Notes for High-Dimensional Probability Second Edition by Roman Vershynin

Gallant Tsao

June 26, 2025

Contents

0 Appetizer: Using Probability to Cover a Set **2**

1 Convex Sets and Functions **4**

2 Concentration of Sums of Independent Random Variables **5**

2.1 Why Concentration Inequalities? 5

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.

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

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

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

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

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

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

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

We can immediately get that the expected value of Z is

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

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

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

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

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

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

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

Then, we get that

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

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

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

□

Covering Geometric Sets

1 Convex Sets and Functions

Definition 1.0.1. 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.

Norms and Inner Products

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

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

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.

Hoeffding Inequality

Definition 2.1.4. 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 2.1.5 (Hoeffding Inequality)**.** Let $X_1, \dots, X_N$ be independent Rademacher random variables, and let $a = (a_1, \cdots, a_n) \in \mathbb{R}^n$ be fixed. Then for any $t \geq 0$,

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

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

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

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

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

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

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

Next we will use the following inequality:

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

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

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

Substituting this inequality into what we have above gives

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

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

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

Then the second derivative test gives

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

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

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

□

Remark 2.1.6 (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 2.1.7 (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 2.1.8 (The probability of $\frac{3}{4}N$ heads)**.** Using Hoeffding, returning back to example 2.1.1 and bound the probabiltiy of at least $\frac{3}{4}N$ heads in N tosses of a fair coin. Since $Y \sim$ Bernoulli(1/2), $2Y - 1$ is Rademacher. Since S_N is a sum of N independent Be(1/2) random variables, $2S_N - N$ is

a sum of N independent Rademacher random variables. Hence

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

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

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

Theorem 2.1.9 (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 $t \geq 0$,

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