Notes for High-Dimensional Probability Second Edition by Roman Vershynin

Gallant Tsao

July 20, 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.

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

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 $x \in \text{conv}(T)$ 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 $x \in \text{conv}(T)$ 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 $P \subset \text{conv}(P) \subset \text{conv}(T)$ where $T = \{\text{Vertices of } P\}$. Then we apply Theorem 0.0.2 to any point $x \in P \subseteq \text{conv}(T)$ 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

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

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

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

Rearranging the terms above gives

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

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

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

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 ($\mathrm{Vol}(P) \leq \mathrm{Vol}(B)$). □

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:

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

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

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

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 $1, \ldots, n$ 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 $a_1, \ldots, a_n \in \mathbb{R}$, 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 $X_1, \ldots, X_n$. 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 $X_1, \ldots, X_{n-1}$, we get that

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

The inequality holds because X_n is independent of $X_1, \ldots, X_{n-1}$, 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

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

□

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:

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

Then as $N \to \infty$,

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

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

Example 1.7.4 (Bernoulli and binomial distributions)**.** When $X_i \sim \text{Ber}(p)$, $S_N \sim \text{Binom}(N,p)$. 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 $X \sim \text{Pois}(\lambda)$, 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 $S_N = X_1 + Partofnegotiations. \cdots + X_N$, 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 $Y \sim$ 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 $X_1, \dots, X_N$ be independent Rademacher random variables, and let $a = (a_1, \cdots, a_N) \in \mathbb{R}^N$ be fixed. Then for any $t \geq 0$,

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

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 $X_1, \dots, X_N$ 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 $X_i \sim \text{Ber}(p_i)$ 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 $X_i \sim \text{Ber}(p_i)$,

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

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,

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

Moreover,

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

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 (add link), $S_N \sim \text{Pois}(\mu)$. 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 $t = (1 + \delta)\mu$,

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

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,

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

Same as before, expanding the expression into Taylor series gives

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

Plugging the bound gives

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

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 $N = 200,\ \mu = 10$:

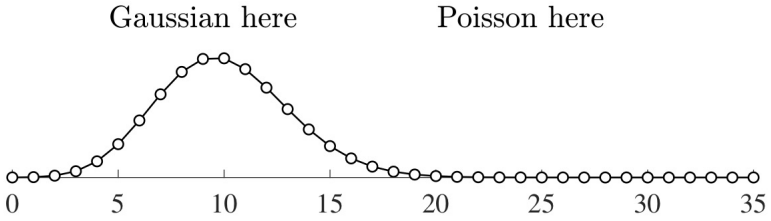

Figure 2.1 The probability mass function of the distribution $\mathrm{Binom}(N, \mu/N)$ with $N = 200$ and $\mu = 10$. It is approximately normal near the mean μ, but it is heavier far from the mean.

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:

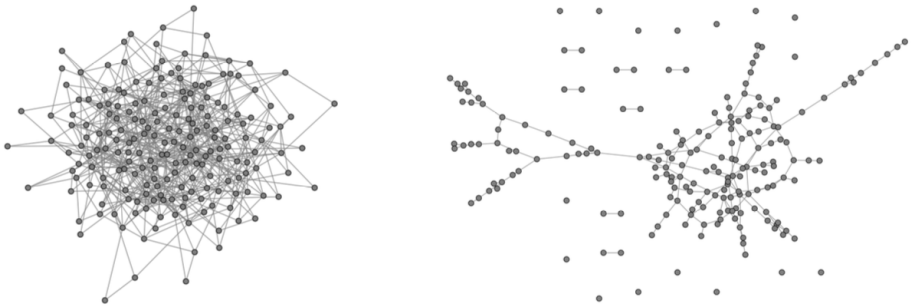

Figure 2.2 Examples of random graphs in the Erdős-Rényi model $G(n,p)$ with $n = 200$ vertices and connection probabilities $p = 0.03$ (left) and $p = 0.01$ (right).

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 $\mathrm{Ber}(p)$ 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. WLOG assume $K_1 = 1$. If not, we can scale X to X/K_1 and our analysis will not be affected. The integrated tail formula (Lemma 1.6.1 + link) for $|X|^p$ gives

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

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 + link, 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 sub-

gaussian 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 vari-

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

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

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

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

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 $N \ge 2$, that are not necessarily independent. Then

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

In particular,

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

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 $t \ge 0$, we have

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

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,

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

Then by Proposition 2.6.6 ($(c) \iff (a)$) we get $\|\max_i X_i\|_{\psi_2} \le C\sqrt{\ln N}$.
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

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

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

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

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:

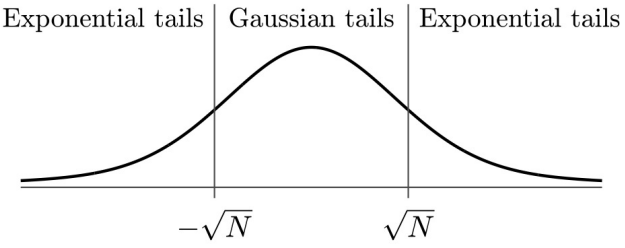

Figure 2.3 Bernstein inequality exhibits a mixture of two tails: gaussian for small deviations and exponential for large deviations.

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

$$\big\|\|x\|_2 - \sqrt{n}\big\|_{\psi_2} \le CK^2$$

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

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

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

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

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

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

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

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

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 $z, \delta \ge 0$:

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

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

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

Then for any $\delta \ge 0$,

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

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

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

This is also true in higher dimensions:

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

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

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

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:

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

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

Proof. If X_i are uncorrelated, 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,

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

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

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

Splitting the event based on whether E_n occurs, we get

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

Hence

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

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

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

The RHS does not depend on v or t, and goes to zero as $n \to \infty$.
We can also show that $P(g_1 \le t) - P(\sqrt{n}\,\langle X, v\rangle \le t)$ 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 $X \sim \mathrm{Unif}(K)$, if its density is $1/\mathrm{Vol}(K)$ on K and zero everywhere else.
The mean of X is

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

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:

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

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 $u_1, \dots, u_N$, 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 $X \sim \text{Unif}\{\sqrt{N}u_1, \dots, \sqrt{N}u_N\}$ is isotropic.

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

Proof. (i) $\Rightarrow$ (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) $\Rightarrow$ (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) $\Rightarrow$ (ii) Multiply both sides by the vector x gives the result.
(iii) $\Rightarrow$ (ii) Taking the inner product with the vector x gives the result. □

Example 3.3.12 (Coordinate distribution)**.** The standard basis $\{e_1, \dots, e_n\}$ in $\mathbb{R}^n$ is a Parseval frame. Therefore, a coordinate random vector

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

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:

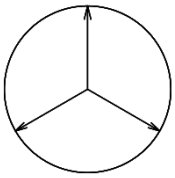

Figure 3.7 A Mercedez-Benz frame: three equispaced points on the circle of radius $\sqrt{2/3}$ form a Parseval frame in $\mathbb{R}^2$.

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 $X = (X_1, \dots, X_n)$ 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 $X = (X_1, \dots, X_n)$ 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 $v = (v_1, \dots, v_n) \in S^{n-1}$. 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 $X \sim \text{Unif}(K)$ 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

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

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

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

Then for any Hilbert space H, we have

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

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

There is nothing random in the statement 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 $K \le 14.1$ in this section, and the other one with $K \le 1.783$ 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

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

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):

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

Thus

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

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

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

We cannot use the same reasoning for $S_\pm$, since the random variable U_i^+ is unbounded. Instead, let us treat the random variables U_i^+ and V_j^- as elements of the Hilbert space L^2 with 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 $\|V_j^-\|_{L^2} \le \|V_j\|_{L^2} = 1$ 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

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

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

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

Then for any Hilbert space H, we have

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

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:

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

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 "$\leq$" by "$\geq$" 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 $\langle B_i, X\rangle \leq b_i$). 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

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

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

Proof. The first bound follows with $X_i = (x_i, 0, \ldots, 0)^T$. 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 $V = \{1, \ldots, n\}$.

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 maxcut(G), is the largest possible cut over all partitions of vertices. The figure below is an illustration:

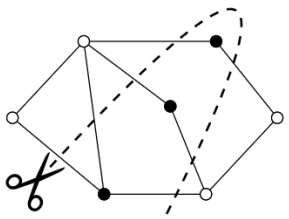

Figure 3.8 The dashed line shows the maximal cut of this graph, splitting the vertices into black and white and giving maxcut$(G) = 7$.

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 $V = \{1, \ldots, n\}$ 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 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

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

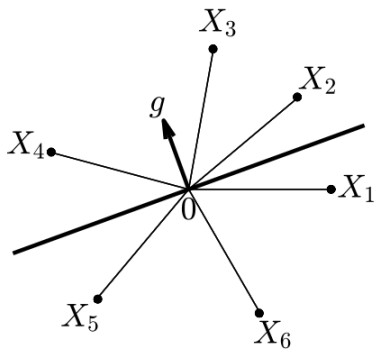

Figure 3.9 We do randomized rounding of these vectors $X_i \in \mathbb{R}^n$ into labels $x_i = \pm 1$ by choosing a random hyperplane with normal vector g (shown in bold) and assigning $x_2 = x_3 = x_4 = 1$ and $x_1 = x_5 = x_6 = -1$.

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).

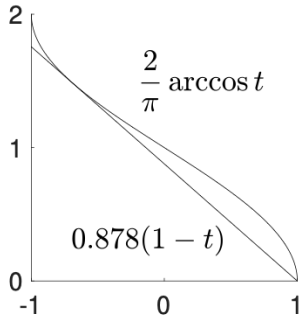

Figure 3.10 The inequality $\frac{2}{\pi}\arccos t \geq 0.878(1-t)$ holds for all $t \in [-1, 1]$.

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

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

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:

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

This would give $K \le \pi/2 \approx 1.57$.
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 $(a_{i_1\dots i_k})$ is an $n_1 \times n_2 \times \cdots \times n_k$ array of real numbers $a_{i_1\dots i_k}$. The canonical inner product on $\mathbb{R}^{n_1\times\cdots\times n_k}$ defines the inner product of tensors: For A, B tensors (of the same dimensions),

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

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 $x_i = \operatorname{sign}\langle g m u_i'\rangle$ and $y_i = \operatorname{sign}\langle g, v_i'\rangle$.

3.7.3 Kernels and Feature Maps

Since the kernel trick worked so well for Grothendieck inequality, we might wonder - what other nonlinearities can it handle? Given a function of two variables $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ on some set $\mathcal{X}$, when can we find a Hilbert space H and a transformation $\Phi : \mathcal{X} \to H$ 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 $(*)$ let's 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 $\sigma_1^2, \ldots, \sigma_n^2$ and corresponding orthonormal eigenvectors $v_1, \ldots, v_n \in \mathbb{R}^n$, so that $A^T A v_i = \sigma_i^2 v_i$. The vectors Av_i are orthogonal:

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

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

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

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 $v_1, \ldots, v_n$ 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

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

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

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

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 $E = \operatorname{span}(u_1, \ldots, u_k)$. 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

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

and we have the upper bound. For the lower bound on λ_k, we need to find $x \in S(E)$ such that $x^T A x \leq \lambda_k$. Here we let the subspace be $F = \operatorname{span}(u_k, \ldots, u_n)$.
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

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

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 $y^TQARx = (Qy)^TA(Rx)$ 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,

$$\|A\| = \max_k |\lambda_k| = \max_{\|x\|=1} |x^TAx|.$$

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 $|\lambda_k| \leq \max_{\|x\|=1} |x^TAx|$, 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. $A^TA = I_n$,

(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) $\|A^TA - I_n\| \le \varepsilon$.

(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

$$\|A^TA - I_n\| = \max_{\|x\|_2=1} |x^T(A^TA - I_n)x| = \max_{\|x\|_2=1} |\|Ax\|_2^2 - 1|,$$

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

$$\|A^TA - I_n\| \le \max(\delta, \delta^2) \implies 1-\delta \le \sigma_n \le \sigma_1 \le 1+\delta.$$

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 $\varepsilon \in \mathcal{N}$. □

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:

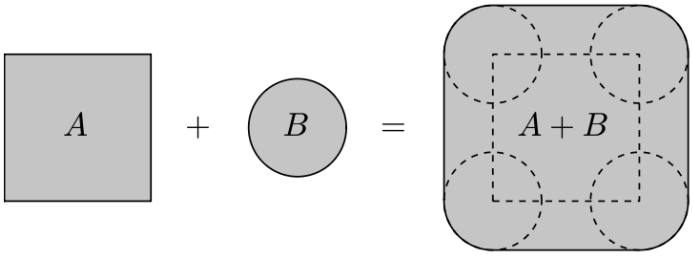

Figure 4.2 Minkowski sum of a square and a circle is a square with rounded corners.

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 $\varepsilon > 0$:

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

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:

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

□

To simplify Corollary 4.2.11, we can divide this into two cases for ε:
For $\varepsilon \in (0, 1]$, we have

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

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:

$$d_H(x, y) := |\{i : \ x(i) \neq y(i)\}|, \ x, y \in \{0,1\}^n.$$

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 $m \in \{0, \ldots, n\}$:

$$\frac{2^n}{\sum_{k=0}^{m} \binom{n}{k}} \leq \mathcal{N}(K, d_H, m) \leq \mathcal{P}(K, d_H, m) \leq \frac{2^n}{\sum_{k=0}^{\lfloor m/2 \rfloor} \binom{n}{k}}.$$

Proof. Use the volumetric argument from above using cardinality instead of the volume (Exercise 4.32). □

4.3 Application: Error Correcting Codes

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

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

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 $u^2 \geq C^2 K^2 (n + m + t^2)$, 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

$$\mathbb{E}[\|A\|] \leq CK(\sqrt{m} + \sqrt{n}).$$

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

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

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 $t > 0$,

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

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 $1 - 4\exp(-t^2)$)

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

By the triangle inequality, $\|A\| \leq \|A^+\| + \|A^-\|$ hence the proof is complete. □

4.5 Application: Community Detection in Networks

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

$$\|\frac{1}{m}A^TA - I_n\| \le K^2 \max(\delta, \delta^2) \text{ where } \delta = C\left(\sqrt{\frac{n}{m}} + \frac{t}{\sqrt{m}}\right).$$

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

$$|\mathcal{N}| \le 9^n.$$

Using Lemma 4.4.2, we can evaluate the operator norm in the equation above on $\mathcal{N}$:

$$\|\frac{1}{m}A^TA - I_n\| \le 2\max_{x\in\mathcal{N}}\left|\left\langle(\frac{1}{m}A^TA - I_n)x, x\right\rangle\right| = 2\max_{x\in\mathcal{N}}\left|\frac{1}{m}\|Ax\|_2^2 - 1\right|.$$

Therefore, to prove the statement, it is enough to show that, with the required probability,

$$\max_{x\in\mathcal{N}}\left|\frac{1}{m}\|Ax\|_2^2 - 1\right| \le \frac{\varepsilon}{2} \text{ where } \varepsilon = K^2\max(\delta, \delta^2).$$

(**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 $\|A_i\|_{\psi_2} \le K$. Thus $X_i = \langle A_i, x\rangle$ are independent subgaussian random variables with $\mathbb{E}[X_i^2] = 1$ and $\|X_i\|_{\psi_2} \le K$. This makes $X_i^2 - 1$ independent, mean zero, and subexponential random variables with

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

Thus we can use Bernstein inequality (Corollary 2.9.2) and obtain

$$\begin{aligned}
P\left(\left|\frac{1}{m}\|Ax\|_2^2 - 1\right| \ge \frac{\varepsilon}{2}\right) &= P\left(\left|\frac{1}{m}\sum_{i=1}^m X_i^2 - 1\right| \ge \frac{\varepsilon}{2}\right) \\
&\le 2\exp\left[-c_1\min\left(\frac{\varepsilon^2}{K^4}, \frac{\varepsilon}{k^2}\right)m\right] \\
&= 2\exp\left(-c_1\delta^2 m\right) \\
&\le 2\exp\left(-c_1C^2(n + t^2)\right).
\end{aligned}$$

The last inequality comes from the definition of δ and using the inequality

$$(a + b)^2 \ge a^2 + b^2 \text{ for } a, b \ge 0.$$

(**Union bound**) Now unfix $x \in \mathcal{N}$. By union bound,

$$P\left(\max_{x\in\mathcal{N}}\left|\frac{1}{m}\|Ax\|_2^2 - 1\right| \ge \frac{\varepsilon}{2}\right) \le 9^n \cdot 2\exp\left(-c_1C^2(n + t^2)\right) \le 2\exp\left(-t^2\right)$$

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:

$$\mathbb{E}\left[\|\frac{1}{m}A^TA - I_n\|\right] \le CK^2\left(\sqrt{\frac{n}{m}} + \frac{n}{m}\right).$$

There is another version of the proof for Theorem 4.6.1 in Exercise 4.46.

4.7 Application: Covariance Estimation and Clustering

5 Concentration Without Independence

This chapter mainly explores other approaches to concentration that do not rely on independence.

5.1 Cencentration 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 $X \sim \text{Unif}(\sqrt{n}S^{n-1})$. Then for any Lipschitz function $f : \sqrt{n}S^{n-1} \to \mathbb{R}$ we have

$$\|f(X) - \mathbb{E}[f(X)]\|_{\psi_2} \leq C\|f\|_{\text{Lip}}.$$

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:

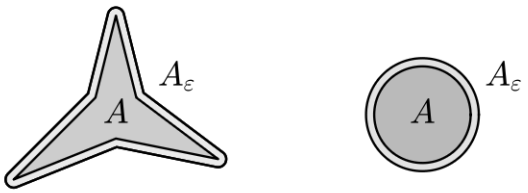

Figure 5.1 The isoperimetric inequality says that among all sets A with a given volume, Euclidean balls minimize the volume of their ε-neighborhood A_ε.

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

$$\sigma(H_t) \geq P(X_1 \leq t/\sqrt{2}) \geq 1 - 2\exp(-ct^2).$$

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

$$P(X \in A_t) \geq 1 - 2\exp(-ct^2).$$

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

$$P(f(X) \leq M + t) \geq 1 - 2\exp(-ct^2).$$

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 $|X_i| \leq 1$.

Theorem 5.2.12 (Talagrand concentration inequality)**.** Consider a random vector in $\mathbb{R}^n$, $X = (X_1, \ldots, X_n)$ whose coordinates are independent and satisfy $|X_i| \leq 1$ almost surely. Then for any Lipschitz function $f : [-1, 1]^n \to \mathbb{R}$,

$$\|f(X) - \mathbb{E}[f(X)]\|_{\psi_2} \leq C\|f\|_{\text{Lip}}.$$

5.3 Application: Johnson-Lindenstrauss Lemma

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 $X_1, \ldots, X_N$ be independent, mean zero, $n \times n$ symmetric random matrices, such that $\|X_i\| \leq K$ almost surely for all i. Then for every $t \geq 0$,

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

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

$$\begin{aligned} E &\leq n\lambda_{\max}(\exp(g(\lambda)Z)) \\ &= m\exp(g(\lambda)\lambda_{\max}(Z)) \\ &= n\exp(g(\lambda)\|Z\|) \quad (\text{Since } Z \succeq 0) \\ &= n\exp(g(\lambda)\sigma^2) \quad (\text{By definition of } \sigma). \end{aligned}$$

Plugging in this bounde for $E = \mathbb{E}[e^{\lambda\cdot\lambda_{\max}(S)}]$ into the original equation gives

$$P(\lambda_{\max}(S) \geq t) \leq n\exp(-\lambda t + g(\lambda)\sigma^2).$$

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 compuiting 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

$$P(\lambda_{\max}(S) \geq t) \leq n\exp\left(-\frac{t^2/2}{\sigma^2 + Kt/3}\right).$$

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

$$\mathbb{E}\left[\left|\sum_{i=1}^N X_i\right|\right] \leq \left(\mathbb{E}\left[\left|\sum_{i=1}^N X_i\right|^2\right]\right)^{1/2} = \left(\sum_{i=1}^N \mathbb{E}[X_i^2]\right)^{1/2}.$$

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 $\varepsilon_1, \ldots, \varepsilon_N$ be independent Rademacher random variables and $A_1, \ldots A_N$ be any (fixed) symmetric $n \times n$ matrices. Then for any $t > 0$,

$$P\left(\|\sum_{i=1}^N \varepsilon_i A_i\| \geq t\right) \leq 2n\exp\left(-\frac{t^2}{2\sigma^2}\right)$$

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.

5.6 Application: Covariance Estimation for General Distributions

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 $X = (X_1, \dots, X_N)$ 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 $X = (X_1, \ldots, X_n)$ 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 $X' = (X'_1, \ldots, X'_n)$ 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 $F : \mathbb{R} \to \mathbb{R}$,

$$\mathbb{E}[F(X^T A X)] \leq \mathbb{E}[F(4 X^T A X')].$$

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 $I \subset \{1, \ldots, n\}$ 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 $\delta_1, \ldots, \delta_n \sim_{iid} \mathrm{Ber}(1/2)$. 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

$$X^T AX = \sum_{i \neq j} a_{ij} X_i X_j = 4\mathbb{E}_\delta \left[\sum_{i \neq j} \delta_i (1 - \delta_j) a_{ij} X_i X_j \right] = 4\mathbb{E}_I \left[\sum_{(i,j \in I \times I^C)} a_{ij} X_i X_j \right].$$

(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 $\delta = (\delta_1, \ldots, \delta_n)$, 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

$$\mathbb{E}_X[F(X^T AX)] \leq \mathbb{E}_I \left[\mathbb{E}_X \left[F \left(4 \sum_{(i,j) \in I \times I^c} a_{ij} X_i X_j \right) \right] \right].$$

It follows that there exists a realization of a subset I such that

$$\mathbb{E}_X[F(X^T AX)] \leq \mathbb{E}_X \left[F \left(4 \sum_{(i,j) \in I \times I^c} a_{ij} X_i X_j \right) \right].$$

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

$$\mathbb{E}_X[F(X^T AX)] \leq \mathbb{E} \left[F \left(4 \sum_{(i,j) \in I \times I^c} a_{ij} X_i X'_j \right) \right].$$

(**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

$$\mathbb{E} \left[F \left(4 \sum_{(i,j) \in I \times I^c} a_{ij} X_i X'_j \right) \right] \leq \mathbb{E} \left[F \left(4 \sum_{(i,j) \in [n] \times [n]} a_{ij} X_i X'_j \right) \right]$$

where $[n] = \{1, \ldots, n\}$. To do this, we can decompose the sum on the right side as

$$\sum_{(i,j) \in [n] \times [n]} a_{ij} X_i X'_j = \underbrace{\sum_{(i,j) \in I \times I^c} a_{ij} X_i X'_j}_{Y} + \underbrace{\sum_{(i,j) \in I \times I} a_{ij} X_i X'_j + \sum_{(i,j) \in I^c \times [n]} a_{ij} X_i X'_j}_{Z}$$

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

$$\mathbb{E}\left[F\left(\sum_{i\neq j} a_{ij}X_iX_j\right)\right] \leq \mathbb{E}\left[F\left(4\sum_{i,j} a_{ij}X_iX'_j\right)\right]$$

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 $\|X\|_{\psi_2} \leq K$. Then for every $t \geq 0$,

$$P(\|X\|_2 \geq CK(\sqrt{n} + t)) \leq e^{-t^2}.$$

Proof. WLOG, we can assume that $K = 1$. □

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 $X_1, \ldots, X_N$ be independent, mean zero random vectors in a

normed space, and let $\varepsilon_1, \dots, \varepsilon_N$ be independent Rademacher random variables. Then

$$\frac{1}{2}\mathbb{E}\left[\left\|\sum_{i=1}^N \varepsilon_i X_i\right\|\right] \le \mathbb{E}\left[\left\|\sum_{i=1}^N X_i\right\|\right] \le 2\mathbb{E}\left[\left\|\sum_{i=1}^N \varepsilon_i X_i\right\|\right].$$

Proof. (**Upper bound**) Let (X_i') be an independent copy of (X_i). Since $\sum_{i=1} X_i'$ has mean zero, we have

$$p := \mathbb{E}\left[\left\|\sum_i X_i\right\|\right] \le \mathbb{E}\left[\left\|\sum_i X_i - \sum_i X_i'\right\|\right] = \mathbb{E}\left[\left\|\sum_i (X_i - X_i')\right\|\right].$$

The inequality above comes from the fact that for independent random vectors Y and Z,

$$\mathbb{E}[Z] = 0 \implies \mathbb{E}[\|Y\|] \le \mathbb{E}[\|Y + Z\|].$$

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

$$\begin{aligned} p &\le \mathbb{E}\left[\left\|\sum_i \varepsilon_i (X_i - X_i')\right\|\right] \\ &\le \mathbb{E}\left[\left\|\sum_i \varepsilon_i X_i\right\|\right] + \mathbb{E}\left[\left\|\sum_i \varepsilon_i X_i'\right\|\right] \quad \text{(Triangle inequality)} \\ &= 2\mathbb{E}\left[\left\|\sum_i \varepsilon_i X_i\right\|\right] \quad \text{(The two terms are identically distributed).} \end{aligned}$$

(**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 $e_1, \dots, e_n$, 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

6.6 Contraction Principle

There is one more useful inequality the text covers in the chapter:

Theorem 6.6.1 (Contraction principle)**.** Let $x_1, \dots, x_N$ be any vectors in a normed space, $(a_1, \dots, a_N) \in \mathbb{R}^N$, and $\varepsilon_1, \dots, \varepsilon_N$ be independent Rademacher random variables. Then

$$\mathbb{E}\left[\left\|\sum_{i=1}^N a_i \varepsilon_i x_i\right\|\right] \le \|a\|_\infty \cdot \mathbb{E}\left[\left\|\sum_{i=1}^N \varepsilon_i x_i\right\|\right].$$

Proof. WLOG, assume that $\|a\|_\infty \le 1$. 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 $\|a\|_\infty \le 1$, 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

$$f(a) \le \mathbb{E}\left[\left\|\sum_{i=1}^N \varepsilon_i x_i\right\|\right] \text{ whenever } \|a\|_\infty \le 1,$$

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 $X_1, \dots, X_N$ be independent, mean zero random vectors in a normed space. Let $g_1, \dots, g_N \sim N(0,1)$ be independent Gaussian random variables, which are also independent of X_i. Then

$$\frac{c}{\sqrt{\log N}} \mathbb{E}\left[\left\|\sum_{i=1}^N g_i X_i\right\|\right] \le \mathbb{E}\left[\left\|\sum_{i=1}^N X_i\right\|\right] \le 3\mathbb{E}\left[\left\|\sum_{i=1}^N g_i X_i\right\|\right].$$

Proof. **(Upper bound)** By symmetrization (Lemma 6.3.2), we have

$$E := \mathbb{E}\left[\left\|\sum_{i=1}^N X_i\right\|\right] \le 2\mathbb{E}\left[\left\|\sum_{i=1}^N \varepsilon_i X_i\right\|\right].$$

To interject Gaussian random variables, recall that $\mathbb{E}[|g_i|] = \sqrt{2/\pi}$. Then we can continue the bound as follows:

$$\begin{aligned} E &\le 2\sqrt{\frac{\pi}{2}} \mathbb{E}_X\left[\left\|\sum_{i=1}^N \varepsilon_i \mathbb{E}_g[|g_i|] X_i\right\|\right] \\ &\le 2\sqrt{\frac{\pi}{2}} \mathbb{E}\left[\left\|\sum_{i=1}^N \varepsilon_i |g_i| X_i\right\|\right] \quad \text{(Jensen inequality)} \\ &= 2\sqrt{\frac{\pi}{2}} \mathbb{E}\left[\left\|\sum_{i=1}^N g_i X_i\right\|\right]. \end{aligned}$$

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

$$\begin{aligned} \mathbb{E}\left[\left\|\sum_{i=1}^N g_i X_i\right\|\right] &= \mathbb{E}\left[\left\|\sum_{i=1}^N \varepsilon_i g_i X_i\right\|\right] \quad \text{(Symmetry of } g_i\text{)} \\ &\le \mathbb{E}_g\left[\mathbb{E}_X\left[\|g\|_\infty \mathbb{E}_\varepsilon\left[\left\|\sum_{i=1}^N \varepsilon_i X_i\right\|\right]\right]\right] \quad \text{(Theorem 6.6.1)} \\ &= \mathbb{E}_g\left[\|g\|_\infty \mathbb{E}_\varepsilon\left[\mathbb{E}_X\left[\left\|\sum_{i=1}^N \varepsilon_i X_i\right\|\right]\right]\right] \quad \text{(Independence)} \\ &\le 2\mathbb{E}_g\left[\|g\|_\infty \mathbb{E}_X\left[\left\|\sum_{i=1}^N X_i\right\|\right]\right] \quad \text{(Lemma 6.3.2)} \\ &= 2\mathbb{E}[\|g\|_\infty] \cdot \mathbb{E}\left[\left\|\sum_{i=1}^N X_i\right\|\right] \quad \text{(Independence).} \end{aligned}$$

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

8 Chaining

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.