Notes for High-Dimensional Probability Second Edition by Roman Vershynin

Gallant Tsao

July 7, 2025

Contents

0 Appetizer: Using Probability to Cover a Set

Definition 0.0.1. A convex combination of points $z_1, \ldots, z_m \in \mathbb{R}^n$ is a linear combination with coefficients that are nonnegative and sum to 1, i.e. it is a sum of the form

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

Definition 0.0.2. The convex hull of a set $T \in \mathbb{R}^n$ is the set of all convex combinations of all finite collections of points in T, i.e.

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

Theorem 0.0.3 (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 0.0.4 (Approximate Caratheodory Theorem). Consider a set $T \subseteq \mathbb{R}^n$ that is contained in the unit Euclidean ball. Then, for every point $x \in \operatorname{conv}(T)$ and every $k \in \mathbb{N}$, one can find points $x_1, \ldots, x_k \in T$ such that

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

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

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

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 $P \subset \mathrm{conv}(P) \subset \mathrm{conv}(T)$ where $T = \{\text{Vertices of } P\}$. Then we apply theorem 0.0.4 to any point $x \in P \subseteq \mathrm{conv}(T)$ 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

Definition 1.1.1. 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

Definition 1.2.1. The Euclidean norm of a vector $x \in \mathbb{R}^n$ is

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

Definition 1.2.2. The inner product (dot product) of two vectors $x, y \in \mathbb{R}^n$ is

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

Definition 1.2.3. For $p \in [1,\infty]$, the ℓ^p norm of a vector $x \in \mathbb{R}^n$ is

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

Theorem 1.2.4 (Minkowski's inequality)**.** 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 $p \in [1,\infty]$.

Theorem 1.2.5 (Cauchy-Schwartz inequality)**.** For all vectors $x, y \in \mathbb{R}^n$,

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

Theorem 1.2.6 (Hölder's inequality)**.** 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:

Definition 1.3.1. The expectation (mean) of a random variable X is

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

Its variance is

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

The expectation is linear:

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

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

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

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

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

Its expectation is given by

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

Definition 1.3.2. The moment generating function (mgf) of a random variable X is given by

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

Definition 1.3.3. 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:

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

Definition 1.3.4. The standard deviation of a random variable X is

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

The covariance of two random variables X and Y is

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

Definition 1.3.5. 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

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

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

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

Definition 1.5.1. Given a probability space, the conditional probability of an event E given an event F is

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

1.6 Probabilistic Inequalities

Theorem 1.6.1 (Jensen's Inequality)**.** 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

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

Theorem 1.6.2 (Inequalities for random variables)**.** Minkowski inequality states that for any $p \in [1, \infty]$ and any random variables $X, Y \in L^p$,

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

1.7 Limit Theorems

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

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

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

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

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

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

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

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

Then as $N \to \infty$,

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

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

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

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

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

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

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

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

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

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

Assume that as $N \to \infty$,

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

Then as $N \to \infty$,

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

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

Lemma 1.7.7 (Stirling approximation)**.**

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

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

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

Of course, there are also non-asymptotic results:

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

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

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

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

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

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

Exponentiating and rearranging gives the upper bound. □

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

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

Repeated integration by parts gives

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

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

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

2 Concentration of Sums of Independent Random Variables

2.1 Why Concentration Inequalities?

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

Example 2.1.1. Toss a fair coin N times. What is the probability that we get at least $\frac{3}{4}$ heads? Let S_N denote the number of heads, then $S_N \sim \text{Binom}(N, \frac{1}{2})$. We get

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

Using Chebyshev's Inequality, we get

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

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

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

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

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

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

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

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

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

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

□

Therefore 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 2.1.3 (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

Definition 2.2.1. 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 2.2.2 (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:

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

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

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

Substituting this inequality into what we have above gives

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

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

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

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

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

□

Remark 2.2.3 (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 2.2.4 (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 2.2.5 (The probability of $\frac{3}{4}N$ heads)**.** Using Hoeffding, returning back to example 2.1.1 and bound the probabiltiy of at least $\frac{3}{4}N$ heads in N tosses of a fair coin. Since $Y \sim \text{Bernoulli}(1/2)$, $2Y - 1$ is Rademacher. Since S_N is a sum of N independent $\text{Be}(1/2)$ 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 2.2.6 (Hoeffding inequality, two-sided)**.** Let $X_1, \ldots, X_N$ be independent Rademacher random variables, and let $a = (a_1, \cdots, a_N) \in \mathbb{R}^N$ be fixed. Then for any $t \geq 0$,

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

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

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

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

Hoeffding inequality can be also be applied to general bounded random variables:

Theorem 2.2.7 (Hoeffding inequality for bounded random variables)**.** Let $X_1, \ldots, X_N$ be independent random variables such that $X_i \in [a_i, b_i]$ for every i. Then for any $t > 0$, we have

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

Proof. Done in Exercise 2.10. □

2.3 Chernoff Inequality

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

Theorem 2.3.1 (Chernoff inequality)**.** Let $X_i \sim \text{Ber}(p_i)$ 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 2.2.2 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,

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

Moreover,

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

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

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

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

Proof. Done in Exercise 2.11. □

Remark 2.3.3 (Poisson tails)**.** When p_i is small for the Bernoulli random variables, by the Poisson Limit Theorem (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 $t = (1+\delta)\mu$,

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

Expanding the expression inside the exponent via Taylor series,

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

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

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

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

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

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

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

Same as before, expanding the expression into Taylor series gives

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

Plugging the bound gives

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

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

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

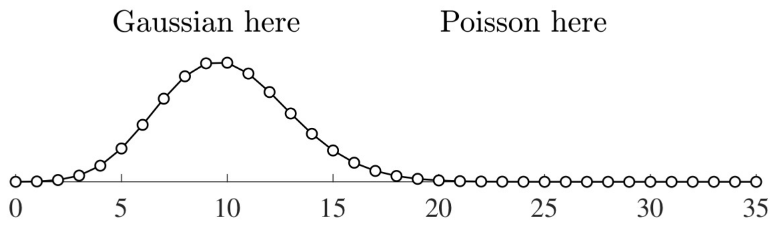

Figure 2.1 The probability mass function of the distribution $\text{Binom}(N, \mu/N)$ with $N = 200$ and $\mu = 10$. It is approximately normal near the mean μ, but it is heavier far from the mean.

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.

2.5 Application: Degrees of Random Graphs

2.6 Subgaussian Distributions

Standard form for Hoeffding Inequality (including subgaussian distributions):

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

Definition 2.6.1. 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 2.6.2 (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 2.6.3 (Zero mean)**.** For property (d) above, $\mathbb{E}[X]$ is a necessary and sufficient condition (Exercise 2.23)!

Remark 2.6.4 (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 2.6.5. A random variable X is called subgaussian if it satisfies any of the equivalent properties in proposition 2.6.2. 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 2.6.6. 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 2.6.2 with those having the subgaussian norm:

Proposition 2.6.7 (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 $X_1, \dots, X_N$ be independent mean zero subgaussian random variables. Then

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

where C is an absolute constant.

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

$$\begin{aligned}
\mathbb{E}[\exp(\lambda S_N)] &= \prod_{i=1}^{N} \mathbb{E}[\exp(\lambda X_i)] \quad \text{(independence)} \\
&\leq \prod_{i=1}^{N} \exp(C\lambda^2 \|X_i\|_{\psi_2}^2) \quad \text{(proposition 2.6.7 (d))} \\
&= \exp(\lambda^2 K^2), K^2 = C\sum_{i=1}^{N} \|X_i\|_{\psi_2}^2.
\end{aligned}$$

Then by proposition 2.6.2, $(d) \Rightarrow (c)$ 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 $X_1, \dots, X_N$ be independent, mean zero, subgaussian random varirables. Then for every $t \geq 0$,

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

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

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

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

2.7.2 Subgaussian Khintchine Inequality

Below is a two-sided bound on the L^p norms of sums of independent random variables:

Theorem 2.7.5 (Khintchine Inequality)**.** Let $X_1, \ldots, X_N$ be independent subgaussian random variables with zero means with unit variances. Let $a_1, \ldots, a_n \in \mathbb{R}$. 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 2.6.7.

2.7.3 Maximum of Subgaussians

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

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

In particular,

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

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

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

where the last inequality comes from (a) of proposition 2.6.7. 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,

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

Then by proposition 2.6.7 ($(c) \iff (a)$) 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 $Z = \max_{i=1,\ldots,N} |X_i|$. Then

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

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 2.6.7 for $p = 1$. □

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

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

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

2.7.4 Centering

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

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

There is a similar phenomenon for the subgaussian norm:

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

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

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

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

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

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

where the last step comes from (b) of proposition 2.6.7 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

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

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

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

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

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:

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

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

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

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

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

We also have that

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

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

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

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 $\lambda \ge 1$!

2.8.2 The Subexponential Norm

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

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

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

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

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

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

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

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

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

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

Example 2.8.7. The following random variables are subexponential:

(a) Any subgaussian random variable,

(b) The square of any subgaussian random variable,

(c) Exponential,

(d) Poisson,

(e) Geometric,

(f) Chi-squared,

(g) Gamma.

The Cauchy the Pareto distributions are *not* subexponential.

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

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

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

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

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

Quantitatively,

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

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

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

2.9 Bernstein Inequality

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

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

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

where $c > 0$ is an absolute constant.

Proof. By using the exponential moment method,

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

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

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

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

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

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

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

Plugging this optimal λ^* back we get

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

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

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

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

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

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

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

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

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

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

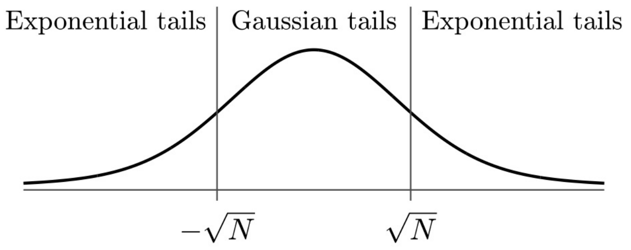

Figure 2.3 Bernstein inequality exhibits a mixture of two tails: gaussian for small deviations and exponential for large deviations.

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

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

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

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

Proof. Exercise 2.47. □

3 Random Vectors in High Dimensions

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

3.1 Concentration of the Norm

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

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

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

Proof. Using proposition 2.6.7, we can rewrite the above as

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

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

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

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

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

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

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

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

$$1 = \|X_1\|_{L^2} \leq C\|X_1\|_{\psi_2} \leq CK \text{ by } \textit{proposition } 2.6.7.$$

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

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

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

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

Then for any $\delta \geq 0$,

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

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

Remark 3.1.2 (Thin shell phenomenon)**.** The theorem above shows that random vectors in $\mathbb{R}^n$ mostly stay in a shell of constant thickness around the sphere of radius $\sqrt{n}$. This might seem surprising, but here's an intuitive explanation:
The square of the norm, $\|X\|_2^2$, has a chi-squared distribution with n degrees of freedom. Hence its mean is n, and standard deviation $\sqrt{2n}$. Thus it makes sense for $\|X\|_2$ to deviate by $O(1)$ around $\sqrt{n}$ because

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

3.2 Covariance Matrices and PCA

Definition 3.2.1. 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 3.2.2. 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:

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

□

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 3.2.3. 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 3.2.4 (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

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

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 3.2.5 (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-dimensonsal, 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 3.2.6. 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 3.2.2 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.

Remark 3.2.7 (Standardizing)**.** In one dimension, a random variable X can be standardized to a zero mean, unit variance random variable Z by doing

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

This is also true in higher dimensions:

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

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

Definition 3.3.1. 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 3.3.2 (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 3.3.3 (Name)**.** 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 3.3.4 (Sum of independent normals is normal)**.** Consider independent normal random variables $X_i \sim N(\mu_i, \sigma_i^2)$. Then,

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

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

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

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

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

□

3.3.2 General Normal

Definition 3.3.5. 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 3.3.6 (Uniqueness of normal)**.** The distribution of X is uniquely determined by μ and Σ. Specifically, X has the same distribution as

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

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

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

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

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

Proposition 3.3.7. 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 3.3.8 (Name)**.** Random variables $X_1, \dots, X_n$ are jointly normal if the random vector $X = (X_1, \dots, X_n)$ is normally distributed. Jointly normal random variables are independent if and only if they are uncorrelated.

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

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

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

3.3.3 Uniform on the Sphere

Proposition 3.3.9 (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 3.3.9. By (c) from proposition 3.2.2,

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

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

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

In fact, the CDF converges uniformly:

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

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

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

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

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

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

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

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

Splitting the event based on whether E_n occurs, we get

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

Hence

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

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

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

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

Remark 3.3.11 (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

Definition 3.3.12. Let $K \subset \mathbb{R}^n$ be a convex set. A random variable X is uniformly distributed in K, denoted $X \sim \text{Unif}(K)$, if its density is $1/\text{Vol}(K)$ on K and zero everywhere else.

The mean of X is

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

which is the center of gravity of K. If Σ is the covaraince matrix of K, then the standard score $Z := \Sigma^{-1/2}(X - \mu)$ is an isotropic random vector from remark 3.2.7. In fact, Z is uniformly distributed in the affinely transformed copy of K:

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

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

3.3.5 Frames

A frame extends the concept of a basis, but drops the requirement of linear independence. Frames are intimately connected to discrete isotropic distributions:

Proposition 3.3.13 (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) ⇒ (iv) The identity for (i) can be written as

$$\|x\|_2^2 = \frac{1}{N}\sum_{i=1}^N \left\langle \sqrt{N}u_i, x\right\rangle^2 = \mathbb{E}[\langle X, x\rangle^2].$$

Since this holds for all $x \in \mathbb{R}^n$, the random vector is isotropic.
(iv) ⇒ (iii) Since X is isotropic,

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

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

Example 3.3.14 (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 3.3.15 (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:

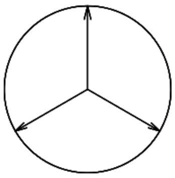

Figure 3.7 A Mercedez-Benz frame: three equispaced points on the circle of radius $\sqrt{2/3}$ form a Parseval frame in $\mathbb{R}^2$.

Here are two more examples of isotropic distributions:

Example 3.3.16 (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 3.3.17 (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 3.3.10) 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 (Name)**.** Let $X \sim \mathrm{Unif}(S^{n-1})$. 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

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

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

Example 3.4.7 (Coordinate distribution)**.** Let $X \sim \text{Unif}\{\sqrt{n}e_1, \dots, \sqrt{n}e_n\}$. 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 3.3.13) 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

3.6 Application: Maximum Cut for Graphs

3.7 Kernel Trick and Tightening of Grothendieck Inequaltity

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:

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

Therefore, there exist orthonormal vectors $u_1, \ldots, u_n \in \mathbb{R}^n$ 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 $u_1, \dots, u_k$ 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 $\underbrace{1, \dots, 1}_{k}, \underbrace{0, \dots, 0}_{n-k}$.

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 $E = \operatorname{span}(u_1, \dots, u_k)$. 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 $F = \operatorname{span}(u_k, \dots, u_n)$.
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 $\sigma_1 \geq \cdots \geq \sigma_n$,

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

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)

$$\sigma_n \|x - y\|_2 \leq \|Ax - Ay\|_2 \leq \sigma_1 \|x - y\|_2.$$

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 $m \geq n$, the following are equivalent:

(a) The columns of A are orthonormal, i.e. $A^T A = I_n$,

(b) A is an isometry,

(c) All singular values of A are 1.

There is a stronger result where the properties hold approximately instead of exactly (useful when dealing with random matrices):

Lemma 4.1.17 (Approximate isometries)**.** Let $A \in \mathbb{R}^{m \times n}$ with $m \geq n$ and let $\varepsilon \geq 0$. The following are equivalent:

(a) $\|A^T A - I_n\| \leq \varepsilon$.

(b) $(1 - \varepsilon)\|x\|_2^2 \leq \|Ax\|_2^2 \leq (1 + \varepsilon)\|x\|_2^2$ for any $x \in \mathbb{R}^n$.

(c) $1 - \varepsilon \leq \sigma_n^2 \leq \sigma_1^2 \leq 1 + \varepsilon$.

Proof. (a) ⇔ (b) By rescaling, we can assume that $\|x\|_2 = 1$ in (b). Then we have

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

The above being bounded by ε is equivalent to (b) for all unit vectors x.
(b) ⇔ (c) follows from the relationship for singular values distorting space from above. □

Remark 4.1.18. Here is a more handy version of (a) ⇒ (c) in lemma 4.1.17. For $z \in \mathbb{R}$ and $\delta \geq 0$,

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

Then substituting $\varepsilon = \max(\delta, \delta^2)$, we get

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

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.