

Notes for High-Dimensional Probability Second Edition by
Roman Vershynin

Gallant Tsao

June 29, 2025

Contents

0	Appetizer: Using Probability to Cover a Set	2
1	A Quick Refresher on Analysis and Probability	4
1.1	Convex Sets and Functions	4
1.2	Norms and Inner Products	4
2	Concentration of Sums of Independent Random Variables	5
2.1	Why Concentration Inequalities?	5
2.2	Hoeffding Inequality	6
2.3	Chernoff Inequality	8
2.4	Application: Median-of-means Estimator	11
2.5	Application: Degrees of Random Graphs	11
2.6	Subgaussian Distributions	11
2.6.1	The Subgaussian Norm	13
2.7	Subgaussian Hoeffding and Khintchine Inequalities	13
2.7.1	Subgaussian Hoeffding Inequality	14
2.7.2	Subgaussian Khintchine Inequality	14
2.7.3	Maximum of Subgaussians	15
2.7.4	Centering	16
2.8	Subexponential Distributions	16
2.8.1	Subexponential Properties	16
2.9	Bernstein Inequality	16

0 Appetizer: Using Probability to Cover a Set

Definition 0.0.1. A convex combination of points $z_1, \dots, z_m \in \mathbb{R}^n$ 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 $T \subseteq \mathbb{R}^n$ is the set of all convex combinations of all finite collections of points in T , i.e.

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

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

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

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 $m - 1$ points $\{x_2 - x_1, \dots, x_m - x_1\}$ have to be linearly dependent because we have at least $n + 1$ points in a subspace of \mathbb{R}^n . Let $b_2, \dots, b_m \in \mathbb{R}$ 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, \dots, c_n such that

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

Define $I = \{i \in \{1, 2, \dots, n\} : c_i > 0\}$. 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. \square

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 $x \in \text{conv}(T)$ and every $k \in \mathbb{N}$, one can find points $x_1, \dots, x_k \in T$ such that

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

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

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

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

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

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, \dots, Z_k with the same distribution as Z . The strong law of large numbers tells us that

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

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, \dots, Z_k such that

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

□

Covering Geometric Sets

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 $K \subseteq \mathbb{R}^n$ be a convex subset. A function $f : K \rightarrow \mathbb{R}$ 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

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

Theorem 1.2.4 (Minkowski's Inequality). For any vector $x, y \in \mathbb{R}^n$,

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

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 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

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

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 $\text{Be}(\frac{1}{2})$ 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 $S_N = X_1 + \dots + X_N$, and let

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

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 $a = (a_1, \dots, a_N) \in \mathbb{R}^N$ be fixed. Then for any $t \geq 0$,

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

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

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

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

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

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

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

Next we will use the following inequality:

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

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

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

Substituting this inequality into what we have above gives

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

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

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

Then the second derivative test gives

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

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

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

□

Remark 2.2.3 (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 $N \rightarrow \infty$. 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 probability 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, \dots, X_N be independent Rademacher random variables, and let $a = (a_1, \dots, a_N) \in \mathbb{R}^N$ 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. \square

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, \dots, X_N be independent random variables such that $X_i \in [a_i, b_i]$ for every i . Then for any $t > 0$, we have

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

Proof. Done in Exercise 2.10. \square

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

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

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. \square

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. \square

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,

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

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

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

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,

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

Same as before, expanding the expression into Taylor series gives

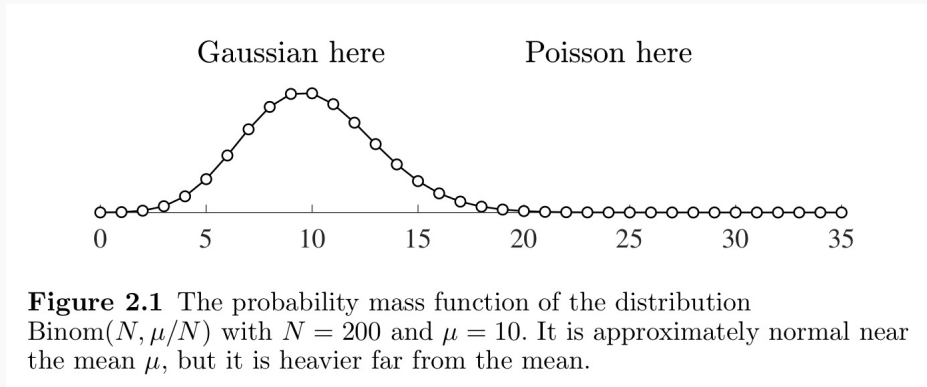
$$\begin{aligned} (1 - \delta) \ln(1 - \delta) + \delta &= (1 - \delta) \left(-\delta - \frac{\delta^2}{2} - \frac{\delta^3}{3} - \dots \right) + \delta \\ &= \left(-\delta - \frac{\delta^2}{2} - \frac{\delta^3}{3} - \dots \right) + (\delta^2 + \frac{\delta^3}{2} + \frac{\delta^4}{3} + \dots) + \delta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

Plugging the bound gives

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

Summing up both bounds via union bound gives the result. \square

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



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, \dots, X_N be independent copies of X (representing a sample). We want an estimator $\hat{\mu}(X_1, \dots, X_N)$ 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 properties 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

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

(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

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

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

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

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

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 $x = n + t$, $1/2 \leq t < 1$,

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

Then taking the p th 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. \square

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 to 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 $\lambda \in \mathbb{R}$,

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

Then by proposition 2.6.2, $(d) \Rightarrow (c)$ hence

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

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. \square

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

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

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

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

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

□

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

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

We then get the factor of \sqrt{p} in the final result from (b) of proposition 2.6.7.

2.7.3 Maximum of Subgaussians

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

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

In particular,

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

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

Proof. Two proof methods are provided in the book.

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

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

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

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

Then by proposition 2.6.7 ((c) \iff (a)) we get $\max_i \|X_i\|_{\psi_2} \leq C \sqrt{\ln N}$.

Method 2: Maximum with sum. Again, assume that $\max_i \|X_i\|_{\psi_2} = 1$ and denote $Z = \max_{i=1, \dots, N} |X_i|$. Then

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

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

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

Then $\|Z\|_{\psi_2} \leq M = \sqrt{2 \ln(2N)}$, proving the first statement. The second statement follows from the first statement via (b) of proposition 2.6.7 for $p = 1$. \square

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

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

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

2.7.4 Centering

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

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

There is a similar phenomenon for the subgaussian norm:

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

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

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

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

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

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

where the last step comes from (b) of proposition 2.6.7 with $p = 1$. Substituting this back into the equation for the triangle inequality and we are done. \square

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

2.9 Bernstein Inequality