Notes for High-Dimensional Probability Second Edition by Roman Vershynin

Gallant Tsao

July 10, 2025

Contents

2 Concentration of Sums of Independent Random Variables

2.1 Why Concentration Inequalities?

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

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

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

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

Using Chebyshev's Inequality, we get

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

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

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

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

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

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

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

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

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

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

The lower bound is proven in Exercise 2.2. □

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

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

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

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

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

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

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

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

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

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

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

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

2.2 Hoeffding Inequality

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

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

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

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

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

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

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

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

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

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

Next we will use the following inequality:

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

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

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

Substituting this inequality into what we have above gives

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

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

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

Then the second derivative test gives

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

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

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

□

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

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

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

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

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

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

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

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

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

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

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

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

Theorem 2.2.6 (Hoeffding inequality for bounded random variables)**.** Let $X_1, \dots, X_N$ be independent random variables such that $X_i \in [a_i, b_i]$ for every i. Then for any $t > 0$, we have

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

Proof. Done in Exercise 2.10. □

2.3 Chernoff Inequality

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

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

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

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

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

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

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

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

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

Substituting back to the original equation gives

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

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

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

Moreover,

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

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

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

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

Proof. Done in Exercise 2.11. □

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

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

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

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

Corollary 2.3.4 (Chernoff inequality: small deviations)**.** In the setting of Theorem 2.3.1,

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

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

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

Expanding the expression inside the exponent via Taylor series,

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

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

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

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

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

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

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

Same as before, expanding the expression into Taylor series gives

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

Plugging the bound gives

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

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

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

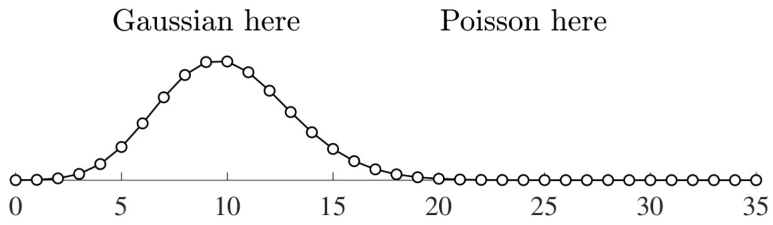


Figure 2.1 The probability mass function of the distribution $\text{Binom}(N, \mu/N)$ with $N = 200$ and $\mu = 10$. It is approximately normal near the mean μ, but it is heavier far from the mean.

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.

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

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

The expected value and the variance of this estimator is

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

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

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

For example, the error is at most $10\sigma/\sqrt{N}$ with at least 99% probability, which is an acceptable solution to the mean estimation problem.
Is the solution above **optimal** though? Could the probability decay quicker than the rate of $1/t^2$?
For the Gaussian distribution, the answer is yes.

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

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

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

For example, the error is at most $3\sigma/\sqrt{N}$ with at least 99% probability. We might think that Gaussian tail decay requires Gaussian distributions, but surprisingly, a mean estimator exists with Gaussian tail decay that works for **any** distirbution with finite variance!

Theorem 2.4.1 (Median-of-means estimator)**.** Let X be a random variable with mean μ and variance σ^2, and let $X_1, \dots, X_N$ be independent copies of X. For any $0 \leq t \leq \sqrt{N}$, there exists an estimator $\hat{\mu} = \hat{\mu}(X_1, \dots, X_N)$ that satisfies

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

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

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

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

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

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

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

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

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

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

where $c_0 > 0$ is some absolute constant.
Similarly, the probability $P\left(\mu_b \geq \mu - \frac{t\sigma}{\sqrt{N}}\right)$ has the same bound. Combining the two bounds above completes the proof.
Notice that we assumed B must be an integer that divides N. The choice above, $B = t^2/t$, only ensures that $0 \leq B \leq N$ by the assumption on t. This issue can be fixed (Exercise 2.16). □

2.5 Application: Degrees of Random Graphs

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

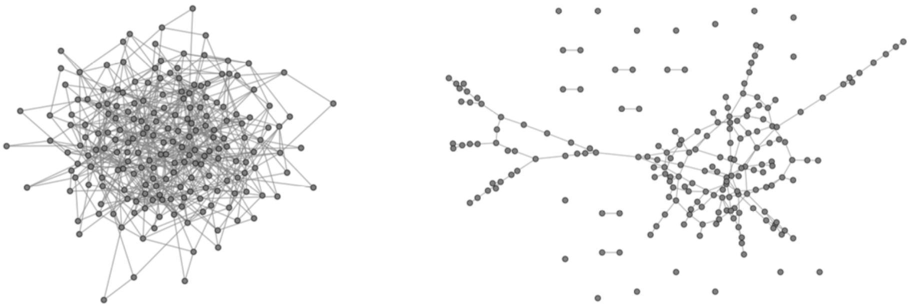


Figure 2.2 Examples of random graphs in the Erdős-Rényi model $G(n, p)$ with $n = 200$ vertices and connection probabilities $p = 0.03$ (left) and $p = 0.01$ (right).

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

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

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

Proposition 2.5.1 (Dense graphs are almost regular)**.** There is an absolute constant C such that the following holds:
Consider a random graph $G \sim G(n, p)$ with expected degree satisfying $d \geq C \log n$. Then with probability at least 0.99, all vertices of G have degrees between $0.9d$ abd $1.1d$.

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

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

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

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

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

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

and the proof is complete. □

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

2.6 Subgaussian Distributions

Standard form for Hoeffding Inequality (including subgaussian distributions):

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

A random variable X has a subgaussian distribution if

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

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

Proposition 2.6.1 (Subgaussian properties)**.** Let X be a random variable. The following peoperties are equivalent, with the parameters K_i differing by at most an absolute constant factor, i.e. There exists an absolute constant C such that property i implies property j with parameter $K_j \leq CK_i$ for any two properties i, j.

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

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

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

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

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

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

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

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

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

Proof. The proof is all about transforming one type of information about random variables into another. $(a) \Rightarrow (b)$ Assume (a) holds. 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

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

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

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

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

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

(b) guarantees that $\mathbb{E}[X^{2p}] \leq (2p)^p$, and $p! \geq (p/e)^p$ by lemma 1.7.8 + link, hence substituting these bound in, we get

$$\mathbb{E}[\exp(\lambda^2 X^2)] \leq 1 + \sum_{p=1}^{\infty} \frac{(2\lambda^2 p)^p}{(p/e)^p} = \sum_{p=0}^{\infty} (2e\lambda^2)^p = \frac{1}{1 - 2e\lambda^2} = 2$$

if we choose $\lambda = 1/2\sqrt{e}$. This means we get (c) with $K_3 = 2\sqrt{e}$.
$(c) \Rightarrow (a)$ WLOG assume that $K_3 = 1$ and property (c) holds. By exponentiating and using Markov's inequality,

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

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

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

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

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

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

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

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

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

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

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

Then by summing up these probabilities,

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

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

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

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

2.6.1 The Subgaussian Norm

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

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

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

Example 2.6.5. The following random variables are subgaussian:

(a) Normal,

(b) Rademacher,

(c) Bernoulli,

(d) Binomial,

(e) Any bounded random variable.

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

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

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

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

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

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

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

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

2.7 Subgaussian Hoeffding and Khintchine Inequalities

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

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

There is a similar weaker property for the subgaussian norm:

Proposition 2.7.1 (Subgaussian norm of a sum)**.** Let $X_1, \dots, X_N$ 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 $\lambda \in \mathbb{R}$,

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

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

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

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. □

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

2.7.1 Subgaussian Hoeffding Inequality

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

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

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

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

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

2.7.2 Subgaussian Khintchine Inequality

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

Theorem 2.7.5 (Khintchine Inequality)**.** Let $X_1, \ldots, X_N$ be independent subgaussian random variables with zero means with unit variances. Let $a_1, \dots, a_n \in \mathbb{R}$. Then for every $p \in [2, \infty)$, we have

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

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

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

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

□

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

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

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

2.7.3 Maximum of Subgaussians

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

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

In particular,

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

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

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

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

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

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

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

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

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

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

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

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

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

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

2.7.4 Centering

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

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

There is a similar phenomenon for the subgaussian norm:

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

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

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

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

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

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

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

2.8 Subexponential Distributions

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

2.8.1 Subexponential Properties

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

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

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

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

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

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

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

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

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

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

Proof. The equivalence of (i)-(iii) is done in Exercise 2.41. (iii)⇒(iv) and (iv)⇒(i) are a bit different and will be done here.
(iii)⇒(iv) Assume that (iii) holds, and WLOG assume $K_3 = 1$. We'll use again the inequality coming from Taylor's theorem with Lagrange form remainder:

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

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

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

Then property (iv) is true with $K_4 = 2$.
(iv)⇒(i) Assume that (iv) holds, and WLOG assume $K_4 = 1$. Exponentiating, applying Markov inequality, and using (iv) for $\lambda = 1$, we get

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

We also have that

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

Combining the two equations above vis union bound, we get $P(|X| >= t) <= 2e^{1-t}$. There are now two cases:
Case 1: $t \geq 2$. Then the $2e^{1-t} \leq 2e^{-t/2}$ hence we are done.
Case 2: $t < 2$. Then $2e^{-t/2} \geq 1$ hence the probability is trivially bounded, we are done.
Therefore we get property (i) with $K_1 = 2$. □

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

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

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

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

2.8.2 The Subexponential Norm

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

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

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

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

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

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

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

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

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

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

Example 2.8.7. The following random variables are subexponential:

(a) Any subgaussian random variable,

(b) The square of any subgaussian random variable,

(c) Exponential,

(d) Poisson,

(e) Geometric,

(f) Chi-squared,

(g) Gamma.

The Cauchy the Pareto distributions are *not* subexponential.

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

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

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

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

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

Quantitatively,

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

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 $X_1, \ldots, X_N$ be indepependent, mean zero, subexponential random variables. Then for every $t \geq 0$,

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

where $c > 0$ is an absolute constant.

Proof. By using the exponential moment method,

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

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

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

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

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

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:

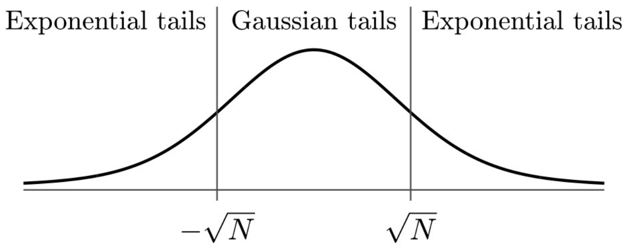


Figure 2.3 Bernstein inequality exhibits a mixture of two tails: gaussian for small deviations and exponential for large deviations.

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. □

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

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

where the first max/min is taking with respect to all subspaces of a fixed dimension, and $S(E)$ denotes the Euclidean unit sphere of E, i.e. the set of all unit vectors in E.

Proof. Let us focus on the first equation. To prove the upper bound on λ_k, we need to find a k-dimensional subspace E such that

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

To find the set E, take the spectral decomposition $A = \sum_{i=1}^{n} \lambda_i u_i u_i^T$ and pick the subspace $E = \operatorname{span}(u_1, \dots, u_k)$. 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

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

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

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

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

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

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

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

with the same notation as Section 4.1.2.

4.1.3 Frobenius and Operator Norms

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

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

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

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

Equivalently,

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

From the Frobenius norm, we can get that

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

Also, from above we can get

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

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,

$$\begin{aligned}\|QAR\|_F &= \operatorname{tr}(R^TATQ^TQAR)\\ &= \operatorname{tr}(R^TA^TAR)\\ &= \operatorname{tr}(RR^TA^TA)\\ &= \operatorname{tr}(A^TA)\\ &= \|A\|_F^2.\end{aligned}$$

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

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

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

Proof. For the Frobenius norm, by orthogonal invariance (Lemma 4.1.10),

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

which directly gives us the result.
The result for the operator norm directly follows from Corollary 4.1.7 with $k = 1$. □

Remark 4.1.12 (Symmetric matrices)**.** For a symmetric matrix A with eigenvalues λ_k,

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

The first equality becomes Lemma 4.1.11 since the singular values of A are $|\lambda_k|$. The min-max theorem (Section 4.1.2) gives $|\lambda_k| \leq \max_{\|x\|=1} |x^T Ax|$, proving the upper bound in the equation above. The lower bound can be proven by taking $x - y$ in the definition of the operator norm (Definition 4.1.8).

4.1.5 Low-rank Approximation

For a given matrix A, what is the closest approximation to it for a given matrix of rank k? The answer is just truncating the SVD of A:

Theorem 4.1.13 (Eckart-Young-Mirski theorem)**.** Let $A = \sum_{i=1}^{n} \sigma_i u_i v_i^T$. Then for any $1 \leq k \leq n$,

$$\min_{\operatorname{rank}(B)=k} \|A - B\| = \sigma_{k+1}.$$

The minimum is attained at $B = \sum_{i=1}^{k} \sigma_i u_i v_i^T$.

Proof. If $B \in \mathbb{R}^{m\times n}$ has rank k, $\dim \ker(B) = n - k$. Then the min-max theorem (Corollary 4.1.7) for $k + 1$ instead of k gives

$$\|A - b\| \geq \max_{x\in S(E)} \|(A - B)x\|_2 = \max_{x\in S(E)} \|Ax\|_2 \geq \sigma_{k+1}.$$

In the opposite direction, setting $B = \sum_{i=1}^{k} \sigma_i u_i v_i^T$ gives $A - b = \sum_{i=k+1}^{n} \sigma_i u_I v_i^T$. The maximal singular value of this matrix σ_{k+1}, which is the same as its operator norm by Lemma 4.1.11. □

4.1.6 Perturbation Theory

We can also study how eigenvalues/eigenvectors change under matrix perturbations:

Lemma 4.1.14 (Weyl inequality)**.** The k-th largest eigenvalue of symmetric matrices A, B satisfy

$$|\lambda_k(A) - \lambda_k(B)| \leq \|A - B\|.$$

Similarly, the k-th largest singular values of general rectangular matrices satisfy

$$|\sigma_k(A) - \sigma_k(B)| \leq \|A - B\|.$$

A similar result holds for eigenvectors, however we have to track the same eigenvector before and after the perturbation. If the eigenvalues are too close, a small perturbation can swap them, leading to huge error since their eigenvectors are orthogonal and far apart.

Theorem 4.1.15 (Davis-Kahan inequality)**.** Consider two symmetric matrices A, B with spectral decompositions

$$A = \sum_{i=1}^{n} \lambda_i u_i u_i^T, \;\; B = \sum_{i=1}^{n} \mu_i v_i v_i^T,$$

where the eigenvalues are weakly decreasing. Assume the the k-th largest eigenvalue of A is δ-seperated from the rest:

$$\min_{i \neq k} |\lambda_k - \lambda_i| = \delta > 0.$$

Then the angle between the eigenvectors u_k and v_k satisfies

$$\sin \angle u_k, v_k \leq \frac{2\|A - B\|}{\delta}.$$

The theorem above can be derived via a stronger result of Davis-Kahan focusing on spectral projections - the orthogonal projections onto the span of some subset of eigenvectors:

Lemma 4.1.16 (Davis-Kahan inequality for spectral projections)**.** Consider A, B as in Theorem 4.1.15. Let I, J be two δ-seperated subsets of $\mathbb{R}$, with I being an interval. Then the spectral projections

$$P = \sum_{i:\lambda_i \in I} u_i u_i^T \text{ and } Q = \sum_{j:\lambda_j \in J} v_j v_j^T \text{ satisfy } \|QP\| \leq \frac{\|A - B\|}{\delta}.$$

Proof. WLOG, assume I is finite and closed. Adding the same multiple of Identity to A and B, we can center I as $[-r, r]$, so that $|\lambda_i| \leq r$ for $i \in I$ and $|\mu_j| \geq r + \delta$ for $\mu_j \in J$. The idea is to see how P and Q interact through $H := B - A$:

$$\|H\| \geq \|QHP\| = \|QBP - QAP\| \geq \|QBP\| - \|QAP\|.$$

The spectral projection A commutes with B, hence

$$\|QBP\| \geq \|BQP\| \geq (r + \delta)\|QP\|.$$

To see the last inequality, the image of Q is spanned by orthogonal vectors v_j with $|\mu_j| \geq r + \delta$. The matrix B maps each such vector v_j to $\mu_j v_j$, hence scaling it by at least $r + \delta$. Thus B expands the norm of any vector in the image of Q by at least $r + \delta$ so

$$\|BQPx\|_2 \geq (r + \delta)\|QPx\|_2 \text{ for any } x.$$

Taking the supremum over all unit vectors gives the result with the operator norm.
Also, $AP = PAP = \sum_{i:\lambda_i \in I} \lambda_i u_i u_i^T$ so

$$\|QAP\| = \|QPAP\| \leq \|QP\| \cdot \|AP\| \leq r\|AP\|,$$

because $\|AP\| = \max_{i:\lambda_i \in I} |\lambda_i| \leq r$. Putting the two bounds together we get

$$\|H\| = \|B - A\| \geq \delta\|QP\|,$$

which completes the proof. □

Proof for Theorem 4.1.15. Since the LHS is a trig angle, we can assume that $\varepsilon := \|A - B\| \leq \delta/2$ or else the inequality holds trivially. By Weyl inequality (Lemma 4.1.14), $|\lambda_j - \mu_j| \leq \varepsilon$ for each j hence

$$\min_{j:j \neq k} |\lambda_k - \mu_k| \geq \min_{j:j \neq k} |\lambda_k - \lambda_j| - \varepsilon = \delta - \varepsilon \geq \delta/2.$$

Apply Lemma 4.1.16 for the $\delta/2$-seperated subsets $I = \{\lambda_k\}$ and $J = \{\mu_j : j \neq k\}$ to get $\|QP\| \leq 2\varepsilon/\delta$. Since P and $I_n - Q$ are the orthogonal projections on the directions of u_k and v_k respectively,

$$\|QP\| = \max_{\|x\|=1} \|QPx\|_2 = \|Qu_k\|_2 = \sin \angle(u_k, v_k).$$

Combining this with the inequality on $\|QP\|$ above completes the proof. □

4.1.7 Isometries

The singular values of a matrix A satisfy (by the min-max theorem)

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

The extreme singular values set the limits on how the linear map A distorts space.
A matris is an isometry if

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

Notice that A need not be a square matrix. T
For $A \in \mathbb{R}^{m\times n}$ with $m \ge n$, the following are equivalent:

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

(b) A is an isometry,

(c) All singular values of A are 1.

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

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

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

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

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

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

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

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

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

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

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

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

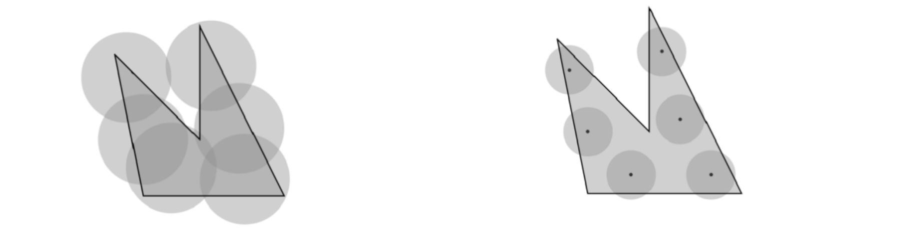
4.2 Nets, Covering, and Packing

The ε-net argument is useful for analysis of random matrices. It is also connected to ideas like covering, packing, entropy, volume, and coding.

Definition 4.2.1. Let (T,d) be a metric space. Consider $K \subset T$ and $\varepsilon > 0$. A subset $\mathcal{N} \subset T$ is called an ε-net of K is every point in K is within distance ε of some point in $\mathcal{N}$, i.e.

$$\forall x \in K \exists x_0 \in \mathcal{N} : \ d(x, x_0) \le \varepsilon.$$

Equivalently, $\mathcal{N}$ is an ε-net of K if the balls of radius ε centered at points in $\mathcal{N}$ cover K, like in the figure below:



(a) This covering of a polygon K by six ε-balls shows that $\mathcal{N}(K,\varepsilon) \leq 6$.

(b) $\mathcal{P}(K,\varepsilon) \geq 6$ means that there exist six ε-separated points in K; the $\varepsilon/2$-balls centered at these points are disjoint.

Figure 4.1 Covering and packing

Definition 4.2.2. The smallest cardinality of an ε-net of K is called the covering number of K, and is denoted $\mathcal{N}(K,d,\varepsilon)$.

Remark 4.2.3 (Compactness)**.** An important result in real analysis says that a subset K of a complete metric space (T,d) is precompact (i.e. the closure of K is compact) if and only if

$$N(K,d,\varepsilon) < \infty \text{ for every } \varepsilon > 0.$$

We can think about the covering numbers as a quantitative measure of how compact K is.

Definition 4.2.4. A subset $\mathcal{N}$ of a metric space (T,d) is ε-seperated if

$$d(x,y) > \varepsilon \text{ for any distinct points } x,y \in \mathcal{N}.$$

The largest possible cardinality of an ε-seperated subset of a given $K \subset T$ is called the packing number of K and is denoted $\mathcal{P}(K,d,\varepsilon)$.

Remark 4.2.5 (Packing balls into K)**.** If $\mathcal{N}$ is ε-seperated, the closed $\varepsilon/2$-balls centered at points in $\mathcal{N}$ are disjoint by the triangle inequality, hence we can always pack into K at least $\mathcal{P}(K,d,\varepsilon)$ disjoint $\varepsilon/2$-balls.

Lemma 4.2.6 (Nets from seperated sets)**.** Let $\mathcal{N}$ be a maximal ε-seperated subset of K, i.e. adding any new point to $\mathcal{N}$ destroys the seperation property. Then $\mathcal{N}$ is an ε-net of K.

Proof. Let $x \in K$. We want to show that there exists $x_0 \in \mathcal{N}$ such that $d(x,x_0) \leq \varepsilon$. If $x \in \mathcal{N}$, the conclusion is trivial by choosing $x_0 = x$. Suppose $x \notin \mathcal{N}$. The maximality assumption implies that $\mathcal{N} \cup \{x\}$ is not ε-seperated, meaning $d(x,x_0) \leq \varepsilon$ for some $\varepsilon \in \mathcal{N}$. □

Remark 4.2.7 (Constructing a net)**.** The lemma above (Lemma 4.2.6) gives an iterative algorithm to construct an ε-net for a given set K. Pick $x_1 \in K$ arbitrarily, then pick $x_2 \in K$ that is farther than ε from x_1, then pick x_3 that it is farther than ε from both x_1 and x_2, and so on. If K is compact, then the process will stop in a finite number of iterations!

Lemma 4.2.8 (Equivalence of covering and packing numbers)**.** For any set $K \subset T$ and $\varepsilon > 0$,

$$\mathcal{P}(K,d,2\varepsilon) \leq \mathcal{N}(K,d,\varepsilon) \leq \mathcal{P}(K,d,\varepsilon).$$

Proof. The upper bound follows from Lemma 4.2.6 because the packing number is exactly the number that makes $\mathcal{N}$ a maximal ε-seperated set.
For the lower bound, take any 2ε-seperated subset $\mathcal{P} = \{x_i\}$ in K and any ε-net $\mathcal{N} = \{y_j\}$ of K. By definition, each point x_i is in the ε-ball centered at some point y_j. Since any closed ε ball cannot contain two 2ε-seperated points, each ε-ball centered at y_j can contain at most one x_i. The pigeonhole principle gives $|\mathcal{P}| \leq |\mathcal{N}|$. Since $\mathcal{P}$ and $\mathcal{N}$ are arbitrary, the bound follows. □

4.2.1 Covering Numbers and Volume

This sections is about covers with $T = \mathbb{R}^n$ with the Eudlidean metric

$$d(x, y) = \|x - y\|_2.$$

Therefore, we can omit the metric when denoting the covering and packing numbers:

$$\mathcal{N}(K, \varepsilon) = \mathcal{N}(K, d, \varepsilon).$$

How do the covering numbers relate to the most classical measure, the volume of K in $\mathbb{R}^n$?

Definition 4.2.9 (Minkowski sum)**.** Let $A, B \subseteq \mathbb{R}^n$. The Minkowski sum is defined as

$$A + B := \{A + B : \ a \in A, b \in B\}.$$

Below is an example of the Minkowski sum of two sets on the plane:

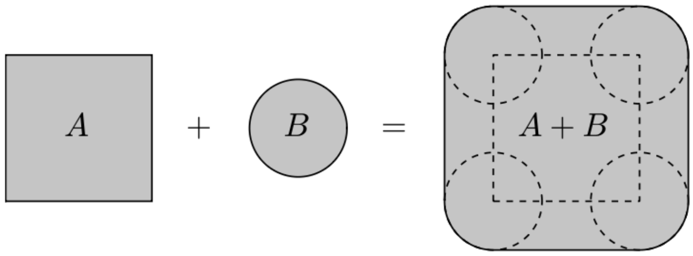


Figure 4.2 Minkowski sum of a square and a circle is a square with rounded corners.

Proposition 4.2.10 (Covering numbers and Volume)**.** Let $K \subset \mathbb{R}^n$ and $\varepsilon > 0$. Then

$$\frac{\mathrm{Vol}(K)}{\mathrm{Vol}(\varepsilon B_2^n)} \leq \mathcal{N}(K, \varepsilon) \leq \mathcal{P}(K, \varepsilon) \leq \frac{\mathrm{Vol}(K + (\varepsilon/2)B_2^n)}{\mathrm{Vol}((\varepsilon/2)B_2^n)},$$

where B_2^n denotes the unit ball in $\mathbb{R}^n$.

Proof. The middle inequality was already proven in Lemma 4.2.8, hence we focus on the left and right bounds.
(**Lower bound**) Let $N := \mathcal{N}(K, \varepsilon)$. Then K can be covered by N balls with radii ε. Comparing the volumes,

$$\mathrm{Vol}(K) \leq N \cdot \mathrm{Vol}(\varepsilon B_2^n),$$

which gives the lower bound.
(**Upper bound**) Let $N := \mathcal{P}(K, \varepsilon)$. Then we can find N disjoint closed $\varepsilon/2$-balls with centers $x_i \in K$. While these balls may not fit entirely in K (Figure 4-1), they do fit in a slightly inflated set, namely $K + (\varepsilon/2)B_2^n$ (Basically putting balls at the boundary of K). Comparing the volume gives

$$N \cdot \mathrm{Vol}((\varepsilon/2)B_2^n) \leq \mathrm{Vol}(K + (\varepsilon/2)B_2^n),$$

which completes the upper bound. □

An important consequence of the volumetric bound is that the covering (hence packing) numbers are typically *exponential* in the dimension n:

Corollary 4.2.11 (Covering numbers of the Euclidean ball)**.** The covering numbers of the unit Euclidean ball B_2^n satisfy the following for any $\varepsilon > 0$:

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

Proof. The lower bound immediately follows from Proposition 4.2.10, since the volumd in $\mathbb{R}^n$ scale as $\text{Vol}(\varepsilon B_2^n) = \varepsilon^n \text{Vol}(B_2^n)$.
The upper bound follows from Proposition 4.2.10 as well:

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

□

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

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

In the other case where $\varepsilon > 1$, one ε-ball covers the unit ball hence $\mathcal{N}(B_2^n, \varepsilon) = 1$.

Remark 4.2.12 (Volume of the ball)**.** The proof of Corollary 4.2.11 works with the volume of the Euclidean ball but never actually calculates it! We can compute the volume geometrically, probabilistically, and analytically (Exercises 4.27-4.29), and also extend this notion of volume to ℓ^p balls (Exercise 4.30).

Remark 4.2.13 (How to construct a net?)**.** We have an algorithm to construct nets already (Remark 4.2.7), but for the Euclidean ball, we can also use a scaled integer lattice (Exercise 4.31), or just use random points (Exercise 4.39).

We can also use covering/packing notions for other objects via volumetric arguments, here is another example:

Definition 4.2.14. The Hamming cube $\{0, 1\}^n$ consists of all binary strings of length n. To turn it into a metric space, we define the hamming distance as the number of bits where the strings x and y differ:

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

Proposition 4.2.15 (Covering and packing numbers of the Hamming cube)**.** The covering and packing numbers of the Hamming cube $K = \{0, 1\}^n$ satisfy the following for any integer $m \in \{0, \ldots, n\}$:

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

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

4.3 Application: Error Correcting Codes

4.4 Upper Bounds on Subgaussian Random Matrices

This section is mostly concerned with non-asymptotic theory of random matrices. Most of the questions here will be about the distributions of singular values, eigenvalues, and eigenvectors.
But before that, we need to know how ε-nets can help compute the operator norm of a matrix.

4.4.1 Computing the Norm on an εnet

To evaluate $\|A\|$, we need to control $\|Ax\|_2$ uniformly over the sphere S^{n-1}. However, we'll show that instead of the entire sphere, it is enough to control just an ε-net of the sphere (in Euclidean metric).

Lemma 4.4.1 (Operator norm on a net)**.** Let $A \in \mathbb{R}^{m\times n}$ and $\varepsilon \in (0, 1]$. Then for any ε-net $\mathcal{N}$ of the sphere S^{n-1}m we have

$$\sup_{x\in\mathcal{N}} \|Ax\|_2 \leq \|A\| \leq \frac{1}{1-\varepsilon} \sup_{x\in\mathcal{N}} \|Ax\|_2.$$

Proof. The lower bound is true since $\mathcal{N} \subset S^{n-1}$.
To prove the upper bound, fix a vector $x \in S^{n-1}$ for which $\|A\| = \|Ax\|_2$ and choose $x_0 \in \mathcal{N}$ such that $\|x - x_0\|_2 \leq \varepsilon$. By the definition of the operator norm, this implies

$$\|Ax - Ax_0\|_2 \leq \|A(x - x_0)\|_2 \leq \|A\| \|x - x_0\|_2 \leq \varepsilon\|A\|.$$

By the triangle inequality,

$$\|Ax_0\|_2 \geq \|Ax\|_2 - \|Ax - Ax_0\|_2 \geq \|A\|0\varepsilon\|A\| = (1 - \varepsilon)\|A\|.$$

Dividing by $1 - \varepsilon$ gives the result. □

There is alsoa version for quadratic forms from the way the operator norm is characterized. Since

$$\|A\| = \max_{x\in S^{n-1}, y\in S^{m-1}} |\langle Ax, y\rangle|,$$

we can take $x = y$ and use the spheres' corresponding nets:

Lemma 4.4.2 (Maximizing quadratic forms on a net)**.** Let $A \in \mathbb{R}^{m\times n}$ and $\varepsilon \in [0, 1)$. Then for any ε-net $\mathcal{N}$ of the sphere S^{n-1} and any ε-net $\mathcal{M}$ of the sphere S^{m-1},

$$\sup_{x\in\mathcal{N}, y\in\mathcal{M}} |\langle Ax, y\rangle| \leq \|A\| \leq \frac{1}{1-2\varepsilon} \sup_{x\in\mathcal{N}, y\in\mathcal{M}} |\langle Ax, y\rangle|.$$

Moreover, if $m = n$, A is symmetric, and $\mathcal{N} = \mathcal{M}$, we can take $x = y$.

Proof. There are two methods - one by modifying the proof of Lemma 4.4.1 (Exercise 4.36), and a different method using ε-net expansions (Exercise 4.34). □

4.4.2 The Norms of Subgaussian Random Matrices

Theorem 4.4.3. Let $A \in \mathbb{R}^{m\times n}$ be a random matrix with independent, mean zero, subgaussian entries A_{ij}. Then for any $t > 0$,

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

with probability at least $1 - 2\exp(-t^2)$. Here $K = \max_{i,j} \|A_{ij}\|_{\psi_2}$.

Proof. The proof is an example of an *ε-net argument.* We need to control $\langle Ax, y\rangle$ for all x, y in the unit sphere. To this end, we will discretize the sphere using a net (Approximation), establish a tight control of $\langle Ax, y\rangle$ for fixed vectors x, y from the net (Concentration), and finish by taking a union bound over all x, y in the net.
(**Approximation**). Choose $\varepsilon = 1/4$. Using the result from Corollary 4.2.11, we can find respective ε-nets $\mathcal{N}, \mathcal{M}$ of S^{n-1}, S^{m-1} with cardinalities

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

By Lemma 4.4.2, the norm of A can be bounded using these nets as

$$\|A\| \leq 2 \sup_{x\in\mathcal{N}, y\in\mathcal{M}} |\langle Ax, y\rangle| .$$

(**Concentration**). Fix $x \in \mathcal{N}, y \in \mathcal{M}$. The quadratic form

$$\langle Ax, y\rangle = \sum_{i=1}^{n}\sum_{j=1}^{m} A_{ij}x_i y_j$$

is a sum of independent, subgaussian random variables. By Proposition 2.7.1, the sum is subgaussian, and

$$\begin{aligned}
\|\langle Ax, y\rangle\|_{\psi_2}^2 &\leq C\sum_{i=1}^{n}\sum_{j=1}^{m}\|A_{ij}x_i y_j\|_{\psi_2}^2 \\
&\leq CK^2\sum_{i=1}^{n}\sum_{j=1}^{m} x_i^2 y_j^2 \\
&= CK^2\left(\sum_{i=1}^{n} x_i^2\right)\left(\sum_{j=1}^{m} y_j^2\right) \\
&= CK^2.
\end{aligned}$$

Using (i) from Proposition 2.6.6, we an restate this as a tail bound:

$$P(|\langle Ax, y\rangle| \geq u) \leq 2\exp\left(-cu^2/K^2\right), \ u \geq 0.$$

(**Union bound**) Next, we unfix x and y and use a union bound. The event

$$\max_{x\in\mathcal{N}, y\in\mathcal{M}} |\langle Ax, y\rangle| \geq u \implies \exists x \in \mathcal{N}, y \in \mathcal{M} \text{ such that } |\langle Ax, y\rangle| \geq u.$$

Therefore union bound gives

$$P\left(\max_{x\in\mathcal{N}, y\in\mathcal{M}} |\langle Ax, y\rangle| \geq u\right) \leq \sum_{x\in\mathcal{N}, y\in\mathcal{M}} P\left(|\langle Ax, y\rangle| \geq u\right).$$

Using the tail bound above and the estimates on $|\mathcal{N}|$ and $|\mathcal{M}|$, the probability is bounded above by

$$9^{n+m}\cdot 2\exp\left(-cu^2/K^2\right) \quad (*)$$

Choose

$$u = CK(\sqrt{n} + \sqrt{m} + t).$$

Then $u^2 \geq C^2K^2(n + m + t^2)$, and if the constnat C is chosen sufficiently large, the exponent in (*) is large enough, say $cu^2/K^2 \geq 3(n + m) + t^2$. Thus

$$P\left(\max_{x\in\mathcal{N}, y\in\mathcal{M}} |\langle Ax, y\rangle| \geq u\right) \leq 9^{n+m}\cdot 2\exp\left(-3(n + m) - t^2\right) \leq 2\exp\left(-t^2\right).$$

Combining with the approximation step,

$$P(\|A\| \geq 2u) \leq 2\exp\left(-t^2\right).$$

By the choice of u that we had, the proof is complete. □

Remark 4.4.4 (Expectation bounds). High-probability bounds like Theorem 4.4.3 can be usually turned into simpler but less informative expectatiom bounds using the integrated tail formula (Lemma 1.6.1). In Exercise 4.41, we get

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

Remark 4.4.5 (Optimality)**.** Theorem 4.4.3 is typically tight since the matrix's operator norm is bounded below by the Euclidean norm of any row/column of the matrix (Exercise 4.7). For example, if A has Rademacher entries, its columns have norm $\sqrt{m}$ and rows $\sqrt{n}$, so

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

with probability 1. There is also a fully general lower bound (Exercise 4.42).

Remark 4.4.6 (Relaxing independence)**.** The independence assumption in Theorem 4.4.3 can be relaxed: We just need the rows (or columns) of A to be independent, even with dependent entries (Exercise 4.43).

4.4.3 Symmetric Matrices

Theorem 4.4.3 also extends to symmetric matrices:

Corollary 4.4.7. Let $A \in \mathbb{R}^{n \times n}$ be a symmetric random matric with independent, mean zero, subgaussian entries A_{ij} on and above the diagonal. Then for any $t > 0$,

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

with probability at least $1 - 4\exp(-t^2)$. Here $K = \max_{i,j} \|A_{ij}\|_{\psi_2}$.

Proof. Split A into the upper triangular-part A^+ and the lower-triangular part A^-. The diagonal can go either way, so let's just assume it's in A^+. Then $A = A^+ + A^-$.
Applying Theorem 4.4.3 to A^+ and A^- gives (each with probability at least $1 - 4\exp(-t^2)$)

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

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

4.5 Application: Community Detection in Networks

4.6 Two-sided Bounds on Subgaussian Matrices

Theorem 4.4.3 gives an upper bound on the singular values of an $\mathbb{R}^{m \times n}$ subgaussian random matrix A, which says

$$\sigma_1 \leq \|A\| \leq C(\sqrt{m} + \sqrt{n})$$

with high probability.
In fact, there is a sharper two-sided bound on the **entire spectrum** of A:

$$\sqrt{m} - C\sqrt{n} \leq \sigma_i \leq \sqrt{m} + C\sqrt{n}.$$

In other words, the below shows that a tall random matrix $\frac{1}{\sqrt{m}}A$ with $m \gg n$ is an approximate isometry.

Theorem 4.6.1 (Name)**.** Let $A \in \mathbb{R}^{m \times n}$ be a random matrix with independent, mean zero, subgaussian, isotropic tows A_i. Then for any $t \geq 0$ we have

$$\sqrt{m} - CK^2(\sqrt{n} + t) \leq \sigma_n \leq \sigma_1 \leq \sqrt{m} + CK^2(\sqrt{n} + t)$$

with probability at least $1 - 2\exp(-t^2)$. Here $K = \max_i \|A_i\|_{\psi_2}$.

Proof. We'll prove a slightly stronger conclusion than the theorem statement, namely

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

Proving this implies proving the theorem (I haven't checked yet).
Again, we'll apply an ε-net argument, but use Bernstein inequality for the concentration step instead of Hoeffding which we did in Theorem 4.4.3.
(**Approximation**) Using Corollary 4.2.11, we can find an $\frac{1}{4}$-net $\mathcal{N}$ of the unit sphere S^{n-1} with cardinality

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

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

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

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

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

(**Concentration**) Fix $x \in \mathcal{N}$ and express $\|Ax\|_2^2$ as a sum of independent random variables:

$$\|Ax\|_2^2 = \sum_{i=1}^m \langle A_i, x\rangle^2 =: \sum_{i=1}^m X_i^2.$$

By assumption, the wors A_i are independent, isotropic, and subgaussian random vectors with $\|A_i\|_{\psi_2} \leq K$. Thus $X_i = \langle A_i, x\rangle$ are independent subgaussian random variables with $\mathbb{E}[X_i^2] = 1$ and $\|X_i\|_{\psi_2} \leq K$. This makes $X_i^2 - 1$ independent, mean zero, and subexponential random variables with

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

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

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

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

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

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

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

if we choose the constant C to be large enough. Then by the necessary condition in the approximation step, the proof is complete. □

Remark 4.6.2 (Expectation bound)**.** From Remark 4.4.4, we can convert high-probability bounds to expectation bounds. Exercise 4.41 yields the following form for Theorem 4.6.1:

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

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

4.7 Application: Covariance Estimation and Clustering