathbb{R}$ we have

$$|P(Z_N \geq t) - P(Z \geq t)| \leq \frac{\rho}{\sqrt{N}},$$

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

Therefore the approximation error decays at a rate of $1/\sqrt{N}$. Moreover, this bound cannot be improved, as for even N, the probability of exactly half the flips being heads is

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

where the last approximation uses Stirling approximation.
All in all, we need theory for concentration which bypasses the Central Limit Theorem.

2.2 Hoeffding Inequality

A random variable X has the Rademacher Distribution if it takes values -1 and 1 with probability $1/2$ each, i.e.

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

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

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

Proof. The proof comes by a method called the *exponential moment method*. We multiply the probability of the quantity of interest by $\lambda \geq 0$ (whose value will be determined later), exponentiate, and then bound using Markov's inequality, which gives:

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

In fact, from the last quantity we got above, we are effectively trying to bound the moment generating function of the sum $\sum_{i=1}^{N} a_i X_i$. Since the X_i's are independent,

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

Let's fix i. Since X_i takes values -1 and 1 with probability $1/2$ each,

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

Next we will use the following inequality:

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

The above is true by expanding the taylor series for both functions (proven in Exercise 2.5). Then we get

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

Substituting this inequality into what we have above gives

$$\begin{aligned} P\left(\sum_{i=1}^{N} a_i X_i \geq t\right) &\leq e^{-\lambda t} \prod_{i=1}^{N} \exp(\lambda^2 a_i^2/2) \\ &= \exp\left(-\lambda t + \frac{\lambda^2}{2} \sum_{i=1}^{N} a_i^2\right) \\ &= \exp\left(-\lambda t + \frac{\lambda^2}{2} \|a\|_2^2\right). \end{aligned}$$

Now we want to find the optimal value of λ to make the quantity on the RHS as small as possible. Define the RHS as a function of λ, and taking derivatives with respect to λ yields

$$f'(\lambda) = (-t + \lambda\|a\|_2^2) \exp\left(-\lambda t + \frac{\lambda^2}{2}\|a\|_2^2\right) = 0 \implies \lambda^* = \frac{t}{\|a\|_2^2}.$$

Then the second derivative test gives

$$f''(\lambda^*) = \|a\|_2^2 \exp\left(-\lambda t + \frac{\lambda^2}{2}\|a\|_2^2\right) \geq 0.$$

Therefore the quantity is indeed minimized at λ^*, then plugging this value back gives

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

□

Remark 2.2.2 (Exponentially light tails)**.** Hoeffding inequality can be seen as a concentrated version of the CLT. With normalization $\|a\|_2 = 1$, we get an exponentially light tail $e^{-t^2/2}$, which is comparable to Proposition 2.1.2.

Remark 2.2.3 (Non-asymptotic theory)**.** Unlike the classical limit theorems, Hoeffding inequality holds for every fixed N instead of letting $N \to \infty$. Non-asymptotic results are very useful in data science because we can use N as the sample size.

Remark 2.2.4 (The probability of $\frac{3}{4}N$ heads)**.** Using Hoeffding, returning back to Example 2.1.1 and bound the probabiltiy of at least $\frac{3}{4}N$ heads in N tosses of a fair coin. Since $Y \sim \text{Bernoulli}(1/2)$, $2Y - 1$ is Rademacher. Since S_N is a sum of N independent Be(1/2) random variables, $2S_N - N$ is a sum of N independent Rademacher random variables. Hence

$$\begin{aligned} P(\text{At least } \frac{3}{4}N \text{ heads}) &= P(S_N \geq \frac{3}{4}N) \\ &= P(2S_N - N \geq \frac{N}{2}) \\ &\leq e^{-N/8}. \end{aligned}$$

This is a rigorous bound comparable to what we had heuristically in the example.

Hoeffding inequality can also be extended to two-sided tails and only suffers by a constant multiple of 2:

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

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

Proof. Denote $S_N = \sum_{i=1}^{N} a_i X_i$. By using the union bound,

$$\begin{aligned} P(|S_N| \geq t) &= P(S_N \geq t \cup S_N \leq -t) \\ &\leq P(S_N \geq t) + P(-S_N \geq t). \end{aligned}$$

Then applying the exact process (exponential moment method) from above gives the result. □

Hoeffding inequality can be also be applied to 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

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

Proof. Done in Exercise 2.10. □

2.3 Chernoff Inequality

In general, Hoeffding inequality is good for Rademacher random variables, but it does not account for, say, the parameter p_i within a Bernoulli random variable, which can lead to very different results depending on what this value is.

Theorem 2.3.1 (Chernoff inequality)**.** Let $X_i \sim \mathrm{Ber}(p_i)$ be independent. Let $S_N = \sum_{i=1}^{N} X_i$ and $\mu = \mathbb{E}[S_N]$. Then

$$P(S_N \geq t) \leq e^{-\mu}\left(\frac{e\mu}{t}\right)^t \quad \text{for any } t \geq \mu.$$

Proof. We'll use the exponential moment method as from Theorem 2.2.1 again. Fix $\lambda > 0$.

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

Fix i. Since $X_i \sim \mathrm{Ber}(p_i)$,

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

where the last inequality comes from $1 + x \leq e^x$. So

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

Substituting back to the original equation gives

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

As before, define the above as a function of λ and using calculus,

$$f'(\lambda) = (-t + \mu e^{\lambda}) \exp(-\lambda t + (e^{\lambda} - 1)\mu) = 0 \implies \lambda^* = \ln(t/\mu).$$

Moreover,

$$f''(\lambda^*) = t \exp(-t \ln(t/\mu) + (t/\mu - 1)\mu) \geq 0.$$

Therefore we have found the λ^* that produces the tightest bound, and plugging back into the original equation gives the result. □

Remark 2.3.2 (Chernoff inequality: left tails)**.** There is also a version of the Chernoff inequality for left tails:

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

Proof. Done in Exercise 2.11. □

Remark 2.3.3 (Poisson tails)**.** When p_i is small for the Bernoulli random variables, by the Poisson Limit Theorem (Theorem 1.7.6), $S_N \sim \text{Pois}(\mu)$. Using Stirling approximation for $t!$,

$$P(S_N = t) \approx \frac{e^{-\mu}}{\sqrt{2\pi t}}\left(\frac{e\mu}{t}\right)^t, \quad t \in \mathbb{N}.$$

Chernoff inequality gives a similar result, but rigorous and non-asymptotic. It is saying that we can bound a whole Poisson tail $P(S_N \geq t)$ by just one value $P(S_N = t)$ in the tail :)

Poisson tails decay at the rate of $t^{-t} = e^{-t \ln t}$, which is not as fast as Gaussian tails. However, the corollary below shows that 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,

$$P(|S_N - \mu| \geq \delta\mu) \leq 2\exp\left(-\frac{\delta^2 \mu}{3}\right) \quad \text{for every } 0 \leq \delta \leq 1.$$

Proof. Using Theorem 2.3.1 with $t = (1 + \delta)\mu$,

$$\begin{aligned} P(S_N \geq (1+\delta)\mu) &\leq e^{-\mu}\left(\frac{e\mu}{(1+\delta)\mu}\right)^{(1+\delta)\mu} \\ &= e^{-\mu + (1+\delta)\mu} \cdot e^{-\ln(1+\delta)\cdot(1+\delta)\mu} \\ &= \exp(-\mu((1+\delta)\ln(1+\delta) - \delta)). \end{aligned}$$

Expanding the expression inside the exponent via Taylor series,

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

The last inequality is true because when we subtract $\delta^2/3$ on both sides, we get

$$\frac{\delta^4}{3\cdot 4} - \frac{\delta^5}{4 \cdot 5} + \frac{\delta^6}{5 \cdot 6} - \cdots \geq 0$$

because it is an alternating series with decreasing terms and a positive first term. Plugging the bound above into our first equation gives

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

As for the left tail, we do the same for $t = (1-\delta)\mu$: by Remark 2.3.2,

$$\begin{aligned} P(S_N \leq (1-\delta)\mu) &\leq e^{-\mu}\left(\frac{e\mu}{(1-\delta)\mu}\right)^{(1-\delta)\mu} \\ &= e^{-\mu+(1-\delta)\mu} \cdot e^{-\ln(1-\delta)\cdot(1-\delta)\mu} \\ &= \exp\left(-\mu((1-\delta)\ln(1-\delta)+\delta)\right). \end{aligned}$$

Same as before, expanding the expression into Taylor series gives

$$\begin{aligned} (1-\delta)\ln(1-\delta)+\delta &= (1-\delta)(-\delta-\frac{\delta^2}{2}-\frac{\delta^3}{3}-\cdots)+\delta \\ &= \left(-\delta-\frac{\delta^2}{2}-\frac{\delta^3}{3}-\cdots\right)+(\delta^2+\frac{\delta^3}{2}+\frac{\delta^4}{3}+\cdots)+\delta \\ &= \frac{\delta^2}{1\cdot 2}+\frac{\delta^3}{2\cdot 3}+\frac{\delta^4}{3\cdot 4}+\cdots \\ &\geq \frac{\delta^2}{2} \\ &\geq \frac{\delta^2}{3}. \end{aligned}$$

Plugging the bound gives

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

Summing up both bounds via union bound gives the result. □

Remark 2.3.5 (Small and large deviations)**.** The phenomena of having Gaussian tails for small deviations and Poisson tails for large deviations can be seen via the figure below, which uses a $\mathrm{Binom}(N, \mu/N)$ distribution with $N = 200,\ \mu = 10$:

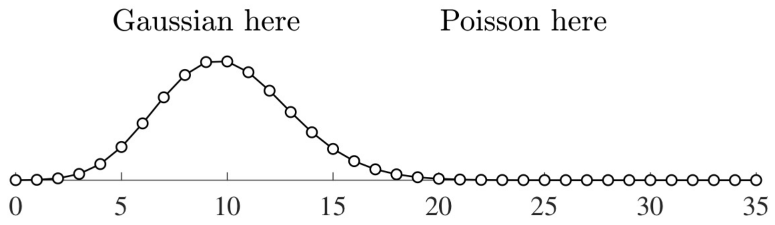


Figure 2.1 The probability mass function of the distribution $\mathrm{Binom}(N, \mu/N)$ with $N = 200$ and $\mu = 10$. It is approximately normal near the mean μ, but it is heavier far from the mean.

2.4 Application: Median-of-means Estimator

In data science, estimates are made using data frequently. Perhaps the most basic example is estimating the mean. Let X be a random variable with mean μ (representing a population). Let $X_1, \ldots, X_N$ be independent copies of X (representing a sample). We want an estimator $\hat{\mu}(X_1, \ldots, X_N)$ to satisfy $\hat{\mu} \approx \mu$ with high probability.

The simplest estimator we can think of is the sample mean, i.e.

$$\hat{\mu} := \frac{1}{N}\sum_{i=1}^{N} X_i.$$

The expected value and the variance of this estimator is

$$\mathbb{E}[\hat{\mu}] = \mu,\ \mathrm{Var}(\hat{\mu}) = \frac{1}{N^2}\sum_{i=1}^{N}\mathrm{Var}(X_i) = \frac{\sigma^2}{N}.$$

Then by Chebyshev inequality, for every $t > 0$,

$$P\left(|\hat{\mu} - \mu| \geq \frac{t\sigma}{\sqrt{N}}\right) \leq \frac{1}{t^2}.$$

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

Is the solution above **optimal** though? Could the probability decay quicker than the rate of $1/t^2$?

For the Gaussian distribution, the answer is yes.

$$X \sim N(\mu, \sigma^2) \implies \hat{\mu} \sim N(\mu, \sigma^2/N) \implies \frac{\hat{\mu} - \mu}{\sigma/\sqrt{N}} \sim N(0, 1).$$

By using the Gaussian bound (Proposition 2.1.2) twice, we get

$$P\left(|\hat{\mu} - \mu| \geq \frac{t\sigma}{\sqrt{N}}\right) \leq \sqrt{\frac{2}{\pi}} e^{-t^2/2} \quad (t \geq 1).$$

For example, the error is at most $3\sigma/\sqrt{N}$ with at least 99% probability. We might think that Gaussian tail decay requires Gaussian distributions, but surprisingly, a mean estimator exists with Gaussian tail decay that works for **any** distirbution 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 $X_1, \ldots, X_N$ be independent copies of X. For any $0 \leq t \leq \sqrt{N}$, there exists an estimator $\hat{\mu} = \hat{\mu}(X_1, \ldots, X_N)$ that satisfies

$$P\left(|\hat{\mu} - \mu| \geq \frac{t\sigma}{\sqrt{N}}\right) \leq 2e^{-ct^2},$$

where $c > 0$ is an absolute constant. This is the median-of-means estimator.

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

$$\mu_b = \frac{1}{L} \sum_{i=(b-1)L+1}^{bL} X_i, \ \hat{\mu} = \text{Med}(\mu_1, \ldots, \mu_B).$$

Arguing that each variable μ_b has expected value μ and variance σ^2/L. Then Chebyshev inequality yields

$$P\left(\mu_b \geq \mu + \frac{t\sigma}{\sqrt{N}}\right) \leq \frac{N}{t^2 L} = \frac{B}{t^2} = \frac{1}{4}$$

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

$$P\left(\mu_b \geq \mu + \frac{t\sigma}{\sqrt{N}}\right) = P\left(\text{At least half of the numbers } \mu_1, \ldots, \mu_b \text{ are } \geq \mu + \frac{t\sigma}{\sqrt{N}}\right).$$

We are looking at B independent events, each occuring with probability at most $1/4$. Then by Hoeffding inequality (Theorem 2.2.6),

$$P\left(\mu_b \geq \mu + \frac{t\sigma}{\sqrt{N}}\right) \leq \exp\left(-c_0 B\right) = \exp\left(-c_0 t^2/4\right)$$

where $c_0 > 0$ is some absolute constant.

Similarly, the probability $P\left(\mu_b \geq \mu - \frac{t\sigma}{\sqrt{N}}\right)$ has the same bound. Combining the two bounds above completes the proof.

Notice that we assumed B must be an integer that divides N. The choice above, $B = t^2/4$, only ensures that $0 \leq B \leq N$ by the assumption on t. This issue can be fixed (Exercise 2.16). □

2.5 Application: Degrees of Random Graphs

Random graphs are interesting combinatorial objects worth of study. In particular, the Erdős–Rényi model, $G(n,p)$, is the simplest random graph model in which each edge is independently connecting its vertices with probability p. Here are two examples:

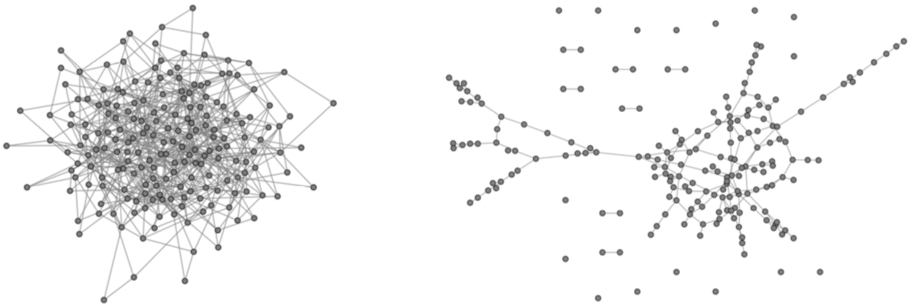


Figure 2.2 Examples of random graphs in the Erdős-Rényi model $G(n,p)$ with $n = 200$ vertices and connection probabilities $p = 0.03$ (left) and $p = 0.01$ (right).

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

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

We can use the concentration inequalities (namely Chernoff) to prove some interesting properties of random graphs:

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, all vertices of G have degrees between $0.9d$ abd $1.1d$.

Proof. We'll use a combination of concentration and union bound. Let's fix a vertex i on the graph G. The degree of i, denoted d_i, is a sum of $n-1$ independent $\text{Ber}(p)$ random variables. Then by Chernoff inequality (Corollary 2.3.4),

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

The bound above holds for each vertex i. Next, we can unfix i by taking the union bound (Lemma 1.4.1) for all n vertices:

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

If $d >= C \log n$ for sufficiently large C, the probability is bounded by 0.01. This means that with probability 0.99, the complementary event occurs:

$$P(\forall i \leq n : |d_i - d| \leq 0.1d) \geq 0.99$$

and the proof is complete. □

Remark 2.5.2 (Sparse random graphs are far from regular)**.** The condition $d \geq C \log N$ in Proposition 2.5.1 is indeed optimal. If $d < (1-\varepsilon)\ln n$, an isolated vected appears (Exercise 1.10), making the minimum degree zero.

2.6 Subgaussian Distributions

Standard form for Hoeffding Inequality (including subgaussian distributions):

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

A random variable X has a subgaussian distribution if

$$P(|X_i| > t) \leq 2e^{-ct^2} \text{ for all } t \geq 0.$$

There are also other equivalent representations of subgaussian distributions due to their importance, and they all convey the same meaning: The distribution is bounded by a normal distribution.

Proposition 2.6.1 (Subgaussian properties)**.** Let X be a random variable. The following peoperties are equivalent, with the parameters K_i differing by at most an absolute constant factor, i.e. 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.

(a) (Tails) $\exists K_1 > 0$ such that

$$P(|X| > t) \leq 2\exp\left(t^2/K_1^2\right) \text{ for all } t \geq 0.$$

(b) (Moments) $\exists K_2 > 0$ such that

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

(c) (MGF of X^2) $\exists K_3 > 0$ such that

$$\mathbb{E}[\exp\left(X^2/K_3^2\right)] \leq 2.$$

Additionally, if $\mathbb{E}[X] = 0$, then the properties above are equivalent to

(d) (MGF) $\exists K_4 > 0$ such that

$$\mathbb{E}[\exp\left(\lambda X\right)] \leq \exp\left(K_4^2\lambda^2\right) \text{ for all } \lambda \in \mathbb{R}.$$

Proof. The proof is all about transforming one type of information about random variables into another. $(a) \Rightarrow (b)$ Assume (a) holds. Without loss of generality, assume $K_1 = 1$ since it only affects the other constants we obtain by a constant factor, so we can just scale everything. The integrated tail formula (Lemma 1.6.1) for $|X|^p$ gives

$$\begin{aligned}
\mathbb{E}[|X|^p] &= \int_0^\infty P(|X|^p \geq u)\, du \\
&= \int_0^\infty P(|X| \geq t) pt^{p-1}\, dt (\text{ Change of variables } u = t^p) \\
&\leq \int_0^\infty 2e^{-t^2} pt^{p-1}\, dt (\text{ By } (a) \text{ }) \\
&= p\Gamma(p/2) \text{ (Set } t = s \text{ and use Gamma function)} \\
&\leq 3p(p/2)^{p/2}.
\end{aligned}$$

Where the last inequality uses the fact that $\Gamma(x) \leq 3x^x$ for all $x \geq 1/2$: If we let $x = n + t,\ 1/2 \leq t < 1$,

$$\begin{aligned}
\Gamma(x) &= (x-1)\Gamma(n-1+t) \\
&= \cdots \\
&= (x-1)\cdots x(x-(n-1))\Gamma(t) \\
&\leq x \cdot x \cdots x \cdot 3 \\
&= 3x^x.
\end{aligned}$$

Then taking the pth root of the first bound gives (b) with $K_2 \leq 3$.
$(b) \Rightarrow (c)$ Again, WLOG we can assume that $K_2 = 1$ and property (b) holds. By the Taylor series expansion of the exponential function,

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

(b) guarantees that $\mathbb{E}[X^{2p}] \leq (2p)^p$, and $p! \geq (p/e)^p$ by Lemma 1.7.8, hence substituting these bound in, 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 means we get (c) with $K_3 = 2\sqrt{e}$.
$(c) \Rightarrow (a)$ WLOG assume that $K_3 = 1$ and property (c) holds. By exponentiating and using Markov's inequality,

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

This gives (a) with $K_1 = 1$.
Now assume that additionally $\mathbb{E}[X] = 0$.
$(c) \Rightarrow (d)$ Assume WLOG $K_3 = 1$ and property (c) holds. We'll use the following inequality which follows from Taylor's Theorem with Lagrange remainder:

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

Replace the above with $x = \lambda X$ and taking expectations, we get

$$\begin{aligned}
\mathbb{E}[e^{\lambda X}] &\leq 1 + \frac{\lambda^2}{2}\mathbb{E}[X^2 e^{|\lambda X|}] \quad (\mathbb{E}[X] = 0) \\
&\leq 1 + \frac{\lambda^2}{2}e^{\lambda^2/2}\mathbb{E}[e^{X^2}] \quad (x^2 \leq e^{x^2/2} \text{ and } |\lambda x| \leq \lambda^2/2 + x^2/2) \\
&\leq (1+\lambda^2)e^{\lambda^2/2} \quad (\mathbb{E}[e^{X^2}] \leq 2 \text{ by } (c)) \\
&\leq e^{3\lambda^2/2} \quad (1 + x \leq e^x).
\end{aligned}$$

Then we get property (d) with $K_4 = \sqrt{3/2}$.
$(d) \Rightarrow (a)$ WLOG assume $K_4 = 1$ and property (d) holds. By the exponential moment method (Hi again :]), let $\lambda > 0$ to be chosen.

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

Optimizing the above gives $\lambda^* = t/2$, and plugging back in gives

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

By using the exponential moment method again for $-X$,

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

Then by summing up these probabilities,

$$P(|x| \geq t) \leq 2e^{-t^2/4}.$$

Hence property (a) is true with $K_1 = 2$, and the proof is complete. □

Remark 2.6.2 (Zero mean). For property (d) above, $\mathbb{E}[X]$ is a necessary and sufficient condition (Exercise 2.23)!

Remark 2.6.3 (On constant factors)**.** The constant '2' in properties (a) and (c) don't have any special meaning. Any absolute constant greater than 1 works!

2.6.1 The Subgaussian Norm

Definition 2.6.4. A random variable X is called subgaussian if it satisfies any of the equivalent properties in Proposition 2.6.1. Its subgaussian norm is

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

This represents how quickly the tails of X decays compared to a normal distribution.

Example 2.6.5. The following random variables are subgaussian:

(a) Normal,

(b) Rademacher,

(c) Bernoulli,

(d) Binomial,

(e) Any bounded random variable.

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

We can replace the results from 2.6.1 with those having the subgaussian norm:

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

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

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

(c) (MGF of X^2) $\mathbb{E}[\exp(X^2/\|X\|_{\psi_2}^2)] \leq 2$.

(d) (MGF) If additionally $\mathbb{E}[X] = 0$ then $\mathbb{E}[\exp(\lambda X)] \leq \exp(C\lambda^2\|X\|_{\psi_2}^2)$ for all $\lambda \in \mathbb{R}$.

There are a number of other equivalent ways to describe subgaussian random variables (Exercise 2.26-2.28, 2.39). Moreover, there is a sharper way do define the subgaussian norm such that we won't lose any absolute constant factors (Exercise 2.40)!

2.7 Subgaussian Hoeffding and Khintchine Inequalities

From exercise 0.3, we have shown that for independent mean zero random variables,

$$\left\|\sum_{i=1}^{N} X_i\right\|_{L^2}^2 = \sum_{i=1}^{N}\|X_i\|_{L^2}^2.$$

There is a similar weaker property 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

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

where C is an absolute constant.

Proof. We can compute the MGF of the sum $S_N = \sum_{i=1}^{N} X_i$. For any $\lambda \in \mathbb{R}$,

$$\begin{aligned} \mathbb{E}[\exp(\lambda S_N)] &= \prod_{i=1}^{N} \mathbb{E}[\exp(\lambda X_i)] \quad \text{(independence)} \\ &\leq \prod_{i=1}^{N} \exp(C\lambda^2 \|X_i\|_{\psi_2}^{2}) \quad \text{(Proposition 2.6.6 (d))} \\ &= \exp(\lambda^2 K^2), K^2 = C \sum_{i=1}^{N} \|X_i\|_{\psi_2}^{2}. \end{aligned}$$

Then by Proposition 2.6.1, $(d) \Rightarrow (c)$ hence

$$\mathbb{E}[\exp(xS_N^2/K^2)] \leq 2$$

where $c > 0$ is some constant. Then by the definition of the subgaussian norm, $\|S_N\|_{\psi_2} \leq K/\sqrt{c}$, and we are done. □

Remark 2.7.2 (Reverse bound)**.** The inequality in Proposition 2.7.1 can be reversed, but only if X_i are identically distributed (Exercise 2.33, 2.34).

2.7.1 Subgaussian Hoeffding Inequality

Theorem 2.7.3 (Subgaussian Hoeffding Inequality)**.** Let $X_1, \ldots, X_N$ be independent, mean zero, subgaussian random varirables. Then for every $t \geq 0$,

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

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

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

This is exactly the Hoeffding inequality for the Rademacher distribution but with the constant c instead of $1/2$. We can recover the general form of Hoeffding inequality for bounded random variables from this method, again up to an absolute constant (Exercise 2.29).

2.7.2 Subgaussian Khintchine Inequality

Below is 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 with unit variances. Let $a_1, \dots, a_n \in \mathbb{R}$. Then for every $p \in [2, \infty)$, we have

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

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

Proof. For $p = 2$, we have an equality, since the Pythagorean identity with unit variance assumption gives

$$\left\|\sum_{i=1}^N a_i X_i\right\|_{L^2} = \left(\sum_{i=1}^N a_i^2 \|X_i\|_{\psi_2}^2\right)^{1/2} = \left(\sum_{i=1}^N a_i^2\right)^{1/2}$$

□

The lower bound in the theorem follows from the monotonicity of the L^p norms. For the upper bound, we use Proposition 2.7.1 to get

$$\left\|\sum_{i=1}^N a_i X_i\right\|_{\psi_2} \leq C\left(\sum_{i=1}^N a_i^2 \|X_i\|_{\psi_2}^2\right)^{1/2} \leq CK\left(\sum_{i=1}^N a_i^2\right)^{1/2}.$$

We then get the factor of $\sqrt{p}$ in the final result from (b) of Proposition 2.6.6.

2.7.3 Maximum of Subgaussians

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

$$\|\max_{i=1,\dots,N} X_i\|_{\psi_2} \leq C\sqrt{\ln N} \max_{i=1,\dots,N} \|X_i\|_{\psi_2}.$$

In particular,

$$\mathbb{E}[\max_{i=1,\dots,N} X_i] \leq CK\sqrt{\ln N}$$

where $K = \max_i \|X_i\|_{\psi_2}$. The same bounds obviously hold for $\max_i |X_i|$.

Proof. Two proof methods are provided in the book.
Method 1: Union bound. WLOG, we can assume that $\max_i \|X_i\|_{\psi_2} = 1$. This is because we can just scale down all the random variables if needed. For any $t \geq 0$, we have

$$P(\max_{i=1,\dots,N} X_i \geq t) \leq \sum_{i=1}^N P(X_i \geq t) \leq 2N \exp(-ct^2)$$

where the last inequality comes from (a) of Proposition 2.6.6. If $N < \exp(ct^2/2)$, then the probability above is bounded by $2\exp(-ct^2/2)$, which is stronger than needed. If $N > \exp(ct^2/2)$, the probability of any event is bounded by $2\exp(ct^2/3\ln N)$ as by definition this quantity is greater than 1. Then in either case,

$$P(\max_{i=1,\dots,N} X_i \geq t) \leq 2\exp\left(-\frac{ct^2}{3\ln N}\right) \text{ for any } t \geq 0.$$

Then by Proposition 2.6.6 ($(c) \iff (a)$) we get $\|\max_i X_i\|_{\psi_2} \leq C\sqrt{\ln N}$.
Method 2: Maximum with sum. Again, assume that $\max_i \|X_i\|_{\psi_2} = 1$ and denote $Z = \max_{i=1,\dots,N} |X_i|$. Then

$$\mathbb{E}[e^{Z^2}] = \mathbb{E}[\max_{i=1,\dots,N} e^{X_i^2}] \leq \mathbb{E}\left[\sum_{i=1}^N e^{X_i^2}\right] = \sum_{i=1}^N \mathbb{E}[e^{X_i^2}] \leq 2N.$$

Let $M := \sqrt{2 \ln 2N} \geq 1$. Then Jensen's inequality yields

$$\mathbb{E}[e^{Z^2/M^2}] \leq (\mathbb{E}[e^{Z^2}])^{1/M^2} \leq (2N)^{1/2\ln(2N)} = \sqrt{e} < 2.$$

Then $\|Z\|_{\psi_2} \leq M = \sqrt{2\ln(2N)}$, proving the first statement. The second statement follows from the first statement via (b) of Proposition 2.6.6 for $p = 1$. □

Remark 2.7.7 (Gaussian samples have no outliers)**.** The factor $\sqrt{\ln N}$ in Proposition 2.7.6 is unavoidable. In Exercise 2.38, we prove that i.i.d random $N(0,1)$ samples Z_i satisfy

$$\mathbb{E}[\max_{i=1,\ldots,N} |Z_i|] \approx \sqrt{2\ln N}.$$

However, not all hope is lost as logarithmic functions grow slowly. This means for sampling, it helps prevent extreme outliers. On average, the farthest point in an N-point sample from a normal distribution is approximately $\sqrt{2\ln N}$ away from the mean!

2.7.4 Centering

From exercise 0.2, we see that centering reduces the L^2 norm:

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

There is a similar phenomenon for the subgaussian norm:

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

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

Proof. From Exercise 2.42, we know that $\|\cdot\|_{\psi_2}$ is a norm hence the triangle inequality gives

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

We only need to bound the second term. From part (b) of exercise 2.24, for any constant random variable a, $\|a\|_{\psi_2} \lesssim |a|$. Then using $a = \mathbb{E}[X]$ and Jensen's inequality for $f(x) = |x|$, we get

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

where the last step comes from (b) of Proposition 2.6.6 with $p = 1$. Substituting this back into the equation for the triangle inequality and we are done. □

2.8 Subexponential Distributions

Main idea: Subgaussian distributions cover a wide range of distributions already, but leaves out some more heavy-tailed distributions. For tails behaving like exponential distributions, we cannot use conclusions from before like Hoeffding inequality, as the distributions are not subgaussian.

2.8.1 Subexponential Properties

Proposition 2.8.1 (Subexponential properties)**.** Let X be a random variable. The following are equivalent, with $K_i > 0$ differing by at most a constant factor:

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

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

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

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

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

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

Moreover, if $\mathbb{E}[X] = 0$ then properties (i)-(iii) are equivalent to

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

$$\mathbb{E}[\exp(\lambda X)] \leq \exp(K_4^2 \lambda^2) \text{ for all } |\lambda| \leq \frac{1}{K_4}.$$

Proof. The equivalence of (i)-(iii) is done in Exercise 2.41. (iii)⇒(iv) and (iv)⇒(i) are a bit different and will be done here.

(iii)⇒(iv) Assume that (iii) holds, and WLOG assume $K_3 = 1$. We'll use again the inequality coming from Taylor's theorem with Lagrange form remainder:

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

Assume that $|\lambda| \leq 1/2$ and substitute the above with $x = \lambda X$ to get

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

Then property (iv) is true with $K_4 = 2$.

(iv)⇒(i) Assume that (iv) holds, and WLOG assume $K_4 = 1$. Exponentiating, applying Markov inequality, and using (iv) for $\lambda = 1$, we get

$$P(X \geq t) = P(e^X \geq e^t) \leq e^{-t} \mathbb{E}[e^X] \leq e^{1-t}.$$

We also have that

$$P(-X \geq t) = P(e^{-X} \geq e^t) \leq e^{-t} \mathbb{E}[e^{-X}] \leq e^{1-t}.$$

Combining the two equations above vis union bound, we get $P(|X| >= t) <= 2e^{1-t}$. There are now two cases:

Case 1: $t \geq 2$. Then the $2e^{1-t} \leq 2e^{-t/2}$ hence we are done.

Case 2: $t < 2$. Then $2e^{-t/2} \geq 1$ hence the probability is trivially bounded, we are done.

Therefore we get property (i) with $K_1 = 2$. □

Remark 2.8.2 (MGF near the origin)**.** It may be surprising that the bound for subgaussian and subexponential distributions have the same bound on the MGFs near the origin. However, it is expected for any random variable X with mean zero. To see why, assume X is bounded and has unit variance. Then the MGF is approximately

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

as $\lambda \to 0$. For $N(0,1)$, the appxomation becomes an equality. For subgaussian distributions, the above holds for all $\lambda \in \mathbb{R}$, while for subexponential distributions, the above holds only for small λ.

Remark 2.8.3 (MGF far from the origin)**.** For subexponentials, the MGF bound is only guaranteed near zero. For example, the MGF of an Exp(1) random variable is infinite for $\lambda \geq 1$!

2.8.2 The Subexponential Norm

Definition 2.8.4. A random variable X is subexponential if it satisfies any of (i)-(iii) in Proposition 2.8.1. Its subexponential norm is

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

$\|\cdot\|_{\psi_1}$ defines a norm on the space of subexponential random variables (Exercise 2.42). Subgaussian and Subexponential distributions are closely connected:

Lemma 2.8.5. X is subgaussian if and only if X^2 is subexponential, and

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

Lemma 2.8.6. If X and Y are subgaussian then XY is subexponential, and

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

Proof. WLOG, we can assume that $\|X\|_{\psi_2} = \|Y\|_{\psi_2} = 1$. By definition, this implies that $\mathbb{E}[e^{X^2}] \leq 2$ and $\mathbb{E}[e^{Y^2}] \leq 2$. Then

$$\begin{aligned}
\mathbb{E}[\exp(|XY|)] &\leq \mathbb{E}\left[\exp\left(\frac{X^2}{2}\right) + \exp\left(\frac{Y^2}{2}\right)\right] \quad (|ab| \leq \frac{a^2}{2} + \frac{b^2}{2}) \\
&= \mathbb{E}\left[\left(\frac{X^2}{2}\right)\left(\frac{Y^2}{2}\right)\right] \\
&\leq \frac{1}{2}\mathbb{E}[\exp(X^2) + \exp(Y^2)] \\
&\leq \frac{1}{2}(2+2) \\
&= 2.
\end{aligned}$$

By definition, $\|XY\|_{\psi_1} \leq 1$ and we are done. □

Example 2.8.7. The following random variables are subexponential:

(a) Any subgaussian random variable,

(b) The square of any subgaussian random variable,

(c) Exponential,

(d) Poisson,

(e) Geometric,

(f) Chi-squared,

(g) Gamma.

The Cauchy the Pareto distributions are *not* subexponential.

Many properties of subgaussian distributions extend to subexponentials, such as centering (Exercise 2.44):

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

There are a lot of norms that are being discussed, and here is their relationship:

Remark 2.8.8 (All the norms!)**.**

$$\begin{aligned} X \text{ is bounded almost surely } &\implies X \text{ is subgaussian} \\ &\implies X \text{ is subexponential} \\ &\implies X \text{ has moments of all orders} \\ &\implies X \text{ has finite variance} \\ &\implies X \text{ has finite mean.} \end{aligned}$$

Quantitatively,

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

The above holds for any $p \in [2, \infty)$, where the $\lesssim$ sign hides an $O(p)$ factor in one of the inequalities and absolute constant factors in the other two inequalities.

Remark 2.8.9 (More general: ψ_α and Orlics norms)**.** Subgaussian and subexponential distributions are part of a broader family of ψ_α distributions. The general framework is provided by Orlicz spaces and norms (Exercise 2.42, 2.43).

2.9 Bernstein Inequality

Below is a version of Hoeffding inequality that works for subexponential distributions:

Theorem 2.9.1 (Subexponential Bernstein Inequality)**.** Let $X_1, \dots, X_N$ be indepependent, mean zero, subexponential random variables. Then for every $t \ge 0$,

$$P\left(\left|\sum_{i=1}^N X_i\right| \ge t\right) \le 2\exp\left(-c\min\left(\frac{t^2}{\sum_{i=1}^N \|X_i\|_{\psi_1}^2}, \frac{t}{\max_i \|X_i\|_{\psi_1}}\right)\right).$$

where $c > 0$ is an absolute constant.

Proof. By using the exponential moment method,

$$\begin{aligned} P(S_N \ge t) &= P(\exp(\lambda S_N) \ge e^{\lambda t}) \\ &\le e^{-\lambda t}\mathbb{E}[\exp(\lambda S_N)] \\ &= e^{-\lambda t}\prod_{i=1}^N \mathbb{E}[\exp(\lambda X_i)]. \end{aligned}$$

Fix i. To bound the MGF of X_i, by (iv) in Proposition 2.8.1, if λ is small enough, i.e.

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

then $\mathbb{E}[\exp(\lambda X_i)] \le \exp(C\lambda^2\|X_i\|_{\psi_1}^2)$. Substituting this back into the inequality above, we get

$$P(S_N \ge t) \le \exp(-\lambda t + C\lambda^2\sigma^2),\ \sigma^2 = \sum_{i=1}^N \|X_i\|_{\psi_1}^2.$$

When we minimize the expression above in terms of λ subject to the constaint (*), then the optimal chocie that we get is

$$\lambda^* = \min\left(\frac{t}{2C\sigma^2}, \frac{c}{\max_i \|X_i\|_{\psi_1}}\right).$$

Plugging this optimal λ^* back we get

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

Repeating the exponential moment method for $-X_i$ instead of X_i gives the same result, hence also have the same bound for $P(-S_N \geq t)$. Combining the two bounds gives the result. □

Of course, we can apply the argument to $\sum_{i=1}^N a_i X_i$ as well:

Corollary 2.9.2 (Simpler subexponential Bernstein inequality)**.** Let $X_1, \ldots, X_N$ be independent, mean zero, subexponential random variables, and $a_i \in \mathbb{R}$. Then for every $t \geq 0$, we have that

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

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

Remark 2.9.3 (Why two tails?)**.** Unlike Hoeffding inequality (Theorem 2.7.3), Bernstein inequality has two tails - gaussian and exponential. The gaussian tail comes from what we would expect from the CLT. The exponential tail is also there because there can be one term X_i having a heavy exponential tail, which is strictly heavier than a gaussian tail. The cool thing is that Bernstein inequality says that if you have some number of random variables with exponential tails, only the one with the largest subexponential norm matters!

Remark 2.9.4 (Small and large deviations)**.** Normalizing the sum in Corollary 2.9.2 like in the CLT, we get

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

In the small deviations range we have a gaussian tail bound. This range grows at the rate of $\sqrt{N}$, reflecting the increasing strength of the CLT. For the large deviations range, we have an exponential tail bound driven by a single term X_i, shown in the figure below:

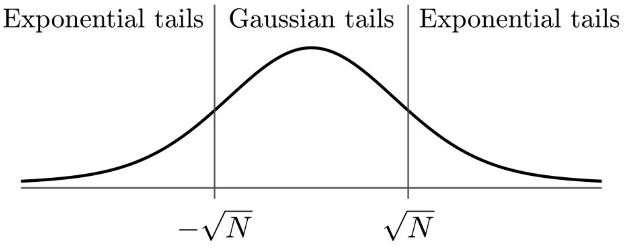


Figure 2.3 Bernstein inequality exhibits a mixture of two tails: gaussian for small deviations and exponential for large deviations.

There is also a version of Bernstein inequality that uses the variances of the terms X_i. However, we need a stronger assumption that the terms X_i are bounded almost surely:

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

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

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

Proof. Exercise 2.47. □

3 Random Vectors in High Dimensions

This chapter mainly deals with the curse of dimensionality, and how vectors interact in these high-dimensional settings.

3.1 Concentration of the Norm

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

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

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

Proof. Using Proposition 2.6.6, we can rewrite the above as

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

We can prove the bound using Bernstein inequality. If we consider the quantity

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

the above is a sum of independent, mean zero random variables. Moreover, since XX_i are subgaussian, $X_i^2 - 1$ are subexponential. Then by the centering lemma (Lemma 2.7.8), we have that

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

Applying Bernstein inequality ($N = n$ and $a_i = 1/n$), we get that for any $u \geq 0$,

$$\begin{aligned} P\left(\left|\frac{1}{n}\|X\|_2^2 - 1\right| \geq u\right) &\leq 2\exp\left[-c_1 \min\left(\frac{u^2 n}{K^4}, \frac{un}{K^2}\right)\right] \\ &\leq 2\exp\left[-\frac{cn}{K^4}\min(u^2, u)\right]. \end{aligned}$$

where in the last step, we used the fact that K is bounded below by an absolute constant, since

$$1 = \|X_1\|_{L^2} \leq C\|X_1\|_{\psi_2} \leq CK \text{ by Proposition 2.6.6.}$$

We'll now use the concentration inequality for $\|X\|_2^2$ to deduce one for $\|X\|_2$. We'll use the following propery for all $z, \delta \geq 0$:

$$|z - 1| \geq \delta \implies |z^2 - 1| \geq \max(\delta, \delta^2).$$

This is because since $z \geq 0$, $|z + 1| = z + 1 \geq 1$ and $|z + 1| \geq |z - 1|$. Therefore

$$\begin{aligned} |z^2 - 1| &= |z - 1||z + 1| \\ &\geq |z - 1| \max(|z - 1|, 1) \\ &\geq \max(\delta, \delta^2). \end{aligned}$$

Then for any $\delta \geq 0$,

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

Changing variables with $t = \delta\sqrt{n}$ gives the subgaussian tail. □

Remark 3.1.2 (Thin shell phenomenon)**.** The theorem above 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, but here's an intuitive explanation:
The square of the norm, $\|X\|_2^2$, has a chi-squared distribution with n degrees of freedom. Hence its mean is n, and standard deviation $\sqrt{2n}$. Thus it makes sense for $\|X\|_2$ to deviate by $O(1)$ around $\sqrt{n}$ because

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

3.2 Covariance Matrices and PCA

The covariance matrix of a random vector X taking values in $\mathbb{R}^n$ is

$$\mathrm{Cov}(X) = \mathbb{E}[(X-\mu)(X-\mu)^T] = \mathbb{E}[XX^T] - \mu\mu^T,\ \mu = \mathbb{E}[X].$$

The second moment matrix of X is

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

By translation, the covariance and the second moment matrices are the same, hence many problems can first be reduced into the mean zero case.

3.2.1 Learning from the Covariance Matrix

The covariance matrix can tell us much more than just the covariance of X's coordinates:

Proposition 3.2.1. Let X be a random vector in $\mathbb{R}^n$ with second moment matrix $\Sigma = \mathbb{E}[XX^T]$. Then

(a) (1D marginals) For any fixed vector $v \in \mathbb{R}^n$,

$$\mathbb{E}[\langle X, v\rangle^2] = v^T \Sigma v.$$

(b) (Norm) $\mathbb{E}[\|X\|_2^2] = \mathrm{tr}(\Sigma)$.

(c) If Y is an independent copy of X, then

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

Proof. (a) Using the linearity of expectation,

$$\mathbb{E}[\langle X, v\rangle^2] = \mathbb{E}[v^T XX^T v] = v^T \mathbb{E}[XX^T] v = v^T \Sigma v.$$

(b) The diagonal entries of the second moment matrix are $\Sigma_{ii} = \mathbb{E}[X_{ii}^2]$. Then

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

(c) Since the trace of a matrix is a linear operator, it can be swapped with the expectation:

$$\begin{aligned}
\mathbb{E}[\langle X, v\rangle^2] &= \mathbb{E}[X^T YY^T X] \\
&= \mathbb{E}[\mathrm{tr}(X^T YY^T X)] \\
&= \mathbb{E}[\mathrm{tr}(YY^T XX^T)] \\
&= \mathrm{tr}(\mathbb{E}[X^T XY^T Y]) \\
&= \mathrm{tr}(\mathbb{E}[X^T X]\mathbb{E}[Y^T Y]) \\
&= \mathrm{tr}(\Sigma^2) \\
&= \|\Sigma\|_F^2.
\end{aligned}$$

□

3.2.2 Principle Component Analysis

Since the covariance matrix Σ is symmetric, it has a spectral decomposition:

$$\Sigma = \sum_{i=1}^{n} \lambda_i v_i v_i^T.$$

Here λ_i are the real eigenvalues, and v_i are the corresponding random vectors. There is a nice interpretation for eigenvalues from an optimization perspective:

Proposition 3.2.2. Let Σ be an $n \times n$ symmetric matrix with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n$ and corresponding unit eigenvectors $v_1, \ldots, v_n$. Then for every $k = 1, \ldots, n$, we have

$$\lambda_k = \max_{v \perp \{v_1, \ldots, v_{k-1}\}, \|v\|_2 = 1} v^T \Sigma v.$$

Proof. Consider any unit vector $v \in \mathbb{R}^n$ that is orthogonal to $\{v_1, \ldots, v_{k-1}\}$. Using the spectral decomposition, we get

$$\begin{aligned} v^T \Sigma v &= v^T \left(\sum_{i=1}^{n} \lambda_i v_i v_i^T \right) \\ &= \sum_{i=1}^{n} \lambda_i (v^T v_i)(v_i^T v) \\ &= \sum_{i=k}^{n} \lambda_i \langle v, v_i \rangle^2 \quad \text{(Orthogonality)} \\ &\leq \lambda_k \sum_{i=k}^{n} \langle v, v_i \rangle^2 \\ &\leq \lambda_k. \end{aligned}$$

We also have that $v_k^T \Sigma v_k = v_k^T(\lambda_k v_k) = \lambda_k$, which reaches the minimal value, hence the proof is complete. □

Therefore we have the following corollary:

Corollary 3.2.3 (PCA)**.** Let X be a random vector in $\mathbb{R}^n$ whose covariance matrix has eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n \geq 0$ and eigenvectors $v_1, \ldots, v_n$. Then

$$\lambda_k = \max_{v \perp \{v_1, \ldots, v_{k-1}, \|v\|_2 = 1} \mathrm{Var}(\langle X, v \rangle).$$

The maximum is attained at v_k.

For a random vector $X \in \mathbb{R}^n$ representing data, the top eigenvector of the covariance matrix gives the first *principle component*, indicating the direction with has the largest spread, with λ_1 as the variance in that direction.

Remark 3.2.4 (Dimensionality reduction)**.** It often happens with real data that only a few eigenvalues are large and informative, while the rest are small and treated as noise. Therefore even if the data comes in high-dimensionsal, it is basically low-dimensional hence you just have to project onto the lower dimensional subspace to perform PCA.

3.2.3 Isotropic Distributions

Definition 3.2.5. A random vector X in $\mathbb{R}^n$ is called isotropic if

$$\mathbb{E}[XX^T] = I_n$$

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

Proposition 3.2.1 implies that X is isotropic if and only if

$$\mathbb{E}[\langle X, v\rangle^2] = \|v\|_2^2 \text{ for any fixed vector } v \in \mathbb{R}^n.$$

The above implies that isotropic distributions spread equally in all directions, because the RHS of the equation does not depend on the direction of v.

Note (Standardizing)**.** In one dimension, a random variable X can be standardized to a zero mean, unit variance random variable Z by doing

$$Z = \frac{X - \mu}{\sqrt{\operatorname{Var}(X)}} \implies X = \mu + \operatorname{Var}(X)^{1/2} Z.$$

This is also true in higher dimensions:

$$Z = \operatorname{Cov}(X)^{-1/2}(X - \mu) \implies X = \mu + \operatorname{Cov}(X)^{1/2} Z.$$

Moreover, the idea still holds even if the covariance matrix is not invertible (Exercise 3.10)!

3.3 Examples of High-dimensional Distributions

3.3.1 Standard Normal

A random vector Z has the standard normal distribution in $\mathbb{R}^n$ if its coordinates are independent standard normal variables. Its density is

$$f_Z(z) = \frac{1}{(2\pi)^{n/2}} e^{-\|z\|_2^2/2}, z \in \mathbb{R}^n.$$

The standard normal distribution is isotropic. Moreover, it is *rotation-invariant*:

Proposition 3.3.1 (Rotation invariance)**.** Consider a random vector $Z \sim N(0, I_n)$ and a fixed orthogonal matrix U. Then

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

In particular, by looking at the first coordinate of UZ, we get

$$(UZ)_1 = \langle U_1, Z\rangle \, (0, 1)$$

where U_1 is the first row of U. Since this is an arbitrary unit vector, all 1D marginals of the multivariate standard normal distribution are $N(0, 1)$. More generally:

Corollary 3.3.2 (1D marginals of the standard normal distribution)**.** Consider $Z \sim N(0, I_n)$ and any fixed $v \in \mathbb{R}^n$. Then

$$\langle Z, v\rangle \sim N(0, \|v\|_2^2).$$

From the above, we get

Corollary 3.3.3 (Sum of independent normals is normal)**.** Consider independent normal random variables $X_i \sim N(\mu_i, \sigma_i^2)$. Then,

$$\sum_{i=1}^n X_i \sim N(\mu, \sigma^2),\ \mu = \sum_{i=1}^n \mu_i, \sigma^2 = \sum_{i=1}^n \sigma_i^2.$$

Proof. We can write $X_i = \mu_i + \sigma_i Z_i$, where Z_i are independent standard normal random variables. Then

$$\sum_{i=1}^n X_i = \mu + \sum_{i=1}^n \sigma_i Z_i = \mu + \langle Z, v \rangle \text{ where } v = (\sigma_1, \dots, \sigma_n).$$

Then by Corollary 3.3.3, $\langle Z, v \rangle \sim N(0, \sigma^2)$ hence

$$\mu + \langle Z, v \rangle \sim N(\mu, \sigma^2).$$

□

3.3.2 General Normal

Definition 3.3.4. A random vector X in $\mathbb{R}^n$ is normally distribute if it can be obtained via an affine transformation of a standard normal random vector $Z \sim I(0, I_k)$, i.e.

$$X = \mu + AZ,\ \mu \in \mathbb{R}^n,\ A \in \mathbb{R}^{n \times k}.$$

Here X has mean μ and covariance matrix $\Sigma = AA^T$.

Proposition 3.3.5 (Uniqueness of normal)**.** The distribution of X is uniquely determined by μ and Σ. Specifically, X has the same distribution as

$$Y = \mu + \Sigma^{1/2} Z',\ \Sigma = AA^T,\ Z' \sim N(0, I_n).$$

Proof. We'll use a version of the *Cramer-Wold device*, which says that the distributions of all 1D marginals uniquely determine the distribution in $\mathbb{R}^n$. This means if X, Y are random vectors in $\mathbb{R}^n$ and $\langle X, u \rangle$ and $\langle Y, u \rangle$ have the same distribution for all $u \in \mathbb{R}^n$, then X and Y have the same distribution.
We check that AZ and $\Sigma^{1/2} Z'$ have the same distribution:

$$\langle AZ, v \rangle = \left\langle Z, A^T v \right\rangle \sim N(0, \|A^T v\|_2^2),\ \text{and } \left\langle \Sigma^{1/2} Z', v \right\rangle \sim N(0, \|\Sigma^{1/2} v\|_2^2).$$

From the above, $\|A^T v\|_2^2 = \|\Sigma^{1/2} v\|_2^2$ since $\Sigma = AA^T$. Therefore the proof is complete. □

If Σ is invertible, the density has the formula below:

Proposition 3.3.6. If Σ is invertible, the PDF of a multivariate normal distribution is

$$f(x) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right), x \in \mathbb{R}^n.$$

Proof. Exercise 3.15. □

A special property for normal distributions is that independence and uncorrelation are equivalent, which it not true generally:

Corollary 3.3.7 (Jointly normal random variables)**.** Random variables $X_1, \dots, X_n$ are jointly normal if the random vector $X = (X_1, \dots, X_n)$ is normally distributed. Jointly normal random variables are independent if and only if they are uncorrelated.

Proof. If X_i are uncorrelated, Σ is diagonal. Then the density function can be factored into marginals, i.e.

$$f(x) = f_1(x) \times \cdots \times f_n(x) \text{ for all } x \in \mathbb{R}^n.$$

The joint density of random variables X_i factors if and only if X_i are independent, hence we're done. □

3.3.3 Uniform on the Sphere

Proposition 3.3.8 (A sphere is isotropic)**.** The uniform distribution on S^{n-1} with radius $\sqrt{n}$ is isotropic.

Proof. Let $X \sim \text{Unif}(S^{n-1})$. By symmetry, for distinct i, j, (X_i, X_j) has the same distribution as $(-X_i, X_j)$. Therefore

$$\mathbb{E}[X_i X_j] = -\mathbb{E}[X_i X_j] \implies \mathbb{E}[X_i X_j] = 0.$$

Moreover, since $\|X\|_2 = 1$,

$$1 = \mathbb{E}[\|X\|_2^2] = \mathbb{E}[X_1^2] + \cdots + \mathbb{E}[X_n^2].$$

The X_i are identically distributed, hence $\mathbb{E}[X_i^2] = 1/n$, hence the coordinates of $\sqrt{n}X$ are uncorrelated with second moment equal to 1, hence $\sqrt{n}X$ is isotropic. □

Note (Isotropic Vectors are almost Orthogonal)**.** In the high-dimensional world, pick two random points, and they most likely will be orthogonal!
Consider $X, Y \sim \text{Unif}(S^{n-1})$. Then $\sqrt{n}X, \sqrt{n}Y$ are i.i.d. and isotropic by Proposition 3.3.8. By (c) from Proposition 3.2.1,

$$\mathbb{E}[\langle \sqrt{n}X, \sqrt{n}Y \rangle^2] = \text{tr}(I_n) = n.$$

Fividing the above by n^2 we obtain

$$\mathbb{E}[\langle X, Y \rangle^2] = \frac{1}{n}.$$

Then applying Markov's inequality, we get

$$|\langle X, Y \rangle| = O(1/\sqrt{n}) \text{ with high probability.}$$

Note (Gaussian and spherical distributions are similar)**.** Both $N(0, I_n)$ and $\text{Unif}(S^{n-1})$ are isotropic and rotation-invariant.

$$g \sim N(0, I_n) \implies \frac{g}{\|g\|_2} \sim \text{Unif}(S^{n-1}).$$

Informally, we can say that

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

This defies the low-dimensional intuition. This is because there is almost no volume near the origin in high dimensions.

To say this in rigorous terms:

Theorem 3.3.9 (Projective CLT)**.** Let $X \sim \text{Unif}(S^{n-1})$. Then

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

In fact, the CDF converges uniformly:

$$\sup_{v \in S^{n-1}} \sup_{t \in \mathbb{R}} |P(\sqrt{n}\,\langle X, v \rangle \le t) - P(g_1 \le t)| \to 0$$

where $g_1 \sim N(0,1)$.

Proof. We can assume $X = g/\|g\|_2$ with $g \sim N(0, I_n)$ from above. By rotation invariance, the distribution of $\langle X, v\rangle$ is the same for all $v \in \mathbb{R}^n$. Therefore we can choose $v = e_1$ and get

$$\langle X, e_1\rangle = \frac{g_1}{\|g\|_2}.$$

We'll decompose into a "good event" and a "bad event" that has probability decaying to zero. By the gaussian decay tail in Theorem 3.1.1,

$$E_n := \{|\|g\|_2 - \sqrt{n}| \leq \ln n\} \text{ is likely: } p_n := P(E_n^c) \to 0.$$

If E_n occurs and $t \geq 0$ (which we can assume because of symmetry), then the event of interest $\sqrt{n}\,\langle X, e_1\rangle \leq t$ implies

$$g_1 \leq \frac{t\|g\|_2}{\sqrt{n}} \leq t\left(1 + \frac{\ln n}{\sqrt{n}}\right) =: t_n.$$

Splitting the event based on whether E_n occurs, we get

$$\begin{aligned} P(\sqrt{n}\,\langle X, v\rangle \leq t) &\leq P(\sqrt{n}\,\langle X, v\rangle \leq t \text{ and } E_n) + P(E_n^c) \\ &\leq P(g_1 \leq t_n) + p_n. \end{aligned}$$

Hence

$$P(\sqrt{n}\,\langle X, v\rangle \leq t) - P(g_1 \leq t) \leq P(g_1 \in [t, t_n]) + p_n.$$

The density of g_1 on $[t, t_n]$ is bounded by $e^{-t^2/2}$, so

$$P(g_1 \in [t, t_n]) + p_n \leq e^{-t^2/2}(t_n - t) + p_n = e^{-t^2/2} t \frac{\ln n}{\sqrt{n}} + p_n \leq \frac{C \ln n}{\sqrt{n}} + p_n.$$

The RHS does not depend on v or t, and goes to zero as $n \to \infty$.
We can also show that $P(g_1 \leq t) - P(\sqrt{n}\,\langle X, v\rangle \leq t)$ also goes to zero. Combining the two bounds completes the proof. □

Remark 3.3.10 (Density of 1D marginals of the sphere)**.** The density of the 1D marginals of the uniform distribution on the sphre of radius $\sqrt{n}$ can be computed. It is in fact proportional to $(1 - x^2/n)^{\frac{n-3}{2}}$ (Exercise 3.27). For large n, this approximates $e^{-x^2/2}$, which is exactly the Gaussian limit.

3.3.4 Uniform on a Convex Set

Let $K \subset \mathbb{R}^n$ be a convex set. A random variable X is uniformly distributed in K, denoted $X \sim \text{Unif}(K)$, if its density is $1/\text{Vol}(K)$ on K and zero everywhere else.
The mean of X is

$$\mu = \mathbb{E}[X] = \frac{1}{\text{Vol}(K)} \int_K dx,$$

which is the center of gravity of K. If Σ is the covaraince matrix of K, then the standard score $Z := \Sigma^{-1/2}(X - \mu)$ is an isotropic random vector from Definition 3.2.5. In fact, Z is uniformly distributed in the affinely transformed copy of K:

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

Therefore there is an affine transformation T which makes $T(K)$ isotropic. In convex geometry, we can consider $T(K)$ as a well-conditioned version of K, which makes algorithms like finding the volume work better.

3.3.5 Frames

A frame 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 $u_1, \ldots, u_N$, 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) $I_n = \sum_{i=1}^N u_i u_i^T$.

(iv) (Isotropy) The ranodm vector $X \sim \text{Unif}\{\sqrt{N}u_1, \ldots, \sqrt{N}u_N\}$ is isotropic.

A set of vectors satisfying these equivalent properties is called a Parseval frame.

Proof. (i) ⇒ (iv) The identity for (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 is isotropic.
(iv) ⇒ (iii) Since X is isotropic,

$$I_n = \mathbb{E}[XX^T] = \frac{1}{N}\sum_{i=1}^N \left(\sqrt{N}u_i\right)\left(\sqrt{N}u_i\right)^T = \sum_{i=1}^N u_i u_i^T.$$

(iii) ⇒ (ii) Multiply both sides by the vector x gives the result.
(iii) ⇒ (ii) Taking the inner product with the vector x gives the result. □

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

$$X \sim \text{Unif}\{\sqrt{n}e_1, \ldots, \sqrt{n}e_n\}$$

is isotropic. Among all high-dimensional distributions, Gaussian is often the best to work with and the coordinate distribution is the worst.

Example 3.3.13 (Mercedes-Benz frame)**.** An example of a Parseval frame that is not linearly independent is the set of N equispaced points on the circle of radius $\sqrt{2/N}$, shown below:

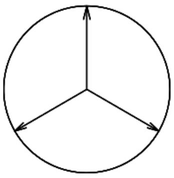


Figure 3.7 A Mercedez-Benz frame: three equispaced points on the circle of radius $\sqrt{2/3}$ form a Parseval frame in $\mathbb{R}^2$.

Here are two more examples of isotropic distributions:

Example 3.3.14 (Uniform on the discrete cube)**.** Let X be a Rademacher random vector, that is,

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

Then X is isotropic.

Example 3.3.15 (Product distributions)**.** 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.

3.4 Subgaussian Distributions in High Dimensions

Definition 3.4.1. 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 the marginals over all unit vectors:

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

Below are some examples :)

3.4.1 Gaussian, Rademacher, and More

Lemma 3.4.2 (Distributions with independent subgaussian coordinates)**.** Let $X = (X_1, \ldots, X_n)$ be a random vector in $\mathbb{R}^n$ with independent, mean zero, subgassian 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 $v = (v_1, \ldots, v_n) \in S^{n-1}$. Then

$$\begin{aligned}
\|\langle X, v\rangle\|_{\psi_2}^2 &= \|\sum_{i=1}^n v_i X_i\|_{\psi_2}^2 \\
&\le C \sum_{i=1}^n \|v_i X_i\|_{\psi_2}^2 \quad \text{By Proposition 2.7.1} \\
&= C \sum_{i=1}^n v_i^2 \|X_i\|_{\psi_2}^2 \\
&\le C \max_{i \le n} \|X_i\|_{\psi_2}^2.
\end{aligned}$$

Since v is arbitrary, the proof is complete. □

Example 3.4.3 (Rademacher)**.** We can immediately get from the above that a Rademacher normal random vector 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)**.** We can also get from the above that if $X \sim N(0, I_n)$, then X is subgaussian. Moreover, $Y \sim N(0, \Sigma)$ is also subgaussian (Exercise 3.38).

3.4.2 Uniform on the Sphere

The projective CLT (Theorem 3.3.9) tells us that the uniform distribution on $\sqrt{n}S^{n-1}$ has approximately Gaussian 1D marginals. In fact, these marginals ar subgaussian:

Theorem 3.4.5 (Uniform distribution on the sphere is subgaussian)**.** Let $X \sim \text{Unif}(S^{n-1})$. Then for any $v \in S^{n-1}$ and $t \geq 0$, we have

$$P(\langle X, v\rangle \geq t) \leq 2\exp\left(-\frac{t^2 n}{2}\right).$$

In particular, X is subgaussian, and $\|X\|_{\psi_2} \leq C/\sqrt{n}$.

Proof. By rotational invariance, we can assume

$$X = \frac{g}{\|g\|_2} \text{ where } g \sim N(0, I_n).$$

Again, the distribution of $\langle X, v\rangle$ does not depend on v hence we can choose $v = e_1$ to get $\langle X, v\rangle = X_1$. This the inequality $\langle X, v\rangle \geq t$ becomes $g_1 \geq t\|g\|_2$. By squaring both sides, moving g_1^2 to the LHS and simplifying, we get

$$g_1 \geq s\|\bar{g}\|_2, \quad s = \frac{t}{\sqrt{1-t^2}} \text{ and } \bar{g} = (g_2, , g_n).$$

To find the probability of the event above, we fix $\|\bar{g}\|_2$ by conditioning on $\bar{g}$, which does not alter the distribution of g since g and $\bar{g}$ are independent. Then we uncondition by taking the expectation over $\bar{g}$. By the tower property,

$$P(\langle X, v\rangle \geq t) = P(g_1 \geq s\|\bar{g}\|_2) = \mathbb{E}[P(g_1 \geq s\|\bar{g}\|_2) \mid \bar{g}] \quad (*).$$

After conditioning, the conditional probability above reduces to a gaussian tail. By exercise 2.6, we get that

$$\mathbb{E}[P(g_1 \geq s\|\bar{g}\|_2)|\bar{g}] \leq \mathbb{E}[\exp\left(-\frac{s^2\|g\|_2^2}{2}\right)] = \left[\mathbb{E}[\exp\left(-\frac{s^2 g_1^2}{2}\right)]\right]^{n-1}.$$

where the last equality comes from the fact that g_i are i.i.d. $N(0,1)$ random variables, and

$$\|\bar{g}\|_2^2 = g_2^2 + \cdots + g_n^2.$$

For the expression above,

$$\begin{aligned}
\mathbb{E}[\exp(-s^2 g_1^2/2)] &= \int_{-\infty}^{\infty} \exp(-s^2x^2/2) \cdot \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(\sqrt{1+s^2}x)^2}{2}\right) dx \\
&= \frac{1}{\sqrt{1+s^2}} \int_{-\infty}^{\infty} e^{-v^2/2}\, dv \quad (v = \sqrt{1+s^2}x) \\
&= \frac{1}{\sqrt{1+s^2}}.
\end{aligned}$$

Thus the expression above becomes

$$\left(\frac{1}{1+s^2}\right)^{\frac{n-1}{2}} = (1-t^2)^{\frac{n-1}{2}} \leq \exp\left(-\frac{t^2(n-1)}{2}\right)$$

since $1 - x \leq e^{-x}$ for all $x \in \mathbb{R}$.
For the expression (*), the probability is zero for $t \geq 1$ since $\langle X, v\rangle \leq \|X\|_2\|v\|_2 = 1$, while for $t \leq 1$,

$$\exp(-t^2(n-1)/2) \leq e^{1/2}\exp(-t^2 n/2) \leq 2\exp(-t^2 n/2)$$

and we are done. □

3.4.3 Non-examples

Some distributions in $\mathbb{R}^n$ are subgaussian, but their subgaussian norm is huge, therefore it is impractical to work with them. Below are a few examples.

Example 3.4.6 (Uniform on a convex body)**.** Let $K \subset \mathbb{R}^n$ be convex, and $X \sim \mathrm{Unif}(K)$ be isotropic. Qualitatively, X is subgaussian since K is bounded. But quantitatively what is it like? Is it bounded by some constant C?
This is true for some isotropic convex bodies like the unit cube $[-1,1]^n$ (Lemma 3.4.2) and the Euclidean ball of radius $\sqrt{n+2}$ (Exercise 3.25 & 3.42). However, for other convex bodies like the ball in the ell^1 norm, the subgaussian norm can grow with n (Exercise 3.44).
Even so, a weaker result holds: X has subexponential marginals, and

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

for all unit vectors v, which comes from C. Borell's lemma, which follows from the Brunn-Minkowski inequality.

Example 3.4.7 (Coordinate distribution)**.** Let $X \sim \mathrm{Unif}\{\sqrt{n}e_1, \ldots, \sqrt{n}e_n\}$. X is subgaussian as it takes on finitely many values. However, from Exercise 3.43,

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

Therefore it is not 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. However, they must take exponentially many values (Exercise 3.46). In particular, this prevents frames (Proposition 3.3.11) as good subgaussian distributions as they take way too many values and are mostly useless in practice.

3.5 Application: Grothendieck Inequality and Semidefinite Programming

In this section, we will used high-dimensional Gaussians to tackle problems that are seemingly not related to probability at all. We first present the Grothendieck inequality.

Theorem 3.5.1 (Grothendieck inequality)**.** Consider $a \in \mathbb{R}^{m\times n}$. Assume that

$$\left|\sum_{i,j} a_{ij}x_iy_j\right| \leq 1 \text{ for any numbers } x_i, y_j \in \{-1,1\}.$$

Then for any Hilbert space H, we have

$$\left|\sum_{i,j} a_{ij}\left\langle u_i, v_j\right\rangle\right| \leq K \text{ for any unit vectors } u_i, v_j \in H.$$

Here $K \leq 1.783$ is an absolute constant.

There is nothing random in the statement above, but we'll approach it using probabilistic reasoning. In fact, there will be two proofs for Grothendieck inequality, one with a much worse bound of $K \leq 14.1$ in this section, and the other one with $K \leq 1.783$ in section 3.7. Before going into the first argument, there is a simple observation that we state here.

Remark 3.5.2 (Homogeneous form of Grothendieck inequality)**.** The assumption of Grothendieck inequality can be equivalently stated as

$$\left|\sum_{i,j} a_{ij}x_iy_j\right| \leq \max_i |x_i| \cdot \max_j |y_j|$$

for any real numbers x_i and y_j (Exercise 3.47). The conclusion of Grothendieck inequality can be equivalently stated as

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

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

Proof of Theorem 3.5.1 with worse bound. (**Step 1: Reductions**) Note that Grothendieck inequality becomes trivial if we allow the value of K to depende on the matrix $A = (a_{ij})$. For example, $K = \sum_{i,j} |a_{ij}|$ would work! Let $A(K)$ be the smallest number that makes the conclusion in Remark 3.5.2 holds 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 both the matrix A and the dimensions m and n.

WLOG, we may show this for a specific Hilbert space H, namely for $\mathbb{R}^N$ equipped with the Euclidean norm $\|\cdot\|_2$. This is because we can 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 fact that all N-dimensional Hilbers 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 bases of $\mathbb{R}^N$.

By the definition of $K = K(A)$, there exist vectors $u_i, v_j \in \mathbb{R}^N$ satisfying

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

(**Step 2: Introducing randomness**) The key idea of the proof is to express the vectors u_i, v_j using Gaussian random variables

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

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 (Corollary 3.3.2 and Exercise 3.9):

$$\mathbb{E}\left[U_i V_j\right] = \langle u_i, v_j \rangle.$$

Thus

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

Suppose for a moment that the random variables $|U_i|$ and $|V_j|$ were to be almost surely bounded by some constant, say R. Then from the assumption in Remark 3.5.2,

$$\left|\sum_{i,j} a_{ij} U_i V_j\right| \leq R^2 \text{ almost surely.}$$

Plugging this into the equation above will give $K \leq R^2$, completing the proof.

(**Step 3: Truncation**) The above is flawed, because the Gaussian random variables U_i, V_j 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 $R \geq 1$ and split the random variables like this:

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

We similarly decompose $V_j = V_j^- + V_j^+$. Nor 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: by Exercise 2.4 (b), a Gaussian tail bound gives

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

A similar bound holds for V_j^+.

(**Step 4: Breaking up the sum**) Replacing U_iV_j with $(U_i^- + U_i^+)(V_j^- + V_j^+)$ in $(*)$ and expanding the sum, we get

$$K = \underbrace{\mathbb{E}\left[\sum_{i,j} a_{ij} U_i^- V_j^-\right]}_{S_-} + \underbrace{\mathbb{E}\left[\sum_{i,j} a_{ij} U_i^+ V_j^-\right]}_{S_\pm} + \underbrace{\mathbb{E}\left[\sum_{i,j} a_{ij} U_i^- V_j^+\right]}_{S_\mp} + \underbrace{\mathbb{E}\left[\sum_{i,j} a_{ij} U_i^+ V_j^+\right]}_{S_+}.$$

Let's bound each term! S_- is the easiest to bound: by construction, $|U_i|$ and $|V_j|$ are bounded by R, so from step 2 we can directly get that

$$S_- \leq R^2.$$

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 the inner product $\langle X, Y\rangle_{L^2} = \mathbb{E}[XY]$. Thus write

$$S_\pm = \sum_{i,j} a_{ij} \left\langle U_i^+, V_j^- \right\rangle_{L^2}.$$

We have $\|U_i^+\|_{L^2} < 2/R$ by $(**)$, and $\|V_j^-\|_{L^2} \leq \|V_j\|_{L^2} = 1$ by construction. Then, applying the conclusion from Remark 3.5.2 for the Hilbert space $H = L^2$, we find that

$$S_\pm = K \cdot \frac{2}{R}.$$

(It might seem odd that we are using the inequality that we are trying to prove. However, we picked $K = K(A)$ at that start to be the smallest value to make Grothendieck inequality work. That is the K that we are using here).
The last two terms, $S_\mp$ and S_+, can be bounded just like the above (Check).
(**Step 5: Putting everything together**) Plugging the bounde on all four terms, we conclude that

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

Setting $R = 12$ and rearranging the terms gives $K \leq 288$. A litte finer analysis, skipping the rough $4/R^2$ bound in $(**)$ yields $K \leq 14.1$ (Exercise 3.48). □

Remark 3.5.3 (Quadratic Grothendieck)**.** We can often relax Grothendieck inequality by taking $x_i = y_i$, bounding a quadratic instead of a bilinear form. The statement becomes: Let $A \in \mathbb{R}^{n\times n}$ be symmetric PSD or diagonal-free. Assume that

$$\left|\sum_{i,j} a_{ij} x_i x_j\right| \leq 1 \text{ for any numbers } x_i \in \{-1, 1\}.$$

Then for any Hilbert space H, we have

$$\left|\sum_{i,j} a_{ij} \langle u_i, v_j\rangle\right| \leq 2K \text{ for any unit vectors} u_i, v_j \in H.$$

Here K is the the absolute constant from Grothendieck inequality.

Proof. Exercises 3.49 & 3.50. □

3.5.1 Semidefinite Programming

Some hard computational problems can be relazed into easier, more computationally tractable programs via semidefinite programming, and Grothendieck inequality can help guarantee its quality.

Definition 3.5.4. A semidefinite program (SDP) is an optimization problem of the following type:

$$\text{maximize } \langle A, X\rangle : \; X \succeq 0, \; \langle B_i, X\rangle \le b_i \text{ for } i = 1, \dots, N.$$

Here A, B are given $n \times n$ matrices, and b_i are given numbers. The variable X is an $n \times n$ symmetric PSD matrix, indicated by the notation $X \succeq 0$. The inner product is the standard one on the space of $n \times n$ matrices:

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

Note that if we *minimize* instead of maximize, we still get a semidefinite program. Same goes for replacing any signs "$\le$" by "$\ge$" or "$=$".

Remark 3.5.5 (An SDP program is a convex program)**.** Every SDP is a convex program because it involves maximizing a *linear* function $\langle A, X\rangle$ over a convex set of matrices (the set of PSD matrices is convex, 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 *algorithmically tractable*, i.e. there are efficient solvers for general convex programs, and specifically for SDPs.

Semidefinite Relaxations

SDPs can provide efficient relaxations of computationally hard problems, such as

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

where A is a given $n \times n$ matrix. This is a *quadratic integer optimization problem*, whose feasible set consists of 2^n vectors $x \in \{-1, 1\}^n$. Finding the maximum via brute force takes exponential time. Moreover, there is probably not a smarter way because it is a computationally hard problem (NP-hard). However, we can relax the problem into a SDP program than approximates the maximum within a constant factor. To do this, we replace the numbers $x_i = \pm 1$ by random variables X_i in $\mathbb{R}^n$. We get

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

Proposition 3.5.6 (The relaxation is an SDP)**.** The optimization problem above is equivalent to the following SDP:

$$\text{maximize } \langle A, Z\rangle : \; Z \succeq 0, Z_{ii} = 1 \text{ for } i = 1, \dots, n.$$

Proof. Recall that the Gram matrix of vectors $X_1, \dots, X_n$ 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 PSD, and (b) conversely, any symmetric PSD matrix is a Gram matric of some set of vectors (Exercise 3.51). □

The guarantee of relaxation

Let's show that the probabilistic SDP approximates the exact SDP within a constant factor:

Theorem 3.5.7. Let $A \in \mathbb{R}^{n \times n}$ be symmetric PSD. Let $\int(A)$ denote the maximum in the integer optimization problem, and $\operatorname{sdp}(A)$ denote the maximum in the probabilistic SDP. Then

$$\int(A) \le \operatorname{sdp}(A) \le 2K \cdot \int(A)$$

where $K \le 1.783$ is the constant in Grothendieck inequality.

Proof. The first bound follows with $X_i = (x_i, 0, \dots, 0)^T$. The second comes directly from the quadratic Grothendieck inequality in Remark 3.5.3. □

Although Theorem 3.5.7 helps us approximate the maximum value in the integer SDP, it is not obvious how to find the actual solution that attain this maximum value. Can we convert the vectors X_i that give a solution to the probabilistic SDP into labels $x_i = \pm 1$ that approximately solve the integer SDP? The answer is yes! But we need some knowledge of max-cuts first. After reading, we can create an even better approximation constant than $2K$ (Exercise 3.58).

3.6 Application: Maximum Cut for Graphs

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 - no loops or multiple edges. For convenience, label the vertices by integers, setting $V = \{1, \dots, n\}$.

Definition 3.6.1. If we partition th evertices of a graph G into two disjoint subsets, the cut is the number of edges between them. Then maximum cut of G, denoted $\text{maxcut}(G)$, is the largest possible cut over all partitions of vertices. The figure below is an illustration:

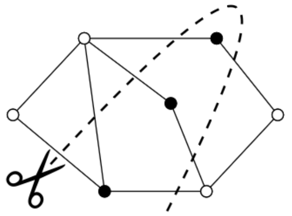


Figure 3.8 The dashed line shows the maximal cut of this graph, splitting the vertices into black and white and giving $\text{maxcut}(G) = 7$.

Finding the maximum cut is generally a computationally hard problem (NP-hard).

3.6.1 A Simple 0.5-approximation Algorithm

We can relax the maximum cut problem into a SDP, but we have to translate the problem into linear algebra first.

Definition 3.6.2. 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 x_i belongs to. For example, in Figure 3.8 from above, the three black vertices might have $x_i = 1$ and the four white vertices $x_i = -1$. The cut for G for this partition is simply the number of edges between vertices with opposite labels:

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

The maximum cut is found by maximizing $\text{cut}(G, x)$ over all partitions x:

$$\text{maxcut}(G) = \frac{1}{4} \max \left\{ \sum_{i,j=1}^{n} A_{ij}(1 - x_i x_j) : \ x_i = \pm 1 \forall i \right\}.$$

Let's start with a simple 0.5-approximation algorithm for the maximum cut - one that finds a vut 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 $0.5\text{maxcut}(G)$.

Proof. A random cut is generated by a Rademacher random vector x. Then, in the formula for $\text{cut}(G,x)$ we hvae $\mathbb{E}[x_i x_j] = 0$ for $i \neq j$ and $A_{ij} = 0$ for $i = j$ since the graph has no loops. Thus, by linearity of expectation,

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

□

3.6.2 Semidefinite Relaxation

We can get a 0.878-approximation algorithm due to Goemans and Williamson. It is based on a semidefinite relaxation of the NP-hard problem. We consider the SDP

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

We'll show that the $\text{sdp}(G)$ approximates $\text{maxcut}(G)$ 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 below). More formally, consider a standard normal random vector $g \sim N(0, I_n)$ and define

$$x_i := \text{sign}\langle X_i, g\rangle, \ i = 1, \ldots, n.$$

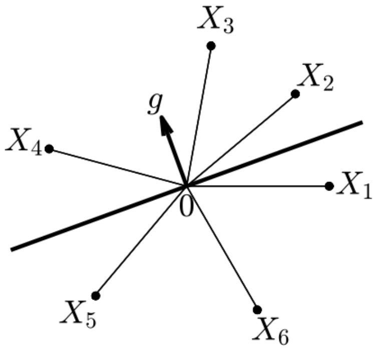


Figure 3.9 We do randomized rounding of these vectors $X_i \in \mathbb{R}^n$ into labels $x_i = \pm 1$ by choosing a random hyperplane with normal vector g (shown in bold) and assigning $x_2 = x_3 = x_4 = 1$ and $x_1 = x_5 = x_6 = -1$.

Theorem 3.6.4. Let G be a graph with adjacency matrix A. Let (X_i) be a solution of the SDP, and $x = (x_i)$ be the result of a randomized rounding of (X_i). Then

$$\mathbb{E}[\text{cut}(G,x)] \geq 0.878\text{sdp}(G) \geq 0.878\text{maxcut}(G).$$

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:

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

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

Proof. Exercise 3.53. □

A downside of the Grothendieck inequality is the nonlinear function arcsin, which is hard to work with. We can replace it with a linear bound using the inequality

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

which can be checked easily with software (shown below).

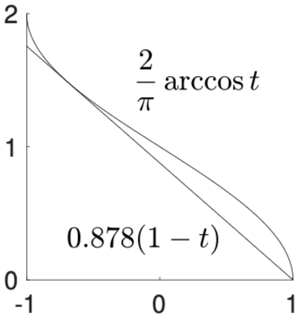


Figure 3.10 The inequality $\frac{2}{\pi}\arccos t \geq 0.878(1-t)$ holds for all $t \in [-1, 1]$.

Proof of Theorem 3.6.4. By the formula for $\operatorname{cut}(G, x)$ and linearity of expectation, we have

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

The definiton of the labels x_I in the rounding step gives

$$\begin{aligned} 1 - \mathbb{E}\left[x_i x_j\right] &= 1 - \mathbb{E}\left[\operatorname{sign}\langle X_i, g\rangle \operatorname{sign}\langle X_j, g\rangle\right] \\ &= 1 - \frac{2}{\pi}\arcsin\langle X_i, X_j\rangle \quad \text{(Lemma 3.6.5)} \\ &\geq 0.878(1 - \langle X_i, X_j\rangle). \end{aligned}$$

Therefore

$$\mathbb{E}\left[\operatorname{cut}(G, x)\right] \geq 0.878 \cdot \frac{1}{4}\sum_{i,j=1}^{n} A_{ij}(1 - \langle X_i, X_j\rangle) = 0.878\mathrm{SDP}(G).$$

This proves the first inequality. The second inequality is trivial since $\operatorname{sdp}(G) \geq \operatorname{maxcut}(G)$. □

3.7 Kernel Trick and Tightening of Grothendieck Inequality

This section takes a different approach to the proof of Grothendieck inequality for (almost) the best known bound: $K \leq 1.783$.

We'll again use Grothendieck identity (Lemma 3.6.5), but the nonlinearity of the arcsin function is a big challenge. If there were no nonlinearity, we would have

$$\mathbb{E}\left[\operatorname{sign}\langle g, u\rangle \operatorname{sign}\langle g, v\rangle\right] = \frac{2}{\pi}\langle u, v\rangle,$$

of which Grothendieck inequality would easily follow:

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

This would give $K \leq \pi/2 \approx 1.57$.
The above is obviously wrong because of nonlinearity. To handle the nonlinear function, of an inner product $\langle u, v\rangle$, we can use a remarkably powerful trickL rewrite it as a (linear) inner product $\langle u', v'\rangle$ for some other vectors u', v' in some Hilbert space H. In the literature of machine learning, this is known as the *kernel trick*.
We will explicitly construct the nonlinear transformations $u' = \Phi(u)$, $v' = \Psi(v)$ that will do the job. The best way to describe them is to use tensors, which generalize matrices to higher dimensions.

3.7.1 Tensors

Definition 3.7.1. An order k tensor $(a_{i_1\ldots i_k})$ is an $n_1 \times n_2 \times \cdots \times n_k$ array of real numbers $a_{i_1\ldots i_k}$. The canonical inner product on $\mathbb{R}^{n_1\times\cdots\times n_k}$ defines the inner product of tensors: For A, B tensors (of the same dimensions),

$$\langle A, B\rangle := \sum_{i_1,\ldots,i_k} a_{i_1\ldots i_k} b_{i_1\ldots i_k}.$$

Example 3.7.2 (Vectors and matrices)**.** Vectors are order 1 tensors, and matrices are order 2 tensors. The inner product for tensors generalized the inner product for vectors and matrices.

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

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

Lemma 3.7.4 (Powers)**.** For any vectors $u, v \in \mathbb{R}^n$ and $k \in \mathbb{N}$, we have

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

Proof. For $n = 3$:

$$\begin{aligned}\left\langle u^{\otimes 3}, v^{\otimes 3}\right\rangle &= \sum_{i,j,k=1}^n (u_i u_j u_k)(v_i v_j v_k)\\ &= \left(\sum_{i=1}^n u_i v_i\right)\left(\sum_{i=1}^n u_i v_i\right)\left(\sum_{i=1}^n u_i v_i\right)\\ &= \langle u, v\rangle^3.\end{aligned}$$

The general case is similar to the above. □

Lemma 3.7.4 reveals something interesting: non-linear expressions like $\langle u, v\rangle^k$ can be written as a stardard *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

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

In fact we can take $H = \mathbb{R}^{n^k}$, the space of k-th order tensors and $\Phi(u) = u^{\otimes k}$.
Now, we can move to general nonlinearities:

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

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

We can take

$$\Phi(u) = (\sqrt{2} u \otimes u) \oplus (\sqrt{5} u \otimes u \otimes u)$$

where $\oplus$ denotes concatenation. So, the target space is $\mathbb{R}^{n^2+n^3}$.

Example 3.7.6 (General polynomials)**.** Polynomials with negative coefficients can make our task impossible since $\langle \phi(u), \Phi(v) \rangle$ is always nonnegative. But here is a neat workaround: we can find *two* transformations, possibly different, such that

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

In this case, we can take

$$\Phi(u) = (\sqrt{u} \otimes u) \oplus (\sqrt{5} u \otimes u \otimes u), \Psi(v) = (\sqrt{2} v \otimes v) \oplus (-\sqrt{5} v \otimes v \otimes v).$$

Note that the transformations keep the lengths of vectors under control. For any unit vector u,

$$\|\Phi(u)\|_2^2 = \|\Psi(u)\|_2^2 = 2 \langle u, u \rangle^2 + 5 \langle u, u \rangle^3 = 2 + 5 = 7,$$

which is just the sum of the absolute values of the coefficients.

Following this approach, we can handle any polynomial $f(x) = \sum_{k=1}^{N} a_k x^k$. Moreover, by taking limits on polynomials, we can handle even more functions:

Lemma 3.7.7 (Real analytic functions)**.** Consider a function $f(x) = \sum_{k=0}^{\infty} a_k x^k$ 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

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

Moreover, for any unit vector u, we have

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

Proof. Exercise 3.55. □

Example 3.7.8 (Sine function)**.** 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

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

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're going to prove Grothendieck inequality (Theorem 3.5.1) with constant

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

We can assume WLOG that $u_i, v_j \in \mathbb{R}^N$ with $N = n+m$, just like in the 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 satisfying

$$\langle u_i', v_j' \rangle = \sin(c \langle u_i, v_j \rangle) \text{ for all } i, j.$$

Again, we can assume WLOG that $H = R^N$. Applying Grothendieck identity (Lemma 3.6.5), we get

$$\mathbb{E}\left[\operatorname{sign}\langle g, u_i'\rangle \operatorname{sign}\langle g, v_j'\rangle\right] = \frac{2}{\pi}\arcsin\langle u_i', v_j'\rangle = \frac{2c}{\pi}\langle u_i, v_j\rangle.$$

Thys

$$\frac{2c}{\pi}\sum_{i,j} a_{ij}\langle u_i, v_j\rangle = \mathbb{E}\left[\sum_{i,j} a_{ij} \underbrace{\operatorname{sign}\langle g, u_i'\rangle}_{X_i} \underbrace{\operatorname{sign}\langle g, v_j'\rangle}_{Y_i}\right] \leq 1.$$

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 and unit vectors u_i, v_j and finds labels $x_i, y_j \in \{-1, 1\}$ satisfying

$$\mathbb{E}\left[\sum_{i,j} a_{ij} x_i y_j\right] \geq \frac{1}{K}\sum_{i,j} a_{ij}\langle u_i, v_j\rangle.$$

Here is how it works: First, find unit vectors $u_i', v_j' \in \mathbb{R}^{n+m}$ with prescribed inner products. Then use randomized rounding: pick $g \sim N(0, I_n)$ and set $x_i = \operatorname{sign}\langle gmu_i'\rangle$ 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 (very possibly nonlinear) 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 $\Phi : \mathcal{X} \to H$ such that

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

The answer is given by Mercer's Theorem, and more precisely, the 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 kernel matrix $(K(u_i, u_j))_{i,j=1}^N$ has to be symmetric PSD. The transformation Φ is called a *feature map*, and the Hilbert space H is called a *Reproducing Kernel Hilbert Space* (RKHS). Popular PSD kernels in machine learning include the Gaussian and polynomial kernels, given by:

$$K(u, v) = \exp\left(-\frac{\|u - v\|_2^2}{2\sigma^2}\right), \; K(u, v) = (\langle u, v\rangle + r)^k, \; u, v \in \mathbb{R}^n,$$

where $\sigma > 0, r > 0, k \in \mathbb{N}$ are hyperparameters. The kernel trick, which expresses a kernel $K(u, v)$ as an inner product, is widely used in machine learning because it lets us handle nonlinear models (determined by K) with techniques designed for linear models (e.g. Kernel Ridge Regression). The exact details of the Hilbert space H and feature map Φ are usually not needed. Moreover, to compute the inner product $\langle \Phi(u), \Phi(v)\rangle$ in H, we don't even need to know Φ - the identity above $(*)$ lets us calculate $K(u, v)$ directly!

4 Random Matrices

This chapter mostly focuses on the theory regarding random matrices - nets, covering and packing numbers. Applications include community detection, covariance estimation, and spectral clustering.

4.1 A Quick Refresher on Linear Algebra

4.1.1 Singular Value Decomposition

Theorem 4.1.1 (SVD)**.** Any $m \times n$ matrix A with real entries can be written as

$$A = \sum_{i=1}^{r} \sigma_i u_i v_i^T \text{ where } r = \min(m, n).$$

Here $\sigma_i > 0$ are 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. WLOG, we can assume that $m \geq n$ or else we can just take the transpose. Since $A^T A \in \mathbb{R}^{n\times n}$ is a symmetric positive semidefinite matrix, the spectral theorem tells us that its eigenvalues are $\sigma_1^2, \dots, \sigma_n^2$ and corresponding orthonormal eigenvectors $v_1, \dots, v_n \in \mathbb{R}^n$, so that $A^T A v_i = \sigma_i^2 v_i$. The vectors Av_i are orthogonal:

$$\langle Av_i, Av_j \rangle = \langle A^T A v_i, v_j \rangle = \sigma_i^2 \langle v_i, v_j \rangle = \sigma_i^2 \delta ij.$$

Therefore, there exist orthonormal vectors $u_1, \dots, u_n \in \mathbb{R}^n$ such that

$$Av_i = \sigma_i u_i, \quad i = 1, \dots, n.$$

For the above, for all i with $\sigma_i \neq 0$, the vectors u_i are uniquely defined and ensures that they are orthonormal. If $\sigma_i = 0$, then $Av_i = 0$ holds triviall. In this case, we can pick any u_i while keeping orthonormality.

Since $v_1, \dots, v_n$ form an orthonormal basis of $\mathbb{R}^n$, we can write $I_n = \sum_{i=1}^n v_i v_i^T$. Multiplying by A on the left and plugging the equation above gives

$$A = \sum_{i=1}^{n} (Av_i) v_i^T = \sum_{i=1}^{n} \sigma_i u_i v_i^T.$$

□

Remark 4.1.2 (Geometric interpretation)**.** SVD gives a geometric view of matrices: it stretches the orthogonal direction of v_i by σ_i, then rotates the space, mapping the orthonormal basis v_i to u_i.

Remark 4.1.3 (SVD matrix form)**.** We can set $\sigma_i = 0$ for $i > r$ and arrange them in weakly decreasing order. Then by extending $\{u_i\}$ and $\{v_i\}$ to orthonormal bases in $\mathbb{R}^m$ and $\mathbb{R}^n$, we get

$$A = U \Sigma V^T$$

where U is the $m \times m$ 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 the singular values σ_i on the diagonal. If A is symmetric, we get the spectral decomposition instead:

$$A = U \Lambda U^T.$$

Remark 4.1.4 (Spectral decomposition v.s. SVD)**.** The spectral and singular value decompositions are tightly connected. Since

$$AA^T = \sum_{i=1}^{r} \sigma_i^2 u_i u_i^T \text{ and } A^T A = \sum_{i=1}^{r} \sigma_i^2 v_i v_i^T$$

the left singular vectors u_i of A are the eigenvectors of AA^T, while the right singular vectors v_i of A are the eigenvectors of $A^T A$, and the singular values σ_i of A are

$$\sigma_i(A) = \sqrt{\lambda_i(AA^T)} = \sqrt{\lambda_i(A^T A)}.$$

Remark 4.1.5 (Orthogonal projection)**.** 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

$$P = \sum_{i=1}^{k} u_i u_i^T = UU^T$$

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

There is another optimization-based description of eigenvalues:

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

$$\lambda_k(A) = \max_{\dim E = k} \min_{x \in S(E)} x^T Ax = \min_{\dim E = n-k+1} \max_{x \in S(E)} x^T Ax,$$

where the first max/min is taking with respect to all subspaces of a fixed dimension, 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 λ_k, we need to find a k-dimensional subspace E such that

$$x^T Ax \geq \lambda_k \text{ for all } x \in S(E).$$

To find the set E, take the spectral decomposition $A = \sum_{i=1}^{n} \lambda_i u_i u_i^T$ and pick the subspace $E = \text{span}(u_1, \ldots, u_k)$. The eigenvectors form an orthonormal basis of E, so any vector $x \in S(E)$ can be written as $x = \sum_{i=1}^{k} a_i u_i$. Then by orthonormality of u_i and monotonicity of λ_i, we get

$$x^T Ax = \sum_{i=1}^{k} \lambda_i a_i^2 \leq \lambda_k \sum_{i=1}^{k} a_i^2 = \lambda_k$$

and we have the upper bound. For the lower bound on λ_k, we need to find $x \in S(E)$ such that $x^T Ax \leq \lambda_k$. Here we let the subspace be $F = \text{span}(u_k, \ldots, u_n)$.
Since $\dim E + \dim F = n + 1$, the intersection of E and F is nontrivial hence there is a unit vector $x \in E \cap F$. Writing $x = \sum_{i=k}^{n} a_i u_i$, we get

$$x^T Ax = \sum_{i=k}^{n} \lambda_i a_i^2 \geq \lambda_k \sum_{i=k}^{n} a_i^2 = \lambda_k.$$

Then we get the lower bound, and hence the first equality is done.
The second equality is by applying the same technique to $-A$ and reversing the eigenvalues. □

Applying Section 4.1.2 to $A^T A$ and using Remark 4.1.4, we get

Corollary 4.1.7 (Min-max theorem for singular values)**.** Let $A \in \mathbb{R}^{m\times n}$ with singular values $\sigma_1 \geq \cdots \geq \sigma_n \geq 0$. Then

$$\sigma_k(A) = \max_{\dim E = k} \min_{x\in S(E)} \|Ax\|_2 = \min_{\dim E = n-k+1} \max_{x\in S(E)} \|Ax\|_2$$

with the same notation as Section 4.1.2.

4.1.3 Frobenius and Operator Norms

Definition 4.1.8. For a matrix $A \in \mathbb{R}^{m\times n}$, the Frobenius norm is

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

The operator norm of A is the smallest number K such that

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

Equivalently,

$$\|A\| = \max_{x\neq 0} \frac{\|Ax\|_2}{\|x\|_2} = \max_{\|x\|_2\leq 1} \|Ax\|_2 = \max_{\|x\|_2=1} \|Ax\|_2 = \max_{\|x\|_2=\|y\|_2=1} |y^T Ax|.$$

From the Frobenius norm, we can get that

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

Also, from above we can get

$$\|A\|_F^2 = \langle A, A\rangle = \operatorname{tr}(A^T A).$$

For the operator norm, the first three equations follows by rescaling, and the last one comes from the duality formula:

$$\|Ax\| = \max_{\|y\|_2=1} \langle Ax, y\rangle.$$

Here the absolute sign does not matter.

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 operator norms (Exercise 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 A and orthogonal matrices Q, R with proper dimensions, we have

$$\|QAR\|_F = \|A\|_F \text{ and } \|QAR\| = \|A\|.$$

Proof. For the Frobenius norm, by one of the formulas above,

$$\begin{aligned}
\|QAR\|_F &= \operatorname{tr}(R^T ATQ^T QAR) \\
&= \operatorname{tr}(R^T A^T AR) \\
&= \operatorname{tr}(RR^T A^T A) \\
&= \operatorname{tr}(A^T A) \\
&= \|A\|_F^2.
\end{aligned}$$

For the spectral norm, by an equivalent characterization, $\|QAR\|$ is obtained by maximizing the bilinear form $y^T QARx = (Qy)^T A(Rx)$ over all unit vectors x, y. Since Q, R are orthogonal, Qy and Rx also range over all unit vectors, so we just get $\|A\|$ as a result. □

Lemma 4.1.11 (Matrix norms via singular values)**.** For any $A \in \mathbb{R}^{m \times n}$ with singular values $\sigma_1 \geq \cdots \geq \sigma_n$,

$$\|A\|_F = \left(\sum_{i=1}^{n} \sigma_i^2 \right)^{1/2} \quad \text{and} \quad \|A\| = \sigma_1.$$

Proof. For the Frobenius norm, by orthogonal invariance (Lemma 4.1.10),

$$\|A\|_F = \|U \Sigma V^T\|_F = \|\Sigma\|_F$$

which directly gives us the result.
The result for the operator norm directly follows from Corollary 4.1.7 with $k = 1$. □

Remark 4.1.12 (Symmetric matrices)**.** For a symmetric matrix A with eigenvalues λ_k,

$$\|A\| = \max_k |\lambda_k| = \max_{\|x\|=1} |x^T Ax|.$$

The first equality becomes Lemma 4.1.11 since the singular values of A are $|\lambda_k|$. The min-max theorem (Section 4.1.2) gives $|\lambda_k| \leq \max_{\|x\|=1} |x^T Ax|$, proving the upper bound in the equation above. The lower bound can be proven by taking $x - y$ in the definition of the operator norm (Definition 4.1.8).

4.1.5 Low-rank Approximation

For a given matrix A, what is the closest approximation to it for a given matrix of rank k? The answer is just truncating the SVD of A:

Theorem 4.1.13 (Eckart-Young-Mirski theorem)**.** Let $A = \sum_{i=1}^{n} \sigma_i u_i v_i^T$. Then for any $1 \leq k \leq n$,

$$\min_{\text{rank}(B)=k} \|A - B\| = \sigma_{k+1}.$$

The minimum is attained at $B = \sum_{i=1}^{k} \sigma_i u_i v_i^T$.

Proof. If $B \in \mathbb{R}^{m \times n}$ has rank k, $\dim \ker(B) = n - k$. Then the min-max theorem (Corollary 4.1.7) for $k + 1$ instead of k gives

$$\|A - b\| \geq \max_{x \in S(E)} \|(A - B)x\|_2 = \max_{x \in S(E)} \|Ax\|_2 \geq \sigma_{k+1}.$$

In the opposite direction, setting $B = \sum_{i=1}^{k} \sigma_i u_i v_i^T$ gives $A - b = \sum_{i=k+1}^{n} \sigma_i u_I v_i^T$. The maximal singular value of this matrix σ_{k+1}, which is the same as its operator norm by Lemma 4.1.11. □

4.1.6 Perturbation Theory

We can also study how eigenvalues/eigenvectors change under matrix perturbations:

Lemma 4.1.14 (Weyl inequality)**.** The k-th largest eigenvalue of symmetric matrices A, B satisfy

$$|\lambda_k(A) - \lambda_k(B)| \leq \|A - B\|.$$

Similarly, the k-th largest singular values of general rectangular matrices satisfy

$$|\sigma_k(A) - \sigma_k(B)| \leq \|A - B\|.$$

A similar result holds for eigenvectors, however we have to track the same eigenvector before and after the perturbation. If the eigenvalues are too close, a small perturbation can swap them, leading to huge error since their eigenvectors are orthogonal and far apart.

Theorem 4.1.15 (Davis-Kahan inequality)**.** Consider two symmetric matrices A, B with spectral decompositions

$$A = \sum_{i=1}^{n} \lambda_i u_i u_i^T, \; B = \sum_{i=1}^{n} \mu_i v_i v_i^T,$$

where the eigenvalues are weakly decreasing. Assume the the k-th largest eigenvalue of A is δ-seperated from the rest:

$$\min_{i \neq k} |\lambda_k - \lambda_i| = \delta > 0.$$

Then the angle between the eigenvectors u_k and v_k satisfies

$$\sin \angle u_k, v_k \leq \frac{2\|A - B\|}{\delta}.$$

The theorem above can be derived via a stronger result of Davis-Kahan focusing on spectral projections - the orthogonal projections onto the span of some subset of eigenvectors:

Lemma 4.1.16 (Davis-Kahan inequality for spectral projections)**.** Consider A, B as in Theorem 4.1.15. Let I, J be two δ-seperated subsets of $\mathbb{R}$, with I being an interval. Then the spectral projections

$$P = \sum_{i:\lambda_i \in I} u_i u_i^T \text{ and } Q = \sum_{j:\lambda_j \in J} v_j v_j^T \text{ satisfy } \|QP\| \leq \frac{\|A - B\|}{\delta}.$$

Proof. WLOG, assume I is finite and closed. Adding the same multiple of Identity to A and B, we can center I as $[-r, r]$, so that $|\lambda_i| \leq r$ for $i \in I$ and $|\mu_j| \geq r + \delta$ for $\mu_j \in J$. The idea is to see how P and Q interact through $H := B - A$:

$$\|H\| \geq \|QHP\| = \|QBP - QAP\| \geq \|QBP\| - \|QAP\|.$$

The spectral projection A commutes with B, hence

$$\|QBP\| \geq \|BQP\| \geq (r + \delta)\|QP\|.$$

To see the last inequality, the image of Q is spanned by orthogonal vectors v_j with $|\mu_j| \geq r + \delta$. The matrix B maps each such vector v_j to $\mu_j v_j$, hence scaling it by at least $r + \delta$. Thus B expands the norm of any vector in the image of Q by at least $r + \delta$ so

$$\|BQPx\|_2 \geq (r + \delta)\|QPx\|_2 \text{ for any } x.$$

Taking the supremum over all unit vectors gives the result with the operator norm.
Also, $AP = PAP = \sum_{i:\lambda_i \in I} \lambda_i u_i u_i^T$ so

$$\|QAP\| = \|QPAP\| \leq \|QP\| \cdot \|AP\| \leq r\|AP\|,$$

because $\|AP\| = \max_{i:\lambda_i \in I} |\lambda_i| \leq r$. Putting the two bounds together we get

$$\|H\| = \|B - A\| \geq \delta\|QP\|,$$

which completes the proof. □

Proof for Theorem 4.1.15. Since the LHS is a trig angle, we can assume that $\varepsilon := \|A - B\| \leq \delta/2$ or else the inequality holds trivially. By Weyl inequality (Lemma 4.1.14), $|\lambda_j - \mu_j| \leq \varepsilon$ for each j hence

$$\min_{j:j \neq k} |\lambda_k - \mu_k| \geq \min_{j:j \neq k} |\lambda_k - \lambda_j| - \varepsilon = \delta - \varepsilon \geq \delta/2.$$

Apply Lemma 4.1.16 for the $\delta/2$-seperated subsets $I = \{\lambda_k\}$ and $J = \{\mu_j : j \neq k\}$ to get $\|QP\| \leq 2\varepsilon/\delta$. Since P and $I_n - Q$ are the orthogonal projections on the directions of u_k and v_k respectively,

$$\|QP\| = \max_{\|x\|=1} \|QPx\|_2 = \|Qu_k\|_2 = \sin \angle(u_k, v_k).$$

Combining this with the inequality on $\|QP\|$ above completes the proof. □

4.1.7 Isometries

The singular values of a matrix A satisfy (by the min-max theorem)

$$\sigma_n \|x-y\|_2 \le \|Ax-Ay\|_2 \le \sigma_1 \|x-y\|_2.$$

The extreme singular values set the limits on how the linear map A distorts space.
A matris is an isometry if

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

Notice that A need not be a square matrix. T
For $A \in \mathbb{R}^{m\times n}$ with $m \ge n$, the following are equivalent:

(a) The columns of A are orthonormal, i.e. $A^T A = I_n$,

(b) A is an isometry,

(c) All singular values of A are 1.

There is a stronger result where the properties hold approximately instead of exactly (useful when dealing with random matrices):

Lemma 4.1.17 (Approximate isometries)**.** Let $A \in \mathbb{R}^{m\times n}$ with $m \ge n$ and let $\varepsilon \ge 0$. The following are equivalent:

(a) $\|A^T A - I_n\| \le \varepsilon$.

(b) $(1-\varepsilon)\|x\|_2^2 \le \|Ax\|_2^2 \le (1+\varepsilon)\|x\|_2^2$ for any $x \in \mathbb{R}^n$.

(c) $1-\varepsilon \le \sigma_n^2 \le \sigma_1^2 \le 1+\varepsilon$.

Proof. (a) $\Leftrightarrow$ (b) By rescaling, we can assume that $\|x\|_2 = 1$ in (b). Then we have

$$\|A^T A - I_n\| = \max_{\|x\|_2=1} |x^T(A^T A - I_n)x| = \max_{\|x\|_2=1} |\|Ax\|_2^2 - 1|,$$

The above being bounded by ε is equivalent to (b) for all unit vectors x.
(b) $\Leftrightarrow$ (c) follows from the relationship for singular values distorting space from above. □

Remark 4.1.18. Here is a more handy version of (a) $\Rightarrow$ (c) in Lemma 4.1.17. For $z \in \mathbb{R}$ and $\delta \ge 0$,

$$|z^2 - 1| \le \max(\delta, \delta^2) \implies |z-1| \le \delta.$$

Then substituting $\varepsilon = \max(\delta, \delta^2)$, we get

$$\|A^T A - I_n\| \le \max(\delta, \delta^2) \implies 1-\delta \le \sigma_n \le \sigma_1 \le 1+\delta.$$

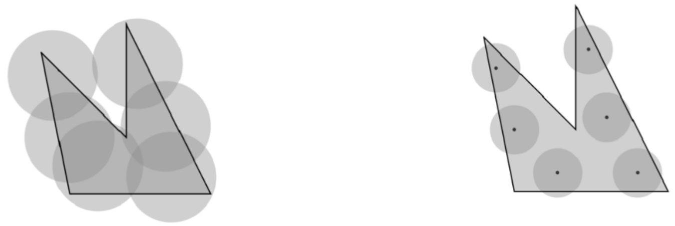
4.2 Nets, Covering, and Packing

The ε-net argument is useful for analysis of random matrices. It is also connected to ideas like covering, packing, entropy, volume, and coding.

Definition 4.2.1. Let (T,d) be a metric space. Consider $K \subset T$ and $\varepsilon > 0$. A subset $\mathcal{N} \subset T$ is called an ε-net of K is every point in K is within distance ε of some point in $\mathcal{N}$, i.e.

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

Equivalently, $\mathcal{N}$ is an ε-net of K if the balls of radius ε centered at points in $\mathcal{N}$ cover K, like in the figure below:



(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

Definition 4.2.2. The smallest cardinality of an ε-net of K is called the covering number of K, and is denoted $\mathcal{N}(K,d,\varepsilon)$.

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

$$N(K,d,\varepsilon) < \infty \text{ for every } \varepsilon > 0.$$

We can think about the covering numbers as a quantitative measure of how compact K is.

Definition 4.2.4. A subset $\mathcal{N}$ of a metric space (T,d) is ε-seperated if

$$d(x,y) > \varepsilon \text{ for any distinct points } x,y \in \mathcal{N}.$$

The largest possible cardinality of an ε-seperated subset of a given $K \subset T$ 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-seperated, the closed $\varepsilon/2$-balls centered at points in $\mathcal{N}$ are disjoint by the triangle inequality, hence we can always pack into K at least $\mathcal{P}(K,d,\varepsilon)$ disjoint $\varepsilon/2$-balls.

Lemma 4.2.6 (Nets from seperated sets)**.** Let $\mathcal{N}$ be a maximal ε-seperated subset of K, i.e. adding any new point to $\mathcal{N}$ destroys the seperation property. 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 $d(x,x_0) \leq \varepsilon$. If $x \in \mathcal{N}$, the conclusion is trivial by choosing $x_0 = x$. Suppose $x \notin \mathcal{N}$. The maximality assumption implies that $\mathcal{N} \cup \{x\}$ is not ε-seperated, meaning $d(x,x_0) \leq \varepsilon$ for some $x_0 \in \mathcal{N}$. □

Remark 4.2.7 (Constructing a net)**.** The lemma above (Lemma 4.2.6) gives an iterative algorithm to construct an ε-net for a given set K. Pick $x_1 \in K$ 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 from both x_1 and x_2, and so on. If K is compact, then the process will stop in a finite number of iterations!

Lemma 4.2.8 (Equivalence of covering and packing numbers)**.** For any set $K \subset T$ and $\varepsilon > 0$,

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

Proof. The upper bound follows from Lemma 4.2.6 because the packing number is exactly the number that makes $\mathcal{N}$ a maximal ε-seperated set.
For the lower bound, take any 2ε-seperated 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-seperated points, each ε-ball centered at y_j can contain at most one x_i. The pigeonhole principle gives $|\mathcal{P}| \leq |\mathcal{N}|$. Since $\mathcal{P}$ and $\mathcal{N}$ are arbitrary, the bound follows. □

4.2.1 Covering Numbers and Volume

This sections is about covers with $T = \mathbb{R}^n$ with the Eudlidean metric

$$d(x, y) = \|x - y\|_2.$$

Therefore, we can omit the metric when denoting the covering and packing numbers:

$$\mathcal{N}(K, \varepsilon) = \mathcal{N}(K, d, \varepsilon).$$

How do the covering numbers relate to the most classical measure, the volume of K in $\mathbb{R}^n$?

Definition 4.2.9 (Minkowski sum)**.** Let $A, B \subseteq \mathbb{R}^n$. The Minkowski sum is defined as

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

Below is an example of the Minkowski sum of two sets on the plane:

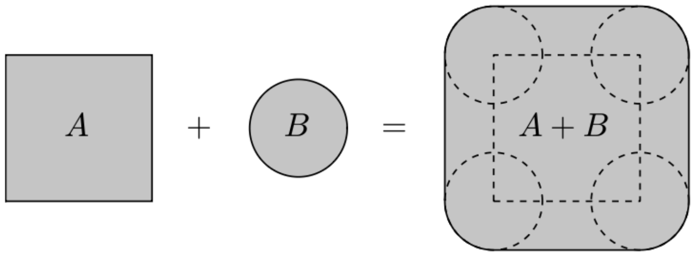


Figure 4.2 Minkowski sum of a square and a circle is a square with rounded corners.

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

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

where B_2^n denotes the unit ball in $\mathbb{R}^n$.

Proof. The middle inequality was already proven in Lemma 4.2.8, hence we focus on 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,

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

which gives the lower bound.
(**Upper bound**) Let $N := \mathcal{P}(K, \varepsilon)$. Then we can find N disjoint closed $\varepsilon/2$-balls with centers $x_i \in K$. While these balls may not fit entirely in K (Figure 4-1), they do fit in a slightly inflated set, namely $K + (\varepsilon/2)B_2^n$ (Basically putting balls at the boundary of K). Comparing the volume gives

$$N \cdot \text{Vol}((\varepsilon/2)B_2^n) \leq \text{Vol}(K + (\varepsilon/2)B_2^n),$$

which completes the upper bound. □

An important consequence of the volumetric bound is that the covering (hence 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 $\varepsilon > 0$:

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

Proof. The lower bound immediately follows from Proposition 4.2.10, since the volumd in $\mathbb{R}^n$ scale as $\mathrm{Vol}(\varepsilon B_2^n) = \varepsilon^n \mathrm{Vol}(B_2^n)$.
The upper bound follows from Proposition 4.2.10 as well:

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

□

To simplify Corollary 4.2.11, we can divide this into two cases for ε:
For $\varepsilon \in (0, 1]$, we have

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

In the other case where $\varepsilon > 1$, one ε-ball covers the unit ball hence $\mathcal{N}(B_2^n, \varepsilon) = 1$.

Remark 4.2.12 (Volume of the ball)**.** The proof of Corollary 4.2.11 works with the volume of the Euclidean ball but never actually calculates it! We can compute the volume geometrically, probabilistically, and analytically (Exercises 4.27-4.29), and also extend this notion of volume to ℓ^p balls (Exercise 4.30).

Remark 4.2.13 (How to construct a net?)**.** We have an algorithm to construct nets already (Remark 4.2.7), but for the Euclidean ball, we can also use a scaled integer lattice (Exercise 4.31), or just use random points (Exercise 4.39).

We can also use covering/packing notions for other objects via volumetric arguments, here is another example:

Definition 4.2.14. 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 as the number of bits where the strings x and y differ:

$$d_H(x, y) := |\{i : \ x(i) \ne y(i)\}|, \ x, y \in \{0, 1\}^n.$$

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 $m \in \{0, \dots, n\}$:

$$\frac{2^n}{\sum_{k=0}^{m} \binom{n}{k}} \le \mathcal{N}(K, d_H, m) \le \mathcal{P}(K, d_H, m) \le \frac{2^n}{\sum_{k=0}^{\lfloor m/2 \rfloor} \binom{n}{k}}.$$

Proof. Use the volumetric argument from above using cardinality instead of the volume (Exercise 4.32). □

4.3 Application: Error Correcting Codes

Covering and packing arguments frequently appear in applications in *coding theory*. We'll give two examples below.

4.3.1 Metric Entropy and Complexity

Intuitively, the covering and packing numbers measure the *complexity* of a set K. Now we'll look at another concept related to that: The logarithm of the covering numbers $\log_2 \mathcal{N}(K,\varepsilon)$ is often called the *metric entropy* of the set 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

$$\log_2 \mathcal{N}(K,d,\varepsilon) \leq \mathcal{C}(K,d,\varepsilon) \leq \lfloor \log_2 \mathcal{N}(K,d,\varepsilon/2) \rfloor .$$

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 (Figure 4.3). Each subset has diameter at most ε, so it can be comvered by a a ball centered in K and with radius ε. 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.

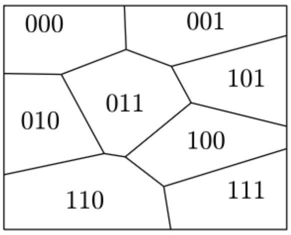


Figure 4.3 Encoding points in K as N-bit strings induces a partition of K into at most 2^N subsets.

(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 w^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 arbitrary accuracy ε. If both x and y are encoded by the same x_0 then by the triangle inequality,

$$d(x,y) \leq d(x,x_0) + d(x_0,y) \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

This shows that $\mathcal{C}(K,d,\varepsilon) \leq N$, which completes the proof. □

4.3.2 Error Correcting Codes

Suppose Alice wants to send to Bob a message with k letters, e.g.

$$x := \text{“fill the glass”}.$$

Suppose further that an adversary can corrupt Alice's message by changing up to r letters. For example, Bob may receive

$$y := \text{“bill the class”}$$

if $r = 2$. Is there a way to protect the communication channel, a method that can correct adversarial errors?

A common approach uses *redundancy.* This means 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 there are r errors in it.

Example 4.3.2 (Repetition code)**.** Alica can just repeat her message several times, e.g. sending the message

$$E(x) := \text{“fill the glass fill the glass fill the glass fill the glass fill the glass”}.$$

Bob can use *majority decoding*: he checks the received copies of each letter in $E(x)$ and picks one that appears most often. If the original message x is repeated $2r+1$ times, majority decoding will recover x exactly, even if r letters in $E(x)$ are corrupted.

The problem with majority decoding is that it is inefficient: it uses

$$n = (2r+1)k$$

letters to encode a k-letter message. As we will see shortly, there exist error correcting codes with much smaller n. But let's define what that is first.

Definition 4.3.3. An error correcting code that encodes k-bit strings into n-bit strings and and correct r 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.

Now we relate error correction to packing numbers of the Hamming cube $\{0,1\}^n$ 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 (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 decider (breaking 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's 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. Then, applying 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 $n-k$ that is nearly linear in r (ignoring a logarithmic term). This is way more efficient than the repetition code, which is optimal (Exercise 4.33).

4.4 Upper Bounds on Subgaussian Random Matrices

This section is mostly concerned with non-asymptotic theory of random matrices. Most of the questions here will be about the distributions of singular values, eigenvalues, and eigenvectors.
But before that, we need to know how ε-nets can help compute the operator norm of a matrix.

4.4.1 Computing the Norm on an εnet

To evaluate $\|A\|$, we need to control $\|Ax\|_2$ uniformly over the sphere S^{n-1}. However, we'll show that instead of the entire sphere, it is enough to control just an ε-net of the sphere (in Euclidean metric).

Lemma 4.4.1 (Operator norm on a net)**.** Let $A \in \mathbb{R}^{m\times n}$ and $\varepsilon \in (0,1]$. Then for any ε-net $\mathcal{N}$ of the sphere S^{n-1}m we have

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

Proof. The lower bound is true 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 $x_0 \in \mathcal{N}$ such that $\|x - x_0\|_2 \le \varepsilon$. By the definition of the operator norm, this implies

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

By the triangle inequality,

$$\|Ax_0\|_2 \ge \|Ax\|_2 - \|Ax - Ax_0\|_2 \ge \|A\|0\varepsilon\|A\| = (1-\varepsilon)\|A\|.$$

Dividing by $1-\varepsilon$ gives the result. □

There is alsoa version for quadratic forms from the way the operator norm is characterized. Since

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

we can take $x = y$ and use the spheres' corresponding nets:

Lemma 4.4.2 (Maximizing quadratic forms on a net)**.** Let $A \in \mathbb{R}^{m\times n}$ and $\varepsilon \in [0,1)$. 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},

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

Moreover, if $m = n$, A is symmetric, and $\mathcal{N} = \mathcal{M}$, we can take $x = y$.

Proof. There are two methods - one by modifying the proof of Lemma 4.4.1 (Exercise 4.36), and a different method using ε-net expansions (Exercise 4.34). □

4.4.2 The Norms of Subgaussian Random Matrices

Theorem 4.4.3. Let $A \in \mathbb{R}^{m\times n}$ be a random matrix with independent, mean zero, subgaussian entries A_{ij}. Then for any $t > 0$,

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

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

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

(**Approximation**). Choose $\varepsilon = 1/4$. Using the result from Corollary 4.2.11, we can find respective ε-nets $\mathcal{N}, \mathcal{M}$ of S^{n-1}, S^{m-1} with cardinalities

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

By Lemma 4.4.2, the norm of A can be bounded using these nets as

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

(**Concentration**). Fix $x \in \mathcal{N}, y \in \mathcal{M}$. 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. By Proposition 2.7.1, the sum is subgaussian, and

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

Using (i) from Proposition 2.6.6, we an restate this as a tail bound:

$$P(|\langle Ax, y\rangle| \geq u) \leq 2\exp\left(-cu^2/K^2\right),\ u \geq 0.$$

(**Union bound**) Next, we unfix x and y and use a union bound. The event

$$\max_{x\in\mathcal{N}, y\in\mathcal{M}} |\langle Ax, y\rangle| \geq u \implies \exists x \in \mathcal{N}, y \in \mathcal{M} \text{ such that } |\langle Ax, y\rangle| \geq u.$$

Therefore union bound gives

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

Using the tail bound above and the estimates on $|\mathcal{N}|$ and $|\mathcal{M}|$, the probability is bounded above by

$$9^{n+m}\cdot 2\exp\left(-cu^2/K^2\right) \quad (*)$$

Choose

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

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

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

Combining with the approximation step,

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

By the choice of u that we had, the proof is complete. □

Remark 4.4.4 (Expectation bounds)**.** High-probability bounds like Theorem 4.4.3 can be usually turned into simpler but less informative expectatiom bounds using the integrated tail formula (Lemma 1.6.1). In Exercise 4.41, we get

$$\mathbb{E}[\|A\|] \le CK(\sqrt{m} + \sqrt{n}).$$

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/column of the matrix (Exercise 4.7). For example, if A has Rademacher entries, its columns have norm $\sqrt{m}$ and rows $\sqrt{n}$, so

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

with probability 1. There is also a fully general lower bound (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 also extends to symmetric matrices:

Corollary 4.4.7. Let $A \in \mathbb{R}^{n\times n}$ be a symmetric random matric with independent, mean zero, subgaussian entries A_{ij} on and above the diagonal. Then for any $t > 0$,

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

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 the lower-triangular part A^-. The diagonal can go either way, so let's just assume it's in A^+. Then $A = A^+ + A^-$.
Applying Theorem 4.4.3 to A^+ and A^- gives (each with probability at least $1 - 4\exp(-t^2)$)

$$\|A^+\| \le CK(\sqrt{n} + t) \text{ and } \|A^-\| \le CK(\sqrt{n} + t).$$

By the triangle inequality, $\|A\| \le \|A^+\| + \|A^-\|$ hence the proof is complete. □

4.5 Application: Community Detection in Networks

Random matrix theory has many applications, one of them being network analysis
Real-world networks often have *communites*, 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

We'll explore the stochastic block model, a straightforward extension of the Erdős–Rényi model we discussed in Section 2.5.

Definition 4.5.1. 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 a if they are in different communities. This random graph model is called the stochastic block model, denoted $G(n, p, q)$.

When $p = q$ we just get the Erdős–Rényi model $G(n, p)$. But when $p > q$, edges are more likely to form within communities than between them, creating a community structure (Figure 4.4).

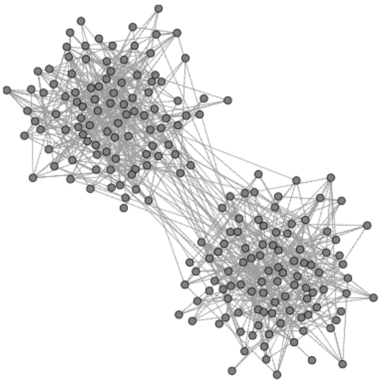


Figure 4.4 A random graph following the stochastic block model $G(n, p, q)$ with $n = 200$, $p = 1/20$ and $q = 1/200$.

4.5.2 The Expected Adjacency Matrix Holds the Key

A graph G can be conveniently represented by its adjacency matrix A (Definition 3.6.2). For a random graph $G \sim G(n, p, q)$, the adjacency matrix A is a random matrix, which we can use the tools we developed in this chapter to analyze.

Let's split A into deterministic and random parts:

$$A = D + R \text{ where } D = \mathbb{E}[A]$$

and think of D as the *signal* and R as *noise*.

To see why D is informative, let's compute its eigenstructure. The entries 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 looks like this:

$$D = \mathbb{E}[A] = \begin{bmatrix} p & p & q & q \\ p & p & q & q \\ q & q & p & p \\ q & q & p & p \end{bmatrix}.$$

Take a look at the simpler matrix

$$\begin{bmatrix} p & q \\ q & p \end{bmatrix}.$$

Its eigenvalues are $\frac{p+q}{2}$ and $\frac{p-q}{2}$, with eigenvectors $(1,1)^T$ and $(1,-1)^T$. 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:

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

The key object here is the *second* eigenvector of $u_2(D)$, which holds all the information about the community structure. If we know $u_2(D)$, we could identify the communities by the signs of its coefficients.

4.5.3 The Actual Adjacency Matrix is a Good Approximation

Unfortunately, we don't know the expected adjacency matrix $D = \mathbb{E}[A]$. 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:

$$\|R\| \le C\sqrt{n} \text{ with probability at least } 1 - 4e^{-n}.$$

So for large n, the noise R is much smaller then the signal D. It means that A is close to D, so we can just use the matrix A instead of D to extract the community inforemation. Let's justify this using the matrix perturbation theory developed earlier.

4.5.4 Perturbation Theory

We'll apply the David-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-seperated 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 $R = A - D$, 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):

$$\sin\angle(\bar{u}_2(D), \bar{u}_2(A)) \le \frac{2\|R\|}{\delta} \lesssim \frac{\sqrt{n}}{\mu n} \lesssim \frac{1}{\mu\sqrt{n}}.$$

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 was not a unit vector - it had norm $\sqrt{n}$. Multipkying both sides by $\sqrt{n}$, we get

$$\|u_2(D) - \theta u_2(A)\|_2 \lesssim \frac{1}{\mu}.$$

This implies that the *signs* of most coefficients of $u_2(D)$ and $\theta u_2(A)$ 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 $u(D)$ are ±1, we see that each disagreement contributes 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 $\lesssim \frac{1}{\mu^2}$.

4.5.5 Spectral Clustering

In summary, we use $u_2(A)$ to estimate $u_2(D)$, whose coefficients are ±1 and identify the two communities. This method for community detection is known as *spectral clustering*:

Algorithm 1 Spectral Clustering

```
Input: graph $G$
Output: a partition of the vertices $G$ into two communities
  Compute the adjacency matrix $A$ of the graph.
  Compute the eigenvector $v_2(A)$ for the second largest eigenvalue of $A$.
  Split vertices into two communities based on the signs of $v_2(A)$'s coefficients.
```

In fact, we have proved above that:

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 ($q \geq$ const) and the probabilities of within- and across- community edges are well seperated ($p - q \geq$ const) . The condition $q \geq$ const is in fact not essential (Exercise 4.45).

Remark 4.5.3 (Sparsity)**.** The sparsest graphs for which Theorem 4.5.2 is nontrivial, meaning $C/\mu^2 \leq n$, have expected average degree

$$\frac{n(p+q)}{2} \asymp \sqrt{n}.$$

With more tools, we will handle much more sparser graphs in Section 5.5.

4.6 Two-sided Bounds on Subgaussian Matrices

Theorem 4.4.3 gives an upper bound on the singular values of an $\mathbb{R}^{m\times n}$ subgaussian random matrix A, which says

$$\sigma_1 \leq \|A\| \leq C(\sqrt{m} + \sqrt{n})$$

with high probability.
In fact, there is a sharper two-sided bound on the **entire spectrum** of A:

$$\sqrt{m} - C\sqrt{n} \leq \sigma_i \leq \sqrt{m} + C\sqrt{n}.$$

In other words, the below shows that a tall random matrix $\frac{1}{\sqrt{m}}A$ with $m \gg n$ is an approximate isometry.

Theorem 4.6.1 (Name)**.** Let $A \in \mathbb{R}^{m\times n}$ be a random matrix with independent, mean zero, subgaussian, isotropic tows A_i. Then for any $t \geq 0$ we have

$$\sqrt{m} - CK^2(\sqrt{n} + t) \leq \sigma_n \leq \sigma_1 \leq \sqrt{m} + CK^2(\sqrt{n} + t)$$

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

Proof. We'll prove a slightly stronger conclusion than the theorem statement, namely

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

Proving this implies proving the theorem (I haven't checked yet).
Again, we'll apply an ε-net argument, but use Bernstein inequality for the concentration step instead of Hoeffding which we did in Theorem 4.4.3.
(**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 the equation above on $\mathcal{N}$:

$$\|\frac{1}{m}A^TA - I_n\| \leq 2\max_{x\in\mathcal{N}}\left|\left\langle (\frac{1}{m}A^TA - I_n)x, x\right\rangle\right| = 2\max_{x\in\mathcal{N}}\left|\frac{1}{m}\|Ax\|_2^2 - 1\right|.$$

Therefore, to prove the statement, it is enough to show that, with the required probability,

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

(**Concentration**) Fix $x \in \mathcal{N}$ and express $\|Ax\|_2^2$ as a sum of independent random variables:

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

By assumption, the wors 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 with

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

Thus we can use Bernstein inequality (Corollary 2.9.2) and obtain

$$\begin{aligned} P\left(\left|\frac{1}{m}\|Ax\|_2^2 - 1\right| \geq \frac{\varepsilon}{2}\right) &= P\left(\left|\frac{1}{m}\sum_{i=1}^{m} X_i^2 - 1\right| \geq \frac{\varepsilon}{2}\right) \\ &\leq 2\exp\left[-c_1 \min\left(\frac{\varepsilon^2}{K^4}, \frac{\varepsilon}{k^2}\right) m\right] \\ &= 2\exp\left(-c_1 \delta^2 m\right) \\ &\leq 2\exp\left(-c_1 C^2 (n + t^2)\right). \end{aligned}$$

The last inequality comes from the definition of δ and using the inequality

$$(a+b)^2 \geq a^2 + b^2 \text{ for } a, b \geq 0.$$

(**Union bound**) Now unfix $x \in \mathcal{N}$. By union bound,

$$P\left(\max_{x \in \mathcal{N}}\left|\frac{1}{m}\|Ax\|_2^2 - 1\right| \geq \frac{\varepsilon}{2}\right) \leq 9^n \cdot 2\exp\left(-c_1 C^2 (n + t^2)\right) \leq 2\exp\left(-t^2\right)$$

if we choose the constant C to be large enough. Then by the necessary condition in the approximation step, the proof is complete. □

Remark 4.6.2 (Expectation bound)**.** From Remark 4.4.4, we can convert high-probability bounds to expectation bounds. Exercise 4.41 yields the following form for Theorem 4.6.1:

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

There is another version of the proof for Theorem 4.6.1 in Exercise 4.46.

4.7 Application: Covariance Estimation and Clustering

When analyzing high-dimensional data, i.e. given points $X_1, \ldots, X_m \in \mathbb{R}^n$ sampled from an unknown distribution, PCA is a basic way to analyze the data from Section 3.2.2.

PCA finds the "principle components" as top eigenvectors of the data's covariance matrix. Although we do not know the covariance matrix of the underlying distribution, we can approximately estimate using the sample. The David-Kahan theorem (Theorem 4.1.15) then helps us estimate the principle 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, and let's denote its covariance matrix by

$$\Sigma = \mathbb{E}\left[XX^T\right].$$

To estimate Σ, we can use the sample covariance matrix:

$$\Sigma_m = \frac{1}{m}\sum_{i=1}^{m} X_i X_i^T.$$

Since X and X_i are identically distributed, this estimate is unbiased:

$$\mathbb{E}\left[\Sigma_m\right] = \Sigma.$$

By the Law of Large Numbers (Theorem 1.7.1) applied to each entry of Σ, we get

$$\Sigma_m \to \Sigma \text{ almost surely as } m \to \infty.$$

This leads to the quantitative problem: How big does the sample size m need to be so that

$$\Sigma_m \approx \Sigma \text{ with high probability?}$$

For dimension reasons, we need at least $m \gtrsim n$ sample points. 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 speficically, assume that there exists $K \geq 1$ such that

$$\|\langle X, x\rangle\|_{\psi_2} \leq K\|\langle X, x\rangle\|_{L^2} = Kx^T\Sigma x \text{ for any } x \in \mathbb{R}^n.$$

Then for every positive integer m, we have

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

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 (This can be dropped like in Exercise 3.10). Setting

$$Z = Z_i = \Sigma^{-1/2}X,$$

then $Z, Z_1, \ldots, Z_m$ are independent and isotropic random vectors satisfying

$$X = \Sigma^{1/2}Z \text{ and } X_i = \Sigma^{1/2}Z_i.$$

The assumption in the theorem then implies that (by Definition 3.4.1)

$$\begin{aligned}
\|Z\|_{\psi_2}^2 &= \sup_{s\in S^{n-1}} \|\langle Z, x\rangle\|_{\psi_2}^2 \\
&= \sup_{x\in S^{n-1}} \|\langle \Sigma^{-1}X, x\rangle\|_{\psi_2}^2 \\
&\leq \Sigma^{-1} \sup_{x\in S^{n-1}} \|\langle X, x\rangle\|_{\psi_2}^2 \\
&\leq K^2\Sigma^{-1} \sup_{x\in S^{n-1}} x^T\Sigma x \\
&= K^2.
\end{aligned}$$

Therefore $\|Z\|_{\psi_2} \leq K$ and $\|Z_i\|_{\psi_2} \leq K$. Then

$$\|\Sigma_m - \Sigma\| = \|\Sigma^{1/2}R_m\Sigma^{1/2}\| \leq \|R_m\|\|\Sigma\| \text{ where } R_m := \frac{1}{m}\sum_{i=1}^m Z_iZ_i^T - I_n.$$

Consider the $m \times n$ random matrix A whose rows are Z_i^T. Then

$$\frac{1}{m}A^TA - I_n = \frac{1}{m}\sum_{i=1}^m Z_iZ_i^T - I_n = \mathbb{R}_m.$$

Applying Theorem 4.6.1, we get

$$\mathbb{E}\left[R_m\right] \leq CK^2\left(\sqrt{\frac{n}{m}} + \frac{n}{m}\right)$$

(see Remark 4.6.2). Substituting this into the bound for $\|\Sigma_m - \Sigma\|$, 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 matric with a small relative error:

$$\mathbb{E}\left[\|\Sigma_m - \Sigma\|\right] \leq \mathbb{E}\left[\|\Sigma\|\right],$$

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)**.** Out argument also gives a high-probability bound: for any $u \geq 0$, we have

$$\|\Sigma_m - \Sigma\| \leq CK^2 \left(\sqrt{\frac{n+u}{m}} + \frac{n+u}{m} \right) \|\Sigma\|$$

with probability at least $1 - 2e^{-u}$. See Exercise 4.49.

4.7.1 Application: Clustering of Point Sets

Let's illustrate Theorem 4.7.1 with an application to clustering. In particular, we'll do this for Gaussian mixture models (GMMs):

Definition 4.7.4. 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.

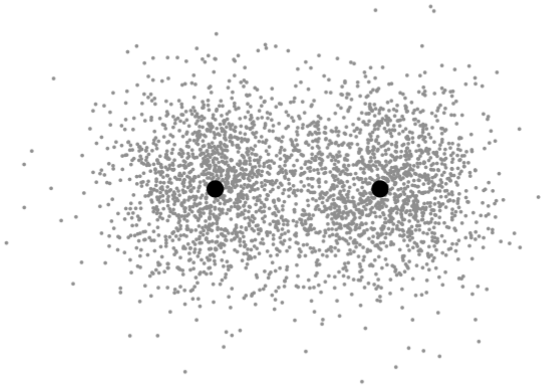


Figure 4.5 $m = 3000$ points drawn from the Gaussian mixture model with means $-\mu$ and μ shown as two big black dots, for $\mu = (-1.6, 0)$.

Given m sample points from a Gaussian mixture model, we would like to know which points belong to which cluster. To do this, we can use a variant of the *spectral clustering* algorithm.

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 via PCA. Next, we can project the data points onto the principle component and classify them based on which side of the origin they lie on. Here is the algorithm:

Algorithm 2 Spectral Clustering for GMMs

```
Input: points $X_1, \dots, X_m \in \mathbb{R}^n$
Output: a partition of the points into two clusters
  1 Compute the sample covariance matrix $\Sigma_m = \frac{1}{m}\sum_{i=1}^m X_i X_i^T$.
  2 Compute the top eigenvector $v = v_1(\Sigma_m)$.
  3 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 GMMs)**.** Take points $X_1, \dots, X_m \in \mathbb{R}^n$ from the Gaussian mixture model with means $\pm\mu$. 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.

Proof. Exercise 4.51. □

It is remarkable that accurate classification is possible even when the cluster seperation $\|\mu\|_2$ is much smaller than their diameter, which is $\asymp \sqrt{n}$.

5 Concentration Without Independence

This chapter mainly explores other approaches to concentration that do not rely on independence.

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.

$$f(X) \approx \mathbb{E}[f(X)] \text{ with high probability?}$$

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 for general *nonlinear* functions f? We can't expect good concentration for any f, but if f does not oscillate too wildly, we might get good concentration. Namely, we'll use Lipschitz functions to rule out these oscillations:

5.1.1 Lipschitz Functions

Definition 5.1.1. 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

$$d_Y(f(u), f(v)) \leq L \cdot d_X(u, v) \text{ for every } u, v \in X.$$

The infimum of all L in this definition is called the Lipschitz norm because of f and is denoted $\|f\|_{\text{Lip}}$.

If $\|f\|_{\text{Lip}} \leq 1$, f is called a contraction.

(**Important**) Technically the Lipschits norm is only a seminorm, since it vanishes on nonzero constant functions. It's called a norm in the book for brevity.

The class of Lipschitz functions sits between differentiable and uniformly continuous:

$$f \text{ is differentiable } \implies f \text{ is Lipschitz } \implies f \text{ if uniformly continuous.}$$

Moreover, from Exercise 5.1,

$$\|F\|_{\text{Lip}} \leq \sup_{x \in \mathbb{R}^n} \|\nabla f(x)\|_2.$$

Example 5.1.2. Vectors, matries, and norms define natural Lipschitz functions:

(a) For a fixed vector $\theta \in \mathbb{R}^n$, the linear functional

$$f(x) = \langle x, \theta \rangle \text{ has Lipschitz norm } \|f\|_{\text{Lip}} = \|\theta\|_2.$$

(b) More generally, any $m \times n$ matrix A, the linear operator

$$f(x) = Ax \text{ has Lipschitz norm } \|F\|_{\text{Lip}} = \|A\|.$$

(c) For any norm $\|\cdot\|$ on $\mathbb{R}^n$, the function

$$f(x) = \|x\|$$

has Lipschitz norm equal to the smallest L such that

$$\|x\| \leq L\|x\|_2 \text{ for all } x \in \mathbb{R}^n.$$

Proof. Exercise 5.2. □

5.1.2 Concentration via Isoperimetric Inequalities

Any Lipschitz function on the Euclidean sphere $S^{n-1} = \{x \in \mathbb{R}^n : \|x\|_2 = 1\}$ concentrates:

Theorem 5.1.3. Let $X \sim \mathrm{Unif}(\sqrt{n}S^{n-1})$. Then for any Lipschitz function $f : \sqrt{n}S^{n-1} \to \mathbb{R}$ we have

$$\|f(X) - \mathbb{E}[f(X)]\|_{\psi_2} \leq C\|f\|_{\mathrm{Lip}}.$$

The theorem above works for the geodesic distance metric as well (Exercise 5.4).
Theorem 5.1.3 has been proved already for linear functions f. Theorem 3.4.5 tells us that X is a subgaussian random vectos, and this by definition means that any lienar 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. We'll use the aread of their sublevel sets - regions of the sphere where $f(x) \leq a$ for a given level a. To do this, we'll use the *isoperimetric inequality*, namely the one for subsets on $\mathbb{R}^n$:

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

$$A_\varepsilon = \{x \in \mathbb{R}^n : \exists y \in A \text{ such that } \|x-y\|_2 \leq \varepsilon\} = A + \varepsilon B_2^n.$$

The figure below illustrates the isoperimetric inequality:

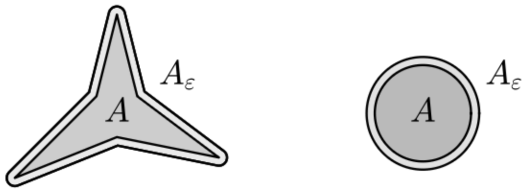


Figure 5.1 The isoperimetric inequality says that among all sets A with a given volume, Euclidean balls minimize the volume of their ε-neighborhood A_ε.

A similar isoperimetric inequality holds for subsets on S^{n-1}, and in this case the minimizers are the spherical caps - neighborhoods of a single point. To state this principle, let σ_{n-1} denote the normalized are 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 subsets $A \subset S^{n-1}$ with given area $\sigma_{n-1}(A)$, the spherical caps minimizer the area of the neighborhood $\sigma_{n-1}(A_\varepsilon)$, where

$$A_\varepsilon := \{x \in \mathbb{R}^n : \exists y \in S^{n-1} \text{ such that } \|x-y\|_2 \leq \varepsilon\}.$$

5.1.3 Blow-up of Sets on the Sphere

The isoperimetric inequality leads to a remarkable and counterintuitive 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. To simplify things in view of Theorem 5.1.3, we'll operate on the sphere with radius $\sqrt{n}$.

Lemma 5.1.6 (Blow-up)**.** Let $A \subset \sqrt{n}S^{n-1}$, and let σ denote the normalized are on that sphere. If $\sigma(A) \geq 1/2$, then for every $t \geq 0$,

$$\sigma(A_t) \geq 1 - 2\exp(-ct^2).$$

Proof. Consider the hemisphere defined by the first coordinate:

$$H := \{x \in \sqrt{n}S^{n-1} : x_1 \leq 0\}.$$

By assumption, $\sigma(A) \geq 1/2 = \sigma(H)$, hence the isoperimetric inequality (Theorem 5.1.5) implies that

$$\sigma(A_t) \geq \sigma(H_t).$$

The neighborhood H_t of the hemisphere H is a spherical cap (a portion of a sphere cut off by a plane), and we could compute its area directly, but it is easier to use Theorem 3.4.5 instead, which states that a random vector $X \sim \text{Unif}(\sqrt{n}S^{n-1})$ is subgaussian, and $\|X\|_{\psi_2} \leq C$. Since σ is the uniform probability measure on the sphere, it follows that

$$\sigma(H_t) = P(X \in H_t).$$

Now, the definition of the neighborhood implies that

$$\{x \in \sqrt{n}S^{n-1} : \ x_1 \leq t/\sqrt{2}\} \subset H_t.$$

Thus

$$\sigma(H_t) \geq P(X_1 \leq t/\sqrt{2}) \geq 1 - 2\exp\left(-ct^2\right).$$

The last inequality holds because $\|X_1\|_{\psi_2} \leq \|X\|_{\psi_2} \leq C$. Then the lemma is proved because $\sigma(A_t) \geq \sigma(H_t)$. □

Remark 5.1.7 (A 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 (Exercise 5.3)!

Remark 5.1.8 (A zero-one law)**.** The blow-up phenomenen we just saw can be quite counterintuitive at first. However, this is a typical p 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 zero or one.

5.1.4 Proof of Theorem 5.1.3

WLOG, we can assume that $\|f\|_{\text{Lip}} = 1$. Let M denote the median of $f(X)$, which by definition satisfies

$$P(f(X) \leq M) \geq \frac{1}{2} \text{ and } P(f(X) \geq M) \geq \frac{1}{2}.$$

Consider the sublevel set

$$A := \{x \in \sqrt{n}S^{n-1} : \ f(x) \leq M\}.$$

Since $P(X \in A) \geq \frac{1}{2}$, Lemma 5.1.6 implies that

$$P(X \in A_t) \geq 1 - 2\exp\left(-ct^2\right).$$

On the other hand, we claim that

$$P(X \in A_t) \leq P(f(X) \leq M + t).$$

Indeed, if $X \in A_t$ then $\|X - y\|_2 \leq t$ for some point $y \in A$. By definition, $f(y) \leq M$. Since f is Lipschitz with $\|f\|_{\text{Lip}} = 1$, it follows that

$$f(X) \leq f(y) + \|X - y\|_2 \leq M + t.$$

Combining the two bounds above, we conclude that

$$P(f(X) \leq M + t) \geq 1 - 2\exp\left(-ct^2\right).$$

Repeating the argument for $-f$, we obtain a similar bound for the probability that $f(x) \geq M - t$ (do). Combining the two, we get a similar bound for the probability that $|f(X) - M| \leq t$, and conclude that

$$\|f(X) - M\|_{\psi_2} \leq C.$$

Then we can replace the median by the mean, which follows by centering (Lemma 2.7.8). Therefore the proof is complete. □

5.2 Concentration on Other Metric Measure Spaces

We can extend concentration from the sphere to other spaces as well. The proof of Theorem 5.1.3 relied on two ingredients:

(a) an isoperimetric inequality,

(b) a blow-up of its minimizers.

There are not unique to the sphere - many spaces satusfy them hence we can derive similar concentration results.

Remark 5.2.1. Concentration keeps the mean, median, and L^p norms close. Therefore, 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 (Exercise 5.10).

5.2.1 Gaussian Concentration

The Gaussian measure of a Borel set $A \subset \mathbb{R}^n$ is defined as

$$\gamma_n(A) := P(X \in A) = \frac{1}{(2\pi)^{n/2}} \int_A e^{-\|x\|_2^2/2}\, dx$$

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 $\gamma_n(A_\varepsilon)$.

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

$$\|f(X) - \mathbb{E}[f(X)]\|_{\psi_2} \leq C\|f\|_{\text{Lip}}.$$

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).

5.2.2 Hamming Cube

The method based on isoperimetry also works on the Hamming cube $(\{0,1\}^n, d, \mathbb{P})$ (Definition 4.2.14), where $d(x,y)$ is the normalized Hamming distance:

$$d(x,y) = \frac{1}{n}|\{i : x_i \neq y_i\}|.$$

The measure $\mathbb{P}$ is the uniform probability measure on the cube:

$$\mathbb{P}(A) = \frac{|A|}{2^n} \text{ for any } A \subset \{0,1\}^n.$$

Theorem 5.2.5 (Concentration on the Hamming cube)**.** Consider a random vector $X \sim \{0,1\}^n$. Then for any function $f : \{0,1\}^n \to \mathbb{R}$ we have

$$\|f(X) - \mathbb{E}[f(X)]\|_{\psi_2} \leq \frac{C\|f\|_{\text{Lip}}}{\sqrt{n}}.$$

5.2.3 Symmetric Group

A similar result holds for the symmetric group S_n, a set of all $n!$ permutations of $\{1, \dots, n\}$. We can view the symmetric group as a metric measure space $(S_n, d, \mathbb{P})$, where $d(\pi, \rho)$ 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:

$$\mathbb{P}(A) = \frac{|A|}{n!} \text{ for any } A \subset S_n.$$

Theorem 5.2.6 (Concentration on the symmetric group)**.** Consider a random permutation $X \sim \text{Unif}(S_n)$ and a function $f : S_n \to \mathbb{R}$. Then

$$\|f(X) - \mathbb{E}[f(X)]\|_{\psi_2} \leq \frac{C\|f\|_{\text{Lip}}}{\sqrt{n}}.$$

5.2.4 Riemannian Manifolds with Strictly Positive Curvature

(Feel free to skip this if not familiar with differential geometry)
A compact connected Riemannian manifold (M, g) comes with the geodesic distance $d(x, y)$, which is the shortest length of a curve connecting the points. Then we can define 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 $c(M) > 0$, then it can be proved that

$$\|f(X) - \mathbb{E}[f(X)]\|_{\psi_2} \leq \frac{C\|f\|_{\text{Lip}}}{\sqrt{c(M)}}$$

for any Lipschitz function $f : M \to \mathbb{R}$.
To give an example, $c(S^{n-1}) = n - 1$. Then the above gives another approach for the concentration inequality of the sphere.

5.2.5 Special Orthogonal Group

The special orthogonal group $\text{SO}(n)$ consists of all $n \times n$ orthogonal matrices with determinant 1. We can treat it as a metric measure space $(\text{SO}(n), \|\cdot\|_F, \mathbb{P})$, with distance given by the Frobenius norm $\|A - B\|_F$ and $\mathbb{P}$ as the uniform probability measure.

Theorem 5.2.7 (Concentration on the special orthogonal group)**.** Consider a random orthogonal matrix $X \sim \text{Unif}(\text{SO}(n))$ and a function $f : \text{SO}(n) \to \mathbb{R}$. Then

$$\|f(X) - \mathbb{E}[f(X)]\|_{\psi_2} \leq \frac{C\|f\|_{\text{Lip}}}{\sqrt{n}}.$$

The result above can be deduced from the concentration on general Riemannian manifolds.

Remark 5.2.8 (Haar measure)**.** To generate a random orthogonal matrix $X \sim \text{Unif}(\text{SO}(n))$, one way is to start with an $n \times n$ Gaussian random matrix G with $N(0, 1)$ independent entries, and compute its SVD $G = U\Omega V^T$. Then the matrix of left singular vectors is uniformly distributed in $\text{SO}(n)$.
The uniform probability distribution on $\text{SO}(n)$ is given by

$$\mu(A) := P(X \in A) \text{ for } A \subset \text{SO}(n).$$

This is the unique rotation-invariant probaility measure on $\text{SO}(n)$, called the Haar measure.

5.2.6 Grassmannian

The Grassmannian 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}. Therefore the concentration on the Grassmannian includes the concentration on the sphere.
We can treat $G_{n,m}$ as a metric space $(G_{n,m}, d, \mathbb{P})$, where the distance between subspaces is given by the operator norm

$$d(E, F) = \|P_E - P_F\|$$

where P_E and P_F are the orthogonal projections onto the subspaces. The probability measure is the Haar measure (Remark 5.2.8). A random subspace E can hence be computed by computing the image of the random $n \times m$ Gaussian random matric with i.i.d. $N(0, 1)$ entries.

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

$$\|f(X) - \mathbb{E}[f(X)]\|_{\psi_2} \leq \frac{C\|f\|_{\mathrm{Lip}}}{\sqrt{n}}.$$

Proof. The proof is a bit involved: Express the Grassmannian as the quotient via the special orthogonal group:

$$G_{n,m} = \mathrm{SO}(n)/(\mathrm{SO}(m) \times \mathrm{SO}(n-m))$$

and use the fact that concentration carries over to quotients. □

5.2.7 Continuous Cube and Euclidean Ball

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

$$\|f(X) - \mathbb{E}[f(X)]\|_{\psi_2} \leq C\|f\|_{\mathrm{Lip}}.$$

Proof. Exercises 5.12 & 5.13. □

5.2.8 Densities of the Form $e^{-U(x)}$

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 $U : \mathbb{R}^n \to \mathbb{R}$. For example, $X \sim N(0, I_n)$, the normal density gives $U(x) = \|x\|_2^2 + c$ where c is constant (dependent on n but not on x), and Gaussian concentration holds for X.
In general, we would expect that if U has curvature at least like $\|x\|_2^2$, then there would be at least Gaussian concentration. As the theorem below shows, this depends on the Hessian of U:

Theorem 5.2.11. Consider a random vector X in $\mathbb{R}^n$ whose density has the form $e^{-U(x)}$ for some function $U : \mathbb{R}^n \to \mathbb{R}$. Assume there exists $\kappa > 0$ such that

$$\nabla^2 U(x) \succcurlyeq \kappa I_n \text{ for all } x \in \mathbb{R}^n.$$

Then any Lipschitz function $f : \mathbb{R}^n \to \mathbb{R}$ satisfies

$$\|f(X) - \mathbb{E}[f(X)]\|_{\psi_2} \leq \frac{C\|f\|_{\mathrm{Lip}}}{\sqrt{\kappa}}.$$

Proof. THe proof uses semigroup methods, which are not covered in the text. □

5.2.9 Random Vectors with Independent Bounded Coordinates

There is a remarkable partial generalization of Theorem 5.2.10 for random vectors X with independent coordinates that have arbitrary bounded distributions (not just uniform). By scaling, we can assume WLOG that $|X_i| \le 1$.

Theorem 5.2.12 (Talagrand concentration inequality)**.** Consider a random vector in $\mathbb{R}^n$, $X = (X_1, \dots, X_n)$ whose coordinates are independent and satisfy $|X_i| \le 1$ almost surely. Then for any Lipschitz function $f : [-1,1]^n \to \mathbb{R}$,

$$\|f(X) - \mathbb{E}[f(X)]\|_{\psi_2} \le C\|f\|_{\text{Lip}}.$$

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 the dimension without losing the geometry of the data? The simplest way is to project onto a low-dimensional subspace

$$E \subset \mathbb{R}^n, \ \dim(E) := m \ll n,$$

see Figure 5.2 below. How shall we choose the subspace E, and how small should its dimension m be?

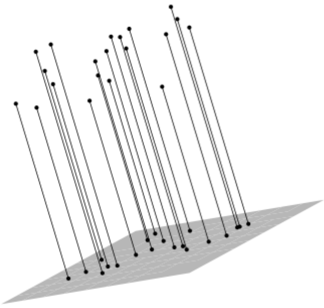


Figure 5.2 Johnson-Lindenstrauss Lemma reduces dimension of the data by random projection onto a low-dimensional subspace.

The Johnson-Lindenstrauss Lemma states that the geometry of data is well preserved if we choose E to be a *random subspace* of dimension

$$m \asymp \log N.$$

Here we say that E is a random m-dimensional subspace in $\mathbb{R}^n$ uniformly distributed in $G_{n,m}$, i.e.

$$E \sim \text{Unif}(G_{n,m}),$$

if E is a random m-dimensional subspace of $\mathbb{R}^n$ whose distribution is rotation invariant, i.e.

$$P(E \in \mathcal{E}) = P(U(E) \in \mathcal{E})$$

for any fixed subset $\mathcal{E} \in G_{n,m}$ 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

$$m \ge C\varepsilon^{-2} \log N.$$

Let P be the orthogonal projection in $\mathbb{R}^n$ onto a random m-dimensional subspace $E \sim \text{Unif}(G_{n,m})$. Then, with probability at least $1 - 2\exp(-c\varepsilon^2 m)$, the scaled projection $Q = \sqrt{n/m}P$ is an approximate isometry on $\mathcal{X}$:

$$(1-\varepsilon)\|x-y\|_2 \le \|Qx - Qy\|_2 \le (1+\varepsilon)\|x-y\|_2 for all x, y \in \mathcal{X}.$$

The proof will be based on concentration of Lipschitz functions on the sphere in Section 5.1 We use it to examine how the random projection P acts on the fixed vector $x - y$, then take the 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) $(\mathbb{E}\left[\|Pz\|_2^2\right])^{1/2} = \sqrt{\frac{m}{n}}\|z\|_2$.

(b) With probability at least $1 - 2\exp(-c\varepsilon^2 m)$, we have

$$(1-\varepsilon)\sqrt{\frac{m}{n}}\|z\|_2 \le \|Pz\|_2 \le (1+\varepsilon)\sqrt{\frac{m}{n}}\|z\|_2.$$

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 rotate $\mathbb{R}^m$, we cna fix the 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 has the same distribution!

(a) Since P is the projection onto the first m coordinates in $\mathbb{R}^n$,

$$\mathbb{E}\left[\|Pz\|_2^2\right] = \mathbb{E}\left[\sum_{i=1}^m z_i^2\right] = \sum_{i=1}^m \mathbb{E}\left[z_i^2\right] = m\mathbb{E}\left[z_1^2\right],$$

because the coordinates z_i of the random vector $z \sim \mathrm{Unif}(S^{n-1})$ are identically distributed. To compute $\mathbb{E}\left[z_1^2\right]$, note that $\sum_{i=1}^n z_i^2 = 1$. Taking expectations on both sides, we obtain

$$\sum_{i=1}^n \mathbb{E}\left[z_i^2\right] = 1 \implies \mathbb{E}\left[z_1^2\right] = \frac{1}{n}.$$

Then, putting this into the equation above, we have

$$\mathbb{E}\left[\|Pz\|_2^2\right] = \frac{m}{n}.$$

(b) $x \mapsto \|Px\|_2$ is a Lipschitz function on S^{n-1} with Lipschitz norm bounded by 1. Then from Exercise 5.5, the concentration inequality gives

$$P\left(\left|\|Px\|_2 - \sqrt{\frac{m}{n}}\right| \ge t\right) \le 2\exp(-cnt^2).$$

(We replaced $\mathbb{E}[\|x\|_2]$ by $(\mathbb{E}\left[\|x\|_2^2\right]^{1/2})$ in the concentration inequality using Remark 5.2.1). Choosing $t := \varepsilon\sqrt{m/n}$, we complete the proof. □

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 inequality is equivalent to

$$(1-\varepsilon)\sqrt{\frac{m}{n}}\|z\|_2 \le \|Pz\|_2 \le (1+\varepsilon)\sqrt{\frac{m}{n}}\|z\|_2.$$

For any fixed z, Lemma 5.3.2 states that the above 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 the bound above holds simultaneously for all $z \in \mathcal{X} - \mathcal{X}$, with probability at least

$$1 - |\mathcal{X} - \mathcal{X}| \cdot 2\exp(-c\varepsilon^2 m) \ge 1 - N^2 \cdot 2\exp(-c\varepsilon^2 m).$$

If $m \ge C\varepsilon^{-2}\log N$ then this probability is at least $1 - 2\exp(-c\varepsilon^2 m/2)$, as claimed. Hence the proof is done. □

Remark 5.3.3 (Non-adaptive, dimension-free)**.** A remarkable feature of the JL lemma is that the dimension reduction map A is *non-adaptive*, meaning it does not depend on the data. Note also that the ambient dimension n of the data plays no role. With more toold, we will develop more advanced versions of the JH lemma (Exercise 9,37-9.39).

Remark 5.3.4 (Optimality)**.** The JL 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 that we can't - the log dimension is the best we can do, even with nonlinear maps.

5.4 Matrix Bernstein Inequality

We extend generalized concentration inequalities from sums of independent random variables to sums of independent random matrices. We'll make a matrix version of Bernstein inequality (Theorem 2.9.5) by replacing random variables by random matrices, and absolute value by the operator norm. No need for independence of entries, rows, or columns within each random matrix!

Theorem 5.4.1 (Matrix Bernstein inequality)**.** Let $X_1, \dots, X_N$ be independent, mean zero, $n \times n$ symmetric random matrices, such that $\|X_i\| \leq K$ almost surely for all i. Then for every $t \geq 0$,

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

where $\sigma^2 = \|\sum_{i=1}^{N} \mathbb{E}[X_i^2]\|$ is the operator norm of the matrix variance of the sum.

We can rewrite the RHS of the inequality as the mixture of subgaussian and subexponential tail, like in the scalar Bernstein inequality:

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

The proof is similar to that of the scalar version: Repeat the MGF argument, swapping scalars with matrices. However, there is a big problem: Matrix multiplication is not commutative! Therefore we need some matrix calculus knowledge first.

5.4.1 Matrix Calculus

For an $n \times n$ symmetric matrix X, operations such as inversion or squaring only affect eigenvalues. For example, if the spectral decomposition of X is $X = \sum_{i=1}^{n} \lambda_i u_i u_i^T$, then

$$X^{-1} = \sum_{i=1}^{n} \frac{1}{\lambda_i} u_i u_i^T,\ X^2 = \sum_{i=1}^{n} \lambda_i^2 u_i u_i^T,\ 2I_n - 5X^3 = \sum_{i=1}^{n} (2 - 5\lambda_i^3) u_i u_i^T.$$

This suggest that for symmetric matrices, applying arbitrary functions on the matrices is equivalent to applying them to the eigenvalues:

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 as above, define

$$f(X) := \sum_{i=1}^{n} f(\lambda_i) u_i u_i^T.$$

This definition agrees with matrix addition and multiplication, and with Taylor series (Exercise 5.16). Of course, matrices can be compared with each other via a partial ordering:

Definition 5.4.3 (Loewner order)**.** We write $X \succcurlyeq 0$ if X is a symmetric positive semidefinite matrix. We write $X \succeq Y$ and $Y \preceq X$ if $X - y \succeq 0$.

This is a partial ordering because there are matrices for which neither $X \succeq Y$ nor $Y \succeq X$ holds.

Proposition 5.4.4 (Simple properties of Loewner order)**.** We have

(a) (Eigenvalue monotonicity) $X \preceq Y$ implies $\lambda_i(X) \leq \lambda_i(Y)$ for all i.

(b) (Trace monotonicity) For a (weakly) increasing function $f : \mathbb{R} \to \mathbb{R}$,
$$X \preceq Y \implies \operatorname{tr}(f(X)) \leq \operatorname{tr}(f(Y)).$$

(c) (Operator norm) For any $a \geq 0$,
$$\|X\| \leq a \iff -aI_n \preceq X \preceq aI_n.$$

(d) (Upgrading scalar to matrix inequalities) For functions $f, g : \mathbb{R} \to \mathbb{R}$,
$$f(x) \leq g(x) \forall x \text{ with } |x| \leq a \implies f(X) \preceq g(X) \forall X \text{ with } \|X\| \leq a.$$

Proof. (a) If $X \preceq Y$, then $Y - X \succeq 0$ hence all eigenvalues of $Y - X$ are greater than equal to 0, and the result follows.
(b) The eigenvalues of $f(X)$ are $f(\lambda_i(X))$. The same can be said for $f(Y)$. By part (a) and the assumption, $f(\lambda_i(X)) \leq f(\lambda_i(Y))$. Summing these gives the result since the trace is the sum of the eigenvalues.
(c) From Remark 4.1.12, $\|X\| \leq a$ implies $u^T X u \leq a$ for all unit vectors u. Therefore $u^T(aI_n - X)u \geq 0$ for all u, meaning $aI_n - X \succeq 0$, thus $X \preceq aI_n$. For the other inequality, again from Remark 4.1.12, $u^T X u \geq -a$ for all unit vectors u. Following the exact procedure above gives $X \succeq -aI_n$.
(d) By considering $g - f$, we can assume that $f = 0$. If $\|X\| \leq a$, then all eigenvalues of X satisfy $|\lambda_i| \leq a$, which implies $g(\lambda_i) \geq 0$ 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 matric absolute value)**.** (c) of Proposition 5.4.4 looks quite familiar... It is a matrix version of the basic fact about absolute values: for $x \in \mathbb{R}$,
$$|x| \leq a \iff -a \leq x \leq a.$$
This makes the operator norm $\|\cdot\|$ a natural matrix version of the absolute value $|\cdot|$, and that's why it appeares in the matrix Bernstein inequality (Theorem 5.4.1).

Remark 5.4.6 (Matrix monotonicity)**.** Can we strenghten trace monotonicity (Proposition 5.4.4 (b)) to matrix monotonicity, i.e.
$$X \preceq Y \implies f(X) \preceq f(Y) \text{ for any weakly increasing } f : \mathbb{R} \to \mathbb{R}?$$
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 the above holds even for non-commuting matrices:
$$0 \preceq X \preceq Y \implies X^{-1} \succeq Y^{-1} \succeq 0 \text{ and } \log X \preceq \log Y$$
whenever X is invertible (Exercise 5.18).

5.4.2 Trace Inequalities

Here is another identity that works for real numbers but not for matrices in general: $e^{x+y} = e^x e^y$ for scalars, but in Exercise 5.19, there are $n \times n$ symmetric matrices X, Y such that

$$e^{X+Y} \neq e^X e^Y.$$

This is unfortunate, because when using the exponential moment method, we relied on this property to split the MGF via independence.
Nevertheless, there are useful substitutes for the missing identity. In particular, this subsection covers two of them, both belonging to the rich family of *trace inequalities*.

Theorem 5.4.7 (Golden-Thompson inequality)**.** For any $n \times n$ symmetric matrcies A and B,

$$\operatorname{tr}(e^{A+B}) \leq \operatorname{tr}(e^A e^B).$$

Note that this does not hold for three or more matrices (we can find counterexamples)!

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 PSD $n \times n$ symmetric matrices.

If X is a random matrix, then Lieb and Jensen inequalities imply that

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

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

$$\mathbb{E}[\operatorname{tr}(\exp(H + Z))] \leq \operatorname{tr}(\exp(H + \log \mathbb{E}[e^Z])).$$

5.4.3 Proof of Matrix Bernstein Inequality

(**Step 1: Reduction of 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. Consider the largest eigenvalue

$$\lambda_{\max}(S) := \max_i \lambda_i(S)$$

and note that

$$\|S\| = \max_i |\lambda_i(S)| = \max(\lambda_{\max}(S), \lambda_{\max}(-S)).$$

To bound $\lambda_{\max}(S)$, we'll use the exponential moment method again. Fix $\lambda > 0$. Via the typical procedure,

$$P(\lambda_{\max}(S) \geq t) = P(e^{\lambda \cdot \lambda_{\max}} \geq e^{\lambda t}) \leq e^{-\lambda t}\mathbb{E}[e^{\lambda \cdot \lambda_{\max}}].$$

Then by Definition 5.4.2, the eigenvalues of $e^{\lambda S}$ are $e^{\lambda \cdot \lambda_i(S)}$ so

$$E := \mathbb{E}[e^{\lambda \cdot \lambda_{\max}(S)}] = \mathbb{E}[\lambda_{\max}(e^{\lambda S})].$$

Since the the eigenvalues of $e^{\lambda S}$ are all positive, the maximal eigenvalue is bounded by the sum of all eigenvalues, which is the trace. Therefore

$$E \leq \mathbb{E}[\operatorname{tr}(e^{\lambda S})].$$

(**Step 2: Application of Lieb Inequality**) To use the Lieb inequality (Lemma 5.4.9), we seperate the last term from the sum S:

$$E \leq \mathbb{E}\left[\operatorname{tr}\left(\exp\left(\sum_{i=1}^{N-1} \lambda X_i + \lambda X_N\right)\right)\right].$$

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 get

$$E \leq \mathbb{E}[\operatorname{tr}(\exp\left(\sum_{i=1}^{N-1} \lambda X_i + \log \mathbb{E}[e^{\lambda X_N}]\right))].$$

Then we continue the same procedure above: seperate λX_{N-1} and apply Lemma 5.4.9, and do the same thing for N times to get

$$E \leq \operatorname{tr}\left(\exp\left[\sum_{i=1}^{N} \log \mathbb{E}[e^{\lambda X_i}]\right]\right).$$

(**Step 3: MGF of the individual terms**) We'll bound the matrix-values MGF via the following lemma:

Lemma 5.4.10 (Matrix MGF)**.** Let X be an $n \times n$ symmetric mean zero random matrix such that $\|X\| \leq K$ almost surely. Then

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

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

Proof. First, we can bound the (scalar) exponential function by the first few terms via Taylor expansion:

$$e^z \leq 1 + z + \frac{1}{1 - |z|/3} \cdot \frac{z^2}{2}, \ |z| < 3.$$

(To get this inequality, write $e^Z = 1 + z + z^2 \sum_{p=2}^{\infty} z^{p-2}/p!$ and use the bound $p! \geq 2 \cdot 3^{p-2}$). Next, apply this inequality to $z = \lambda x$. If $|x| \leq K$ and $|\lambda| < 3/K$ then we obtain

$$e^{\lambda x} \leq 1 + \lambda x + g(\lambda)x^2,$$

where $g(\lambda)$ is the same as how we defined in the statement.
Then we can upgrade this to a matrix inequality using Proposition 5.4.4 (d). If $\|X\| \leq K$ and $|\lambda| < 3/K$, then

$$e^{\lambda X} \preceq I + \lambda X + g(\lambda)X^2.$$

Taking expectations on both sides, since $\mathbb{E}[X] = 0$,

$$\mathbb{E}[e^{\lambda X}] \preceq I + g(\lambda)\mathbb{E}[X^2].$$

To complete the proof of the lemma, let's use the inequality $1 + z \leq e^z$. We can transform this into a matrix inequality via Proposition 5.4.4 (d) and get $I + Z \preceq e^Z$ holds for all matrices Z, and in particular for $Z = g(\lambda)\mathbb{E}[X^2]$. □

(**Step 4: Completion of the proof**) Using Lemma 5.4.10, we obtain

$$E \leq \operatorname{tr}\left(\exp\left(\sum_{i=1}^{N} \log \mathbb{E}[e^{\lambda X_i}]\right)\right) \leq \operatorname{tr}(\exp(g(\lambda)Z)), \text{ where } Z := \sum_{i=1}^{N} \mathbb{E}[X_i^2].$$

where we used matric monotonicity of $\ln x$ (Remark 5.4.6) to take logarithms on both sides, summed up the results, and used trace monotonicity (Proposition 5.4.4 (b)) to take traces of the exponential of 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, hence

$$\begin{aligned} E &\le n\lambda_{\max}(\exp(g(\lambda)Z)) \\ &= m\exp(g(\lambda)\lambda_{\max}(Z)) \\ &= n\exp(g(\lambda)\|Z\|) \quad \text{(Since } Z \succeq 0\text{)} \\ &= n\exp(g(\lambda)\sigma^2) \quad \text{(By definition of } \sigma\text{)}. \end{aligned}$$

Plugging in this bounde for $E = \mathbb{E}[e^{\lambda\cdot\lambda_{\max}(S)}]$ into the original equation gives

$$P(\lambda_{\max}(S) \ge t) \le n\exp(-\lambda t + g(\lambda)\sigma^2).$$

The above is a bound that holds for any $\lambda > 0$ as long as $|\lambda| < 3/K$, so we can minimize it in λ. Better yet, instead of computing the exact minimum (which can be quite ugly), we can choose the following value: $\lambda = t/(\sigma^2 + Kt/3)$, and substituting this value back gives

$$P(\lambda_{\max}(S) \ge t) \le n\exp\left(-\frac{t^2/2}{\sigma^2 + Kt/3}\right).$$

Repeating the argument for $-S$, we will get the same bound as the above, and summing up the two bounds completes the proof. □

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 Theorem 5.4.1 and the integrated tail formula (Lemma 1.6.1), we can deduce that (Exercise 5.20)

$$\mathbb{E}\left[\|\sum_{i=1}^{N} X_i\|\right] \lesssim \|\sum_{i=1}^{N}\mathbb{E}[X_i^2]\|^{1/2}\sqrt{\log(2n)} + K\log(2n)$$

where the $\lesssim$ symbol hides an absolute constant factor. In the scalar case ($n = 1$), an expectation bound is trivial: the variance of sum formula gives

$$\mathbb{E}\left[\left|\sum_{i=1}^{N} X_i\right|\right] \le \left(\mathbb{E}\left[\left|\sum_{i=1}^{N} X_i\right|^2\right]\right)^{1/2} = \left(\sum_{i=1}^{N}\mathbb{E}[X_i^2]\right)^{1/2}.$$

Remark 5.4.12 (The logarithmic price)**.** For the equation in Remark 5.4.11, the high-dimensional version differs the 1-dimensional one by just a logarithmic factor. This is a surprisingly small price for high dimensions! Moreover, this price is in essentially optimal - Exercise 5.28 gives an example of why we can't get rid of it.

5.4.4 Matrix Hoeffding and Khintchine Inequalities

Theorem 5.4.13 (Matrix Hoeffding inequality)**.** Let $\varepsilon_1, \dots, \varepsilon_N$ be independent Rademacher random variables and $A_1, \dots A_N$ be any (fixed) symmetric $n \times n$ matrices. Then for any $t > 0$,

$$P\left(\|\sum_{i=1}^{N}\varepsilon_i A_i\| \ge t\right) \le 2n\exp\left(-\frac{t^2}{2\sigma^2}\right)$$

where $\sigma^2 = \|\sum_{i=1}^{N} A_i^2\|$.

Proof. Exercise 5.21. □

Theorem 5.4.14 (Matric Khintchine inequality)**.** Let $\varepsilon_1, \dots, \varepsilon_N$ be independent Rademacher random variables and $A_1, \dots A_N$ be any (fixed) symmetric $n \times n$ matrices. Then for every $p \in [1, \infty)$, we have

$$\left(\mathbb{E}\left[\|\sum_{i=1}^{N} \varepsilon_i A_i\|^p\right]\right)^{1/p} \leq C\sqrt{p + \log n}\|\sum_{i=1}^{N} A_i^2\|^{1/2}.$$

Proof. Exercise 5.22. Use the matrix Hoeffding inequality. □

Remark 5.4.15 (Non-symmetric, rectangular matrices)**.** Matrix concentration inequalities easily extend to rectangular matrices using the *Hermitian dilation* introduced in Exercise 4.14. Replace each matrix X_i with the symmetric block matrix

$$\begin{bmatrix} 0 & X_i \\ X_i^T & 0 \end{bmatrix}$$

and apply usual matrix concentration. We can get the matrix Bernstein (Exercise 5.23) and Khintchine (Exercise 5.24) inequalities for rectangular matrices this way.

5.5 Application: Community Detection in Sparse Networks

In section 4.5, the method of *spectral clustering* was introduced, which is a basic method for community detection in networks. We showed that it works for relatively dense networks, where the expected average degree is $\gtrsim \sqrt{n}$. Now, using the 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 $p = a/n, q = b/n$ and $b < a < 3b$. Assume that

$$(a - b)^2 \geq Ca \log n.$$

Then, with probability at least 0.99, the spectral clustering algorithm identifies the communities of G with 99% accuracy, i.e. misclassifying at most $0.01n$ vertices.

Proof. We'll follow that same proof as that from Section 4.5, but with a sharper error bound.
Step 1: Decomposition. Again, let's split A into the deterministic and random parts:

$$A = D + R \text{ where } D = \mathbb{E}\left[A\right].$$

Before, the analysis is mostly on the deterministic matrix D, where the second largest eigenvector has ± 1 coefficients representing community membership. Now we have to analyze the random part

$$R = A - \mathbb{E}\left[A\right].$$

Let's decompose it entry by entry, keeping symmetry in mind. We can write R as a sum of independent, mean-zero random matrices Z_{ij} that isolate entries (i, j) and (j, i):

$$R = \sum_{i \leq j} Z_{ij}, \text{ where } Z_{ij} = \begin{cases} R_{ij}(e_i e_j^T + e_j e_i^T) \text{ if i ¡ j,} \\ R_{ii} e_i e_i^T & \text{if } i = j \end{cases}.$$

Step 2: Bounding the error. Since $A_{ij} \in \{0, 1\}$,

$$|R_{ij}| \leq 1 \implies \|Z_{ij}\| = \|R_{ij}\| \leq 1 \implies \|R_{ij} Z_{ij}\| \leq 1.$$

Then by applying the matrix Bernstein inequality (Remark 5.4.11) combined with Markov's inequality, we obtain with probability at least 0.99:

$$\|R\| \lesssim \sigma\sqrt{\log n} + \log n \text{ where } \sigma^2 = \|\mathbb{E}\left[\sum_{i\leq j} Z_{ij}^2\right]\|.$$

Let's compute σ^2. A quick check shows thaaht Z_{ij}^2 is a diagonal matrix:

$$Z_{ij}^2 = \begin{cases} R_{ij}^2(e_ie_i^T + e_je_j^T) & \text{if } i < j, \\ R_{ii}^2 e_ie_i^T & \text{if } i = j \end{cases}.$$

Then, by symmetry,

$$\sum_{i\leq j} Z_{ij}^2 = \sum_{i\leq j} R_{ij}^2(e_ie_i^T + e_je_j^T) + \sum_i R_{ii}^2 e_ie_i^T = \sum_{i=1}^n \left(\sum_{j=1}^n R_{ij}^2\right) e_ie_i^T.$$

This is a diagonal matrix, and so is its expectation. Thus

$$\sigma^2 = \|\mathbb{E}\left[\sum_{i\leq j} Z_{ij}^2\right]\| = \max_{i=1,\ldots,n} \sum_{j=1}^n \mathbb{E}\left[R_{ij}^2\right]$$

since the operator norm of a diagonal matrix is the maximal absolue 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 Ber(p) or Ber(q). So $\mathbb{E}\left[R_{ij}^2\right] = \mathrm{Var}(A_{ij}) \leq p$ since $p > q$. Thus

$$\sigma^2 \leq np = a,$$

and substituting this into the initial bound for $\|R\|$, we get

$$\|R\| \lesssim \sqrt{a \log n} + \log n \lesssim \sqrt{a \log n}$$

because the assumption implies that $a \gtrsim \log n$.

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 seperation 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 $a \leq 3b$ by assumption. Using the bound on R from the end of Step 2, 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

$$\|\bar{u}_2(D) - \theta\bar{u}_2(A)\|_2 \leq \frac{2\|R\|}{\delta} \leq \frac{C_1\sqrt{a \log n}}{a-b} < \frac{1}{10}$$

if we choose the constant C in the assumption of the theorem to be large enough. Multiply both sides by $\sqrt{n}$ to 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 in $\theta u_2(A)_j$ have the same sign as $u_2(D)_j$, 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 the bound of $O(\sqrt{n})$ that we have acheived previously (Remark 4.5.3)!

5.6 Application: Covariance Estimation for General Distributions

In Section 4.7, we learned how to estimate the covariance metrix of a subgaussian distribution in $\mathbb{R}^n$ from a sample of size $O(n)$. Now, we drop the subgaussian assumption, making this work for much broader distributions, even discrete ones. The trade-odd is just a logarithmic oversampling factor!
For noration, we denote the sample covariance matrix as

$$\Sigma_m = \frac{1}{m}\sum_{i=1}^m X_i X_i^T.$$

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 $\mathbb{R}^n$ $(n \geq 2)$. Assume that for some $K \geq 1$,

$$\|X\|_2 \leq K(\mathbb{E}\left[\|X\|_2^2\right])^{1/2} \text{ almost surely.}$$

Then for every positive integer m, we have

$$\mathbb{E}\left[\|\Sigma_m - \Sigma\|\right] \leq C\left(\sqrt{\frac{K^2 n \log n}{m}} + \frac{K^2 n \log n}{m}\right)\|\Sigma\|.$$

Proof. By Proposition 3.2.1 (b), we have $\mathbb{E}\left[\|X\|_2^2\right] = \mathrm{tr}(\Sigma)$, hence the condition in the theorem becomes

$$\|X\|_2^2 \leq K^2 \mathrm{tr}(\Sigma) \text{ almost surely.}$$

Apply the expected version of the matrix Bernstein inequality (Remark 5.4.11) for the sum of i.i.d. mean zero random matrices $X_i X_i^T - \Sigma$ and get

$$\mathbb{E}\left[\|\Sigma_m - \Sigma\|\right] = \frac{1}{m}\mathbb{E}\left[\|\sum_{i=1}^m (X_i X_i^T - \Sigma)\|\right] \lesssim \frac{1}{m}(\sigma\sqrt{\log n} + M\log n)$$

where

$$\sigma^2 = \|\sum_{i=1}^m \mathbb{E}\left[(X_i X_i^t - \Sigma)^2\right]\| = m\|\mathbb{E}\left[(XX^T - \Sigma)^2\right]\|$$

and M is any number chosen so that

$$\|XX^T - \Sigma\| \leq M \text{ almost surely.}$$

To complete the proof, it remains to bound σ^2 and M.
Let's start with σ^2. By expanding the square,

$$\mathbb{E}\left[(XX^T - \sigma)^2\right] = \mathbb{E}\left[(XX^T)^2\right] - \Sigma^2 \lesssim \mathbb{E}\left[(XX^T)^2\right]. \quad (*)$$

Furthermore, the assumption at the beginning of the proof gives

$$(XX^T)^2 = \|X\|^2 XX^T \lesssim K^2 \mathrm{tr}(\Sigma) XX^T.$$

Taking expectations on both sides, we get

$$\mathbb{E}\left[(XX^T)^2\right] \lesssim K^2 \mathrm{tr}(\Sigma)\Sigma.$$

Substituting this bound into $(*)$, we get a bound for σ^2:

$$\sigma^2 \leq K^2 m \mathrm{tr}(\Sigma)\|\Sigma\|.$$

On the other hand, bounding M is easier:

$$\begin{aligned} \|XX^T - \Sigma\| &\leq \|X\|_2^2 + \|\Sigma\| \quad \text{(By triangle inequality)} \\ &\leq K^2 \text{tr}(\Sigma) + \|\Sigma\| \quad \text{(By assumption)} \\ &\leq 2K^2 \text{tr}(\Sigma) =: M. \quad (\|\Sigma\| \leq \text{tr}(\Sigma) \text{ and } K \geq 1). \end{aligned}$$

Substitute the bounds for σ and M into the overall bound, we get

$$\mathbb{E}\left[\|\Sigma_m - \Sigma\|\right] \leq \frac{1}{m}\left(\sqrt{K^2 m \text{tr}(\Sigma)\|\Sigma\|} \cdot \log n + 2K^2 \text{tr}(\Sigma) \cdot \log n\right).$$

Finally, plugging in the bound $\text{tr}(\Sigma) \leq n\|\Sigma\|$ completes the proof. □

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:

$$\mathbb{E}\left[\|\Sigma_m - \Sigma\|\right] \leq \varepsilon \|\Sigma\|,$$

as long as the sample size is

$$m \asymp \varepsilon^{-2} n \log n.$$

Compared to the sample complexity $m \asymp \varepsilon^{-2} n$ for subgaussan distributions (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 proof of Theorem 5.6.1, we used a rough bound $\text{tr}(\Sigma) \leq n\|\Sigma\|$. But instead, we can express the conclusion via the *effective rank* of Σ:

$$r = r(\Sigma) = \frac{\text{tr}(\Sigma)}{\|\Sigma\|}$$

and get a sharper bound

$$\mathbb{E}\left[\|\Sigma_m - \Sigma\|\right] \leq C\left(\sqrt{\frac{K^2 r \log n}{m}} + \frac{K^2 r \log n}{m}\right)\|\Sigma\|.$$

It shows that a sample of size

$$m \asymp \varepsilon^{-2} r \log n$$

is enough to estimate the covaraince matrix. Since $r \leq n$, this sample size is at least as small as the value that we had estimated above. 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 from Remark 5.6.3 really tell us about a PSD matrix Σ? TO get an idea, write it as the sum of eigenvalues divided by the largest one:

$$r(\Sigma) = \frac{\sum_{i=1}^n \lambda_i(\Sigma)}{\max_i \lambda_i(\Sigma)}.$$

This is always bounded by the actual rank (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 the *stable rank*, defined for any matrix A

$$s(A) = \frac{\|A\|_F^2}{\|A\|^2} = \frac{\sum_{i=1}^n \sigma_i^2(A)}{\max_i \lambda_i(\sigma)} = r(A^T A) = r(AA^T)$$

where σ_i denotes the singular values. Both are “soft” versions of rank that are stable under small changes. For some more intuition, see Exercise 5.26.

Remark 5.6.5 (High-probability guarantees)**.** We covered expectation bounds, but our argument actually gives a more informative high-probability guarantee:

$$\|\Sigma_m - \Sigma\| \leq C \left(\sqrt{\frac{K^2 r(\log n + u)}{m}} + \frac{K^2 r(\log n + u)}{m} \right) \|\Sigma\|$$

with probability at least $1 - 2e^{-u}$. Here $r = \mathrm{tr}(\Sigma)/\|\Sigma\| \leq n$ is the effective rank (Exercise 5.26).

Remark 5.6.6 (Boundedness assumption)**.** The boundedness assumption in Theorem 5.6.1 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 (Exercise 5.27). However, this assumption can still be relaxed (Exercise 6.34). In practice, it is usually enforced by truncation - dropping a small percentage of samples with the largest norm.

5.7 Extra notes

There are lots of other concentration theorems not went over in the text. A very useful one is the McDiarmid inequality, which generalizes the Hoeffding inequality:

Theorem 5.7.1 (McDiarmid inequality)**.** Let $X = (X_1, \ldots, X_N)$ be a random vector with independent entries. 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 of a single coordinate of $x \in \mathbb{R}^n$. Then for any $t > 0$,

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

6 Quadratic Forms, Symmetrization, and Contraction

This section concerns mostly with decoupling, concentration of quadratic forms, symmetrization, and contraction, which are a number of basic toold of high-dimensional probability.

6.1 Decoupling

We'll look at quadratic forms of the form

$$\sum_{i,j=1}^{n} a_{ij} X_i X_j = X^T A X = \langle X, AX \rangle$$

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 quadratic forms are known as chaos.
We can compute the expectation of a chaos. If X_i have zero means and unit variances, then

$$\mathbb{E}[X^T A X] = \sum_{i,j=1}^{n} a_{ij} \mathbb{E}[X_i X_j] = \sum_{i=1}^{n} a_{ii} = \operatorname{tr}(A).$$

However, establishing concentration on a chaos is harder, because the terms of the sum above are not independent. However, we can overcome this difficulty via decoupling. We'll replace the quadratic form above with the bilinear form

$$\sum_{i,j=1}^{n} a_{ij} X_i X_j' = X^T A X' = \langle X, AX' \rangle ,$$

where $X' = (X_1', \dots, X_n')$ is an independent copy of X. Bilinear forms are easier to analyze than quadratic forms as they are linear in X. Therefore if we condition on X', we may treat the bilinear form as a sum of independent random variables

$$\sum_{i=1}^{n} \left(\sum_{j=1}^{n} a_{ij} X_j' \right) X_i = \sum_{i=1}^{n} b_i X_i$$

with fixed coefficients b_i.

Theorem 6.1.1 (Decoupling)**.** Let A be an $n \times n$ diagonal free matrix, i.e. all diagonal entries are zero. 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 $F : \mathbb{R} \to \mathbb{R}$,

$$\mathbb{E}[F(X^T A X)] \leq \mathbb{E}[F(4X^T A X')].$$

Proof. We'll replace the chaos by a partial chaos, which we extend back to the original chaos later via Jensen's inequality. The partial chaos is defined by

$$\sum_{(i,j) \in I \times I^c} a_{ij} X_i X_j$$

where $I \subset \{1, \dots, n\}$ is a randomly chosen subset of indices.
(**Step 1: Randomly selecting a partial sum**) To specify a random subset of indices I, we'll use selectors - independent Bernoulli random variables $\delta_1, \dots, \delta_n \sim_{iid} \operatorname{Ber}(1/2)$. We define the index set

$$I := \{i : \delta_i = 1\}.$$

Condition on X. Since by assumption $a_{ii} = 0$ and

$$\mathbb{E}[\delta_i (1 - \delta_j)] = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4} \text{ for all } i \neq j$$

we may express the chaos as

$$X^TAX = \sum_{i\neq j} a_{ij}X_iX_j = 4\mathbb{E}_\delta\left[\sum_{i\neq j}\delta_i(1-\delta_j)a_{ij}X_iX_j\right] = 4\mathbb{E}_I\left[\sum_{(i,j\in I\times I^C)} a_{ij}X_iX_j\right].$$

(In the expression above, the subscripts δ and I indicate the source of randomness in the conditional expectations. Since X is fixed, the expectations are taken over the random selection of $\delta = (\delta_1, \dots, \delta_n)$, or equivalently, the random index set I).

(**Step 2: Applying** F) Applying the function F to both sides and take expectation over X. By Jensen inequality and Fubini theorem, we get

$$\mathbb{E}_X[F(X^TAX)] \leq \mathbb{E}_I\left[\mathbb{E}_X\left[F\left(4\sum_{(i,j)\in I\times I^c} a_{ij}X_iX_j\right)\right]\right].$$

It follows that there exists a realization of a subset I such that

$$\mathbb{E}_X[F(X^TAX)] \leq \mathbb{E}_X\left[F\left(4\sum_{(i,j)\in I\times I^c} a_{ij}X_iX_j\right)\right].$$

Fix such a realization I until the end of the proof, and drop the subscript X on 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' hence

$$\mathbb{E}_X[F(X^TAX)] \leq \mathbb{E}\left[F\left(4\sum_{(i,j)\in I\times I^c} a_{ij}X_iX_j'\right)\right].$$

(**Step 3: Completing the partial sum**) It remains to complete the sum on the RHS to the sum over all pairs of indices. We want to show that

$$\mathbb{E}\left[F\left(4\sum_{(i,j)\in I\times I^c} a_{ij}X_iX_j'\right)\right] \leq \mathbb{E}\left[F\left(4\sum_{(i,j)\in [n]\times[n]} a_{ij}X_iX_j'\right)\right]$$

where $[n] = \{1, \dots, n\}$. To do this, we can decompose the sum on 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 expectation by $\mathbb{E}'$. This fixes Y, while Z has zero conditional expectation (check). Thus, by Jensen inequality, we get

$$F(4Y) = F(4Y + \mathbb{E}'[4Z]) = F(\mathbb{E}'[4Y + 4Z]) \leq \mathbb{E}'[F(4Y + 4Z)].$$

Finally, taking expectations over all remaining random variables, we get

$$\mathbb{E}[F(4Y)] \leq \mathbb{E}[F(4Y + 4Z)].$$

Hence the proof is complete. □

Remark 6.1.2 (Diagonal-free assumption)**.** The assumption is essential in Theorem 6.1.1, since the conclusion fails for diagonal matrices when $F(x) = x$. But we can include the diagonal on the right hand side: for any $n \times n$ matrix $A = (a_{ij})$, we get

$$\mathbb{E}\left[F\left(\sum_{i\neq j} a_{ij}X_iX_j\right)\right] \le \mathbb{E}\left[F\left(4\sum_{i,j} a_{ij}X_iX_j'\right)\right]$$

This is shown in Exercise 6.1, and there are other variants of decoupling (Exercises 6.2-6.4).

6.2 Hanson-Wright Inequality

If X is a subgaussian random vector in $\mathbb{R}^n$, what can we say about its norm? If X has indepdendent entries, then it concentrated (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 (Norm of 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 $t \ge 0$,

$$P(\|X\|_2 \ge CK(\sqrt{n}+t)) \le e^{-t^2}.$$

Proof. WLOG, we can assume that $K = 1$. Squaring and exponentiating both sides and using Markov's inequality, we get

$$P(c\|X\|_2 \ge \sqrt{n}+t) \le e^{-(n+t^2)}\mathbb{E}\left[\exp\left(c^2\|X\|_2^2\right)\right].$$

Now we will use a **Gaussian replacement** trick: for some absolute constant $c > 0$, we claim that

$$\mathbb{E}\left[\exp\left(c^2\|X\|_2^2\right)\right] \le \mathbb{E}\left[\exp\left(\|g\|_2^2/6\right)\right] \text{ where } g \sim N(0, I_n).$$

To see this, condition on X (treating it as a fixed vector); then $\langle g, X\rangle \sim N(0, \|X\|_2^2)$ by Corollary 3.3.2, so (using the MGF of a normal distribution)

$$\mathbb{E}\left[\exp\left(c^2\|X\|_2^2\right)\right] = \mathbb{E}_g\left[\exp\left(\sqrt{2}c\,\langle g, X\rangle\right)\right],$$

where $\mathbb{E}_g$ denotes the conditional expectation over g. Now takes expectation over X over both sides and apply Fubini:

$$\mathbb{E}_X\left[\exp\left(c^2\|X\|_2^2\right)\right] = \mathbb{E}_X\left[\mathbb{E}_g\left[\exp\left(\sqrt{2}c\,\langle g, X\rangle\right)\right]\right] = \mathbb{E}_g\left[\mathbb{E}_X\left[\exp\left(\sqrt{2}c\,\langle X, g\rangle\right)\right]\right].$$

When we condition on g (treating g as a fixed vector), the subgaussian norm of $\langle X, g\rangle$ is at most 1 by assumption ($K = 1$), so Proposition 2.6.6 (iv) gives

$$\mathbb{E}_X\left[\exp\left(\sqrt{2}c\,\langle X, g\rangle\right)\right] \le \exp\left(\|g\|_2^2/4\right)$$

for some absolute constant $c > 0$. Substitute this into the bound above, then our claim for the Gaussian replacement is proved. After that, substitute into the first equation, and the proof is complete. □

The Gaussian replacement trick will also be useful when we are proving concentration regarding a chaos - namely, the Hanson-Wright inequality:

Theorem 6.2.2 (Hanson-Wright inequality)**.** Let A be an $n \times n$ matrix. Let $X = (X_1, \dots, X_n) \in \mathbb{R}^n$ be a random vector with independent, mean-zero, subgaussian coordinates. Then, for every $t \ge 0$, we have

$$P(|X^TAX - \mathbb{E}\left[X^TAX\right]| \ge t) \le 2\exp\left[-c\min\left(\frac{t^2}{K^4\|A\|_F^2}, \frac{t}{K^2\|A\|}\right)\right],$$

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

The proof will be based on bounding the MGF of X^TAX. Here is the plan:

(a) replace X^TAX by X^TAX' by decoupling;

(b) replace X^TAX' by g^TAg' using Gaussian replacement, for $g \sim N(0, I_n)$;

(c) compute g^TAg' by diagonalizing A using the rotational invariance of $N(0, I_n)$.

We start with part (b):

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 $\|X\|_{\psi_2} \leq K$, and X' be its independent copy. Let $g, g' \sim N(0, I_n)$ be independent. Then for any $\lambda \in \mathbb{R}$,

$$\mathbb{E}\left[\exp\left(\lambda X^TAX'\right)\right] \leq \mathbb{E}\left[\exp\left(CK^2\lambda g^TAg'\right)\right].$$

Proof. Condition on X' and take expectation over X, which we denote $\mathbb{E}_X$. Then the random variable $\langle X, AX'\rangle$ is (conditionally) subgaussian, with subgaussian norm $\leq K\|AX'\|_2$. Then Proposition 2.6.6 (iv) gives

$$\mathbb{E}_X\left[\exp\left(\lambda X^TAX'\right)\right] \leq \exp\left(C\lambda^2K^2\|AX'\|_2^2\right), \;\; \lambda \in \mathbb{R}.$$

Compare the above to the normal MGF formula. Applied to the normal random variable $g^TAX' = \langle g, AX'\rangle$ (still conditionally on X'), it gives

$$\mathbb{E}_g\left[\exp\left(\mu g^TAX'\right)\right] = \exp\left(\mu^2\|AX'\|_2^2/2\right), \;\; \mu \in \mathbb{R}.$$

Setting $\mu = \sqrt{2C}K\lambda$, we match the right hand sides of the two equations above and obtain

$$\mathbb{E}_X\left[\exp\left(\lambda X^TAX'\right)\right] = \mathbb{E}_g\left[\exp\left(\sqrt{2C}K\lambda g^TAX'\right)\right].$$

Then, taking expectation over X' on both sides, we see that we have replace X by g in the chaos, at a cost of the factor $\sqrt{2C}K$. Repeating the same thing for X', we can replace X' with g' and get another factor of $\sqrt{2C}K$. □

We now move to step (c):

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

$$\mathbb{E}\left[\exp\left(\lambda g^TAg'\right)\right] \leq \exp\left(\lambda^2\|A\|_F^2\right) \text{ whenever } |\lambda| \leq \frac{1}{2\|A\|}.$$

Proof. Let's use rotational invariance of the normal distribution do diagonalize A. With its singular value decomposition, $A = U\Sigma V^T$, we can write

$$g^TAg' = (U^Tg)^T\Sigma(V^Tg').$$

By the rotational invariance of the normal distribution (Proposition 3.3.1), U^Tg and V^Tg' are independent standard normal random vectors in $\mathbb{R}^n$. So,

$$g^TAg \sim g^T\Sigma g' = \sum_{i=1}^n \sigma_i g_i g_i^T.$$

This is a sum of independent random variables, so

$$\mathbb{E}\left[\exp\left(\lambda g^TAg'\right)\right] = \mathbb{E}\left[\prod_i \mathbb{E}\left[\exp\left(\lambda\sigma_i g_i g_i^T\right)\right]\right] = \prod_i \mathbb{E}\left[\exp\left(\lambda\sigma_i g_i g_i^T\right)\right].$$

Now, for each i and $t \in \mathbb{R}$, we have

$$\mathbb{E}\left[\exp\left(tg_ig_i'\right)\right] = \mathbb{E}\left[\exp\left(\frac{t^2g_i^2}{2}\right)\right] = \frac{1}{\sqrt{1-t^2}} \leq \exp\left(t^2\right) \text{ if } t^2 \leq \frac{1}{2}.$$

The first identity is done by conditioning on g_i and using the MGF formula for the normal random variable g'; the other steps are just direct calculations. Substituting this bound with $t = \lambda\sigma_i$ into the product, we get

$$\mathbb{E}\left[\exp\left(\lambda g^T A g'\right)\right] \le \exp\left(\lambda^2 \sum_i \sigma_i^2\right) \text{ if } \lambda^2 \le \frac{1}{2\max\sigma_i^2}.$$

Since σ_i are the singular values of A, $\sum_i \sigma_i^2 = \|A\|_F^2$ and $\max_i \sigma_i = \|A\|$, hence the lemma is proved. □

Now we move to the main proof!

Proof of Hanson-Wright inequality. Without loss of generality, assume $K = 1$. As usual, it is enough to bound the one-sided tail

$$p := P(X^T X - \mathbb{E}\left[X^T AX\right] \ge t).$$

This is because we can find the lower tail by just replacing A with $-A$. By combining the two tails, the proof would be complete.
In terms of the entries of $A = (a_{ij})$, we have

$$X^T AX = \sum_{i,j} a_{ij} X_i X_j \text{ and } \mathbb{E}\left[X^T AX\right] = \sum_i a_{ii}\mathbb{E}\left[X_i^2\right],$$

where we used the mean zero assumption and independence. So

$$X^T AX - \mathbb{E}\left[X^T AX\right] = \sum_i a_{ii}(X_i^2 - \mathbb{E}\left[X_i^2\right]) + \sum_{i\neq j} a_{ij} X_i X_j.$$

The problem then reduces to estimating the diagonal and off-diagonal sums:

$$p \le P\left(\sum_i a_{ii}(X_i^2 - \mathbb{E}\left[X_i^2\right]) \ge t/2\right) + P\left(\sum_{i\neq j} a_{ij} X_i X_j \ge t/2\right) := p_1 + p_2.$$

Let's bound these probabilities!
Step 1: Diagonal sum. Since X_i are independent and subgaussian, $X_i^2 - \mathbb{E}\left[X_i^2\right]$ are independent, mean-zero, and subexponential. Also,

$$\|X_i^2 - \mathbb{E}\left[X_i^2\right]\|_{\psi_2} \lesssim \|X_i^2\|_{\psi_1} \lesssim \|X_i\|_{\psi_2}^2 \lesssim 1.$$

(The above follows from centering andcreflem:2.8.5). Then, Bernstein's inequality (Corollary 2.9.2) gives

$$p_1 \le \exp\left[-c\min\left(\frac{t^2}{\sum_i a_{ii}^2}, \frac{t}{\max_i |a_{ii}|}\right)\right] \le \exp\left[-c\min\left(\frac{t^2}{\|A\|_F^2}, \frac{t}{\|A\|}\right)\right].$$

Step 2: Off-diagonal sum. Now we bound the off-diagonal sum

$$S := \sum_{i\neq j} a_{ij} X_i X_j.$$

Let $\lambda > 0$ be a parameter to be determined later. By Merkov's inequality, we have

$$p_2 = P(S \ge t/2) = p(\lambda S \ge \lambda t/2) \le \exp\left(-\lambda t/2\right)\mathbb{E}\left[\exp\left(\lambda S\right)\right].$$

We get

$$\begin{aligned}\mathbb{E}\left[\exp\left(\lambda S\right)\right] &\le \mathbb{E}\left[\exp\left(4\lambda X^T AX'\right)\right] \quad \text{(By decoupling)}\\ &\le \mathbb{E}\left[\exp\left(C_1\lambda g^T Ag'\right)\right] \quad \text{(By Lemma 6.2.3)}\\ &\le \exp\left(C\lambda^2\|A\|_F^2\right) \quad \text{(By Lemma 6.2.4)}\end{aligned}$$

whenever $|\lambda| \le \frac{1}{2\|A\|}$. Putting this bound into the exponential bound above, we obtain

$$p_2 \le \exp\left(-\lambda t/2 + C\lambda^2\|A\|_F^2\right).$$

Optimizing over $0 \le \lambda \le \frac{1}{2\|A\|}$, we conclude that

$$p_2 \le \exp\left[-c\min\left(\frac{t^2}{\|A\|_F^2}, \frac{t}{\|A\|}\right)\right].$$

To summarize, we obtained the desired bounds for the probabilities of the diagonal deviation p_1 and the off-diagonal deviation p_2. Putting them together, we complete the proof of Theorem 6.2.2. □

6.3 Symmetrization

A random variable X is called symmetric if it has the same distribution as $-X$. A basic example is the Rademacher random variable, which takes values -1 and 1 with equal probabilities. Mean-zero normal random variables are also symmetric, while the exponential and Poisson distributions 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:

Lemma 6.3.1 (Constructing symmetric distributions)**.** Let X be a random variable and ξ be an independent Rademacher random variables. Then

(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. We'll check that ξX is symmetric. For any interval $A \subset \mathbb{R}$, the law of total probability gives

$$\begin{aligned} P(\xi X \in A) &= P(\xi X \in A|\ \xi = 1) \cdot \frac{1}{2} + P(\xi X \in A|\ \xi = -1) \cdot \frac{1}{2} \\ &= \frac{1}{2}(P(X \in A) + P(-X \in A)). \end{aligned}$$

Let's also do this for $-\xi X$:

$$\begin{aligned} P(-\xi X \in A) &= P(-\xi X \in A|\ \xi = 1) \cdot \frac{1}{2} + P(-\xi X \in A|\ \xi = -1) \cdot \frac{1}{2} \\ &= \frac{1}{2}(P(-X \in A) + P(X \in a)). \end{aligned}$$

Therefore ξX and $-\xi X$ have the same CDF, meaning they have the same distribution.
The rest of the proof is in Exercise 6.16. □

Lemma 6.3.2 (Symmetrization)**.** Let $X_1, \dots, X_N$ be independent, mean zero random vectors in a normed space, and let $\varepsilon_1, \dots, \varepsilon_N$ be independent Rademacher random variables. Then

$$\frac{1}{2}\mathbb{E}\left[\left\|\sum_{i=1}^{N} \varepsilon_i X_i\right\|\right] \leq \mathbb{E}\left[\left\|\sum_{i=1}^{N} X_i\right\|\right] \leq 2\mathbb{E}\left[\left\|\sum_{i=1}^{N} \varepsilon_i X_i\right\|\right].$$

Proof. (**Upper bound**) Let (X_i') be an independent copy of (X_i). Since $\sum_{i=1} X_i'$ has mean zero, we have

$$p := \mathbb{E}\left[\left\|\sum_i X_i\right\|\right] \leq \mathbb{E}\left[\left\|\sum_i X_i - \sum_i X_i'\right\|\right] = \mathbb{E}\left[\left\|\sum_i (X_i - X_i')\right\|\right].$$

The inequality above comes from the fact that for independent random vectors Y and Z,

$$\mathbb{E}[Z] = 0 \implies \mathbb{E}[\|Y\|] \leq \mathbb{E}[\|Y + Z\|].$$

Since $X_i - X_i'$ are symmetric random vectors, they have the same distribution as $\varepsilon_i(X_i - X_i')$ by Lemma 6.3.1 (b). Then

$$\begin{aligned} p &\leq \mathbb{E}\left[\left\|\sum_i \varepsilon_i (X_i - X_i')\right\|\right] \\ &\leq \mathbb{E}\left[\left\|\sum_i \varepsilon_i X_i\right\|\right] + \mathbb{E}\left[\left\|\sum_i \varepsilon_i X_i'\right\|\right] \quad \text{(Triangle inequality)} \\ &= 2\mathbb{E}\left[\left\|\sum_i \varepsilon_i X_i\right\|\right] \quad \text{(The two terms are identically distributed).} \end{aligned}$$

(**Lower bound**) The argument is similar as the proof for the upper bound:

$$\begin{aligned}
\mathbb{E}\left[\left\|\sum_i \varepsilon_i X_i\right\|\right] &\le \mathbb{E}\left[\left\|\sum_i \varepsilon_i (X_i - X_i')\right\|\right] \\
&= \mathbb{E}\left[\left\|\sum_i (X_i - X_i')\right\|\right] \quad \text{(Same distribution)} \\
&\le \mathbb{E}\left[\left\|\sum_i X_i\right\|\right] + \mathbb{E}\left[\left\|\sum_i X_i'\right\|\right] \quad \text{(Triangle inequality)} \\
&= 2\mathbb{E}\left[\left\|\sum_i X_i\right\|\right] \quad \text{(Identical distribution).}
\end{aligned}$$

Question: Where did we use X_i's independence? Do we need mean zero for both upper and lower bounds? □

There are also other versions of symmetrization lemmas (Exercises 6.19-6.21).

6.4 Random Matrices with non-i.i.d. Entries

A typical application of symmetrization consist 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 the signs ε_i. Hence this reduces the problems to Rademacher random variables. To illustrate this techinique, let's bound the operator norm of a random matric 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

$$\mathbb{E}\left[\max_i \|A_i\|_2\right] \le \mathbb{E}\left[\|A\|\right] \le C\sqrt{\log n} \cdot \mathbb{E}\left[\max_i \|A_i\|_2\right],$$

where A_i denotes the rows of A_i.

Proof. The lower bound is already done in 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, like the proof of Theorem 5.5.1. Denote the standard basis of $\mathbb{R}^n$ by $e_1, \ldots, e_n$, then A can be expressed as a sum of independent, mean zero random matrices:

$$A = \sum_{i \le j} Z_{ij}, \text{ where } Z_{ij} = \begin{cases} A_{ij}(e_i e_j^T + e_j e_i^T) & \text{if } i < j, \\ A_{ii} e_i e_i^T & \text{if } i = j \end{cases}.$$

By applying symmetrization (Lemma 6.3.2), we get

$$\mathbb{E}\left[\|A\|\right] = \mathbb{E}\left[\left\|\sum_{i \le j} Z_{ij}\right\|\right] \le 2\mathbb{E}\left[\left\|\sum_{i \le j} \varepsilon_{ij} Z_{ij}\right\|\right] \quad (*)$$

where ε_{ij} are independent Rademacher random variables.
Condition on (Z_{ij}), apply the matrix Khintchine inequality (Theorem 5.4.14) for $p = 1$, and take expectation over (Z_{ij}) using the law of total expectation, which gives

$$\mathbb{E}\left[\left\|\sum_{i \le j} \varepsilon_{ij} Z_{ij}\right\|\right] \le C\sqrt{\log n}\mathbb{E}\left[\left\|\sum_{i \le j} Z_{ij}^2\right\|^{1/2}\right]. \quad (**)$$

Since (Z_{ij}) is a diagonal matrix,

$$Z_{ij}^2 = \begin{cases} A_{ij}^2(e_i e_j^T + e_j e_i^T) & \text{if } i < j, \\ A_{ii}^2 e_i e_i^T & \text{if } i = j \end{cases}.$$

Therefore,

$$\sum_{i\le j} Z_{ij}^2 = \sum_{i=1}^{n}\left(\sum_{j=1}^{n} A_{ij}^2\right) e_i e_i^T = \sum_{i=1}^{n} \|A_i\|_2^2 e_i e_i^T.$$

In other words, this is a diagonal matrix with diagonal entries equal to $\|A_i\|_2^2$. Since the operator norm of a diagonal matrix is the maximal absolute value of its entries, we get

$$\left\|\sum_{i\le j} Z_{ij}^2\right\| = \max_i \|A_i\|_2^2.$$

Substitute the bound above into (**) then into (*) completes the proof. □

There is more practice on symmetrization as well (Exercises 6.22-6.29).

6.5 Application: Matrix Completion

Matrix completion is the process of 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 recover 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

$$Y_{ij} = \delta_{ij} X_{ij} \text{ where } \delta_{ij} \sim_{i.i.d.} \operatorname{Ber}(p).$$

These δ_{ij} are *selectors.* If

$$p = \frac{m}{n^2}$$

then we observe m entries on average.
The question is, how can we recover X from Y? Although X has small rank r, Y may not have small rank. To fix this, we can pick the best rank r approximation to Y. Properly scaled, this gives a good estimate of the original matrix X:

Theorem 6.5.1 (Matrix completion)**.** Let $\hat{X}$ be a best rank r approximation to $p^{-1}Y$. Then

$$\mathbb{E}\left[\frac{1}{n}\|\hat{X} - X\|_F\right] \le C\sqrt{\frac{rn\log n}{m}}\|X\|_\infty,$$

as long as $m \ge n \log n$. Here $\|X\|_\infty = \max_{i,j} |X_{ij}|$ is the largest entry, NOT the usual matrix infinity norm!

Before we prove this, note that the recovery error

$$\frac{1}{n}\|\hat{X} - X\|_F = \left(\frac{1}{n}\sum_{i,j=1}^{n^2}\sum_{i,j=1}^{n} |\hat{X}_{ij} - X_{ij}|62\right)^{1/2}$$

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 assmumption.

Step 1: Bounding the error in the operator norm. Using the triangle inequality, we can split the error as follows:

$$\|\hat{X} - X\| \leq \|\hat{X} - p^{-1}Y\| + \|p^{-1}Y - X\|.$$

Since we have chosen $\hat{X}$ as a best rank r approximation to $p^{-1}Y$, the second summand dominates, i.e. $\|\hat{X} - p^{-1}Y\| \leq \|p^{-1}Y - X\|$, so we have

$$\|\hat{X} - X\| \leq 2\|p^{-1}Y - X\| = \frac{2}{p}\|Y - pX\|.$$

Note that the matrix $\hat{X}$, which is tricky to handle, is gone in 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, Exercise 6.28), we get

$$\mathbb{E}\left[\|Y - pX\|\right] \leq C\sqrt{\log n}\left(\mathbb{E}\left[\max_{i=1,\dots,n}\|(Y - pX)_{i:}\|_2\right] + \mathbb{E}\left[\max_{j=1,\dots,n}\|(Y - pX)_{:j}\|_2\right]\right). \quad (*)$$

To bound the norms of the rows of $Y - pX$, we write them as

$$\|(Y - pX)_{i:}\|_2^2 = \sum_{j=1}^{n}(\delta_{ij} - p)^2 X_{ij}^2 \leq \sum_{j=1}^{n}(\delta_{ij} - p)^2 \cdot \|X\|_\infty^2,$$

and similarly for columns. These sums of independent random variables can be easily bounded using Bernstein's (or Chernoff's) inequality, which yields (Exercise 6.30)

$$\mathbb{E}\left[\max_{i=1,\dots,n}\sum_{j=1}^{n}(\delta_{ij} - p)^2\right] \lesssim pn.$$

Combining with a similar bound for the columns and substituting into $(*)$, we obtain

$$\mathbb{E}\left[\|Y - pX\|\right] \lesssim \sqrt{pn\log n}\|X\|_\infty.$$

Then, by the bound for $\|\hat{X} - X\|$ from earlier, we get

$$\mathbb{E}\left[\|\hat{X} - X\|\right] \lesssim \sqrt{\frac{n\log n}{p}}\|X\|_\infty.$$

Passing to the Frobenius norm. We have not used the low rank assumption yet, so we'll do this now. Since $\operatorname{rank}(X) \leq r$ by assumption and $\operatorname{rank}(\hat{X}) \leq r$ by construction, we have (Exercise 4.4)

$$\operatorname{rank}(\hat{X} - X) \leq 2r \implies \|\hat{X} - X\|_F \leq \sqrt{2r}\|\hat{X} - X\|.$$

Taking expectations and using the bound on the error in the operator norm from step 1, we get

$$\mathbb{E}\left[\|\hat{X} - X\|_F\right] \lesssim \sqrt{\frac{rn\log n}{p}}\|X\|_\infty.$$

Dividing both sides by n, we can rewrite this bound as

$$\mathbb{E}\left[\frac{1}{n}\|\hat{X} - X\|_F\right] \lesssim \sqrt{\frac{rn\log n}{pn^2}}\|X\|_\infty.$$

From the definition above, $pn^2 = m$ so plugging in finishes the proof. □

Remark 6.5.2 (Extensions)**.** Theorem 6.5.1 can be extended and improved in many ways, such as to rectangular matrices (Exercise 6.31) and matrices with noisy observations (Exercise 6.32). It is less trivial but possible to remove the logarithmic factor from the error bound, and acheive zero eror for noiseless observations!

6.6 Contraction Principle

There is one more useful inequality the text covers in the chapter:

Theorem 6.6.1 (Contraction principle)**.** Let $x_1, \ldots, x_N$ be any vectors in a normed space, $(a_1, \ldots, a_N) \in \mathbb{R}^N$, and $\varepsilon_1, \ldots, \varepsilon_N$ be independent Rademacher random variables. Then

$$\mathbb{E}\left[\left\|\sum_{i=1}^N a_i \varepsilon_i x_i\right\|\right] \leq \|a\|_\infty \cdot \mathbb{E}\left[\left\|\sum_{i=1}^N \varepsilon_i x_i\right\|\right].$$

Proof. WLOG, assume that $\|a\|_\infty \leq 1$. Define the function

$$f(a) := \mathbb{E}\left[\left\|\sum_{i=1}^N a_i \varepsilon_i x_i\right\|\right].$$

Then $f : \mathbb{R}^n \to \mathbb{R}$ is convex (Exercise 6.35).
We want to bound for f the set of points a satisfying $\|a\|_\infty \leq 1$, i.e. on the unit cube $[-1,1]^N$. By the maximum principle (Exercises 1.4 & 1.5), the maximum of a convex function 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

$$\mathbb{E}\left[\left\|\sum_{i=1}^N a_i \varepsilon_i x_i\right\|\right] = \mathbb{E}\left[\left\|\sum_{i=1}^N \varepsilon_i x_i\right\|\right],$$

thus

$$f(a) \leq \mathbb{E}\left[\left\|\sum_{i=1}^N \varepsilon_i x_i\right\|\right] \quad \text{whenever } \|a\|_\infty \leq 1,$$

which completes the proof. □

As an application, we can prove a version of symmetrization 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

$$\frac{c}{\sqrt{\log N}}\mathbb{E}\left[\left\|\sum_{i=1}^N g_i X_i\right\|\right] \leq \mathbb{E}\left[\left\|\sum_{i=1}^N X_i\right\|\right] \leq 3\mathbb{E}\left[\left\|\sum_{i=1}^N g_i X_i\right\|\right].$$

Proof. **(Upper bound)** By symmetrization (Lemma 6.3.2), we have

$$E := \mathbb{E}\left[\left\|\sum_{i=1}^N X_i\right\|\right] \leq 2\mathbb{E}\left[\left\|\sum_{i=1}^N \varepsilon_i X_i\right\|\right].$$

To interject Gaussian random variables, recall that $\mathbb{E}\left[|g_i|\right] = \sqrt{2/\pi}$. Then we can continue the bound as follows:

$$\begin{aligned}
E &\leq 2\sqrt{\frac{\pi}{2}}\mathbb{E}_X\left[\left\|\sum_{i=1}^{N}\varepsilon_i\mathbb{E}_g\left[|g_i|\right]X_i\right\|\right] \\
&\leq 2\sqrt{\frac{\pi}{2}}\mathbb{E}\left[\left\|\sum_{i=1}^{N}\varepsilon_i|g_i|X_i\right\|\right] \quad \text{(Jensen inequality)} \\
&= 2\sqrt{\frac{\pi}{2}}\mathbb{E}\left[\left\|\sum_{i=1}^{N}g_iX_i\right\|\right].
\end{aligned}$$

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) We have

$$\begin{aligned}
\mathbb{E}\left[\left\|\sum_{i=1}^{N}g_iX_i\right\|\right] &= \mathbb{E}\left[\left\|\sum_{i=1}^{N}\varepsilon_ig_iX_i\right\|\right] \quad \text{(Symmetry of } g_i\text{)} \\
&\leq \mathbb{E}_g\left[\mathbb{E}_X\left[\|g\|_\infty\mathbb{E}_\varepsilon\left[\left\|\sum_{i=1}^{N}\varepsilon_iX_i\right\|\right]\right]\right] \quad \text{(Theorem 6.6.1)} \\
&= \mathbb{E}_g\left[\|g\|_\infty\mathbb{E}_\varepsilon\left[\mathbb{E}_X\left[\left\|\sum_{i=1}^{N}\varepsilon_iX_i\right\|\right]\right]\right] \quad \text{(Independence)} \\
&\leq 2\mathbb{E}_g\left[\|g\|_\infty\mathbb{E}_X\left[\left\|\sum_{i=1}^{N}X_i\right\|\right]\right] \quad \text{(Lemma 6.3.2)} \\
&= 2\mathbb{E}\left[\|g\|_\infty\right]\cdot\mathbb{E}\left[\left\|\sum_{i=1}^{N}X_i\right\|\right] \quad \text{(Independence).}
\end{aligned}$$

Moreover, by Proposition 2.7.6,

$$\mathbb{E}\left[\|g\|_\infty\right] \leq C\sqrt{\log N}.$$

Plugging back gives the result. □

Remark 6.6.3 (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.

7 Random Processes

This chapter concerns mostly with random processes - collection random variables $(X_t)_{t\in T}$, which may be dependent. In calssical settings like Brownian motion, t represents time so $T \subset \mathbb{R}$. However, in high-dimensional probability T can be any set, and we'll deal with Gaussian processes a lot.
In this chapter, we'll explore powerful comparison inequalities for Gaussian processes - Slepian, Sudakov-Frenique, and Gordon - by using a new trick: Gaussian interpolation. Then we use these tools to prove a sharp bound on the operator norm of $m \times n$ Gaussian random matrices.
How does a Gaussian process $(X_t)_{t\in T}$ capture the geometry of T? We'll prove a lower bound on the Gaussian width using covering numbers, and link it to other ideas like effective dimension. Moreover, we'll also compute the size of a ranodm projection of any bounded set $T \subset \mathbb{R}^n$, which heavily depends on the Gaussian width.

7.1 Basic Concepts and Examples

Definition 7.1.1. A random process is a collection of random variables $(X_t)_{t\in T}$ on the same probability space, which are indexed by elements t of some index set T.

Example 7.1.2 (Discrete time)**.** If $T = \{1, \dots, n\}$ then the random process

$$(X_1, \dots, X_n)$$

can be identifies as 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 an illustration:

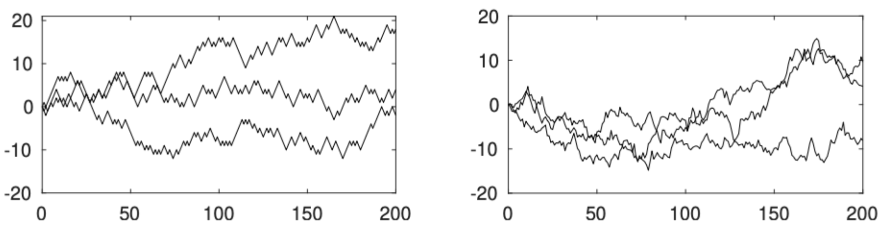


Figure 7.1 A few trials of a random walk (left) and standard Brownian motion (right).

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 above also shows some sample paths of a 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 spatial random process, or *random field.* For example, the water temperature X_t ar the location on Earth that is parameterized by t can be modeled as a spatial random process.

7.1.1 Covariance and Increments

In section 3.2, we introduced the covariance matrix of a random vector. Here we'll define the *covariance function* of a random process $(X_t)_{t\in T}$ in a similar manner. For simplicity, assume the random process has zero mean:

$$\mathbb{E}\left[X_t\right] = 0 \text{ for all } t \in T.$$

The covariance function of the process is defined as

$$\Sigma(t,s) := \operatorname{Cov}(X_t, X_s) = \mathbb{E}\left[X_t X_s\right],\ t,s \in T.$$

The increments of the random process are defined as

$$d(t,s) := \|X_t - X_s\|_{L^2} = \left(\mathbb{E}\left[(X_t - X_s)^2\right]\right)^{1/2},\ t,s \in T.$$

Example 7.1.6. The increments of the standard Brownian motion satisfy

$$d(t,s) = \sqrt{t-s},\ t \geq s$$

by definition. The increments of a random walk of Example 7.1.3 with $\mathbb{E}\left[Z_i^2\right] = 1$ behave similarly:

$$d(n,m) = \sqrt{n-m},\ n \geq m.$$

Remark 7.1.7 (The canonical metric)**.** Even if the index set T has no geometric structure, the increments $d(t,s)$ always define a metric on T, thys automatically turning T into a metric space. 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 v.s. increments)**.** The covariance and the increments contain roughly the same information about the random process. Increments can be writeen using the covariance: Just expand the square to see that

$$d(t,s)^2 = \Sigma(t,t) - 2\Sigma(t,s) + \Sigma(s,s).$$

Vise versa, if the zero random variable belongs to the process, we can also recover the covariance from the increments (Exercise 7.1).

7.1.2 Gaussian Processes

Definition 7.1.9. 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 a 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 (Exercise 3.16).

The notion of Gaussian processes generalized 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 (Proposition 3.3.5). The same goes for a mean-zero Gaussian process: its distribution is determined by the covariance function $\Sigma(t,s)$, or equivalently 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) applies:

Theorem 7.1.11 (Concentration of Gaussian processes)**.** Let $(X_t)_{t\in T}$ be a Gaussian process with finite T. Then

$$\|\sup_{t\in T} X_t - \mathbb{E}\left[\sup_t X_t\right]\|_{\psi_2} \le C \sup_{t\in T} \sqrt{\mathrm{Var}(X_t)}.$$

Proof. Exercise 5.9(b). □

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

$$X_t := \langle g, t\rangle\,,\ t \in T.$$

This guves us a Gaussian process $(X_t)_{t\in T}$ called the *canonical Gaussian process*. The increments match the Euclidean distance:

$$\|X_t - X_s\|_{L^2} = \|t-s\|_2,\ t, s \in T.$$

Actually, one can realize any Gaussian process as the canonical process above because of the lemma below:

Lemma 7.1.12 (Gaussian random vectors)**.** Let X be a mean-zero Gaussian random vector in $\mathbb{R}^n$. Then there exist points $t_1, \ldots, t_n$ such that

$$X \sim (\langle g, t_i\rangle)_{i=1}^n,\text{ where } g \sim N(0, I_n).$$

Proof. IF Σ denotes the covariance matrix of X, then

$$X \equiv \Sigma^{1/2} g \text{ where } g \sim N(0, I_n).$$

The entries of $\Sigma^{1/2}g$ are $\langle t_i, g\rangle$ where the t_i are the rows of $\Sigma^{1/2}$. Done! □

It follows that for any Gaussian process $(X_s)_{x\in S}$, all finite-dimensional marginas $(X_s)_{s\in S_0}$, $|S_0| = n$ can be represented as the canonical Gaussian process indexed in a certain subset $T_0 \subset \mathbb{R}^n$.

7.2 Slepian, Sudakov-Fernique, and Gordon Inequalities

In many applications, it helps to have a *uniform* bound on a random process:

$$\mathbb{E}\left[\sup_{t\in T} X_t\right] = ?$$

Remark 7.2.1 (Making T finite)**.** To avoid measurability issues, let's think of

$$\mathbb{E}\left[\sup_{t\in T} X_t\right] \text{ as shorthand for } \sup_{T_0\subset T} \mathbb{E}\left[\max_{t\in T_0} X_t\right]$$

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

$$\mathbb{E}\left[\sup_{t\le t_0} X_t\right] = \sqrt{\frac{2t_0}{\pi}} \text{ for every } t_0 \ge 0.$$

For general random processes - evern 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

$$\mathbb{E}\left[X_t^2\right] = \mathbb{E}\left[Y_t^2\right] \text{ and } \mathbb{E}\left[(X_t - X_s)^2\right] \le \mathbb{E}\left[(Y_t - Y_s)^2\right].$$

Then $\sup_{t\in T} X_t$ is stochastically dominated by $\sup_{t\in T} Y_t$: For every $\tau \in \mathbb{R}$,

$$P\left(\sup_{t\in T} X_t \ge \tau\right) \le P\left(\sup_{t\in T} Y_t \ge \tau\right).$$

Consequently,

$$\mathbb{E}\left[\sup_{t\in T} X_t\right] \le \mathbb{E}\left[\sup_{t\in T} Y_t\right].$$

We'll provide a proof later in the chapter, as we need some preliminary knowledge on Gaussian interpolation.

7.2.1 Gaussian Interpolation

Assume that T is finite; then we can loot at $X = (X_t)_{t\in T}$ and $Y = (Y_t)_{t\in T}$ as Gaussian random vectors in $\mathbb{R}^n$ with $n = |T|$. We may also assume that X and Y are independent.
Define the Gaussian random vector $Z(u)$ in $\mathbb{R}^n$ that continuously interpolates between $Z(0) = Y$ and $Z(1) = X$:

$$Z(u) := \sqrt{u}X + \sqrt{1-u}Y, \ u \in [0, 1].$$

Then the covariance matrix of $Z(u)$ continuously interpolates linearly between the covariance matrices of Y and X:

$$\Sigma(Z(u)) = u\Sigma(X) + (1-u)\Sigma(Y).$$

This is because

$$\begin{aligned}
\Sigma(Z(u)) &= \mathbb{E}\left[Z(u)Z(u)^T\right] \\
&= \mathbb{E}\left[(\sqrt{u}X + \sqrt{1-u}Y)(\sqrt{u}X + \sqrt{1-u}Y)^T\right] \\
&= u\mathbb{E}\left[(X-\mu_X)(X-\mu_X)^T\right] + \sqrt{u(1-u)}\mathbb{E}\left[(X-\mu_X)(Y-\mu_Y)^T\right] \\
&\quad + \sqrt{u(1-u)}\mathbb{E}\left[(Y-\mu_Y)(X-\mu_X)^T\right] + (1-u)\mathbb{E}\left[(Y-\mu_Y)(Y-\mu_Y)^T\right] \\
&= u\Sigma(X) + 0 + 0 + (1-u)\Sigma(Y) \quad \text{(Independence)} \\
&= u\Sigma(X) + (1-u)\Sigma(Y).
\end{aligned}$$

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'll be able to show that in this case, $\mathbb{E}[f(Z(u))]$ increases in u. This would imply the conclusion of Slepian inequality, since then

$$\mathbb{E}[f(Z(1))] \ge \mathbb{E}[f(Z(0))] \implies P\left(\max_i X_i < \tau\right) \ge P\left(\max_i Y_i < \tau\right)$$

as claimed.
Let's start via 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

$$\mathbb{E}[Xf(X)] = \mathbb{E}[f'(X)].$$

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:

$$\begin{aligned}\mathbb{E}\left[f'(X)\right] &= \int_{\mathbb{R}} f'(x)p(x)\ dx \\ &= \left[f(x)p(x)\right]_{-\infty}^{\infty} - \int_{\mathbb{R}} f(x)p'(x)\ dx \\ &= 0 - \int_{\mathbb{R}} f(x)p'(x)\ dx \\ &= -\int_{\mathbb{R}} f(x)p'(x)\ dx.\end{aligned}$$

We have already proved before (Exercise 2.3) that $p'(x) = -xp(x)$, hence the integral above equals

$$\int_{\mathbb{R}} f(x)p(x)x\ dx = \mathbb{E}\left[Xf(X)\right],$$

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 $X \sim N(0, \sigma^2)$:

$$\mathbb{E}\left[Xf(X)\right] = \sigma^2 \mathbb{E}\left[f'(X)\right].$$

(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

$$\mathbb{E}\left[Xf(X)\right] = \Sigma \cdot \mathbb{E}\left[\nabla f(X)\right]$$

assuming both expectations are finite. In other words,

$$\mathbb{E}\left[X_i f(X)\right] = \sum_{j=1}^{n} \Sigma_{ij} \mathbb{E}\left[\frac{\partial f}{\partial x_j}(X)\right],\ i = 1, \ldots, n.$$

Proof. 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

$$Z(u) := \sqrt{u}X + \sqrt{1-u}Y,\ u \in [0,1].$$

Then for any twice differentiable function $f : \mathbb{R}^n \to \mathbb{R}$, we have

$$\frac{d}{du}\mathbb{E}\left[f(Z(u))\right] = \frac{1}{2}\sum_{i,j=1}^{n} (\Sigma_{ij}^X - \Sigma_{ij}^Y)\mathbb{E}\left[\frac{\partial^2 f}{\partial x_i \partial x_j}(Z(u))\right],$$

assuming all expectations exist and are finite.

Proof. Using the multivariate chain rule,

$$\begin{aligned}\frac{d}{du}\mathbb{E}\left[f(Z(u))\right] &= \sum_{i=1}^{n} \mathbb{E}\left[\frac{\partial f}{\partial x_i}(Z(u))\frac{dZ_i}{du}\right] \\ &= \frac{1}{2}\sum_{i=1}^{n} \mathbb{E}\left[\frac{\partial f}{\partial x_i}(Z(u))\left(\frac{X_i}{\sqrt{u}} - \frac{Y_i}{\sqrt{1-u}}\right)\right].\end{aligned}$$

Let's break the sum above into two, and first compute the contribution of the terms containing X_i. To this end, we condition on Y and express

$$\sum_{i=1}^{n} \frac{1}{\sqrt{u}} \mathbb{E}\left[X_i \frac{\partial f}{\partial x_i}(Z(u))\right] = \sum_{i=1}^{n} \frac{1}{\sqrt{u}} \mathbb{E}\left[X_i g_i(X)\right] \quad (*),$$

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), we get

$$\begin{aligned}\mathbb{E}\left[X_i g_i(X)\right] &= \sum_{j=1}^{n} \Sigma_{ij}^{X} \mathbb{E}\left[\frac{\partial g_i}{\partial x_j}(X)\right] \\ &= \sum_{j=1}^{n} \Sigma_{ij}^{X} \mathbb{E}\left[\frac{\partial^2 f}{\partial x_i \partial x_j}(\sqrt{u}X + \sqrt{1-u}Y)\right] \cdot \sqrt{u}.\end{aligned}$$

Substituting this into $(*)$ to get

$$\sum_{i=1}^{n} \frac{1}{\sqrt{u}} \mathbb{E}\left[X_i \frac{\partial f}{\partial x_i}(Z(u))\right] = \sum_{i,j=1}^{n} \Sigma_{ij}^{X} \mathbb{E}\left[\frac{\partial^2 f}{\partial x_i \partial x_j}(Z(u))\right].$$

Taking expectations on both sides with respect to Y, we left the conditioning on Y.
We can similarly evaluate the other sum (terms containing Y_i) by conditioning on X. Combining the two sums we complete the proof. □

7.2.2 Proof of Slepian Inequality

We'll establish a preliminary, functional form of Spelian's inequality first:

Lemma 7.2.6 (Slepian inequality, functional form)**.** Consider two mean zero Gaussian random vectors X, Y in $\mathbb{R}^n$. Assume that for all $i, j = 1, \ldots, n$, we have

$$\mathbb{E}\left[X_i^2\right] = \mathbb{E}\left[Y_i^2\right] \text{ and } \mathbb{E}\left[(X_i - X_j)^2\right] \leq \mathbb{E}\left[(Y_i - Y_j)^2\right].$$

Consider a twice-differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ such that

$$\frac{\partial^2 f}{\partial x_i \partial x_j} \geq 0 \text{ for all } i, j.$$

Then

$$\mathbb{E}\left[f(X)\right] \geq \mathbb{E}\left[f(Y)\right],$$

assuming both expectations exist and are finite.

Proof. The assumptions imply that the entries of the covariance matrices Σ^X and Σ^Y satisfy

$$\Sigma_{ii}^{X} = \Sigma_{ii}^{Y} \text{ and } \Sigma_{ij}^{X} \geq \Sigma_{ij}^{Y}$$

for all $i, j = 1, \ldots, n$. We can assume that X and Y are independent. Apply Lemma 7.2.5 and using our assumptions, we conclude that

$$\frac{d}{du}\mathbb{E}\left[f(Z(u))\right] \geq 0,$$

so $\mathbb{E}\left[f(Z(u))\right]$ increases in u. Then $\mathbb{E}\left[f(Z(1))\right] = \mathbb{E}\left[f(X)\right]$ is at least as large as $\mathbb{E}\left[f(Z(0))\right] = \mathbb{E}\left[f(Y).\right]$. This completes the proof. □

Now we are ready to prove Slepian's inequality (Theorem 7.2.2). Let's state and prove it in the equivalent form for Gaussian random vectors.

Theorem 7.2.7 (Slepian inequality)**.** Let X, Y be Gaussian random vectors as in Lemma 7.2.6. Then for every $\tau \geq 0$ we have

$$P\left(\max_{i\leq n} X_i \geq \tau\right) \leq P\left(\max_{i\leq n} Y_i \geq \tau\right).$$

Consequently,

$$\mathbb{E}\left[\max_{i\leq n} X_i\right] \leq \mathbb{E}\left[\max_{i\leq n} X_i\right].$$

Proof. Let $h : \mathbb{R} \to [0,1]$ be a twice-differentiable, non-increasing approximation to the indicator function on the interval $(-\infty, \tau)$:

$$h(x) \approx \mathbf{1}_{\{-\infty,\tau\}},$$

like in Figure 7.2 below.

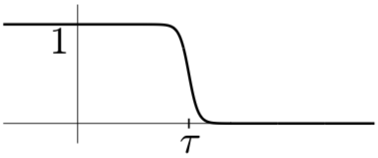


Figure 7.2 The function $h(x)$ is a smooth, non-increasing approximation to the indicator function $\mathbf{1}_{(-\infty,\tau)}$.

Define the function $f : \mathbb{R}^n \to \mathbb{R}$ by

$$f(x) = h(x_1)\cdots h(x_n) = \prod_{i=1}^{n} h(x_i).$$

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 $i \neq j$ we have

$$\frac{\partial^2 f}{\partial x_i \partial x_j} = h'(x_i)h'(x_j) \cdot \prod_{k\notin\{i,j\}} h(x_k).$$

The first two terms are non-positive and the others are nonnegative by assumption, hence the second derivative is nonnegative, as required. It follows that

$$\mathbb{E}\left[f(X)\right] \geq \mathbb{E}\left[f(Y)\right].$$

By approximation, it implies

$$P\left(\max_{i\leq n} X_i < \tau\right) \geq P\left(\max_{i\leq n} Y_i < \tau\right).$$

This proves the first part. The second part follows by using the integrated tail formula in Exercise 1.15 (b):

$$\mathbb{E}\left[f(X)\right] = \int_0^\infty P\left(\max_{i\leq n} X_i \geq \tau\right)\, d\tau \leq \int_0^\infty P\left(\max_{i\leq n} Y_i \geq \tau\right)\, d\tau = \mathbb{E}\left[f(Y)\right].$$

□

7.2.3 Sudakov-Fernique and Gordon Inequalities

Slepian inequality has two assumptions on the processes (X_t) and (Y_t): the equality of varainces and the dominance of increments. We now remove the assumption on the equality of variances:

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

$$\mathbb{E}\left[(X_t - X_s)^2\right] \le \mathbb{E}\left[(Y_t - Y_s)^2\right].$$

Then

$$\mathbb{E}\left[\sup_{t\in T} X_t\right] \le \mathbb{E}\left[\sup_{t\in T} Y_t\right].$$

Proof. It is enough to prove this for Gaussian random vectors X and Y in $\mathbb{R}^n$, just like we did for Slepian's inequality in Theorem 7.2.7.
We again deduce the result from Gaussian Interpolation (Lemma 7.2.5). But this time, we'll approximate $f(x) \approx \max_i x_i$. Let $\beta > 0$ be a parameter and define the function

$$f(x) := \frac{1}{\beta} \log \sum_{i=1}^{n} e^{\beta x_i}.$$

We can check that indeed

$$\lim_{\beta\to\infty} f(x) = \max_{i=1,\ldots,n} x_i.$$

Substituting $f(x)$ into the Gaussian interpolation formula and simplifying shows that (Exercise 7.7)

$$\frac{d}{du}\mathbb{E}\left[f(Z(u))\right] \le 0 \text{ for all } u \in [0,1].$$

Then we can finish the proof just like in Slepian's inequality. □

Gordon's inequality extends the Slepian and Sudakov-Frenique inequalities to the min-max setting:

Theorem 7.2.9 (Gordon's inequality)**.** Let $(X_{ut})_{u\in U, t\in T}$ and $(Y_{ut})_{u\in U, t\in T}$ be two mean-zero Gaussian processes indexed by pairs of points (u,t) in a product set $U \times T$. Assume that

$$\begin{aligned}\mathbb{E}\left[(X_{ut} - X_{us})^2\right] &\le \mathbb{E}\left[(Y_{ut} - Y_{us})^2\right] \text{ for all } u, t, s;\\ \mathbb{E}\left[(X_{ut} - X_{vs})^2\right] &\ge \mathbb{E}\left[(Y_{ut} - Y_{vs})^2\right] \text{ for all } u \ne v \text{ and all } t, s.\end{aligned}$$

Then for every $\tau \ge 0$,

$$P\left(\inf_{u\in U}\sup_{t\in T} X_{ut} \ge \tau\right) \le P\left(\inf_{u\in U}\sup_{t\in T} Y_{ut} \ge \tau\right).$$

Moreover, by the integrated tail formula,

$$\mathbb{E}\left[\inf_{u\in U}\sup_{t\in T} X_{ut}\right] \le \mathbb{E}\left[\inf_{u\in U}\sup_{t\in T} Y_{ut}\right].$$

Proof. The proof under the additional assumption of equal variances is in Exercise 7.9. The proof for this statement is much harder. □

7.3 Application: Sharp Bounds for Gaussian Matrices

Let's pply the Gaussian comparison inequalities to random matrices. In Section 4.6, we used the ε -net argument to bound the expected operator norm like this:

$$\mathbb{E}\left[\|A\|\right] \le \sqrt{m} + C\sqrt{n}$$

where C is a constant (Exercise 4.41). Now, using the the Sudakov-Fernique inequality, we will tighten this bound for *Gaussian* random matrices and make $C = 1$.

Theorem 7.3.1 (Norms of Gaussian random matrices)**.** Let A be an $m \times n$ matrix with independent $N(0,1)$ entries. Then

$$\mathbb{E}\left[\|A\|\right] \leq \sqrt{m} + \sqrt{n}.$$

Proof. Let's write the norm of A as a supremum of Gaussian processes: By Definition 4.1.8,

$$\|A\| = \max_{u \in S^{n-1}, v \in S^{m-1}} \langle Au, v \rangle = \max_{(u,v) \in T} X_{uv}$$

where

$$T = S^{n-1} \times S^{m-1} \text{ and } X_{uv} := \langle Au, v \rangle \sim N(0,1).$$

To apply the 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

$$\begin{aligned}
\mathbb{E}\left[(X_{uv} - X_{wz})^2\right] &= \mathbb{E}\left[(\langle Au, v \rangle - \langle Au, z \rangle)^2\right] \\
&= \mathbb{E}\left[\left(\sum_{i,j} A_{ij}(u_j v_i - w_j z_i)\right)^2\right] \\
&= \sum_{i,j} (u_j v_i - w_j z_i)^2 \quad \text{(By independence, mean zero, variance 1)} \\
&= \|uv^T - wz^T\|_F^2 \\
&\leq \|u - w\|_2^2 - \|v - z\|_2^2 \quad \text{(By Exercise 7.10)}.
\end{aligned}$$

Now, let's define a simpler Gaussian process (Y_{uv}) with similar increments:

$$Y_{uv} := \langle g, u \rangle + \langle h, v \rangle\,, \; (u,v) \in T,$$

where $g \sim N(0, I_n)$ and $h \sim N(0, I_m)$ are independent Gaussian vectors. The increments of this process are

$$\begin{aligned}
\mathbb{E}\left[(Y_{uv} - Y_{wz})^2\right] &= \mathbb{E}\left[(\langle g, u - w \rangle + \langle h, v - z \rangle)^2\right] \\
&= \mathbb{E}\left[\langle g, u - w \rangle^2\right] + \mathbb{E}\left[\langle h, v - z \rangle^2\right] \quad \text{(By independence, mean 0)} \\
&= \|u - w\|_2^2 + \|v - z\|_2^2 \quad \text{(By normality of } g \text{ and } h\text{)}.
\end{aligned}$$

Comparing the increments of the two processes, we see that

$$\mathbb{E}\left[(X_{uv} - X_{wz})^2\right] \leq \mathbb{E}\left[(Y_{uv} - Y_{wz})^2\right] \text{ for all } (u,v), (w,z) \in T,$$

as required in the Sudakov-Fernique inequality. Applying Theorem 7.2.8, we obtain

$$\begin{aligned}
\mathbb{E}\left[\|A\|\right] &= \mathbb{E}\left[\sup_{(u,v) \in T} X_{uv}\right] \\
&\leq \mathbb{E}\left[\sup_{(u,v) \in T} Y_{uv}\right] \\
&= \mathbb{E}\left[\sup_{u \in S^{n-1}} \langle g, u \rangle\right] + \mathbb{E}\left[\sup_{v \in S^{m-1}} \langle h, v \rangle\right] \\
&= \mathbb{E}\left[\|g\|_2\right] + \mathbb{E}\left[\|h\|_2\right] \\
&\leq (\mathbb{E}\left[\|g\|_2^2\right])^{1/2} + (\mathbb{E}\left[\|h\|_2^2\right])^{1/2} \quad \text{(By Exercise 1.11)} \\
&= \sqrt{n} + \sqrt{m} \quad (By Proposition\ 3.2.1(b)).
\end{aligned}$$

□

Theorem 7.3.1 is 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 $m\infty n$ matrix with independent $N(0,1)$ entries. Then for every $t \geq 0$, we have

$$P(\|A\| \geq \sqrt{m} + \sqrt{n} + t) \leq 2\exp\left(-ct^2\right).$$

Proof. Let's combine the bound (Theorem 7.3.1) with Gaussian concentration (Theorem 5.2.3). Think of A as a long random vector in $\mathbb{R}^{n\infty n}$ by concatonating 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 the 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 bounded by 1. Then Theorem 5.2.3 yields

$$P(\|A\| \geq \mathbb{E}\left[\|A\|\right] + t) \leq 2\exp\left(ct^2\right).$$

The bound on $\mathbb{E}\left[\|A\|\right]$ from Theorem 7.3.1 completes the proof. □

Aside from the result above, we have that:
A symmetric Gaussian matric satisfies (Exercise 7.11)

$$\mathbb{E}\left[\|A\|\right] \leq 2\sqrt{n},$$

and the smallest singular value of an $m \times n$ Gaussian matrix A satisfies (Exercise 7.13)

$$\mathbb{E}\left[\sigma_n(A)\right] \geq \sqrt{m} - \sqrt{n}.$$

7.4 Sudakov Inequality

Recall that for a general mean-zero Gaussian process $(X_t)_{t\in T}$ on some index set T, the increments

$$d(t,s) := \|X_t - X_s\|_{L^2} = \left(\mathbb{E}\left[(X_t - X_s)^2\right]\right)^{1/2}$$

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 proces $(X_t)_{t\in T}$ (Remark 7.1.10). So, in theory, we can ask any question about the distribution of the process by understanding the geometry of the metric space (T,d) - studying probability via geometry!
Now the question comes: How can we estimate

$$\mathbb{E}\left[\sup_{t\in T} X_t\right]$$

in terms of the geometry of (T,d)? This is a hard problem we will study from now well into Chapter 8. We'll start with a lower bound in terms of the *metric entropy*, which was introduced in Chapter 4. Recall that for any $\varepsilon > 0$, the *covering number*

$$\mathcal{N}(T,d,\varepsilon)$$

is the samllest cardinality of an ε-net of T in the metric d, or equivalently the smallest number of closed balls of radius ε whose union covers T. The logarithm of the of the covering number, $\log_2 \mathcal{N}(T,d,\varepsilon)$, is called the *metric entropy* ot T.

Theorem 7.4.1 (Sudakov's inequality)**.** Let $(X_t)_{t\in T}$ be a mean-zero Gaussian process. Then, for any $\varepsilon \geq 0$, we have

$$\mathbb{E}\left[\sup_{t\in T} X_t\right] \geq c\varepsilon\sqrt{\log \mathcal{N}(T,d,\varepsilon)}$$

where d is the canonical metric defined above.

Proof. We'll deduce the result from the Sudakov-Frenique comparison inequality (Theorem 7.2.8). Assume that

$$N := \mathcal{N}(T, d, \varepsilon)$$

is finite; the infinite case is in Exercise 7.14. Let $\mathcal{N}$ be a maximal ε-seperated subset of T. Then $\mathcal{N}$ is an ε-net of T (Lemma 4.2.6), and thus

$$|\mathcal{N}| \geq N.$$

Restricting the process to $\mathcal{N}$, we see that it suffices to show that

$$\mathbb{E}\left[\sup_{t\in\mathcal{N}} X_t\right] \geq c\varepsilon\sqrt{\log N}.$$

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:

$$Y_t := \frac{\varepsilon}{\sqrt{2}} g_t \text{ where } g_t \sim_{i.i.d.} N(0,1).$$

To use the 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,

$$\mathbb{E}\left[(X_t - X_s)^2\right] = d(t,s)^2 \geq \varepsilon^2$$

while

$$\mathbb{E}\left[(Y_t - Y_s)^2\right] = \frac{\varepsilon^2}{2}\mathbb{E}\left[(g_t - g_s)^2\right] = \varepsilon^2 \quad (g_t - g_s \sim N(0,2)).$$

This implies that

$$\mathbb{E}\left[(X_t - X_s)^2\right] \geq \mathbb{E}\left[(Y_t - Y_s)^2\right] \text{ for all } t, s \in \mathcal{N}.$$

By applying Theorem 7.2.8, we obtain

$$\mathbb{E}\left[\sup_{t\in\mathcal{N}} X_t\right] \geq \mathbb{E}\left[\sup_{t\in\mathcal{N}} X_t\right] = \frac{\varepsilon}{2}\mathbb{E}\left[\max_{t\in\mathcal{N}} g_t\right] \geq c\varepsilon\sqrt{\log N}.$$

In the last step, we used that the expected maximum of N i.i.d $N(0,1)$ random variables is at least $c\sqrt{\log N}$ (Exercise 2.38 (b)). The proof is complete. □

7.4.1 Application for covering numbers in $\mathbb{R}^n$

Sudakov's 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 $\varepsilon > 0$,

$$\mathbb{E}\left[\sup_{t\in T}\langle g, t\rangle\right] \geq c\varepsilon\sqrt{\log \mathcal{N}(T, \varepsilon)},$$

where $\mathcal{N}(T,\varepsilon)$ just the covering number of T.

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.

$$d(t,s) = \|X_t - X_s\|_{L^2} = \|t - s\|_2 \text{ for any } t, s \in T.$$

Then the corollary directly follows from Sudakov's inequality (Theorem 7.4.1). □

Aside from the bound above, Corollary 7.4.2 is also sharp up to a log factor (Exercise 8.5):

$$\mathbb{E}\left[\sup_{t\in T}\langle g, t\rangle\right] \leq C\log{(n)}\cdot\varepsilon\sqrt{\log \mathcal{N}(T,\varepsilon)}.$$

For a quick application of Sudakov's inequality, let's (roughly) re-derive the boudne on covering numbers of polytopes in $\mathbb{R}^n$ from Corollary 0.1.1:

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

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

Proof. If $x_1, \ldots, X_N$ are the vertices of P, then

$$\mathbb{E}\left[\sup_{t \in P} \langle g, t \rangle\right] \leq \mathbb{E}\left[\sup_{i=1,\ldots,N} \langle g, x_i \rangle\right] \leq C\sqrt{\log N}.$$

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 is due to the maximal inequality from Proposition 2.7.6, as $\langle g, x \rangle \sim N(0, \|x\|_2^2)$ and $\|x\|_2 \leq 1$. Substitute this into Corollary 7.4.2 and simplify completes the proof. □

7.5 Gaussian Width

From the previous subsection, 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

$$w(T) := \mathbb{E}\left[\sup_{t \in T} \langle g, x \rangle\right] \text{ where } g \sim N(0, I_n).$$

Try to 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) $w(T + S) = w(T) + w(S)$ and $w(aT) = aw(T)$ for any $T, S \subset \mathbb{R}^n$ and $a \in \mathbb{R}$.

(e) (Symmetry)

$$w(T) = \frac{1}{2} w(T - T) = \frac{1}{2}\mathbb{E}\left[\sup_{x,y \in T} \langle g, x - y \rangle\right].$$

(f) (Width and diameter)

$$\frac{1}{\sqrt{2\pi}} \cdot \mathrm{diam}(T) \leq w(T) \leq \frac{\sqrt{n}}{2} \cdot \mathrm{diam}(T).$$

(g) (Linear maps) For any $m \times n$ matrix A, $w(AT) \leq \|A\| w(T)$.

Proof. Only the proof for (f) is demonstrated here, with the rest left to 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

$$w(T) \geq \frac{1}{2}\mathbb{E}\left[\max(\langle x - y, g \rangle, \langle y - x, g \rangle)\right] = \frac{1}{2}\mathbb{E}\left[|\langle x - y, g \rangle|\right] = \frac{1}{2}\sqrt{\frac{2}{\pi}}\|x - y\|_2.$$

The last equality holds since $\langle x - y, g \rangle \sim N(0, \|x - y\|_2)$ and $\mathbb{E}\left[|X|\right] = \sqrt{2/\pi}$ for $X \sim N(0, 1)$. Taking the supremum over all $x, y \in T$ gives the result.

For the upper bound, use property (e) again to get

$$w(T) \leq \frac{1}{2}\mathbb{E}\left[\sup_{x,y\in T} \langle g, x-y\rangle\right] \leq \frac{1}{2}\mathbb{E}\left[\sup_{x,y\in T} \|g\|_2\|x-y\|_2\right] \leq \frac{1}{2}\mathbb{E}\left[\|g\|_2\right] \cdot \operatorname{diam}(T).$$

Since $\mathbb{E}\left[\|g\|_2\right] \leq \mathbb{E}\left[\|g\|_2^2\right]^{1/2} = \sqrt{n}$, the proof is complete. □

Remark 7.5.3 (Width and diameter)**.** Both upper and lower bounds in Proposition 7.5.2 (f) are optimal and the $O(\sqrt{n})$ gap between them cannot be improved (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 below), which can be expressed as $\sup_{x,y\in T} \langle \theta, x-y\rangle$.

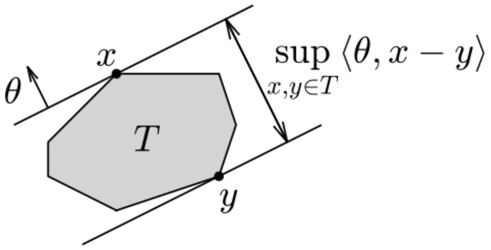


Figure 7.3 The width of a set $T \subset \mathbb{R}^n$ in the direction of a unit vector θ.

If we average the width over all unit directions θ, we get the following definition:

Definition 7.5.4. The spherical width of a set $T \subset \mathbb{R}^n$ is

$$w_s(T) := \mathbb{E}\left[\sup_{x\in T} \langle \theta, x\rangle\right] \quad \text{where } \theta \sim \operatorname{Unif}(S^{n-1}).$$

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 we get

Lemma 7.5.5 (Gaussian v.s. spherical widths)**.** Ths Gaussian width is approximately $\sqrt{n}$ times the spherical width:

$$\left(\sqrt{n} - \frac{C}{\sqrt{n}}\right) w_s(T) \leq w(T) \leq \sqrt{n} w_s(T).$$

Proof. Express the Gaussian vector g throughits length and direction: $g = r\theta$, where $r = \|g\|_2$ and $\theta = g/\|g\|_2$. Now, $\theta \sim \operatorname{Unif}(S^{n-1})$ is independent of r (Exercise 3.22). Thus

$$w(T) = \mathbb{E}\left[\sup_{x\in T} \langle r\theta, x\rangle\right] = \mathbb{E}\left[r\right] \cdot \mathbb{E}\left[\sup_{x\in T} \langle \theta, x\rangle\right] = \mathbb{E}\left[\|g\|_2\right] \cdot w_s(T).$$

Then by using concentration of the norm (Exercise 3.2), this gives

$$\sqrt{n} - \frac{C}{\sqrt{n}} \leq \mathbb{E}\left[\|g\|_2\right] \leq \sqrt{n},$$

which completes the proof. □

7.5.2 Examples

Example 7.5.6 (Euclidean ball and sphere)**.** The Gaussian widths of the unit ball and sphere are

$$w(S^{n-1}) = w(B_2^n) = \mathbb{E}\left[\|g\|_2\right] = \sqrt{n} \pm \frac{C}{\sqrt{n}},$$

where we used concentration of the norm (Exercise 3.2) for the last step. The spherical width of these sets of course equal to 1.

Example 7.5.7 (Cube)**.** The unit ball of the ℓ^∞ norm in $\mathbb{R}^n$ is the cube $[-1,1]^n$. So, by using the duality formula

$$\max\{\langle x, y\rangle : \ y \in B_{p'}^n\} = \|x\|_p,$$

we get

$$w(B_\infty^n) = \mathbb{E}\left[\|g\|_1\right] = \mathbb{E}\left[|g_1|\right] \cdot n = \sqrt{\frac{2}{\pi}} \cdot n.$$

Example 7.5.8 (Cross-polytope)**.** The unit ball of the ℓ^1 norm in $\mathbb{R}^n$ is the cross-polytope

$$B_1^n = \{x \in \mathbb{R}^n : \ \|x\|_1 \leq 1\}.$$

Its Gaussian width satisfies

$$w(B_1^n) \asymp \sqrt{\log N}$$

where the notation $\asymp$ hides the absolute constant factors. This is because

$$w(B_1^n) = \mathbb{E}\left[\|g\|_\infty\right] = \mathbb{E}\left[\max_{i=1,\dots,n} |g_i|\right],$$

where the first equation uses duality. Then the result follows from Exercise 2.38 (b).

Example 7.5.9 (Finite point sets)**.** Any finite set of points $T \subset \mathbb{R}^n$ satisfies

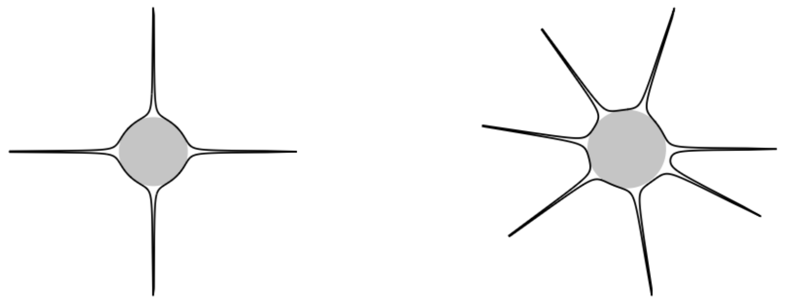
$$w(T) \leq C\sqrt{\log |T|} \cdot \operatorname{diam}(T).$$

To prove this, we can assume that $\operatorname{diam}(T) = 1/2$ (by rescaling), and that T lies in the unit Euclidean ball (by translation). Then the result follows from the bound provided in Corollary 7.4.3.

Remark 7.5.10 (Surprising behavior of width in high dimensions)**.** As we can see from Example 7.5.6 to Example 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* $\frac{1}{\sqrt{n}}B_2^n$, whiich 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-polytops B_1^n only has $2n$ vertices, so a random direction $g \sim N(0, I_n)$ is likely to be far from all of them. The width is not only driven by those lonely $2n$ "spikes" - it's driven by the "bulk". which is roughly the inscribed ball.

Figure 7.4a shows Milmen's *hyperbolic sketch* of B_1^n, highlighting how the bulk (the inscribed ball) dominates since the set has few vertices (spikes). We can make similar sketches to general convex sets too (Figure 7.4b) - they are great for building high-dimensional intuition, even if we lose convexity in the picture.



(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 also a number of helpful cousins of the Gaussian width $w(T)$. Normally, we would take the expected max of $\langle g, t\rangle$, but sometime it's easier to work with L^1 or L^2 averages:

$$w(T) = \mathbb{E}\left[\sup_{x\in T} \langle g, x\rangle\right],\ \gamma(T) := \mathbb{E}\left[\sup_{x\in T} |\langle g, x\rangle|\right],\ h(T) := \left(\mathbb{E}\left[\sup_{x\in T} \langle g, x\rangle^2\right]\right)^{1/2}.$$

where $g \sim N(0, I_n)$. We call $\gamma(T)$ the *Gaussian complexity* of T. Clearly,

$$w(T) \leq (T) \leq h(T).$$

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:

(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 prove the first equivalence here, and we'll leave the second equivalence to Exercise 7.20. We trivially have $\gamma(T) \leq h(T)$. 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

$$\sup_{x\in T}\|x\|_2 = r(T).$$

Then by Gaussian concentration (Theorem 5.2.3),

$$\|\sup_{x\in T} |\langle g, x\rangle| - \gamma(T)\|_{\psi_2} \lesssim r(T).$$

So by the triangle inequality and Proposition 2.6.6 (ii), we get

$$h(T) = \|\sup_{x\in T} |\langle g, x\rangle|\|_{L^2} \lesssim \|\sup_{x\in T} |\langle g, x\rangle|\|_{\psi_2} \lesssim \gamma(T) + r(T) \lesssim \gamma(T)$$

where in the last step, we 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 be susceptible to tiny perturbations of T. Here is a more robust alternative:

Definition 7.5.12. 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 dimensions is bounded above by the linear-algebraic one:

$$d(T) \leq \dim(T),$$

with equality when T is a Euclidean ball in some subspace (Exercise 7.21). Unlike the usual dimension, the effective dimension is stable - small perturbations to T only slightly change its width and diameter.

7.6 Application: Random Projection 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 Grassmannian $G_{m,n}$)? We might view this like dimensionality redution, like in the Johnson-Lindenstrauss lemma. What can we say about the size (diameter) of the projected set PT where P is the random projection?
For a finite set T, the Johnsen-Lindenstrauss lamm (Theorem 5.3.1) says that if $m \gtrsim \log |T|$, the random projection P acts essentially as a scaling of T:

$$\operatorname{diam}(PT) \approx \sqrt{\frac{m}{n}} \operatorname{diam}(T).$$

But if the cardinality of T is too large or infinite, the above may fail. For instance, if $T = B_2^n$ is the unit Euclidean ball, no projection can shrink its size:

$$\operatorname{diam}(PT) = \operatorname{diam}(T).$$

What about for general sets T? The next result states that a random projection cannor shrink T beyond its 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

$$\mathbb{E}\left[\operatorname{diam}(PT)\right] \asymp w_s(T) + \sqrt{\frac{m}{n}} \operatorname{diam}(T),$$

where the notation $\asymp$ hides positive absolute constant factors.

Proof. We'll prove the upper bound here. The lower bound is in Exercise 7.26.
Step 1: Change the model. Let's switch the view just like in the proof of Lemma 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, assume $\operatorname{diam}(T) \leq 1$. 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

$$|\mathcal{N}| \leq 5^m$$

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 of 2 (Execise 4.35):

$$\operatorname{diam}(QT) \leq 2 \sup_{x \in T-T} \max_{z \in \mathcal{N}} \langle Qx, z \rangle = 2 \max_{z \in \mathcal{N}} \sup_{x \in T-T} \left\langle Q^T z, x \right\rangle .$$

Now, here is the plan: we will first bound

$$\sup_{x \in T-T} \langle Q^T z, x \rangle \quad (*)$$

for a fixed $z \in \mathcal{N}$, and then take the union bound over all z.

Step 3: Concentration. Fix $z \in \mathcal{N}$. By construction, $Q^T z$ is uniformly distributed on the sphere: $Q^T z \sim \mathrm{Unif}(S^{n-1})$ (Exercise 7.24). The expectation can be expressed as the spherical width:

$$\mathbb{E}\left[\sup_{x \in T-T} \langle Q^T z, x \rangle\right] = w_s(T-T) = 2w_s(T).$$

(The last equality if just a spherical version of Proposition 7.5.2 (e)). To check that $(*)$ concentrates around its mean, we use the concentration inequality on the sphere (Theorem 5.1.3). Since $\mathrm{diam}(T) \leq 1$ by assumption, the function $z \mapsto \sup_{x \in T-T} \langle z, x \rangle$ on the sphere has Lipschitz norm at most 1. So we get

$$P\left(\sup_{x \in T-T} \langle Q^T z, x \rangle \geq 2w_s(T) + t\right) \leq 2\exp(-cnt^2).$$

Step 4: Union bound. Now we unfix $z \in \mathcal{N}$ by taking the union bound:

$$P\left(\max_{z \in \mathcal{N}} \sup_{x \in T-T} \langle Q^T z, x \rangle \geq 2w_s(T) + t\right) \leq |\mathcal{N}| \cdot 2\exp(-cnt^2).$$

Recall that $|\mathcal{N}| \leq 5^m$. Choosing $t = Cs\sqrt{m/n}$ with constant C large enough, the probability above is bounded by $2e^{-ms^2}$ for any $s \geq 1$. Therefore, we get

$$P\left(\frac{1}{2}\mathrm{diam}(QT) \geq 2w_s(T) + Cs\sqrt{\frac{m}{n}}\right) \leq e^{-ms^2} \text{ for any } s \geq 1.$$

From this, we can bound the expected value of $\mathrm{diam}(QT)$ using the integrated tail formula Lemma 1.6.1, and the proof is complete. □

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 a factor of 2), we can write:

$$\mathrm{diam}(PT) \asymp \max\left[w_s(T), \sqrt{\frac{m}{n}}\mathrm{diam}(T)\right].$$

Let's find the "phase transition" point where these two terms are equal. Set them to be equal and solving for m, we get

$$m = \frac{(\sqrt{n}w_s(T))^2}{\mathrm{diam}(T)^2} \asymp \frac{w(T)^2}{\mathrm{diam}(T)^2} \asymp d(T),$$

using Lemma 7.5.5 and the definition of effective dimension $d(T)$ (Definition 7.5.12). So the take-away:

$$\mathrm{diam}(PT) \asymp \begin{cases} \sqrt{\frac{m}{n}}\mathrm{diam}(T) & \text{if } m \geq d(T) \\ w_s(T) & \text{if } m \leq d(T). \end{cases}$$

See figure 7.5 below.

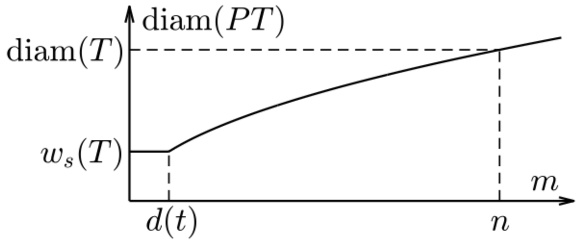


Figure 7.5 The diameter of a random m-dimensional projection of a set T as a function of m.

As we decrease the dimension m of the random projection, initially it shrinks t by roughly $\sqrt{m/n}$ as stated in the Johnson-Lindenstrauss lemma. But once m dips below the effective dimension $d(T)$, the shrinking stops and the diameter stays near the spherical width $w_s(T)$. This is because $\mathrm{conv}(PT)$ looks like a ball of radius $w_s(T)$, as we will see in Section 9.7.2.

8 Chaining

This chapter concerns some of the central methods for bounding random processes (X_t). We'll go over concepts such as chaining, VC theory, generic chaining methods, and bounds such as Talagrand's inequality and Chevet's inequality. We'll apply these to concepts such as Monto Carlo integration, empirical processes, and statistical learning theory.

8.1 Dudley Inequality

For a general Gaussian process $(X_t)_{t \in T}$, Sudakove inequality (Theorem 7.4.1) gives a *lower* bound on

$$\mathbb{E}\left[\sup_{t \in T} X_t\right]$$

in terms of the metric entropy pf T. Now we'll go for an upper bound. Moreover, we generalize from Gaussian processes to subgaussian processes as well.

Definition 8.1.1. A random process $(X_t)_{t \in T}$ on a metric space (T, d) has subgaussian increments if there exists $K > 0$ such that

$$\|X_t - X_s\|_{\psi_2} \leq K d(t,s) \text{ for all } t, s \in T.$$

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:

$$d(t,s) := \|X_t - X_s\|_{L^2}, \ \ t, s \in T,$$

as we explained earlier. With respect to this metric, $(X_t)_{t \in T}$ clearly has 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 in terms of the metric entropy:

Theorem 8.1.3 (Dudley's 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 Definition 8.1.1. Then

$$\mathbb{E}\left[\sup_{t \in T} X_t\right] \leq CK \int_0^\infty \sqrt{\log \mathcal{N}(T, d, \varepsilon)} \, d\varepsilon.$$

Before going to the proof's let's compare Dudley's inequality with Sudakov's inequality (Theorem 7.4.1), which for Gaussian processes, says:

$$\mathbb{E}\left[\sup_{t \in T} X_t\right] \geq c \sup_{\varepsilon > 0} \varepsilon \sqrt{\log \mathcal{N}(T, d, \varepsilon)}.$$

Figure 8.1 below shows both bounds:

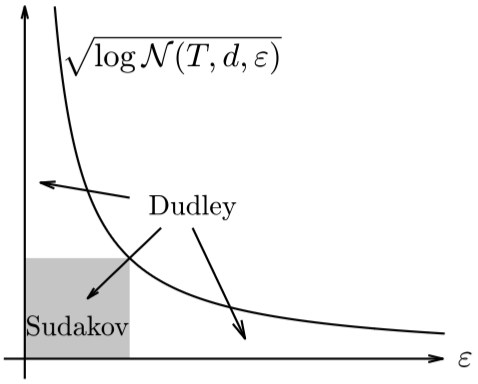


Figure 8.1 Dudley inequality bounds $\mathbb{E}\sup_{t\in T} X_t$ by the area under the curve. Sudakov inequality bounds it below by the largest area of a rectangle under the curve, up to constants.

There is a clear gap between the two bounds, and it turns out that metric entropy alone cannot close it - we will explore this later.

Dudley's inequality hints that $\mathbb{E}\left[\sup_{t\in T} X_t\right]$ is a *multiscale* quantity - to bound it, we need to look at T across all scales ε. That's exactly how the proof works! But let's prove a discrete version using syadic scaled $\varepsilon = 2^{-k}$ like a Riemann sum, then move to the continuous version later.

Theorem 8.1.4 (Discrete Dudley's inequality)**.** Let $(X_t)_{t\in T}$ be a mean-zero random process on a metric space (T,d) with subgaussian increments as from earlier. Then

$$\mathbb{E}\left[\sup_{t\in T} X_t\right] \le CK \sum_{k\in\mathbb{Z}} 2^{-k}\sqrt{\log \mathcal{N}(T,d,\varepsilon)}.$$

The proof uses a technique called *chaining*. It is a multi-scaled version of the ε-net argument that we did in Theorem 4.4.3 and Theorem 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 $d(t,\pi(t)) \le \varepsilon$. Then the increment condition gives

$$\|X_t - X_{\pi(t)}\|_{\psi_2} \le K\varepsilon.$$

This leads to

$$\mathbb{E}\left[\sup_{t\in T} X_t\right] \le \mathbb{E}\left[\sup_{t\in T} X_{\pi(t)}\right] + \mathbb{E}\left[\sup_{t\in T}(X_t - X_{\pi(t)})\right].$$

We can handle the first term via union bound over $|\mathcal{N}| = \mathcal{N}(T,d,\varepsilon)$ points $\pi(t)$. However, the second term is unclear (if we were to use union bound) since there is both t and $\pi(t)$ in the supremum. To fix this, we don't stop at one net, but choose smaller and smaller ε to get better approximations $\pi_1(t), \pi_2(t), \ldots$ to t with finer nets. This is the idea behind *chaining*.

Proof of Theorem 8.1.4. **Step 1: Chaining setup.** Without loss of generality, we may assume that $K = 1$ (because of C) and T is finite (Remark 7.2.1). Define the dyadic scale

$$\varepsilon_k = 2^{-k},\ k \in \mathbb{Z}$$

and choose ε_k-nets T_k of T so that

$$|T_k| = \mathcal{N}(T,d,\varepsilon_k).$$

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

$$T_\kappa = \{t_0\} \text{ for some } t_0 \in T,\ T_K = T.$$

For a point $t \in T$, let $\pi_l(t)$ denote a closest point in T_k, so we have

$$d(t, \pi_k(t)) \le \varepsilon_k.$$

Since $\mathbb{E}[X_{t_0}] = 0$ by assumption,

$$\mathbb{E}\left[\sup_{x\in T} X_t\right] = \mathbb{E}\left[\sup_{x\in T}(X_t - X_{t_0})\right].$$

Let's write $X_t - X_{t_0}$ as a telescoping sum, walking from t_0 to t along a chain (aha!) of points $\pi_k(t)$ that mark progressively finer approximations of t:

$$X_t - X_{t_0} = (X_{\pi_\kappa(t)} - X_{t_0}) + (X_{\pi_{\kappa+1}(t)} - X_{\pi_\kappa(t)}) + \cdots + (X_t - X_{\pi_K(t)}),$$

see Figure 8.2 below for an illustration.

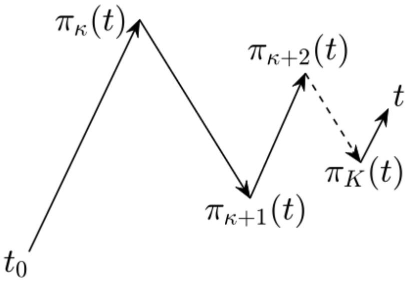


Figure 8.2 Chaining: a walk from a fixed point t_0 to an arbitrary point t in T along elements $\pi_k(T)$ of progressively finer nets of T

The first and last terms of this sum are zero by our definition earlier, so we have

$$X_t - X_{t_0} = \sum_{k=\kappa+1}^{K}(X_{\pi_k(t)} - X_{\pi_{k-1}(t)}).$$

Since the supremum of the sum is bounded by the sum of the suprema, we get

$$\mathbb{E}\left[\sup_{t\in T}(X_t - X_{t_0})\right] \le \sum_{k=\kappa+1}^{K}\mathbb{E}\left[\sup_{t\in T}(X_{\pi_k(t)} - X_{\pi_{k-1}(t)})\right].$$

Step 2: Controlling the increments. In the equation above, 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

$$|T_k| \cdot |T_{k-1}| = |T_k|^2,$$

A number that we can control via covering numbers from above. Moreover, for a fixed t, we can bound the increments in step 1 like this:

$$\begin{aligned}\|X_{\pi_k(t)} - X_{\pi_{k-1}(t)}\|_{\psi_2} &\le d(\pi_k(t), \pi_{k-1}(t)) \quad \text{(By Definition 8.1.1)}\\ &\le d(\pi_k(t), t) + d(t, \pi_{k-1}(t)) \quad \text{(By triangle inequality)}\\ &\le \varepsilon_k + \varepsilon_{k-1} \quad \text{(By definition of)}\pi_k(t)\\ &\le 2\varepsilon_{k-1}.\end{aligned}$$

Recall from Proposition 2.7.6 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:

$$\mathbb{E}\left[\sup_{t\in T}(X_{\pi_k(t)} - X_{\pi_{k-1}(t)})\right] \le C\varepsilon_{k-1}\sqrt{\log|T_k|}.$$

Step 3: Summing up the increments. We have shown that

$$\mathbb{E}\left[\sup_{t\in T}(X_t - X_{t_0})\right] \le C\sum_{k=\kappa+1}^{K}\varepsilon_{k-1}\sqrt{\log|T_k|}.$$

Now plug in the values $\varepsilon_k = 2^{-k}$ and the bounds on $|T_k|$, we get

$$\mathbb{E}\left[\sup_{t\in T}(X_t - X_{t_0})\right] \leq C_1 \sum_{k=\kappa+1}^{K} 2^{-k}\sqrt{\log \mathcal{N}(T,d,2^{-k})}.$$

Hence the theorem is proved. □

Let's now go for the proof for the integral form of Dudley's inequality.

Proof of Dudley's integral inequality (Theorem 8.1.3). To convert the sum from the discrete form 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 the integral, $2^{-k} \geq \varepsilon$, hence $\log \mathcal{N}(T,d,2^{-k}) < \log \mathcal{N}(T,d,2^{-k})$ and the sum is bounded by

$$2\sum_{k\in\mathbb{Z}}\int_{2^{-k-1}}^{2^{-k}} \sqrt{\log \mathcal{N}(T,d,\varepsilon)}\, d\varepsilon = 2\int_{0}^{\infty} \sqrt{\log \mathcal{N}(T,d,\varepsilon)}\, d\varepsilon,$$

and the proof is complete. □

Actually, the discrete and integral Dudley inequalities are equivalent (Exercise 8.3).

8.1.1 Variations and Examples

Remark 8.1.5 (Dudley's inequality: supremum of increments)**.** A quick look at the proof shows that chaining actually gives

$$\mathbb{E}\left[\sup_{t\in T}|X_t - X_{t_0}|\right] \leq CK\int_{0}^{\infty} \sqrt{\log \mathcal{N}(T,d,\varepsilon)}\, d\varepsilon$$

for any fixed $t \in T$. We can combine with the same bound for $X_s - X_{t_0}$, then use the triangle inequality to get

$$\mathbb{E}\left[\sup_{t,s\in T}|X_t - X_s|\right] \leq CK\int_{0}^{\infty} \sqrt{\log \mathcal{N}(T,d,\varepsilon)}\, d\varepsilon.$$

Remark 8.1.6 (Dudley's inequality: a high-probability bound)**.** Dudley's inequality gives only an expectation bound, but chaining actually gives a high-probability bouind. Assuming T is finite (avoid measurability issues), for every $u \geq 0$, the bound

$$\sup_{t,s\in T}|X_t - X_s| \leq CK\left[\int_{0}^{\infty} \sqrt{\log \mathcal{N}(T,d,\varepsilon)}\, d\varepsilon + u\cdot \operatorname{diam}(T)\right]$$

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).

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

$$\mathcal{N}(T,d,\varepsilon) = 1 \implies \log \mathcal{N}(T,d,\varepsilon) = 0.$$

Thus

$$\mathbb{E}\left[\sup_{t\in T} X_t\right] \leq CK\int_{0}^{\operatorname{diam}(T)} \sqrt{\log \mathcal{N}(T,d,\varepsilon)}\, d\varepsilon.$$

If we apply Dudley's inequality for the canonical Gaussian process $\langle g, t\rangle$, just like we did with Sudakov's inequality in Corollary 7.4.2, we get the following:

Theorem 8.1.8 (Dudley's inequality in $\mathbb{R}^n$)**.** The Gaussian width of any bounded set $Y \subset \mathbb{R}^n$ satisfies

$$w(T) \le C \int_0^\infty \sqrt{\log \mathcal{N}(T, \varepsilon)}\ d\varepsilon,$$

where $\mathcal{N}(T, \varepsilon)$ is the smallest number of Euclidean balls with radius ε and centers in T that cover T.

Example 8.1.9 (Dudley's inequality is sharp for the Euclidean ball)**.** Let's test Dudley's inequality for the unit Euclidean ball $T = B_2^n$. From Corollary 4.2.11,

$$\mathcal{N}(B_2^n, \varepsilon) \begin{cases} \le (3/\varepsilon)^n & \text{for } \varepsilon \in (0,1], \\ = 1 & \text{for } \varepsilon > 1 \end{cases}.$$

Then

$$w(B_2^n) \lesssim \int_0^1 \sqrt{n \log{(3/\varepsilon)}}\ d\varepsilon \lesssim \sqrt{n}.$$

This is in fact optimal: as we know from Example 7.5.6, $w(B_2^n) \asymp \sqrt{n}$.

Remark 8.1.10 (Dudley's inequality can be loose - but not too loose)**.** In general, Dudley integral can overestimate the Gaussian width. Here is a bad example:

$$T = \left\{ \frac{e_k}{\sqrt{1 + \log k}},\ k = 1, \dots, n \right\}$$

with e_k being the standard basis in $\mathbb{R}^n$. From exercise 8.4, we can see that

$$w(T) = O(1) \text{ while } \int_0^\infty \sqrt{\log \mathcal{N}(T, d, \varepsilon)}\ d\varepsilon \to \infty$$

as $n \to \infty$. However, the good news:

(a) Dudley equality is tight up to a logarithmic factor (Exercise 8.5);

(b) We will use chaining to remove that logarithmic factor in Section 8.5.

8.2 Application: Empirical Processes

We'll apply Dudley's 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

$$\int_0^1 f(x)\ dx,\ f : [0,1] \to \mathbb{R}$$

(See Figure 8.3a).
We can do this *probabilistically*. Suppose X is a random point in Ω drawn according to μ, i.e. $P(X \in A) = \mu(A)$ for any measurable set $A \subset \Omega$. Then the integral becomes the expectation:

$$\int_\Omega f\ d\mu = \mathbb{E}\left[f(X)\right].$$

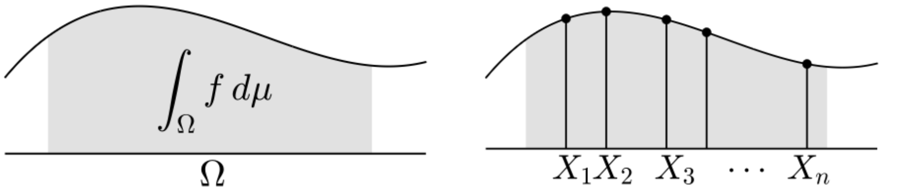
Now take i.i.d. samples $X_1, X_2, \dots$ By the strong law of large numbers (Theorem 1.7.1),

$$\frac{1}{n}\sum_{i=1}^{n} f(X_i) \to \mathbb{E}\left[f(X)\right] \text{ almost surely}$$

as $n \to \infty$. So, we can approximate the integral with just the arithmetic mean:

$$\int_\Omega f \, d\mu \approx \frac{1}{n}\sum_{i=1}^{n} f(X_i)$$

(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 is $O(1/\sqrt{n})$. This comes from the convergence rate in the law of large numbers:

$$\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}\left[f(X)\right]\right|\right] \leq \left[\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} f(X_i)\right)\right]^{1/2} = O\left(\frac{1}{\sqrt{n}}\right).$$

Remark 8.2.2 (Monte Carlo is high-dimensional, agnostic)**.** Monte Carlo works well in high dimensions since the error 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 it is enough. The same is with f - you only need its values at just a few points.

8.2.2 Lipschitz Law of Large Numbers

Can we use the same sample $X_1, \dots, X_n$ 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 totally off.

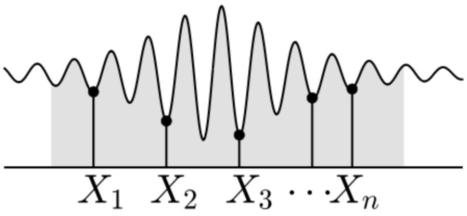


Figure 8.4 One sample $X_1, \dots, X_n$ cannot be used to approximate the integral of *all* functions f.

But what if we stick to function that don't wiggle too much, like Lipschitz functions? Then yes!

Theorem 8.2.3 (Lipschitz Law of Large Numbers)**.** Consider the class of functions

$$\mathcal{F} := \{f : [0,1] \to \mathbb{R},\ \|f\|_{\text{Lip}} \leq L\},$$

where L is any number. Let $X, X_1, \ldots, X_n$ be i.i.d. random variables taking values in $[0,1]$. Then

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}\left[f(X)\right]\right|\right] \leq \frac{CL}{\sqrt{n}}.$$

Remark 8.2.4 (One sample serves all Lipschitz functions)**.** Before the proof, let's iterate the key point: the supremum over $f \in \mathcal{F}$ is *inside* the expectation. Thanks to Markov's 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 $O(1/\sqrt{n})$ error - same rate as the usual law of large numbers with just one function. So, we made the law of large numbers uniform without losing anything!

To make the proof of Theorem 8.2.3 more intuitive, we will also introduce empirical processes:

Definition 8.2.5. 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 $X_1, \ldots, X_n$ be independent copies of X. The random process $(X_f)_{f\in\mathcal{F}}$ defined by

$$X_f := \frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}\left[f(X)\right]$$

is called an empirical process indexed by $\mathcal{F}$m.

Let's go to the proof!

Proof of Theorem 8.2.3. Without loss of generality, it is enough to prove the theorem for the class

$$\mathcal{F} := \{f : [0,1] \to [0,1], \|f\|_{\text{Lip}} \leq 1\}.$$

We would like to bound

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} |X_f|\right]$$

for the empirical process $(X_f)_{f\in\mathcal{F}}$ defined earlier.

Step 1: Checking subgaussian increments. Let's use Dudley's 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

$$\|X_f - X_g\|_{\psi_2} = \frac{1}{n}\|\sum_{i=1}^{n} Z_i\|_{\psi_2} \text{ where } Z_i := (f-g)(X_i) - \mathbb{E}\left[(f-g)(X)\right].$$

Since Z_i are independent, mean-zero random variables, Proposition 2.7.1 gives

$$\|X_f - X_g\|_{\psi_2} \lesssim \frac{1}{n}\left(\sum_{i=1}^{n}\|Z_i\|_{\psi_2}^2\right)^{1/2}.$$

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's inequality. Now apply Dudley's inequality (Theorem 8.1.3), then we get

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}|X_f|\right]=\mathbb{E}\left[\sup_{f\in\mathcal{F}}|X_f-X_0|\right]\lesssim\frac{1}{\sqrt{n}}\int_0^1\sqrt{\log\mathcal{N}(\mathcal{F},\|\cdot\|_{L^\infty},\varepsilon)}\,d\varepsilon.$$

(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). It is not difficult to bound the covering numbers of $\mathcal{F}$ like this (Exercise 8.9):

$$\mathcal{N}(\mathcal{F},\|\cdot\|_{L^\infty},\varepsilon)\leq e^{C/\varepsilon}.$$

Substitute this bound into the integral, we get

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}|X_f|\right]\lesssim\frac{1}{\sqrt{n}}\int_0^1\sqrt{\frac{C}{\varepsilon}}\,d\varepsilon\lesssim\frac{1}{\sqrt{n}}.$$

hence the proof is complete. □

8.2.3 Empirical Measure

For a broader perspective, take one more look at Definition 8.2.5. Given an i.i.d sample $X_1,\dots,X_n$ 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:

$$\mu_n=\frac{1}{n}\sum_{i=1}^n\delta_{X_i}.$$

Here δ_x is the Dirac probability measure at x, i.e.

$$\delta_x(A)=\begin{cases}1 & \text{if } x\in A,\\ 0 & \text{otherwise.}\end{cases}$$

The integral of f with respect to the original measure μ is $\mathbb{E}[f(X)]$, while the integral of f with respect to the empirical measure μ_n is $\frac{1}{n}\sum_{i=1}^n f(X_i)$. The empirical process X_f we defined above tracks the deviation of the population expectation from the empirical expectation.

This deviation, which we bounded 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.*

8.3 VC Dimension

We'll learn about VC dimension, which is a huge part of statistical learning theory. We'll connect it to covering numbers, and then, through Dudley's inequality, to random processes and the uniform law of large numbers. Applications to statistical learning theory is in the next section.

8.3.1 Definition and Examples

VC dimensions measures how complex a class of Boolean functions is, where 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. 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 it on Λ. Formally, this means the restriction of f onto Λ is g, i.e. $f(x)=g(x)$ for all $x\in\Lambda$.
The Vapnik-Chervonenkis dimension (VC dimension) of $\mathcal{F}$, denoted $\mathrm{vc}(\mathcal{F})$, is the largest cardinality of a subset $\Lambda\subseteq\Omega$ that is shattered. If there is no largest one, then $\mathrm{vc}(\mathcal{F})=\infty$.

Let's go through a few examples to make the definition clearer:

Example 8.3.2 (Intervals)**.** Let $\mathcal{F}$ consist of the indicators of all closed intervals in $\mathbb{R}$:

$$\mathcal{F} = \left\{ \mathbf{1}_{[a,b]} : \ a, b \in \mathbb{R}, \ a \leq b \right\}.$$

We claim that

$$\mathrm{vc}(F) = 2.$$

We first show that $\mathrm{vc}(\mathcal{F}) \geq 2$ by finding a two-point set $\Lambda \subset \mathbb{R}$ that is shattered by $\mathcal{F}$. Take, for example, $\Lambda = 3, 5$. 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 $f = \mathbf{1}_{[a,b]}$ ontp Λ. For example, $g(3) = 1, g(5) = 0$ comes from $f = \mathbf{1}_{[2,4]}$. The other three cases are shown in Figure 8.5, so Λ is indeed shattered by $\mathcal{F}$.

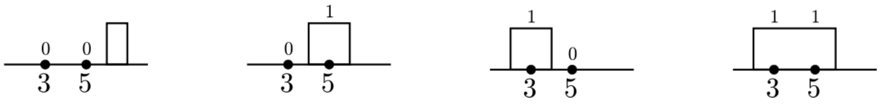


Figure 8.5 The binary function $g(3) = g(5) = 0$ is the restriction of $\mathbf{1}_{[6,7]}$ onto $\Lambda = \{3, 5\}$ (left). The function $g(3) = 0$, $g(5) = 1$ is the restriction of $\mathbf{1}_{[4,6]}$ onto Λ (middle left). The function $g(3) = 1$, $g(5) = 0$ is the restriction of $\mathbf{1}_{[2,4]}$ 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 cannot be a restriction of any indicator interval onto Λ (it is not linearly seperable).

Example 8.3.3 (Half-planes)**.** Let $\{$ consist of the indicators of all closed half-planes in $\mathbb{R}^2$. Then we claim that

$$\mathrm{vc}(\mathcal{F}) = 3.$$

To prove $\mathrm{vc}(\mathcal{F}) \geq 3$, 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 posiition like in Figure 8.6 below. Each of the $2^3 = 8$ binary labelings $g : \Lambda \to \{0, 1\}$ is a restriction of the indicator function of some half-plane. Too see this, attange the half-plane to containe exactly these points where g takes value 1. Thus, Λ is shattered.

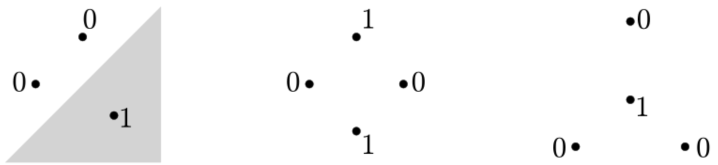


Figure 8.6 The proof that VC(half-planes)= 3 in Example 8.3.3 consists of two steps. To show VC ≥ 3, we find a three-point set Λ on which every binary labeling g is linearly separable (left). To show VC < 4, we demonstrate that on any four-point set Λ there exists a binary labeling g that is not linearly separable (middle and right).

To prove $\mathrm{vc}(\mathcal{F}) < 4$, 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 also in Figure 8.6. In each of the two cases, there exists a binary labeling such that no half-plane can contain exactly the points labeled 1. This means that there 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 the strings 00, 01, 10, 11. Thys, 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

$$\mathrm{vc}(\mathcal{F}) \geq |\Lambda| = 2.$$

On the other hand, the (only) three-point set $\{1,2,3\}$ is not shattered by $\mathcal{F}$, as that would require all eight binary digits of length three to appear in $\mathcal{F}$, which is not true.

Example 8.3.5 (Half-spaces)**.** A half-space in $\mathbb{R}^m$ is a set of the form

$$\{x : \ \langle a, x \rangle \leq b\} \text{ where } a \in \mathbb{R}^n \text{ and } b \in \mathbb{R}.$$

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 v.s. parameter count)**.** The 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 the VC dimensions (Example 8.3.5). This is not a hard rule but rather a useful heuristic.

8.3.2 Pajor's 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

$$2^{\mathrm{vc}(\mathcal{F})} \leq |\mathcal{F}| \leq 2^n.$$

The upper bound is usually loose - most function classes are closer in side 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's lemma)**.** Let $\mathcal{F}$ be a class of Boolearn functions on a finite set Ω. Then

$$|\mathcal{F}| \leq |\{\Lambda \subseteq \Omega : \ \Lambda \text{ is shattered by } \mathcal{F}\}|.$$

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 $\{2,3\}$. Thus the inequalty in Pajor's lemma reads $4 \leq 6$.

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.
For the inductive step, assume the lemma holds for any n-point set Ω. Now look at a set Ω with $|\Omega| = n+1$. Chopping out one (arbitrary) point from the set Ω, we can express it as

$$\Omega = \Omega_0 \cup \{x_0\}, \text{ where } |\Omega_0| = n.$$

The class $\mathcal{F}$ then natually breaks into two subclasses

$$\mathcal{F}_0 := \{f \in \mathcal{F} : \ f(x_0) = 0\} \text{ and } \mathcal{F}_1 := \{f \in \mathcal{F} : \ f(x_0) = 1\}.$$

By the induction hypothesis, the counting function

$$S(\mathcal{F}) = |\{\Lambda \subseteq \Omega : \ \Lambda \text{ is shattered by } \mathcal{F}\}|$$

satisfies (by restricting to Ω_0)

$$S(\mathcal{F}_0) \geq |\mathcal{F}_0| \text{ and } S(\mathcal{F}_1) \geq |\mathcal{F}_1|.$$

To complete the proof, all we need to check is

$$S(\mathcal{F}) \geq S(\mathcal{F}_0) + S(\mathcal{F}_1),$$

for then the inductive hypothesis would give $S(\mathcal{F}) \geq |\mathcal{F}_0| + |\mathcal{F}_1| = |\mathcal{F}|$, as needed.
The inequality above 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 $S(\mathcal{F})$. However, there may be the risk of 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, $\Lambda \cup \{x_0\}$. This set is indeed shattered by $\mathcal{F}$. This establishes the inequality above, and the proof is complete. □

Let's illustrate the proof above via an example:

Example 8.3.8. Let's reuse Example 8.3.4. Following the proof of Pajor's 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

$$\mathcal{F}_0 = \{010, 100\} \text{ and } \mathcal{F}_1 = \{001, 111\}.$$

There are exactly two subsets Λ shattered by $\mathcal{F}_0$, namely $\{1\}$ and $\{2\}$, and the same two 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. 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 we have at least four subsets shattered by $\mathcal{F}$, making the inequality in Pajor's lemma true.

8.3.3 Sauer-Shelah Lemma

We now deduce a remarkable upper bound on the cardinality of a function class in terms of the 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

$$|\mathcal{F}| \leq \sum_{k=0}^{d} \binom{n}{k} \leq \left(\frac{en}{d}\right)^d \text{ where } d = \mathrm{vc}(\mathcal{F}).$$

Proof. Pajor's 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 $d = \mathrm{vc}(\mathcal{F})$, via the definition of VC dimension (Definition 8.3.1). Thus

$$|\mathcal{F}| \leq |\{\Lambda \subseteq \Omega : \ |\Lambda| \leq d\}| = \sum_{k=0}^{d} \binom{n}{k}$$

since the sum oin the right hand side gives the total number of subsets of an n-elements set with cardinalities at most k. This proces the first inequality of the Sauer-Shelah lemma. The second inequality follows from the bound on the binomial sum from Exercise 0.6. The proof is complete. □

Both Pajor's and Sauer-Shelah lemma are generally sharp (Exercise 8.19).

8.3.4 Growth Function

The Sauer-Shelah lemma assumes that the domain Ω is finite. What if the function classes $\mathcal{F}$ are on infinite domains like $\mathbb{R}^n$? It is often convenient to measure the complexity of $\mathcal{F}$ by the growth function:

Definition 8.3.10. 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:

$$\Pi_{\mathcal{F}}(n) = \sup \{ |\mathcal{F}|_{\Lambda}| : \ \Lambda \subset \Omega, \ |\Lambda| = n \}.$$

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

$$2^d \le \Pi_{\mathcal{F}}(n) \le \left(\frac{en}{d}\right)^d \text{ for all } n \ge d$$

if $d = \mathrm{vc}(\mathcal{F}) < \infty$. The lower bound is a restatement from the part before Pajor's lemma, and the upper bound follows from the Sauer-Shelah lemma (Lemma 8.3.9).
To see how the growth function can be useful, let's duduce from above the stability of VC dimension with respect to natural operations.

Proposition 8.3.11 (VC stability)**.** Let $\mathcal{F}, \mathcal{G}$ be two classes of Boolean functions on the same domain. Let

$$\mathcal{F} \wedge \mathcal{G} = \{ f \wedge g : \ f \in F, \ g \in G \}$$

where $f \wedge g$ denotes the pointwise minimum of the functions f and g. Then

$$\mathrm{vc}(\mathcal{F} \wedge \mathcal{G}) \le 10 \max(\mathrm{vc}(\mathcal{F}), \mathrm{vc}(\mathcal{G})).$$

The same holds for the pointwise maximum.

Proof. Assume towards a contradiction that $n := \mathrm{vc}(\mathcal{F} \wedge \mathcal{G}) > 10d$. 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 directly follow from the bounds of the growth function above, and the middle one is true by definition. However, we can calculate and show that $2^n > (en/d)^{2d}$ whenever $n > 10d$, which is a contradiction to the above. □

Proposition 8.3.11 can be extended to any particular way of combining classes of functions (Exercise 8.21). It can be helpful when we want to bound the VC dimension without computing it directly (which can be quite complicated). For example:

Example 8.3.12 (Strips)**.** A strip in $\mathbb{R}^n$ is a set of the form

$$\{x : \ |\langle a, x\rangle - b| \le c\} \text{ where } a \in \mathbb{R}^n \text{ and } b, c \in \mathbb{R}.$$

For an illustration, see Figure 8.7 below:

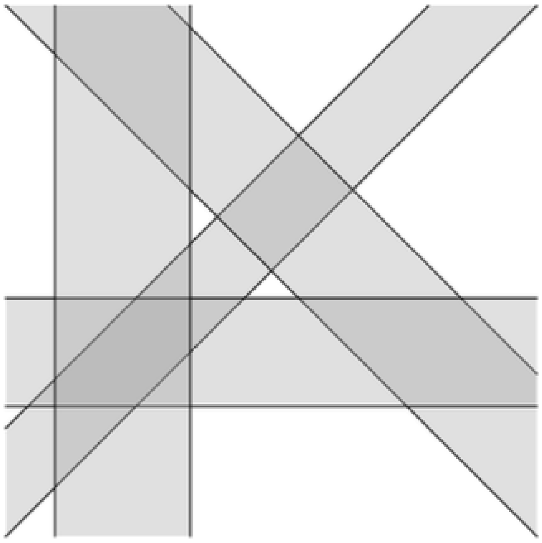


Figure 8.7 Four different strips in $\mathbb{R}^2$

Let $\mathcal{F}$ be the class of indicators of strips. Example 8.3.5 gives

$$\mathrm{vc}(\mathcal{F}) \leq 20(n+1) \leq 40n.$$

Indeed, each strip can be represented as the intersection of two half planes $\{x: \langle a,x\rangle - b \leq c\}$ and $\{x: \langle a,x\rangle - b \geq -c\}$. Thus the indicator of each strip is the pointwise mimimum of the indicators of two hald-spaces. Now apply the VC stability property (Proposition 8.3.11) and the result in Example 8.3.5 to get the bound above.

8.3.5 Covering Numbers via VC Dimension

Covering numbers usually grow exponentially with dimension. Now, let's refine this heuristic by replacing the algebraic dimension with the 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 the distance between functions as

$$d(f,g) = \|f-g\|_{L^2(\mu)} = \left(\mathbb{E}\left[(f-g)(X)^2\right]\right)^{1/2}$$

where X is a random variable with distribution μ. 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 above:

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 $\varepsilon \in (0,1)$,

$$\mathcal{N}(\mathcal{F}, L^2(\mu), \varepsilon 0) \leq \left(\frac{2}{\varepsilon}\right)^{Cd} \quad \text{where } d = \mathrm{vc}(\mathcal{F}).$$

For a first attempt of the proof, let's assume for a moment that Ω is finite, say $|\Omega| = n$. Then the Sauer-Shelah lemma (Lemma 8.3.9) gives

$$\mathcal{N}(\mathcal{F}, L^2(\mu), \varepsilon) \leq |\mathcal{F}| \leq \left(\frac{en}{d}\right)^d.$$

This not quite the result above, but it comes close. To tighten the bound, we need to get rid of n, and we'll do this by shrinking Ω. This lemma would 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-seperated, i.e.

$$\|f-g\|_{L^2(\mu)} > \varepsilon \text{ for all distinct } f, g \in \mathcal{F}.$$

IF $n \geq C\varepsilon^{-4} \log |F|$, then the empirical measure μ_n satisfies the following with probability at least 0.99:

$$\|f-g\|_{L^2(\mu_n)} > \varepsilon/2 \text{ for all distinct } f, g \in \mathcal{F}.$$

By definition of the empirical measure, $\|f-g\|_{L^2(\mu_n)}$ is the same as the metric we defined in the beginning of this subsection, but with the population average replaced by the sample average:

$$\|f-g\|_{L^2(\mu_n)} = \left(\frac{1}{n}\sum_{i=1}^{n}(f-g)(X_i)^2\right)^{1/2},$$

where X_i are i.i.d. copies of X.

Proof of Lemma 8.3.14. The proof is like that of the Johnson-Lindenstrauss lemma - concentration plus a union bound.
Fix a pair of distinct functions $f, g \in \mathcal{F}$, and consider

$$\|f-g\|^2_{L^2(\mu_n)} - \|f-g\|^2_{L^2(\mu)} = \frac{1}{n}\sum_{i=1}^{n} h(X_i) - \mathbb{E}\left[h(X)\right],$$

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

$$P\left(\left|\|f-g\|^2_{L^2(\mu_n)} - \|f-g\|^2_{L^2(\mu)}\right)\right| > \frac{\varepsilon^2}{4} \leq 2\exp\left(-cn\varepsilon^4\right).$$

Therefore, with probability at least $1 - 2\exp\left(-cn\varepsilon^4\right)$, we have

$$\|f-g\|^2_{L^2(\mu_n)} \geq \|f-g\|^2_{L^2(\mu)} - \frac{\varepsilon^2}{4} > \varepsilon^2 - \frac{\varepsilon^2}{4} > \frac{\varepsilon^2}{4},$$

by the 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

$$1 - |\mathcal{F}|^2 \cdot 2\exp\left(-cn\varepsilon^4\right),$$

the bound holds simultaneously for all distinct $f, g \in \mathcal{F}$. By out choise of n, choosing a large enough C yields the quantity above at least 0.99. The proof is complete. □

Proof of Theorem 8.3.13. By the packing-covering equivalence (Lemma 4.2.8), we can find

$$N = \mathcal{N}(\mathcal{F}, L^2(\mu), \varepsilon)$$

functions in $\mathcal{F}$ that are ε-seperated in the $L^2(\mu)$ metric. Set $n = \lfloor C\varepsilon^{-4} \log N \rfloor$ and apply Lemma 8.3.14 to the set of those functions. With positive probability, those functions stay $(\varepsilon/2)$-seperated in the metric $L^2(\mu_n)$ defined earlier, 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 the 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 the Sauer-Shelah lemma (Lemma 8.3.9) for $\mathcal{F}_n$ and Ω_n to get

$$N \leq \left(\frac{en}{d_n}\right)^{d_n} \leq \left(\frac{2C\varepsilon^{-4}\log N}{d_n}\right)^{d_n}$$

where $d_n = \mathrm{vc}(\mathcal{F}_n)$. Simplifying, we get

$$N \leq (2C\varepsilon^{-4})^{2d_n}.$$

Finally, replace $d_n = \mathrm{vc}(\mathcal{F}_n)$ by the larger quantity $d = \mathrm{vc}(\mathcal{F})$ and the proof is complete. □

8.3.6 VC Law of Large Numbers

Any class of Boolean functions with finite VC dimension has a LLN property:

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 $X, X_1, X_2, \ldots, X_n$ be independent random points in Ω with common distribution. Then

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}\left[f(X)\right]\right|\right] \le C\sqrt{\frac{\mathrm{vc}(\mathcal{F})}{n}}.$$

Proof. We will combine Dudley's inequality with the bound on the covering numbers (Theorem 8.3.13). But first, let's symmetrize the process using the empirical version of symmetrization (Exercise 8.11):

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}\left[f(X)\right]\right|\right] \le \frac{2}{\sqrt{n}}\mathbb{E}\left[\sup_{f\in\mathcal{F}}\underbrace{\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i f(X_i)\right|}_{Z_f}\right].$$

Condition on (X_i), leaving all randomness in the random signs $(\varepsilon)_i$. To use Dudley's inequality for the process $(Z_f)_{f\in\mathcal{F}}$, we need to check that the increments are subgaussian. Triangle inequality gives

$$|Z_f - Z_g| \le \frac{1}{\sqrt{n}}\left|\sum_{i=1}^{n}\varepsilon_i (f-g)(X_i)\right|,$$

so using Proposition 2.7.1 and the fact that $\|\varepsilon_i\|_{\psi_2} \lesssim 1$, we get:

$$\begin{aligned}\|Z_f - Z_g\|_{\psi_2} &\le \frac{1}{\sqrt{n}}\left\|\sum_{i=1}^{n}\varepsilon_i (f-g)(X_i)\right\|_{\psi_2}\\ &\lesssim \left(\frac{1}{n}\sum_{i=1}^{n}(f-g)(X_i)^2\right)^{1/2}\\ &= \|f-g\|_{L^2(\mu_n)}\end{aligned}$$

where μ_n is the empirical measure, as mentioned before.
Now use Dudley's inequality (Theorem 8.1.3) conditionallyh on (X_i), then remove the conditioning by taking expectation with respect to (X_i). Check that $\mathrm{diam}(\mathcal{F}) \le 1$. We get

$$\frac{2}{\sqrt{n}}\mathbb{E}\left[\sup_{f\in\mathcal{F}} Z_f\right] \lesssim \frac{1}{\sqrt{n}}\mathbb{E}\left[\int_0^1 \sqrt{\log \mathcal{N}(\mathcal{F}, L^2(\mu_n), \varepsilon)}\, d\varepsilon\right].$$

Finally, we use Theorem 8.3.13 to bound the covering numbers:

$$\log \mathcal{N}(\mathcal{F}, L^2(\mu_n), \varepsilon) \lesssim \mathrm{vc}(\mathcal{F}) \log\left(2/\varepsilon\right).$$

Substituting this into the bound above, we get the integral of $\sqrt{\log\left(2/\varepsilon\right)}$, which is bounded by an absolute constant, leading to

$$\frac{2}{\sqrt{n}}\mathbb{E}\left[\sup_{f\in\mathcal{F}} Z_f\right] \lesssim \sqrt{\frac{\mathrm{vc}(\mathcal{F})}{n}},$$

hence the proof is complete. □

Remark 8.3.16 (Rademacher complexity)**.** If $\mathcal{F}$ is a class of Boolean functions with finite VC dimension, then the expression

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(x_i)\right|\right]$$

is called the *Rademacher complexity* of $\mathcal{F}$ on a given set of points $x_1, \ldots, x_n \in \Omega$. 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 showed that the 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,

$$F(x) = P(X \leq x),$$

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 $X_i \leq x$:

$$F_n(x) := \frac{1}{n}|\{i : \ X_i \leq x\}|.$$

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 CDF F. Then

$$\mathbb{E}\left[\|F_n - F\|_{L^\infty}\right] = \mathbb{E}\left[\sup_{x\in\mathbb{R}} |F_n(x) - F(x)|\right] \leq \frac{C}{\sqrt{n}}.$$

Proof. This is just a restatement of Theorem 8.3.15 for $\Omega = \mathbb{R}$ and the class of indicators of half-infinite intervals

$$\mathcal{F} := \{\mathbf{1}_{(-\infty,x]} : \ x \in \mathbb{R}\},$$

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 distirbution on the unit square $[0,1]^2$, as in Figure 8.8:

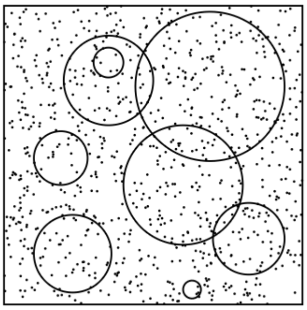


Figure 8.8 The VC law of large numbers implies that the number of points in each circle is proportional to its area with $O(\sqrt{n})$ error.

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 (Exercise 8.13). Then, with high probability, the sample satisfies:

$$\text{fraction of points in } \mathcal{C} = \text{Area}(\mathcal{C}) + O(1/\sqrt{n})$$

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-values functions $\mathcal{F}$ on a set Ω

is called a *uniform Glivenko-Cantelli* class if, for any $\varepsilon > 0$,

$$\lim_{n\to\infty} \sup_{\mu} P\left(\sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}\left[f(X)\right]\right| > \varepsilon\right) = 0,$$

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's inequality implies that any Boolean class with finite VC dimension is uniform Glivenko-Cantelli. The converse is alse true (Exercise 8.27), so in fact the two are equivalent.

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 a few sample points $X_1, \ldots, X_n$ drawn independently from some distribution on Ω. Therefore, our *training data* is

$$(X_i, T(X_i)),\ i = 1, \ldots, n.$$

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).

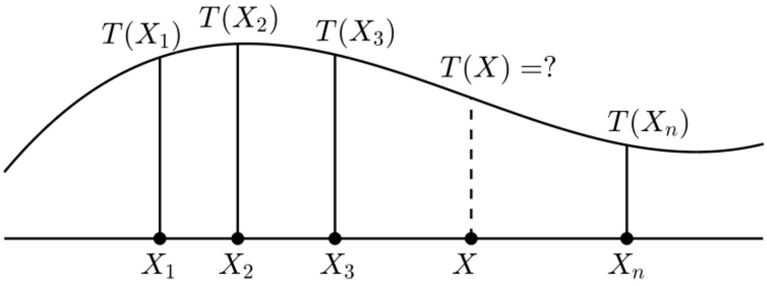


Figure 8.9 We want to learn a function $T : \Omega \to \mathbb{R}$ (a "target function") from its values on the i.i.d. training data $X_1, \ldots, X_n$, so we can predict $T(X)$ for a new random point X.

Example 8.4.1 (Classification)**.** An important type of learning problems is classification, where the function T is Boolean (takes value 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 hralth 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 - 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

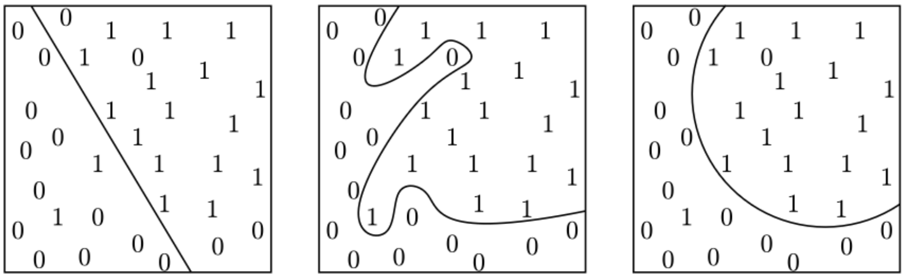
Given the training data, we want to find a function $f : \Omega \to \mathbb{R}$ that approximates T. We aim to minimize the *risk*, defined as

$$R(f) = \mathbb{E}\left[(f(X) - T(X))^2\right].$$

Example 8.4.2. In classification problems where T and f are boolean functions, the risk is the probability of misclassification:

$$R(f) = P(f(X) \neq T(X)).$$

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 don't know this up front, we limit our guesses f to some class of functions $\mathcal{F}$, callled 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* (Figure 8.10a) and miss real patterns. Too complex, we might *overfit*, just memorizing the training data rather than generalizing from it (Figure 8.10b). The sweet spot is a hypothesis class that is just enough to capture the real patterns, without fitting the 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 pick a hypothesis space $\mathcal{F}$, we might just choose the best function f^* in it - one that minimizes the risk:

$$f^* = \arg\min_{f\in\mathcal{F}} R(F).$$

The catch is that we can't actually compute $R(F)$ since we don't have access to the population Ω. Solution? Use the training set and take the expectation.

Definition 8.4.3. For a function $f : \Omega \to \mathbb{R}$, define the empirical risk and empirical minimizer as

$$R_n(f) := \frac{1}{n}\sum_{i=1}^{n}(f(X_i) - T(X_i))^2, \;\; f_n^* = \arg\min_{f\in\mathcal{F}} R_n(f).$$

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.

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

$$\mathbb{E}\left[R(f_n^*)\right] \le R(f^*) + C\sqrt{\frac{\mathrm{vc}(\mathcal{F})}{n}}.$$

Proof. **Step 1: Excess risk.** The following bound holds pointwise:

$$R(f_n^*) - R(f^*) \le 2\sup_{f\in\mathcal{F}} |R_n(f) - R(f)|.$$

To check this, denote $\varepsilon := \sup_{f \in \mathcal{F}} |R_n(f) - R(f)|$ and write

$$\begin{aligned} R(f_n^*) &\leq R_n(f_n^*) + \varepsilon \quad &&(\text{since } f_n^* \in \mathcal{F} \text{ by construction}) \\ &\leq R_n(f^*) + \varepsilon \quad &&(\text{since } f_n^* \text{ minimizes } R_n \text{ in the class} \mathcal{F}) \\ &\leq R(f^*) + 2\varepsilon \quad &&(\text{since } f^* \in \mathcal{F} \text{ by construction}). \end{aligned}$$

Subtracting $R(f^*)$ from both sides gives the claim.

Step 2: Applying VC law of large numbers. Thanks to the claim above, it is enough to show that

$$\mathbb{E}\left[\sup_{f \in \mathcal{F}} |R_n(f) - R(f)|\right] \lesssim \sqrt{\frac{\text{vc}(\mathcal{F})}{n}}.$$

Recalling the definitions of the empirical and population risk, we can rewrite the above as

$$\mathbb{E}\left[\sup_{\ell \in \mathcal{L}} \left|\frac{1}{n}\sum_{i=1}^{n} \ell(X_i) - \mathbb{E}\left[\ell(X)\right]\right|\right] \lesssim \sqrt{\frac{\text{vc}(\mathcal{F})}{n}},$$

where $\mathcal{L} = \{(f - T)^2 : \ f \in \mathcal{F}\}$. A moment's thought reveals that (Exercise 8.29) $\text{vc}(\mathcal{L}) = \text{vc}(\mathcal{F})$. Then, 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" 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.
How well do we do? Since the true circle $\mathcal{C}$ gives zero error, and the VC dimension of circles it 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 decreases 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 $F(f^*)$ comes from an imperfect choice of the hypothesis class (underfitting). 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 $O(\sqrt{\text{vc}(\mathcal{F})/n})$ grows. To keep it in check, we may use more training data (increase n) to avoid overfitting.

8.5 Generic Chaining

Generic chaining improves the loose bound that Dudley's inequality can exhibit sometimes. It is essentially a technique of the chaining method we developed throughout the proof to Dudley's inequality (Theorem 8.1.4).

8.5.1 A Makeover of Dudley's Inequality

Recall the bound we obtained by chaining in Theorem 8.1.4:

$$\mathbb{E}\left[\sup_{t \in T} X_t\right] \lesssim \sum_{k=\kappa+1}^{\infty} \varepsilon_{k-1}\sqrt{\log |T_k|},$$

where $\varepsilon = 2^{-k}$, T_k are smallest ε-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

$$|T_0| = 1, \ |T_k| \leq 2^{2^k}, \ k = 1, 2, \ldots$$

and define

$$\varepsilon_k = \sup_{t\in T} d(t, T_k)$$

where $d(t, T_k)$ denotes the distance from t to the set T_k (the distance between a point t and a set A in a metric space is $d(t, A) = \inf\{d(t, a) : \ a \in A\}$).
Each T_k is then an ε_k-net, and the chaining bound becomes

$$\mathbb{E}\left[\sup_{t\in T} X_t\right] \lesssim \sum_{k=1}^{\infty} 2^{k/2} \sup_{t\in T} d(t, T_{k-1}),$$

or after reindexing,

$$\mathbb{E}\left[\sup_{t\in T} X_t\right] \lesssim \sum_{k=0}^{\infty} 2^{k/2} \sup_{t\in T} d(t, T_k).$$

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 above. The resulting quantity has a name:

Definition 8.5.1. Let (T, d) be a metric space. A sequence of subsets $(T_k)_{k=0}^{\infty}$ of T satisfying

$$|T_0| = 1, \ |T_k| \le 2^{2^k}, \ k = 1, 2, \dots$$

is called an admissible sequence.
The γ_2 functional of T is defined as

$$\gamma_2(T, d) = \inf_{(T_k)} \sup_{t\in T} \sum_{k=0}^{\infty} 2^{k/2} d(t, T_k)$$

where the infimu is over all admissible sequences.

The supremum in the γ_2 functional is outside the sum, hence it is smaller than the Dudley sum above. That might seem like a small change, but it can make a big difference in some cases (Exercise 8.34). Good news: we can improve Dudley's inequality (Theorem 8.1.4) by replacing the Dudley sum (or integral) by 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. Then

$$\mathbb{E}\left[\sup_{t\in T} X_t\right] \le CK\gamma_2(T, d).$$

Proof. We'll use the chaining method from the proof of Dudley's inequality, but more carefully.
Step 1: Chaining setup. As before, without loss of generality assume $K = 1$ and that T is finite, which makes $\gamma_2(T, d)$ finite. Let (T_k) be an admissible sequence of subsets of T which almost attains the supremum in Definition 8.5.1:

$$\sup_{t\in T} \sum_{k=0}^{\infty} 2^{k/2} d(t, T_k) \le 2\gamma_2(T, d) < \infty.$$

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 infinitely many sets T_k, so $d(t, T_k) > \varepsilon$ for all those k and some fixed $\varepsilon > 0$, making the series above diverge.
We walk from t_0 to a general point $t \in T$ along the (finite) chain

$$t_0 = \pi_0(t) \to \pi_1(t) \to \pi_2(t) \to \cdot \to t$$

of points $\pi_k(t) \in T_k$ that are chosen as best approximations to t in T_k, i.e.

$$d(t, \pi_k(t)) = d(t, T_k).$$

Again, the displacement $X_t - X_{t_0}$ can be expressed as a telescoping sum:

$$X_t - X_{t_0} = \sum_{k=1}^{\infty} (X_{\pi_k(t)} - X_{\pi_{k-1}(t)}).$$

Step 2: Controlling the increments. This is where we need to be more caredul. We would like that, with high probability, the following event holds:

$$|X_{\pi_k(t)} - X_{\pi_{k-1}(t)}| \lesssim 2^{k/2} d(t, T_k) \ \forall k \in \mathbb{N}, \forall t \in T.$$

Summing over all k would lead to a desired bound in terms of $\gamma_2(T, d)$.
To prove the above, let's fix k and t first. The subgaussian assumption gives

$$\|X_{\pi_k(t)} - X_{\pi_{k-1}(t)}\|_{\psi_2} \leq d(\pi_k(t), \pi_{k-1}(t)).$$

So for every $u \geq 0$, the event

$$|X_{\pi_k(t)} - X_{\pi_{k-1}(t)}| \leq Cu 2^{k/2} d(\pi_k(t), \pi_{k-1}(t)) \quad (*)$$

holds with probability at least

$$1 - 2\exp(-8u^2 2^k),$$

where we get the constant 8 by choosing C to be big enough.
Now unfix $t \in T$ by taking a union bound over

$$|T_k| \cdot |T_{k-1}| \leq |T_k|^2 \leq 2^{2^{k+1}}$$

pairs $(\pi_k(t), \pi_{k-1}(t))$. Also, unfix k by taking a union bound over all $k \in \mathbb{N}$. Then $(*)$ holds simultaneously for all $t \in T$ and $k \in \mathbb{N}$ with probability at least

$$1 - \sum_{k=1}^{\infty} 2^{2^{k+1}} \cdot 2\exp(-8u^2 2^k) \geq 1 - 2\exp(-u^2).$$

Step 3: Summing up the increments. In the event that the bound $(*)$ does hold for all $t \in T$ and $k \in \mathbb{N}$, we can sum up the inequalities over $k \in \mathbb{N}$ and plug in the result into the chaining sum. We get

$$|X_t - X_{t_0}| \lesssim u \sum_{k=1}^{\infty} 2^{k/2} d(\pi_k(t), \kappa_{k-1}(t)) \quad (**),$$

where the notation $\lesssim$ hides an absolute constant factor. By the triangle inequality,

$$d(\pi_k(t), \pi_{k-1}(t)) \leq d(\pi_k(t), t) + d(t, \pi_{k-1}(t)).$$

Using that bound, reindexing, and plugging in the chaining bound from step 1, we get that the tright-hand side of $(**)$ is at most $Cu\gamma_2(T, d)$, that is

$$|X_t - X_{t_0}| \lesssim u\gamma_2(T, d).$$

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

$$\left\| \sup_{t \in T} |X_t - X_{t_0}| \right\|_{\psi_2} \lesssim \gamma_2(T, d).$$

This quickly that the conclusion of Theorem 8.5.2, and we're done. □

Remark 8.5.3 (Generic chaining: supremum of increments)**.** Similarly to Dudley's inequality (Remark 8.1.5), the generic chaining actually gives

$$\mathbb{E}\left[\sup_{t,s\in T} |X_t - X_s|\right] \leq CK\gamma_2(T,d),$$

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 aseen this before in Remark 8.1.6.
Assuming T is finite, for every $u \geq 0$, the event

$$\sup_{t,s\in T} |X_t - X_s| \leq CK\left[\gamma_2(T,d) + u \cdot \operatorname{diam}(T)\right]$$

holds with probability at least $1 - 2\exp\left(-u^2\right)$ (Exercise 8.35). For Gaussian processes, we can directly deduce this from Gaussian concentration.

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's inequality. But it is often worth the effort - 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 mentioned before. Then

$$c\gamma_2(T,d) \leq \mathbb{E}\left[\sup_{t\in T} X_t\right] \leq C\gamma_2(T,d).$$

Proof. The upper bound directly comes from generic chaining (Theorem 8.5.2). The lower bound is tricker hence not included in the text. □

The upper bound holds not just for Gaussian but also for all subgaussian processes. Therefore, by combining the upper and lower bounds, we can bound any subgaussian processes 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

$$\|X_t - X_s\|_{\psi_2} \leq K\|Y_t - Y_s\|_{L^2} \text{ for all } t,s \in T.$$

Then

$$\mathbb{E}\left[\sup_{t\in T} X_t\right] \leq CK\mathbb{E}\left[\sup_{t\in T} Y_t\right].$$

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 (Theorem 8.5.5) and we get

$$\mathbb{E}\left[\sup_{t\in T} X_t\right] \lesssim K\gamma_2(T,d) \lesssim K\mathbb{E}\left[\sup_{t\in T} Y_t\right].$$

□

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!

We can also 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 $g \sim N(0, I_n)$. From section 7.5,

$$w(T) = \mathbb{E}\left[\sup_{x \in T} \langle g, x \rangle\right]$$

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

$$\|X_x - X_y\|_{\psi_2} \leq K\|x - y\|_2 \text{ for all } x, y \in T.$$

Then

$$\mathbb{E}\left[\sup_{x \in T} X_x\right] \leq CKw(T).$$

Remark 8.5.9 (Subgaussian width Gaussian width)**.** A nice consequence: if X is a subgaussian random vector in $\mathbb{R}^n$, then

$$\mathbb{E}\left[\sup_{t \in T} \langle X, t \rangle\right] \leq CKw(T) \text{ for any bounded set } T \subset \mathbb{R}^n,$$

where $K = \|X\|_{\psi_2}$. Just apply Corollary 8.5.8 to the process $(\langle X, x \rangle)_{x \in T}$, whose increments satisfy

$$\|\langle X, x \rangle - \langle X, y \rangle\|_{\psi_2} = \|\langle X, x - y \rangle\|_{\psi_2} \leq K\|x - y\|_2$$

by definition of a subgaussian random vector.

8.6 Chevet Inequality

Talagrand's comparison inequality (generic chaining) is powerful and works in a wide range of settings. Let's use it on random quadratic forms:

$$\sup_{x \in T, y \in S} \langle Ax, y \rangle \leq ?$$

where A is a random matrix and T, S are bounded sets.
A special case where T, S are Euclidean balls leads to the operator norm of A, which we did some analysis already (Theorem 4.4.3). Here we go for a more general setting, and we'll just use two geometric quantities: the *Gaussian width* $w(T)$, and the *radius*, defined as

$$\text{rad}(T) := \sup_{x \in T} \|x\|_2.$$

Theorem 8.6.1 (Subgaussian Chevet's 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

$$\mathbb{E}\left[\sup_{x \in T, y \in S} \langle Ax, y \rangle\right] \leq CK[w(T)\text{rad}(S) + w(S)\text{rad}(T)]$$

where $K = \max_i \|A_i\|_{\psi_2}$. The same holds if "rows" is replaced by "columns".

Proof. We'll follow the proof for Theorem 7.3.1 (reference here) but with Talagrand's comparison inequality instead of Sudakov-Fernique's.
Without loss of generality, assume K = 1. We need to bound the random process

$$X_{uv} \langle Au, v \rangle, \ u \in T, v \in S.$$

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 the triangle inequality and the subgaussian assumption (Exercise 3.34), we get

$$\begin{aligned}\|X_{uv} - X_{wz}\|_{\psi_2} &\le \|\langle A(u-w), v\rangle\|_{\psi_2} + \|\langle Aw, v-z\rangle\|_{\psi_2} \\ &\lesssim \|u-w\|_2\|v\|_2 + \|v-z\|_2\|w\|_2 \\ &\le \|u-w\|_2 \mathrm{rad}(S) + \|v-z\|_2 \mathrm{rad}(T) \quad (*).\end{aligned}$$

Let's pick a simpler Gaussian process (Y_{uv}) for Talagrand's comparison inequality (Corollary 8.5.6). The increment bound points us to a good choice:

$$Y_{uv} := \langle g, u\rangle \,\mathrm{rad}(S) + \langle h, v\rangle \,\mathrm{rad}(T),$$

where $g \sim N(0, I_n)$ and $h \sim N(0, I_m)$ are independent. The increments of this process are

$$\|Y_{uv} - Y_{wz}\|_{L^2}^2 = \|u-w\|_2^2 \mathrm{rad}(S)^2 + \|v-z\|_2^2 \mathrm{rad}(T)^2.$$

Comparing this to the bound $(*)$, 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's comparison inequality (Corollary 8.5.6), we finish the proof:

$$\begin{aligned}\mathbb{E}\left[\sup_{u\in T, v\in S} X_{uv}\right] &\lesssim \mathbb{E}\left[\sup_{u\in T, v\in S} Y_{uv}\right] \\ &= \mathbb{E}\left[\sup_{u\in T} \langle g, u\rangle\right] \mathrm{rad}(S) + \mathbb{E}\left[\sup_{v\in S} \langle h, v\rangle\right] \mathrm{rad}(T) \\ &= w(T)\mathrm{rad}(S) + w(S)\mathrm{rad}(T).\end{aligned}$$

□

Remark 8.6.2 (Operator norms of random matrices)**.** For the special case $T = S^{n-1}$, $S = S^{m-1}$, Chevet's inequality gives up the familiar sharp bound on the operator norm:

$$\mathbb{E}\left[\|A\|\right] \le CK(\sqrt{n} + \sqrt{m}),$$

which we proved earlier using ε-nets. But this new approach gives more flexibility! For example, picking T, S as ℓ^p balls gives the $\|A\|_{p\to q}$ norm of a random matrix (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's inequality with sharp constant 1:

$$\mathbb{E}\left[\sup_{x\in T, y\in S} \langle Ax, y\rangle\right] \le w(T)\mathrm{rad}(S) + w(S)\mathrm{rad}(T),$$

and a reverse inequality up to a constant (Exercise 8.39). Later, we'll further improve Gaussian Chevet inequality in Section 9.7.1.

9 Deviations of Random Matrices on Sets

The main question in this chapter is: How does an $m \times n$ matrix act on a general set $t \subset \mathbb{R}^n$?

9.1 Matrix Deviation Inequality

Take an $m \times n$ random matrix X with independent, isotropic, and subgaussian rows. The concentration of the norm (Theorem 3.1.1) tells us that for any fixed vector $x \in \mathbb{R}^n$, the approximation

$$\|Ax\|_2 \approx \sqrt{m}\|x\|_2$$

holds with high probability.
Let's ask something more general: Is it true that with high probability, the equation above holds *simultaneously* for many vectors $x \in \mathbb{R}^n$? To quantify how many, pick some 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.

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 $T \subset \mathbb{R}^n$,

$$\mathbb{E}\left[\sup_{x \in T} \left| \|Ax\|_2 - \sqrt{m}\|x\|_2 \right|\right] \leq CK^2\gamma(T),$$

where $\gamma(T)$ is the Gaussian complexity from Section 7.5.3, defined as

$$\gamma(T) = \mathbb{E}\left[\sup_{x \in T} |\langle g, x\rangle|\right], \; g \sim N(0, I_n),$$

and $K = \max_i \|A_i\|_{\psi_2}$.

The plan is to deduce this from Talagrand's comparison inequality (Corollary 8.5.8). To do that, we just have to check the random process

$$Z_x := \|Ax\|_2 - \sqrt{m}\|x\|_2$$

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 Z_x defined above has subgaussian increments:

$$\|Z_x - Z_y\|_{\psi_2} \leq CK^2\|x - y\|_2 \text{ for all } x, y \in \mathbb{R}^n,$$

here $K = \max_i \|A_i\|_{\psi_2}$.

Once we have proved this theorem, we plug it into Talagrand's comparison inequality (Exercise 8.37 (a)) and get

$$\mathbb{E}\left[\sup_{x \in T} |Z_x|\right] \leq CK^2\gamma(T)$$

which directly gives Theorem 9.1.1. So, all we have to do is prove Theorem 9.1.2 - and it is in fact easier since it's for fixed x and y.

Proof of Theorem 9.1.2. This argument will be a bit longer than usual, so we'll (hopefully) make it easier by starting with simpler cases and building up from there.
Step 1: Unit vector x and zero vector y. If $\|x\|_2 = 1$ and $y = 0$, the inequality in the theorem statement becomes

$$\left\| \|Ax\|_2 - \sqrt{m} \right\|_{\psi_2} \leq CK^2.$$

The random vector $Ax \in \mathbb{R}^m$ has independent, subgaussian coordinates $\langle A_i, x\rangle$, which satisfy

$$\mathbb{E}\left[\langle A_i, x\rangle^2\right] = 1$$

by isotropy. So, the equation above follows from the concentration of the 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 the theorem statement becomes

$$\| \|Ax\|_2 - \|Ay\|_2 \|_{\psi_2} \leq CK^2 \|x - y\|_2. \quad (*)$$

Since the *squared* ℓ^2 norm would be simpler to work with (no square roots), let's prove a version of the equation above with squared norms. Here's a good guess to what it should look like: with high probability,

$$\begin{aligned} \|Ax\|_2^2 - \|Ay\|_2^2 &= (\|Ax\|_2 + \|Ay\|_2) \cdot (\|Ax\|_2 - \|Ay\|_2) \\ &\lesssim \sqrt{m} \cdot \|x - y\|_2. \end{aligned}$$

This seems reasonable because $\|Ax\|_2$ and $\|Ay\|_2$ are roughly $\sqrt{m}$ by step 1, and hence we expect $(*)$ to hold.
Let's go ahead and prove this. Expand the matrix-vector product:

$$\|Ax\|_2^2 - \|Ay\|_2^2 = \sum_{i=1}^{m} (\langle A_i, x \rangle^2 - \langle A_i, y \rangle^2) = \sum_{i=1}^{m} \langle A_i, x + y \rangle \langle A_i, x - y \rangle ,$$

then dividing both sides by $\|x - y\|_2$, getting

$$\Delta := \frac{\|Ax\|_2^2 - \|Ay\|_2^2}{\|x - y\|_2^2} = \sum_{i=1}^{m} \langle A_i, u \rangle \langle A_i, v \rangle ,$$

where

$$u := x + y \text{ and } v := \frac{x - y}{\|x - y\|_2}.$$

Our goal is to show that $|\Delta| \lesssim \sqrt{m}$ with high probability.
What do we see in Δ? 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,

$$\mathbb{E}\left[\langle A_i, x \rangle^2 - \langle A_i, y \rangle^2 \right] = 1 - 1 = 0.$$

Moreover, these are *subexponential.* Lemma 2.8.6 and the subgaussian assumption on A_i give

$$\begin{aligned} \|\langle A_i, u \rangle \langle A_i, v \rangle\|_{\psi_1} &\leq \|\langle A_i, u \rangle\|_{\psi_2} \cdot \|\langle A_i, v \rangle\|_{\psi_2} \\ &\leq K\|u\|_2 \cdot K\|v\|_2 \\ &\leq 2K^2 \end{aligned}$$

where in the last step, we used that $\|u\|_2 \leq \|x\|_2 + \|y\|_2 \leq 2$ and $\|v\|_2 = 1$. So we can apply Bernstein's inequality (Theorem 2.9.1) and get

$$P(|\Delta| \geq t\sqrt{m}) \leq 2 \exp\left[-c \min\left(\frac{t^2}{K^4}, \frac{t\sqrt{m}}{K^2} \right) \right] \leq 2 \exp\left(-\frac{c_1 t^2}{K^4} \right)$$

for any $0 \leq t \leq \sqrt{m}$.
Step 3: Unit vector x, y and the original process. Now let's get rid of the squares and prove the original inequality $(*)$ for all unit vectors x and y.
Using the definition of the subgaussian norm (Proposition 2.6.6 (i) and Remark 2.6.3), $(*)$ becomes

$$p(s) := P\left(\frac{|\|Ax\|_2 - \|Ay\|_2|}{\|x - y\|_2^2} \right) \leq 4 \exp\left(-\frac{cs^2}{K^2} \right) \text{ for all } s > 0.$$

(Here the constant 4 instead of 2 will give us a little more room to maneuver.)
Now we have two cases: *Case 1:* $s \leq 2\sqrt{m}$. Let's use the result from step 2. To use it, multiply both sides of the inequality that defines $p(s)$ by $\|Ax\|_2 + \|Ay\|_2$, and recall the definition of Δ to get

$$p(s) = P(|\Delta| \geq s(\|Ax\|_2 + \|Ay\|_2)) \leq P(|\Delta| \geq s\|Ax\|_2).$$

We know from step 2 that $\|Ax\|_2 \approx \sqrt{m}$ with high probability. So it makes sense to consider two cases: THe likely case where $\|Ax\|_2 \geq \sqrt{m}/2$ and thus $|\Delta| \geq s\sqrt{m}/2$, and the unlikely case where $\|Ax\|_2 < \sqrt{m}/2$ (and we deop the clause about Δ, only increasing the probability). This leads to

$$p(s) \leq P\left(|\Delta| \geq \frac{s\sqrt{m}}{2}\right) + P\left(\|Ax\|_2 < \frac{\sqrt{m}}{2}\right) = p_1(s) + p_2(s).$$

The result from Step 2 handles the likely case:

$$p_1(s) \leq 2\exp\left(-\frac{cs^2}{K^4}\right),$$

while the result of Step 1 together with the triangle inequality handle the unlikely case:

$$p_2(s) \leq P\left(|\|Ax\|_2 - \sqrt{m}| > \frac{\sqrt{m}}{2}\right) \leq 2\exp\left(-\frac{cs^2}{K^4}\right).$$

Adding up the two, we get the desired bound:

$$p(s) \leq 4\exp\left(-\frac{cs^2}{K^4}\right).$$

Case 2: $s > 2\sqrt{m}$. By the triangle inequality, $|\|Ax\|_2 - \|Ay\|_2| \leq \|A(x-y)\|_2$, so

$$\begin{aligned} p(s) \leq P(\|Au\|_2 \geq s) \quad \left(u = \frac{x-y}{\|x-y\|_2} \text{ as before}\right) \\ &\leq P(\|Au\|_2 - \sqrt{m} \geq s/2) \quad (\text{Since } s > 2\sqrt{m}) \\ &\leq 2\exp\left(-\frac{cs^2}{K^4}\right) \quad (\text{By step 1}). \end{aligned}$$

Therefore in either case, we get the desired bound.
Step 4: Full generality. Finally, let's show the result for arbitrary $x, y \in \mathbb{R}^n$. By scaling, we can assume without loss of generality that

$$\|x\|_2 = 1 \text{ and } \|y\|_2 \geq 1.$$

Project Y onto the unit sphere, i.e. consider $\bar{y} := y/\|y\|_2$ (See figure 9.1):

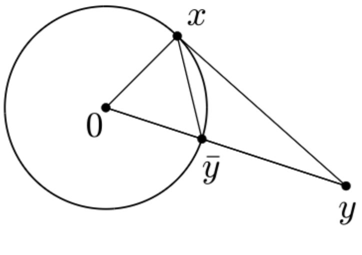


Figure 9.1 The triangle inequality can be approximately reversed for these three vectors: $\|x - \bar{y}\|_2 + \|\bar{y} - y\|_2 \leq \sqrt{2}\|x - y\|_2$.

Then by triangle inequality:

$$\|Z_x - Z_y\|_{\psi_2} \leq \|Z_x - Z_{\bar{y}}\|_{\psi_2} + \|Z_{\bar{y}} - Z_y\|_{\psi_2}.$$

Since both $x, \bar{y}$ are unit vectors, the result of Step 3 handles the first term:

$$\|Z_x - Z_{\bar{y}}\|_{\psi_2} \leq CK^2\|x - \bar{y}\|_2.$$

To handle the second term, note that $\bar{y}$ and y are colinear vectors. So by homogeneity,

$$\|Z_{\bar{y}} - Z_y\|_{\psi_2} = \|\bar{y} - y\|_2 \cdot \|Z_{\bar{y}}\|_{\psi_2}.$$

Now, since $\bar{y}$ is a unit vector, the result of Step 1 gives $\|Z_{\bar{y}}\|_{\psi_2} \leq CK^2$. Conbining the two terms, we conclude that

$$\|Z_x - Z_y\|_{\psi_2} \leq CK^2(\|x - \bar{y}\|_2 + \|\bar{y} - y\|_2).$$

At first this looks bad - we wnat to bound right hand side by $\|x - y\|_2$, but the triangle inequality goes the other way! Luckily, in our case (via the projection of y), the triangle inequality can be approximately reversed (Exercise 9.1):

$$\|x - \bar{y}\|_2 + \|\bar{y} - y\|_2 \leq \sqrt{2}\|x - y\|_2.$$

Plugging this into the bound above, we get the desired bound:

$$\|Z_x - Z_y\|_{\psi_2} \leq \sqrt{2}CK^2\|x - y\|_2,$$

which proves the theorem. □

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 $\mathbb{E}\left[\|Ax\|_2\right]$ (Exercise 9.2).

Remark 9.1.4 (Matrix deviations: a high-probability bound)**.** We only stated Theorem 9.1.1 as an expectation bound, but we can upgrade it to a high-probability bound via the high-probability version of Talagrand's inequality (Exercise 8.37(b)). For any $u \geq 0$, the event

$$\sup_{x \in T} \left| \|Ax\|_2 - \sqrt{m}\|x\|_2 \right| \leq CK^2[w(T) + u \cdot \mathrm{rad}(T)]$$

holds with probability at least $1 - 2\exp(-u^2)$. Can you see why the above implies the expectation bound?

Remark 9.1.5 (Matrix deviations of squares)**.** If you are interested in deviations of the quadratic process $\|Ax\|_2^2$, we can also deduce this from Theorem 9.1.1 (Exercise 9.3):

$$\mathbb{E}\left[\sup_{x \in T} \left| \|Ax\|_2^2 - m\|x\|_2^2 \right|\right] \leq CK^4\gamma(T)^2 + CK^2\sqrt{m}\mathrm{rad}(T)\gamma(T).$$

9.2 Random Matrices, Covariance Estimation, and Johnson-Lindenstrauss

The matrix deviation inequality has lots of useful consequences. We'll go over a few of them in this 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 Chapter 4.
Here is the quick check: since for the sphere we have

$$\mathrm{rad}(T) = 1 \text{ and } w(T) \leq \sqrt{n},$$

the matrix deviation inequality shows that the event

$$\sqrt{m} - CK^2(\sqrt{n} + u) \leq \|Ax\|_2 \leq \sqrt{m} + CK^2(\sqrt{n} + u) \text{ for all } x \in S^{n-1}$$

holds with probability at least $1 - 2\exp(-u^2)$. Then taking the min/max gives

$$\sqrt{m} - CK^2(\sqrt{n} + u) \leq \sigma_n(A) \leq \sigma_1(A) \leq \sqrt{m} + CK^2(\sqrt{n} + u) \text{ for all } x \in S^{n-1},$$

giving Theorem 4.6.1 in a different way.

9.2.2 Random Projections of Sets

From the matrix deviation inequality, we also get a sharper bound on the random projection bound in 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

$$\mathbb{E}\left[\operatorname{diam}(PT)\right] \leq \sqrt{\frac{m}{n}}\operatorname{diam}(T) + CK^2 w_s(T).$$

Here $K = \max_i \|A_i\|_{\psi_2}$ and $w_s(T)$ is the spherical width of T.

Proof. Theorem 9.1.1 implies via triangle inequality:

$$\mathbb{E}\left[\sup_{x \in T}\|Ax\|_2\right] \leq \sqrt{m}\sup_{x \in T}\|x\|_2 + CK^2\gamma(T),$$

which we can rewrite in terms of the radii of AT and T:

$$\mathbb{E}\left[\operatorname{rad}(AT)\right] \leq \sqrt{m}\operatorname{rad}(T) + CK^2\gamma(T).$$

Applying this bound for the difference set $T - T$ instead of T to get

$$\mathbb{E}\left[\operatorname{diam}(AT)\right] \leq \sqrt{m}\operatorname{diam}(T) + 2CK^2 w(T),$$

where we used Lemma 7.5.11 (a) to pass from Gaussian complexity to Gaussian width. Divide both sides by $\sqrt{n}$ completes the proof. □

9.2.3 Covariance Estimation for Low-dimensional Distributions

Let's visit the covariance estimation problem from Section 4.7. We want to estimate the population covariance via the sample covariance matrix $\Sigma_m = \sum_{i=1}^m X_i X_i^T$.
In general, $O(n \log n)$ samples are enough (Section 5.6), but for subgaussian distributions, $m = O(n)$ is enough.
It gets even better for approximately low-dimensional distributions. If a distribution concentrates near a 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 spefically, assume that these exists $K \geq 1$ such that

$$\|\langle X, x\rangle\|_{\psi_2} \leq K\|\langle X, x\rangle\|_{L^2} \text{ for any } z \in \mathbb{R}^n.$$

Then, for every positive integer m,

$$\mathbb{E}\left[\|\Sigma_m - \Sigma\|\right] \leq CK^4\left(\sqrt{\frac{r}{m}} + \frac{r}{m}\right)\|\Sigma\|,$$

where $r = \operatorname{tr}(\Sigma)/\|\Sigma\|$ is the effective rank of Σ.

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

$$\begin{aligned}
\|\Sigma_m - \Sigma\| &= \|\Sigma^{1/2} E_m \Sigma^{1/2}\| \quad (R_m = \frac{1}{m}\sum_{i=1}^{m} Z_i Z_i^T - I_n) \\
&= \max_{x \in S^{n-1}} |x^T \Sigma^{1/2} R_m \Sigma^{1/2} x| \quad \text{(Remark 4.1.12)} \\
&= \max_{x \in T} |x^T R_m x| \quad (T := \Sigma^{1/2} S^{n-1}) \\
&= \max_{x \in T} \left| \frac{1}{m} \sum_{i=1}^{m} \langle Z_i, x \rangle^2 - \|x\|_2^2 \right| \\
&= \frac{1}{m} \max_{x \in T} \left| \|Ax\|_2^2 - m\|x\|_2^2 \right|,
\end{aligned}$$

where A is the $m \times n$ matrix with ros Z_i. As in the proof of Theorem 4.7.1, Z_i are isotropic, and satisfy $\|Z_i\|_{\psi_2} \lesssim 1$. This allows us to apply the matrix deviation inequality for A (in the form given by Exercise 9.3), which gives

$$\mathbb{E}\left[\|\Sigma_m - \Sigma\|\right] \lesssim \frac{1}{m}(\gamma(T)^2 + \sqrt{m}\text{rad}(T)\gamma(T)).$$

The radius and Gaussian complexity of the ellipsoid $T = \Sigma^{1/2} S^{n-1}$ satisfy

$$\text{rad}(T) = \|\Sigma\|^{1/2} \text{ and } \gamma(T) \leq (\text{tr}(\Sigma))^{1/2}.$$

Therefore,

$$\mathbb{E}\left[\|\Sigma_m - \Sigma\|\right] \lesssim \frac{1}{m}(\text{tr}(\Sigma) + \sqrt{m\|\Sigma\|\text{tr}(\Sigma)}).$$

Substituting $\text{tr}(\Sigma) = r\|\Sigma\|$ and simplifying the bound completes the proof. □

Remark 9.2.3 (Covariance estimation: a high-probability guarantee)**.** Just like the versions before (Remark 4.7.3 and Remark 5.6.5), we can upgrade the expectation bound above to a high-probability one. For any $u \geq 0$ we have

$$\|\Sigma_m - \Sigma\| \leq CK^4 \left(\sqrt{\frac{r+u}{m}} + \frac{r+u}{m} \right) \|\Sigma\|$$

with probability at least $1 - 2e^{-u}$.

Proof. Exercise 9.9. □

9.2.4 Johnson-Lindenstrauss Lemma for Infinite Sets

The matrix deviation inequality quickly recovers the Johnson-Lindenstrauss lemma from Section 5.3 - and extends it to general, possibly infinite, sets.

To get a version of the JL lemma from matrix deviation, fix any N-point set $\mathcal{X} \in \mathbb{R}^n$ and consider the normalized differences:

$$T := \left\{ \frac{x-y}{\|x-y\|_2} \; : \; x, y \in \mathcal{X} \text{ distinct} \right\}.$$

The Gaussian complexity of T satisfies

$$\gamma(T) \leq C\sqrt{\log N}.$$

The matrix deviation inequality (Theorem 9.1.1) shows that with high probability,

$$\sup_{x,y \in \mathcal{X}} \left| \frac{\|Ax - Ay\|_2}{\|x-y\|_2} - \sqrt{m} \right| \lesssim \sqrt{\log N}.$$

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.

$$(1-\varepsilon)\|x-y\|_2 \leq \|Qx - Qy\|_2 \leq (1+\varepsilon)\|x-y\|_2 \text{ for all } x, y \in \mathcal{X},$$

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 JL lemma. What we just not gave did not care if $\mathcal{X}$ is finite or not - all that matters is the Gaussian width. So we can extend JL 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 $m \times n$ 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

$$|\|Qx - Qy\|_2 - \|x - y\|_2| \leq \delta \text{ for all } x, y \in \mathcal{X}$$

where $\delta = CK^2 w(\mathcal{X})/\sqrt{m}$ and $K = \max_i \|A_i\|_{\psi_2}$.

Proof. Apply the matrix deviation inequality (Theorem 9.1.1) for the set of differences $T = \mathcal{X} - \mathcal{X}$. Then, with high probability,

$$\sup_{x,y\in\mathcal{X}} \left| \|Ax - Ay\|_2 - \sqrt{m}\|x - y\|_2 \right| \leq CK^2 \gamma(\mathcal{X} - \mathcal{X}) = 2CK^2 w(\mathcal{X}),$$

thanks to Lemma 7.5.11 (a). Divide both sides by $\sqrt{m}$ completes the proof. □

Unlike the classical JL 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 (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 $d(\mathcal{X}) \asymp w(\mathcal{X})^2/\text{diam}(\mathcal{X})^2$. If we choose

$$m \gtrsim \varepsilon^{-2} d(T)$$

(ignoring the dependence on K for simplicity), then we can make $\delta = \varepsilon \text{diam}(\mathcal{X})$, 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 near 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, $\dim(E) \geq n - m$ 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 Grassmannian:

$$E \sim \text{Unif}(G_{n,n-m}).$$

Theorem 9.3.1 (M^* bound)**.** Let $T \subset \mathbb{R}^n$ be a bounded set, and A be an $m \times n$ random matrix with independent, isotropic and subgaussian rows A_i. Then the random subspace $E = \ker A$ satisfies

$$\mathbb{E}[\text{diam}(T \cap E)] \leq \frac{CK^2 w(T)}{\sqrt{m}},$$

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

Proof. Apply Theorem 9.1.1 for $T - T$:

$$\mathbb{E}\left[\sup_{x,y\in T} \left|\|Ax - Ay\|_2 - \sqrt{m}\|x-y\|_2\right| \le CK^2\gamma(T-T) = 2CK^2 w(T),\right]$$

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 both sides by $\sqrt{m}$ to get

$$\mathbb{E}\left[\sup_{x,y\in T\cap \ker A} \|x-y\|_2\right] \le \frac{CK^2 w(T)}{\sqrt{m}},$$

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 id roughly $\sqrt{\log n}$ by Example 7.5.8, we get

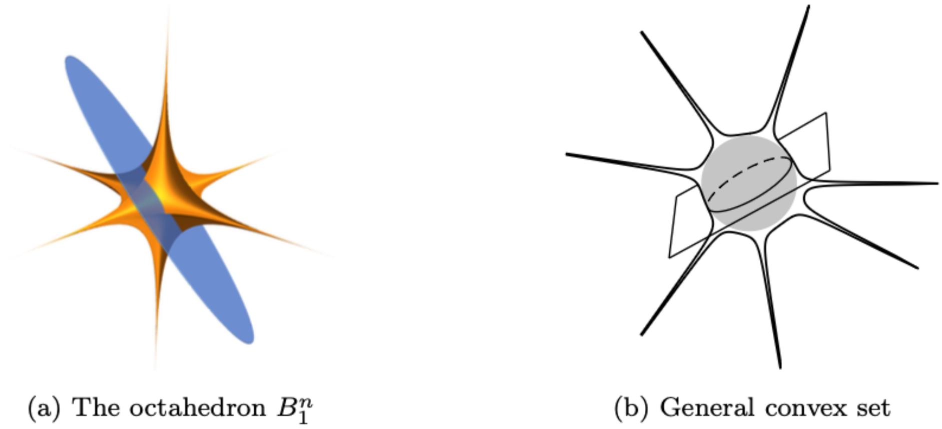
$$\mathbb{E}\left[\operatorname{diam}(B_1^n \cap E)\right] \lesssim \sqrt{\frac{\log n}{m}}.$$

For example, if $m = 0.01n$, then

$$\mathbb{E}\left[\operatorname{diam}(T \cap E)\right] \lesssim \sqrt{\frac{\log n}{n}}.$$

So, a random $0.99n$-dimensional slice of a cross-polytope is tiny!

How can this be? This relates to what we discussed in Remark 7.5.10. 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). Therefore the slice ends up with diameter about $O(1/\sqrt{n})$, maybe with a log factor as shown above. This intuition can be extended to general convex sets as well.



(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 (Definition 7.5.12). The M^* bound shows that slicing shrinks the dimaeter:

$$\mathbb{E}\left[\operatorname{diam}(T \cap E)\right] \leq 0.01 \cdot \operatorname{diam}(T)$$

as long as $m \gtrsim d(T)$. Since $\dim(E) = n - m$, this condition is equivalent to

$$\dim(E) + cd(T) \leq n.$$

That lines up with the linear algebra intuition: If T is a cnetered Euclidean ball in some subspace $F \subset \mathbb{R}^n$, slicing can shrink the diameter of T only when $\dim E + \dim F \leq n$.

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 (Figure 9.3), then it does as long as the codimension of E is not too small:

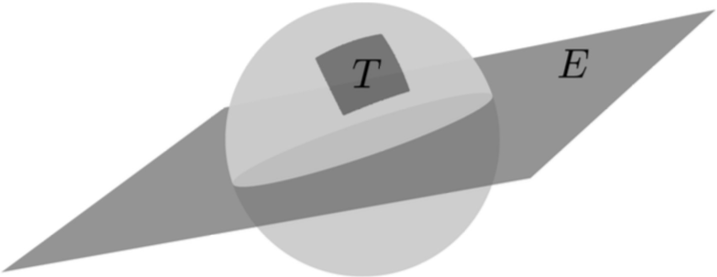


Figure 9.3 The escape theorem quantifies when a random subspace E misses a given subset T of the sphere.

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

$$m \geq CK^4 w(T)^2,$$

then the random subspace $E = \ker A$ satisfies

$$T \cap E = \emptyset$$

with probability ar least $1 - 2\exp(-cm/K^4)$. Here $K = \max_i \|A_i\|_{\psi_2}$.

Proof. Let us use the high-probability version of the matrix deviation inequality (Remark 9.1.4): With probability at least $1 - 2\exp(-u^2)$,

$$\sup_{x \in T} \left| \|Ax\|_2 - \sqrt{m} \right| \leq C_1 K^2 (w(T) + u).$$

Suppose the above occurs. If $T \cap E \neq \emptyset$, then for any $x \in T \cap E$ we have $Ax = 0$, so

$$\sqrt{m} \leq C_1 K^2 (w(T) + u).$$

Set $u = \sqrt{m}/(2C_1K^2)$ and simplify this bound to get

$$\sqrt{m} \leq 2C_1 K^2 w(T),$$

which contradicts the assumption if C is large enough.Therefore, with that choice of u, the event implies $T \cap E = \emptyset$. Done! □

9.4 Application: High-dimensional Linear Models

Let's apply our tools on a classic data science problem: learning a linear model in high dimensions. Let

$$y_i = \langle A_i, x\rangle + w_i,\ \ i = 1, \dots, m.$$

Her $A_i \in \mathbb{R}^n$ are known, and w_i are unknown numbers representing noise (Figure 9.4). In matrix form, we get

$$y = Ax + w.$$

The goal is to recover x from y and A as accurately as possible.

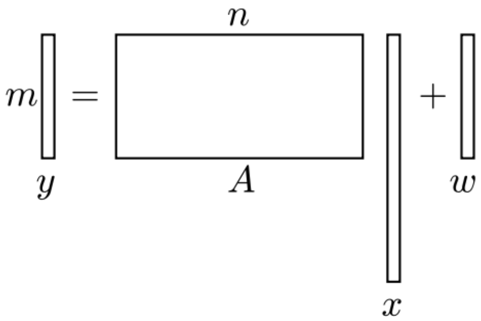


Figure 9.4 A high-dimensional linear model: recover x from $y = Ax + w$.

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.

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 (Figure 9.5).

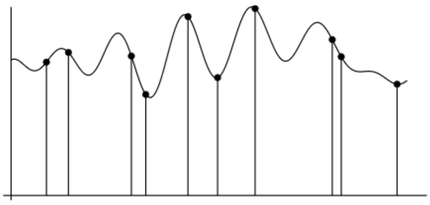


Figure 9.5 Signal recovery problem in audio sampling: recover an audio signal x from its values at m random time points.

Example 9.4.2 (Linear regression)**.** A core porblem 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 that we are trying to learn.

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 genetic study, there might be ∼100 patients, but ∼10000 genes. In this high-dimensional setting, even solving $Ax = y$ becomes impossible, as there are too many possible solutions as they live in a large subspace of dimension at least $m - n$.

Not all hope is lost. If we have some prior information about the structure of x, which we can write this as

$$x \in T$$

for some known set $T \subset \mathbb{R}^n$, then we might be able to recover x. For example, if x is sparse, we can pick T to be the set of all sparse vectors.

9.4.1 Constrained Recovery

Let's solve the noiseless case first:

$$y = Ax, \;\; x \in T.$$

How do we solve this dimensional constrained linear problem?
A simple idea is just to pick any vector $x' \in T$ that matches the observations:

$$\text{find } x' : \;\; y = Ax', \;\; x \in T.$$

If T is convex, this is a convex program, and many algorithms exist to numerically solve it. 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 $\hat{x}$ of the convex linear program satisfies

$$\mathbb{E}\left[\|\hat{x} - x\|_2\right] \leq \frac{CK^2 w(T)}{\sqrt{m}},$$

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

Proof. Since $x, \hat{x} \in T$ and $Ax = A\hat{x} = y$, we have

$$x, \hat{x} \in T \cap E_x, \;\; \text{where } E_x = x + \ker A$$

See Figure 9.6 for an illustration.

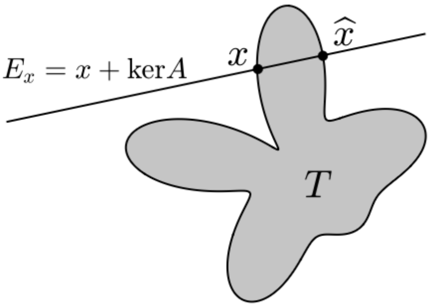


Figure 9.6 The geometry of a constrained high-dimensional linear problem.

Then the M^* bound (in the form of Exercise 9.12) gives

$$\mathbb{E}\left[\|\hat{x} - x\|_2\right] \leq \mathbb{E}\left[\operatorname{diam}(T \cap E_x)\right] \leq \frac{CK^2 w(T)}{\sqrt{m}}.$$

□

Remark 9.4.5 (Effective dimension)**.** To get some intuition, rewrite the accuracy guarantee in

Theorem 9.4.4 using the effective dimension (Definition 7.5.12). We get a nontrivial error bound

$$\mathbb{E}\left[\|\hat{x}-x\|_2\right] \leq 0.01\text{diam}(T)$$

as long as the number of observations satisfies (suppressing K again)

$$m \gtrsim d(T).$$

Since $d(T)$ can be much smaller than the ambient dimension n, recovery is often possible even in the high-dimensional regime.

Remark 9.4.6 (Convex relaxation)**.** If T is not convex, we can just use its convex hull $\text{conv}(T)$. The recovery guarantees from Theorem 9.4.4 do not change, as $w(\text{conv}(T)) = w(T)$ by Proposition 7.5.2 (c).

Remark 9.4.7 (Unconstrained optimization)**.** Forcing strict rules on the solution like $y = Ax'$ and $x' \in T$ can be too rigid - noise or a bad choice for T might mean no solution exists. Instead, we could aim to penalize how much they are broken by solving the unconstrained convex problem

$$\min\|y-Ax\|_2^2 + \lambda\|x\|_T, \; x \in \mathbb{R}^n,$$

where $\|\cdot\|_T$ is any norm you like, and $\lambda > 0$ is a parameter we can adjust to see which summand we want to prioritize more.

We also have the following guarantee from Exercise 9.20: if A is an $m \times n$ random matrix, and λ is chosen well, then the solution $\hat{x}$ satisfies

$$\mathbb{E}\left[\|\hat{x}-x\|_2\right] \lesssim \frac{w(T)\|x\|_T + \|w\|_2}{\sqrt{m}},$$

where T is the unit ball of $\|\cdot\|_T$.

In short, if x is well structured and the noise w is small, then you can recover x accurately from $m \asymp d(T)$ observations, where $d(T)$ is the effective dimension of T.

9.4.2 Example: Sparse Recovery

Sometimes we believe that x is *sparse* - most of its entries are zero or nearly zero. We can quantify the sparsity of $x \in \mathbb{R}^n$ by the number of nonzero entries:

$$\|x\|_0 = |\text{supp}(x)| = |\{i : \; x_i \neq 0\}|,$$

and we say that x is sparse if $\|x\|_0 \leq s$. The "ℓ^0 norm" is not really a norm, but is a limit of ℓ^p norms as $p \to 0$ (Exercise 9.23).

A quick dimension count shows that we can recover x from $y = Ax$ if A is in a general position and we have enough observations: $m \geq 2\|x\|_0$ (Exercise 9.22). Sounds great, as we can recover a sparse vector from only a few observations. The catch? It's computationally hard unless we already know the support of x. Without that, the best way is to brute force - $\binom{n}{s} \geq 2^s$ subsets to check!

To solve this, we can pick the prior using the ℓ^1 norm, the closest ℓ^p that is actually a norm (Exercise 9.23). Since s-sparse vectors with $\|x\|_2 \leq 1$ satisfy $\|x\|_2 \leq \sqrt{s}$, it makes sense to pick the convex set

$$T := \sqrt{s}B_1^n$$

as our prior. The recovery program becomes

$$\text{find } x : \; y = Ax, \; \|x\|_1 \leq \sqrt{s}.$$

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 $\|x\|_2 \leq 1$. Then any solution $\hat{x}$ to the program above satisfies

$$\mathbb{E}\left[\|\hat{x} - x\|_2\right] \leq CK^2 \sqrt{\frac{s \log n}{m}},$$

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 on the Gaussian width of the ℓ^1 ball:

$$w(T) = \sqrt{s} w(B_1^n) \leq C\sqrt{s \log n}.$$

□

Remark 9.4.9 (Observations scale almost linearly with sparsity)**.** Corollary 9.4.8 gives a small error as long as

$$m \gtrsim s \log n$$

(if the hidden constant is appropriately large). That's great news - we can efficiently recover a sparse vector from way fewer observations m than the full dimension n.

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 bit tighter. Instead of using the ℓ^1 ball, use the *truncated* ℓ^1 ball

$$T_{n,s} = \sqrt{s} B_1^n \cap B_2^n.$$

This relaxation is pretty tight - Exercise 9.25 gives

$$\mathrm{conv}(S_{n,s}) \subset T_{n,s} \subset 2\mathrm{conv}(S_{n,s}).$$

This tightening gives a logarithmic improvement on the bound in Corollary 9.4.8, showing that

$$m \gtrsim s \log\left(en/s\right)$$

observations suffice for sparse recovery (Exercise 9.21).

9.4.3 Example: Low-rank Recovery

Here is one more example of a high-dimensional linear problem: recover a $d \times d$ matrix X (instead of a vector) from m linear observations:

$$y_i = \langle A_i, X \rangle, \; i = 1, \ldots, m,$$

where A_i are known, independent random matrices, adn the inner product is

$$\langle A, B \rangle = \mathrm{tr}(A^T B).$$

Normally, we 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 entries, rank counts nonzero singular values.

In section 9.4.2, we took a relaxation by replacing the ℓ^0 norm with the ℓ^1 norm. Here we'll use a convex relaxation by replacing the ℓ^0 norm for the singular values with the ℓ^1 norm, which is also known as the *nuclear norm*:

$$\|X\|_* := \sum_{i=1}^{d} \sigma_i(X).$$

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 $\|X\|_F = 1$ satisfies

$$\|X\|_* \le \sqrt{r}.$$

Therefore, it makes sense to consider $T = \sqrt{r}B_*$ as our prior, where

$$B_* = \{X \in \mathbb{R}^{d\times d} : \ \|X\|_* \le 1\}.$$

Then, the recovery program becomes

$$\text{find } X : \ y_i = \langle A_i, X\rangle \,\forall i; \ \|X\|_* \le \sqrt{r},$$

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. $N(0,1)$ entries. Assume an unknown $d \times d$ matrix X has rank at most r and $\|X\|_F \le 1$. Then any solution $\hat{X}$ of the program satisfies

$$\mathbb{E}\left[\|\hat{X} - X\|_F\right] \le C\sqrt{\frac{rd}{m}}.$$

Proof. Using the duality between the nuclear and operator norms (Exercise 7.18a), we get that for a $d \times d$ Gaussian matrix G with i.i.d. $N(0,1)$ entries:

$$w(B_*) = \mathbb{E}\left[\sup_{\|X\|_* \le 1} \langle G, X\rangle\right] = \mathbb{E}\left[\|G\|\right] \le 2\sqrt{d}$$

by Theorem 7.3.1. Now just apply Theorem 9.4.4 with $T = \sqrt{r}B_*$. □

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 from war fewer observations m then the entirety d^2. This is similar to the matrix completion problem in Section 6.5, where we can recover a low-rank matrix from about $m \asymp rd \log d$ randomly chosen entries.

9.5 Application: Exact Sparse Recovery

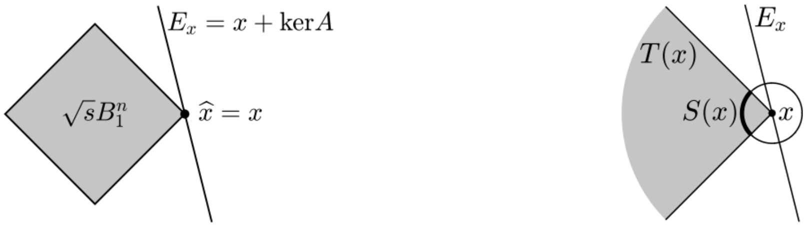
In the noiseless case, we can do even better - we can recover a sparse vector x from $y = Ax$ *exactly* (and algorithmically effective)! We will look at two ways to get this surprising result:

(a) Use the escape theorem (Theorem 9.3.4).

(b) Using the restricted isometry property - then show random matrices satisfy it with high probability.

9.5.1 Exact Recovery Based on the Escape Theorem

Let's look at Figure 9.7 below for some intuition.



(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

Suppose we are trying to recover an unknown s-sparse unit vector x from $y = Ax$ by solving the convex program in Section 9.4.2. A solution $\hat{x}$ lies in the intersection of the prior set $T = \sqrt{s}B_1^n$ 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 (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, hence the solution $\hat{x}$ must be exact:

$$\hat{x} = x.$$

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) comes in. 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 formalize this. We want to recover x from

$$y = Ax$$

by solving the optimization problem

$$\min \|x\|_1 \text{ subject to } y = Ax.$$

Theorem 9.5.1 (Exact sparse recovery)**.** A $m \times n$ 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 satisfies

$$m \geq CK^4 s \log n,$$

then for any s-sparse vector $x \in \mathbb{R}^n$, a solution $\hat{x}$ of the convex program is exact:

$$\hat{x} = x.$$

To prove this, we want to show that the recovery error is 0:

$$h := \hat{x} - x = 0.$$

First, let's show a weaker claim: h has more mass on the spport 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

$$h_{S^c} \leq \|h_S\|_1.$$

Proof. Since $\hat{x}$ is the minimizer in the program, we have

$$\|\hat{x}\|_1 \leq \|x\|_1.$$

But there is also a lower bound

$$\begin{aligned}\|\hat{x}\|_1 &= \|x + h\|_1 \\ &= \|x_S + h_S\|_1 + \|x_{S^c} + h_{S^c}\|_1 \\ &\geq \|x\|_1 - \|h_S\|_1 + \|h_{S^c}\|_1,\end{aligned}$$

where the last line follows by triangle inequality and using $x_S = x$ and $x_{S^c} = 0$. Substituting this into the first equation completes the proof. □

Lemma 9.5.3 (The error is approximately sparse)**.** The error vector satisfies

$$\|h\|_1 \leq 2\sqrt{s}\|h\|_2.$$

Proof. Using Lemma 9.5.2 and then the Cauchy-Schwartz inequality, we get

$$\begin{aligned}\|h\|_1 &= \|h_S\|_1 + \|h_{S^{\subset}1}\|_1 \\ &\leq 2\|h_S\|_1 \\ &\leq 2\sqrt{s}\|h_S\|_2 \\ &\leq 2\sqrt{s}\|h\|_2.\end{aligned}$$

□

Now let's return to the proof for exact sparse recovery.

Proof of Theorem 9.5.1. Assume $h = \hat{x} - x \neq 0$. Lemma 9.5.3 gives

$$\frac{h}{\|h\|_2} \in T_s := \{z \in S^{n-1} : \ \|z\|_1 \leq 2\sqrt{s}\}.$$

Since also $Ah = A\hat{x} - Ax = y - y = 0$, we have

$$\frac{h}{\|h\|_2} \in T_s \cap \ker A.$$

The escape theorem (Theorem 9.3.4) shows that this intersection is empty with high probability as long as $m \geq C_1 K^4 w(T_s)^2$. Now, since $T_s \subset 2\sqrt{s}B_1^n$, we get

$$w(T_s) \leq 2\sqrt{s}w(B_1^n) \leq C_2\sqrt{s \log n}.$$

Thus, if $m \geq CK^4 s \log n$, the intersection $T_s \cap \ker A$ is empty with high probability, which means the inclusion in the first equation cannot hold. Therefore, our assumption that $h \neq 0$ is false with high probablity, and the proof is complete. □

Remark 9.5.4 (Improving the logarithmic factor)**.** By slightly tightening the last equation from Lemma 9.5.3, we can improve the number of sufficient observations in Theorem 9.5.1 to

$$m \geq CK^4 s \log{(en/s)}.$$

This follows from Exercise 9.26.

9.5.2 Restricted Isometries

We'll now find a *deterministic* condition that ensures a matrix A works for sparse recovery, and prove that random matrices satisfy this condition.

Definition 9.5.5. An $m \times n$ matrix A satisfies the restricted isometry property/RIP with parameters α, β, and s if the inequality

$$\alpha \leq \|v\|_2 \leq \|Av\|_2 \leq \beta \|v\|_2$$

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 A_I of A satisfy

$$\alpha \leq \sigma_s(A_I) \leq \sigma_1(A_i) \leq \beta.$$

And if $\alpha \approx \beta \approx 1$, then RIP tells us that all those submatrices are approximate isometries.

Theorem 9.5.6 (RIP implies exact recovery)**.** Suppose a $m \times n$ matrix A satisfies TIP 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 the convex relaxation problem earlier.

Proof. As in the proof of Theorem 9.5.1, we need to show that the error

$$h = \hat{x} - x$$

is zero. To do this, we decompose h in a way ver similar to Exercise 9.25.

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

$$Ah = A\hat{x} - Ax = y - y = 0,$$

the triangle inequality gives

$$0 = \|Ah\|_2 \geq \|A_{I_{01}} h_{I_{01}}\|_2 - \|A_{I_{01}^c} h_{I_{01}^c}\|_2. \quad (*)$$

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 the triangle inequality followed by RIP gives

$$\|A_{I_{01}^c} h_{I_{01}^c}\|_2 \leq \sum_{i \geq 2} \|A_{I_i} h_{I_i}\|_2 \leq \beta \sum_{i \geq 2} \|h_{I_i}\|_2.$$

Plugging into $(*)$ gives

$$\beta \sum_{i \geq 2} \|h_{I_i}\|_2 \geq \alpha \|h_{I_{0,1}}\|_2. \quad (**)$$

Step 3: Summing up. Nextm 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 $i \geq 2$. Thus

$$\|h_{I_i}\|_2 \leq \frac{1}{\sqrt{\lambda s}} \|h_{I_{i-1}}\|_1.$$

Summing up, we get

$$\begin{aligned} \sum_{i \geq 2} \|h_{I_i}\|_2 &\leq \frac{1}{\sqrt{\lambda s}} \sum_{i \geq 1} \|h_{I_i}\|_1 \\ &= \frac{1}{\sqrt{\lambda s}} \|h_{I_0^c}\|_1 \\ &\leq \frac{1}{\sqrt{\lambda s}} \|h_{I_0}\|_1 \quad \text{(Lemma 9.5.2)} \\ &\leq \frac{1}{\sqrt{\lambda}} \|h_{I_0}\|_2 \\ &\leq \frac{1}{\sqrt{\lambda}} \|h_{I_{0,1}}\|_2. \end{aligned}$$

Putting this into (∗∗), we get

$$\frac{\beta}{\sqrt{\lambda}}\|h_{I_{0,1}}\|_2 \geq \alpha\|h_{I_{0,1}}\|_2.$$

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 matries 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

$$m \geq CK^4 s \log{(en/s)}$$

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 paremeters

$$\alpha = 0.9\sqrt{m}, \beta = 1.1\sqrt{m}, \text{ and } s.$$

Proof. We need to check that

$$\alpha \leq \sigma_s(A_I) \leq \sigma_1(A_i) \leq \beta.$$

for all $m \times s$ submatrices A_I. First, fix I. By Theorem 4.6.1, we get

$$0.9\sqrt{m} \leq \sigma_s(A_I) \leq \sigma_1(A_I) \leq 1.1\sqrt{m}$$

with probability at least $1 - 2\exp{(-cm/K^4)}$ (set $t = \sqrt{2cm}/K$ and use the assumption on m, with constants c and C chosen appropriately).
Now take a union bound over all $\binom{n}{s}$ possible s-element subsets $I \subset \{1, \dots, n\}$. Then the above holds with probability at least

$$1 - 2\exp{(-cm/K^4)} \cdot \binom{n}{s} > 1 - 2\exp{(-cm/K^4)},$$

using the bound $\binom{n}{s} \leq \exp{(s\log{(en/s)})}$ from Exercise 0.6 and the assumption on m. Hence the proof is complete. □

In fact, 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. The proof is complete. We even get the logarithmic improvement from Exercise 9.5.4! □

9.6 Deviations of Random Matrices for General Norms

We can 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 triangle inequality is enough.

Definition 9.6.1. A real-valued function f on a linear vector space V is called:

- Positive-homogeneous if $f(\alpha x) = \alpha f(x)$ for all $\alpha \geq 0$ and $x \in V$;
- Subadditive if $f(x + y) \leq f(x) + f(y)$ 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 (i.e. any *linear functional*);

(c) In particular, the function $f(x) = x^T y$ for any fixed vector $y \in \mathbb{R}^m$;

(d) the *support function* of any bounded set $S \subset \mathbb{R}^n$, defined by

$$f(x) := \sup_{y \in S} \langle x, y \rangle, \; x \in \mathbb{R}^m.$$

We can make Theorem 9.1.1 work for all norms (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 $b \in \mathbb{R}$ such that

$$f(x) \leq b\|x\|_2 \text{ for all } x \in \mathbb{R}^n.$$

Then for any subset $T \subset \mathbb{R}^n$,

$$\mathbb{E}\left[\sup_{x \in T} |f(Ax) - \mathbb{E}\left[f(Ax)\right]|\right] \leq Cb\gamma(T),$$

where $\gamma(T)$ is the Gaussian complexity.

With the same logic as in the proof for Theorem 9.1.1, Theorem 9.6.3 would immediately follow from Talagrand's comparison inequality once we show that the random process

$$Z_x := f(Ax) - \mathbb{E}\left[f(Ax)\right]$$

has subgaussian increments. Let's do this :)

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 homogeneous and subadditive function satisfying

$$f(x) \leq b\|x\|_2 \text{ for all } x \in \mathbb{R}^n.$$

Then the random process

$$Z_x := f(Ax) - \mathbb{E}\left[f(Ax)\right]$$

has subgaussian increments:

$$\|Z_x - Z_y\|_{\psi_2} \leq Cb\|x - y\|_2 \text{ for all } x, y \in \mathbb{R}^n.$$

Proof. Without loss of generality, we may assume that $b = 1$. Just like in the proof of Theorem 9.1.2, first assume that

$$\|x\|_2 = \|y\|_2 = 1.$$

In this case, the inequality in the theorem becomes

$$\|f(Ax) - f(Ay)\|_{\psi_2} \leq C\|x - y\|_2.$$

Step 1: Creating independence. Consider the vectors

$$u := \frac{x + y}{2}, \; v := \frac{x - y}{2}.$$

Then $x = u + v$ and $y = u - v$, and thus

$$Ax = Au + Av, \; Ay = Au - Av \quad \text{(See Figure 9.8 below)}$$

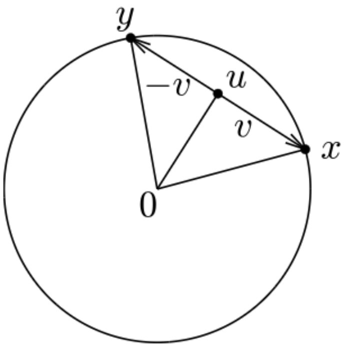


Figure 9.8 Creating a pair of orthogonal vectors u, v out of x, y.

Since u, v are orthogonal, the Gaussian random vectors Au and Av are independent (Exercise 3.20).
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

$$a + Av = a + \|v\|_2 g, \text{ where } g \sim N(0, I_m) \quad \text{(Exercise 3.20)}$$

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

$$\begin{aligned} f(a + \|v\|_2 t) - f(a + \|v\|_2 s) &\le f(\|v\|_2 t - \|v\|_2 s) \\ &= \|v\|_2 f(t - s) \quad \text{(Positive homogeneity)} \\ &\le \|v\|_2 \|t - s\|_2 \quad (b = 1), \end{aligned}$$

proving the claim.
Concentration in the Gauss space (Theorem 5.2.3) then yields

$$\|f(a + Av) - \mathbb{E}_a [f(a + Av)]\|_{\psi_2(a)} 2 \le C\|v\|_2,$$

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 that of $a + Av$, it satisfies the same bound:

$$\|f(a - Av) - \mathbb{E}_a [f(a - Av)]\|_{\psi_2(a)} \le C\|v\|_2,$$

Subtract the bottom equation from the top one, use the triangle inequality and the fact that the expectations are the same gives

$$\|f(a + Av) - f(a - Av)\|_{\psi_2(a)} \le 2C\|v\|_2.$$

This bound holds conditionally for any fixed $a = Au$, Therefore, it holds for the original distribution too:

$$\|f(a + Av) - f(a - Av)\|_{\psi_2} \le 2C\|v\|_2.$$

Passing back the the x, y notation, we obtained the desired inequality.
We proved the theorem for unit vectors x, y. To extend it to the general case, argue exactly as step 4 in the proof of Theorem 9.1.2. □

Remark 9.6.5. It is an open question if Theorem 9.6.3 holds for general subgaussian matrices A.

9.7 Two-sided Chevet Inequality and Dvoretzky-Milman Theorem

9.7.1 Two-sided Chevet's Inequality

Another consequence of general matrix deviation is a sharper version of Chevet's inequality:

Theorem 9.7.1 (Two-sided Chevet's inequality)**.** Let A be an $m \times n$ Gaussian random matrix with i.i.d $N(0,1)$ entries. Let $T \subset \mathbb{R}^n$ and $S \subset \mathbb{R}^m$ be arbitrary bounded sets. Then

$$\mathbb{E}\left[\sup_{x \in T}\left|\sup_{y \in S}\langle Ax, y\rangle - w(S)\|x\|_2\right|\right] \leq C\gamma(T)\text{rad}(S),$$

where $\gamma(T)$ is the Gaussian complexity and $\text{rad}(T)$ is the radius.

Proof. Let's apply Theorem 9.6.3 for the support function of S:

$$f(x) = \sup_{y \in S}\langle x, y\rangle .$$

This is a bounded function, since the Cauchy-Schwartz inequality gives

$$f(x) \leq \sup_{y \in S}\|x\|_2\|y\|_2 = \text{rad}(S)\|x\|_2 \text{ for all } x \in \mathbb{R}^n. \quad (*)$$

Since Ax has the same distribution as $g\|x\|_2$ where $g \sim N(0, I_m)$ from Exercise 3.20, we have that

$$\begin{aligned}
\mathbb{E}\left[f(Ax)\right] = \|x\|_2\mathbb{E}\left[f(g)\right] \quad &\text{(Positive homogeneity)}\\
= \|x\|_2\mathbb{E}\left[\sup_{y \in S}\langle x, y\rangle\right] \quad &\text{(By definition of)}f\\
= \|x\|_2 w(S) \quad &\text{(By definition of Gaussian width)}. \quad (**)
\end{aligned}$$

Substitute $(*)$ and $(**)$ into Theorem 9.6.3 completes the proof. □

9.7.2 Dvoretzky-Milman Theorem

We'll now prove this amazing result: If you project any bounded set in $\mathbb{R}^n$ to a low-dimensional subspace, it will look *approximately round* with high probability (Figure 9.9).

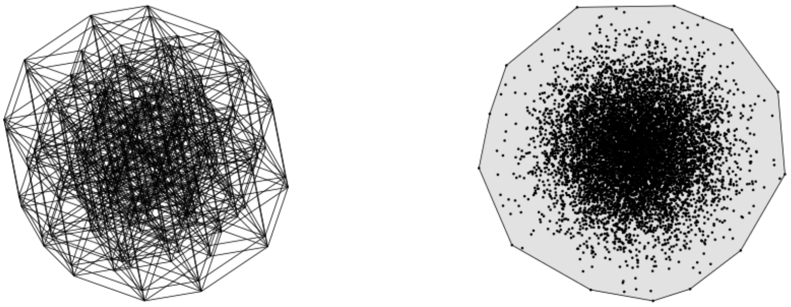


Figure 9.9 A random projection of a 8-dimensional cube (left) and 10^4 Gaussian points (right) onto the plane

It's easier to work with Gaussian projections, where the result says:

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 subset. Then the following holds with probability at least 0.99:

$$r_- B_2^m \subset \text{conv}(AT) \subset r_+ B_2^m$$

where B_2^m denotes the unit Euclidean ball in $\mathbb{R}^m$, and

$$r_\pm = w(T) \pm C\sqrt{m}\mathrm{rad}(T).$$

The left inclusion holds only if r_- is nonnegative; the right inclusion always holds.

Proof. Let's use the two-sided Chevet's inequality (Theorem 9.7.1) in the following form:

$$\mathbb{E}\left[\sup_{y\in S}\left|\sup_{x\in T}\langle Ax, y\rangle - w(T)\|y\|_2\right|\right] \le C\gamma(S)\mathrm{rad}(T),$$

where $T \subset \mathbb{R}^n$ and $S \subset \mathbb{R}^m$. To get this, just apply the theorem to A^T with T and S swapped.
Let S be the sphere S^{m-1}; its Gaussian complexity satisfies $\gamma(T) \le \sqrt{m}$. Then, by Markov's inequality, the following holds with probability at least 0.99:

$$\left|\sup_{x\in T}\langle Ax, y\rangle - w(T)\right| \le C\sqrt{m}\mathrm{rad}(T) \text{ for every } y \in S^{m-1}.$$

By the triangle inequality and the definition of $r_\pm$, this implies

$$r_- \le \sup_{x\in T}\langle Ax, y\rangle \le r_+ \text{ for every } y \in S^{m-1}.$$

Rewriting $\sup_{x\in T}\langle Ax, y\rangle$ as $\sup_{x\in AT}\langle x, y\rangle$ and using homogeneity, we get

$$r_-\|y\|_2 \le \sup_{x\in AT}\langle x, y\rangle \le r_+ \text{ for every } y \in \mathbb{R}^m.$$

By duality (Exercise 9.40), this is the same as the statement of the theorem, and we're done! □

Remark 9.7.3 (The effective dimension)**.** Assume that T is bounded, convex, and contains the origin, and let

$$m \le cd(T)$$

where $d(T) \asymp w(T)^2/\mathrm{rad}(T)^2$ is the effective dimension (Definition 7.5.12). If we pick the absolute constant c to be small enough, we can make $C\sqrt{m}\mathrm{rad}(T) \le 0.01w(T)$, so that the Dvoretzky-Milman theorem (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 Example 7.5.7,

$$w(T) = \sqrt{2/\pi}\cdot n \text{ and } \mathrm{diam}(T) = 2\sqrt{n}$$

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 gives an illustration.

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 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. The Dvoretzky-Milman theorem explains why: PT is now an *approximate round ball* of radius of order $w_s(T)$ (Exercise 9.43), which obviously does not shrink under any projection!