High-Dimensional Probability Scribe: Kiran Vodrahalli
Roman Vershynin, UC Irvine 05/29/2018 — 06/06/2018

Contents

1 Introduction

Most of the material for this course is from my HDP book. We will assume you don't know anything beyond standard probability. The first lecture will be really basic, not as high-dimensional as you would hope. Starting from second lecture, we will do some stuff. The first class is about concentration. The goal of this class, and also more generally of the lecture series, is to get you up to speed with modern probabilistic methods for statistical data analysis.

1.1 Concentration

Concentration is a big collection of tools that are available to control random variables around their means. What we want is for a random variable X, whose expectation is μ – we want to bound the deviation from μ. Typically we want to bound both left and right tails. That is a big goal of concentration methods. We can develop some ideas by asking a very simple question first.

Example 1.1. A sloppy example.

Let us toss a fair coin n times and ask the probability that we get at least $(3/4)n$ heads. In undergrad probability class, we look at the binomial distribution of the sum of indicators (if head, if tail). These are standard Bernoulli random variables. Then we reach $\mathbb{P}\{S_n > 3n/4\} = \mathbb{P}\left\{\frac{S_n - n/2}{\sqrt{n/4}} > \sqrt{n/4}\right\}$. This term is a sum of random variables, and by central limit theorem we can compare to a standard normal variable. Thus we look at $\mathbb{P}\left\{g > \sqrt{n/4}\right\}$ – then we can look at the table or use the fact that the tail of the standard normal distribution decays very fast. This probability gets bounded by $\frac{1}{2\pi}\exp(-n/8)$ using the lemma below.

Lemma 1.2. *Tail of standard normal.*
We have $g \sim \mathcal{N}(0,1)$. Then $\mathbb{P}\{g > t\} \leq \frac{1}{2\pi}e^{-t^2/2}$. You can see the integral of the density is bounded by the density.

Now this argument was sloppy – in the application of the central limit theorem. We were very sloppy about the error. The error in the central limit theorem is about $1/\sqrt{n}$. This error does not decay exponentially fast, and it cannot – you can take simple example of binomial distribution; it will not. So this is far too optimistic. We were far too sloppy here, but we cannot repair it in this argument, since we have to add the error $1/\sqrt{n}$, which kills the exponential decay.

1.2 Hoeffding

So the central limit does not help to get exponential decay. We need to develop an alternative approach – this is a complementary, non-asymptotic treatment of the central limit theorem, if you will. How do we correct?

Theorem 1.3. *Hoeffding's inequality.*
Let X_i be symmetric independent Bernoulli variables (± 1 with equal probability). We claim that $\mathbb{P}\{\sum_{i=1}^n a_i X_i > t\} \leq \exp(-t^2/2\|a\|_2^2)$.

Proof. The proof is due to Bernstein in 1914. It's a very cool and simple idea. We will treat the tail by looking at Chebyshev's inequality. That won't give us e^{-t^2} – that will only give us reciprocal of the linear; not exponential. To recreate e^{-t^2}, we will just exponentiate both sides before applying Markov. We will also pre-multiply both sides by a postiive parameter λ, a handle that we will optimize later.

$$\begin{aligned}\mathbb{P}\left\{\sum_i a_i X_i > t\right\} &= \mathbb{P}\left\{\exp(\lambda \sum_i a_i X_i) > \exp(\lambda t)\right\} \\ &\leq \mathbb{E}\left[e^{-\lambda t}\exp(\lambda \sum_i a_i X_i)\right]\end{aligned} \tag{1}$$

This is just an application of Markov's inequality; and now we are left with the moment-generating-function (MGF). MGFs are good! We can use independence of the X_i to break up the product:

$$\begin{aligned} &= e^{-\lambda t}\prod_{i=1}^{n}\mathbb{E}\left[e^{-\lambda a_i X_i}\right] \\ \mathbb{E}\left[\exp(\lambda a_i X_i)\right] &= \frac{1}{2}(\exp(\lambda a_i) + \exp(-\lambda a_i)) = \cosh(\lambda a_i) \end{aligned} \tag{2}$$

This guy is called the hyperbolic cosine. It's an even function, and is bounded by e^x on the positive reals and e^{-x} on the negative reals. For any x, it holds that $\cosh(x) \leq e^{-x^2/2}$. Therefore, we get

$$\leq e^{\lambda^2 a_i^2/2} \tag{3}$$

Thus the probability we started to bound above is bounded by

$$\begin{aligned} &\leq e^{-\lambda t}\prod_{i=1}^{n}\exp(\lambda^2 a_i^2/2) \\ &= \exp(-\lambda t + \frac{\lambda^2}{2}\sum_{i=1}^{n} a_i^2) = \exp(-\lambda t + \frac{\lambda^2}{2}\|a\|_2^2) \end{aligned} \tag{4}$$

Then we can optimize for $\lambda > 0$. Here, the minimum is where $-t + \lambda\|a\|_2^2 = 0$, so $\lambda = \frac{t}{\|a\|_2^2}$. This yields the desired result. □

Example 1.4. Application of Hoeffding.
Now we can get the probability of getting at least $3n/4$ heads after flipping coin. You have to do the correct scaling of the inequality; this gives you a bound of $\exp(-n/8)$. So, what we guessed before is true! But not because of the central limit theorem. This is a rigorous and non-asymptotic form of central limit theorem; which does not involve the error of the central limit theorem, which can be harmful.

Remark 1.5. We can get both tails by paying a multiplicative factor of 2: $\mathbb{P}\left\{|\sum_i a_i X_i| > t\right\} \leq 2\exp(-t^2/2\|a\|_2^2)$. This follows from union bound.

Remark 1.6. What about lower bounding the probability? Well, it'll be 0 – so you can't here. (In general, this idea is anti-concentration).

Remark 1.7. If we use cosh directly, do we get something better? Maybe locally you can have something better or a better constant – globally you can't, because it must converge to a Gaussian tail. Kind of strange, because this cosh inequality is kind of sloppy. So for large n you can't expect anything better, for small n perhaps you can. What really is going on in the argument is that we are locally approximating cosh. Then, this estimate is not so sloppy – it is actually tight.

1.3 Sub-gaussian random variables

Now, for what random variables can we have an inequality like Hoeffding? Definitely for ±1. But the proof suggests more is possible. We only used it to calculate the MGF, but we bounded it anyways. So we didn't really need to compute it exactly. So what's the largest class of random variables we can calculate something like this for?

Consider that the sum is only one term. For Hoeffding to hold, you must have a "sub-gaussian" tail for this term. Most of the time for modern research problems, people are happy if they prove something for the class of sub-gaussian random variables. We will now describe sub-gaussian distributions more rigorously.

Lemma 1.8. *Sub-gaussian tail behavior.*
Suppose X is a random variable. Then the following properties are equivalent:

(a) $\mathbb{P}\{|x| > t\} \leq 2\exp(-t^2/K^2)$ *for all* $t \geq 0$. *(tails)*

(b) $(\mathbb{E}[|X|^p])^{1/p} = \|X\|_{L^p} \leq K\sqrt{p}$ *for all* $p \geq 1$. *(moments)*

(c) $\mathbb{E}[\exp(\lambda^2 X^2)] \leq \exp(K^2\lambda^2)$ *for all* $|\lambda| \leq 1/K$. *(MGF of X^2)*

(d) If X is mean 0, $\mathbb{E}[\exp(\lambda X)] \leq \exp(K^2\lambda^2)$ *for all* λ. *(MGF of X)*

Note that Gaussians satisfy all of these properties. Now what is the role of K? These statements are equivalent, up to different constant factors of K. More formally, we should write K_1, K_2, K_3, K_4 for each of the statements respectively. The constants K_i differ from each other by an absolute constant factor. That depends on which direction you want to go (e.g., $K_3 \to K_1$ versus $K_1 \to K_4$).

Proof. For an example, let us prove (1) $\implies$ (2). Suppose you know the tails, how do you get the moments? We have $\mathbb{E}[Z] = \int_0^\infty \mathbb{P}\{Z > t\}\, dt$ for all $Z \geq 0$. This is by integration by parts. Now let's prove (1) $\implies$ (2).

$$\begin{aligned}\mathbb{E}[|X|^p] &= \int_0^\infty \mathbb{P}\{|X|^p > t\}\, dt \\ &\leq \text{... use (1) to finish the proof.}\end{aligned} \tag{5}$$

We can now do (2) $\implies$ (3). I don't have an easy proof of this. Suppose we know the moments grow in a perscribed way, then can we compute the MGF? When we don't know what to do, we just do the Taylor expansion of the MGF. We have

$$\begin{aligned}\mathbb{E}\left[\exp(\lambda^2 X^2)\right] &= 1 + \mathbb{E}\left[\sum_{p=1}^\infty \frac{(\lambda^2 X^2)^p}{p!}\right] \\ &= 1 + \sum_{p=1}^\infty \frac{\lambda^{2p}\mathbb{E}[X^{2p}]}{p!} \\ &\leq 1 + \sum_{p=1}^\infty \frac{(2\lambda^2 p)^p}{(p/e)^p}\end{aligned} \tag{6}$$

assuming $K_2 = 1$ and using $p! \geq (p/e)^p$ for anything. Then we get

$$\begin{aligned} &= 1 + \sum_{p=1}^{\infty} (2e\lambda^2)^p = \frac{1}{1-2e\lambda^2} \\ &\leq e^{4e\lambda^2} \end{aligned} \tag{7}$$

for all $|\lambda| \leq 1/2\sqrt{e}$. This holds only for these λ. The MGF is infinite outside for Gaussian case. That's my complicated way for proving (2) $\implies$ (3) – if you have a better way without Taylor series, let me know! □

Definition 1.9. Sub-gaussian and sub-gaussian norm $\|\cdot\|_{\psi_2}$.
We call random variables which satisfy these properties **sub-gaussian**. The smallest of $K_1, \cdots, K_4$ will be defined as the sub-gaussian norm for a random variable X. We have

$$\|X\|_{\psi_2} \leq \inf\left\{t > 0 : \mathbb{E}\left[\exp(X^2/t^2)\right] \leq 2\right\} \tag{8}$$

The standard choice is using the third definition. This is a standard definition in functional analysis called the Orlicz norm. All constants K_1, K_2, K_4 are $\sim \|X\|_{\psi_2}$ (up to an absolute constant).

Remark 1.10. Centering (e.g., mean 0) is only needed for property 4 – for the others, it is not necessary.

Sub-gaussian distributions are very good, and satisfy many properties that Gaussian distributions satisfy, yet they are much more general.

Example 1.11. Examples and counter-examples of sub-gaussian r.v.s.
Gaussians, Bernoullis, bounded random variables are all sub-Gaussian. Exponential random variable, χ^2, and Cauchy are NOT. These are called heavy-tailed.

Now, what can we say about the class of sub-Gaussian random variables? Recall that a sum of independent Gaussian random variables is Gaussian. The same holds for sub-Gaussian random variables.

Lemma 1.12. *Suppose we have independent $X_1, \cdots, X_n$ are mean 0 sub-Gaussian random variables. Then $\sum_{i=1}^n X_i$ is also mean 0 and sub-Gaussian random variable. More quantitatively,*

$$\|\sum_{i=1}^n X_i\|_{\psi_2} \leq C * \sum_{i=1}^n \|X_i\|_{\psi_2}^2 \tag{9}$$

for an absolute constant C.

This is much like the standard formula for the variance for independent random variables. This is a more advanced version of it. The variance controls the second moment of X – now, we're controlling *all the moments*, not just the second moment. The $\|\cdot\|_{\psi_2}$ is a stronger control of the size, and it satisfies the same thing.

Proof. We will use property 4 of the sub-gaussian properties. We are claiming that the sum is $\mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^n X_i\right)\right] = \prod_{i=1}^n \mathbb{E}\left[\exp(\lambda X_i)\right]$ by independence. We can then use property 4 to get this is bounded by

$$\leq \prod_{i=1}^n \exp\left(C\lambda^2 \|X_i\|_{\psi_2^2}\right) \tag{10}$$

where the absolute constant C comes from the fact that the sub-gaussian norm is defined from property 3, and K_3 and K_4 differ by some absolute constant C. This yields

$$= \exp\left(C\lambda^2 \sum_{i=1}^N \|X_i\|_{\psi_2}^2\right) \tag{11}$$

Now by definition of sub-gaussian (property 4), we have

$$\|\sum_{i=1}^n X_i\|_{\psi_2}^2 \leq C \sum \|X_i\|_{\psi_2}^2 \tag{12}$$

□

Now, this is our extended form of the variance equality for independent r.v.s (sum of variances equals variance of sum). But this immediately happens to imply the general Hoeffding's inequality!

Corollary 1.13. *(General Hoeffding's inequality).*
If $X_1, \cdots, X_n$ are independent, mean 0, sub-gaussian, then

$$\mathbb{P}\left\{|\sum_{i=1}^n X_i| \geq t\right\} \leq 2\exp\left(-\frac{Ct^2}{\sum_{i=1}^n \|X_i\|_{\psi_2}^2}\right) \tag{13}$$

where C is an absolute constant.

1.4 Sub-exponential distributions

Now, this doesn't really work if you look at χ^2 – but something like it should. How can we widen the set of distributions? There's a more general class of distributions called sub-exponential. What happens if we take a normal and then square it? Let's look:

$$\mathbb{P}\left\{g^2 > t\right\} = \mathbb{P}\left\{g > \sqrt{t}\right\} \leq e^{-(\sqrt{t})^2/2} = e^{-t/2} \tag{14}$$

So the square of a Gaussian looks approximately exponential. This is heavier tailed. But still you can hope that you can re-do everything, and it is true.

Lemma 1.14. *Sub-exponential equivalencies.*
The following are equivalent:

(a) $\mathbb{P}\{|X| > t\} \leq 2\exp(-t/K)$ *for all* $t \geq 0$.

(b) $(\mathbb{E}[|X|^p])^{1/p} \leq Kp$ *for all* $p \geq 1$.

(c) $\mathbb{E}[\exp(\lambda|X|)] \leq \exp(K\lambda)$ *for all* $|\lambda| \leq 1/K$.

(d) If $\mathbb{E}[X] = 0$*, then* $\mathbb{E}[\exp(\lambda X)] \leq \exp(K^2\lambda^2)$*, for all* $|\lambda| \leq 1/K$*. Note the additional restriction on* λ*! This will end up being the reason for the mixing between the sub-Gaussian and sub-exponential parts of the bound in Bernstein's inequality. The dependency on* λ^2 *doesn't change here – there is a quadratic cup locally regardless of whether you are sub-gaussian or sub-exponential.*

Definition 1.15. Sub-exponential r.v. and norm $\|\cdot\|_{\psi_1}$.
We call such random variables **sub-exponential**. The sub-exponential norm is defined

$$\|X\|_{\psi_1} = \inf\{t > 0 : \mathbb{E}[\exp(|X|/t)] \leq 2\} \tag{15}$$

Again, we use definition 3.

Example 1.16. $\mathcal{N}(0,1)^2$ is sub-exponential, as is the square of every sub-gaussian – this is a triviality by the tail definition (property 1). Poisson is sub-exponential (its tail is $t^{-t} < e^{-t}$), but not sub-Gaussian. Chi-square is sub-exponential. But not Cauchy, as usual. Cauchy is just bad in general.

1.5 Bernstein's inequality

Now we would like to show that something like Hoeffding holds for a larger class – in this case, sub-exponential. Here is the theorem:

Theorem 1.17. *Bernstein's inequality.*
Let X_i*'s be independent, mean* 0*, sub-exponential random variables. Then we can bound the tail*

$$\mathbb{P}\left\{|\sum_{i=1}^n X_i| > t\right\} \leq 2\exp\left(-\min\left\{\frac{t^2}{\sum_{i=1}^n \|X_i\|_{\psi_1}^2}, \frac{t}{\max_i \|X_i\|_{\psi_1}}\right\}\right) \tag{16}$$

Essentially it's a mixture of two tails: A sub-gaussian tail and a sub-exponential tail. The first term comes from Hoeffding/CLT. The second term is contributed by a single term in the sum – is there any heavy tail?

Proof. Exercise: prove Bernstein's inequality just the way we proved Hoeffding's inequality. Exponentiate both sides, apply Markov, find the MGF of each term. For sub-exponential, we'll use the bound from property 4. Now we only have a bound for λ in some neighborhood – we optimize over λ in the neighborhood $|\lambda| \leq 1/K_4$ – same as before, minimize a quadratic, except we have to stay in the neighborhood! That's the reason for the second term in the theorem, just maintaining the boundary. □

Remark 1.18. There must be a sub-Gaussian part of the bound if there exists a second moment – this is by central limit theorem. We can also imagine as follows: Think of central limit theorem, taking $n \to \infty$, larger and larger — the boundary in which we can make the sub-gaussian approximation gets wider and wider as $n \to \infty$, but there is always some correction at the tails which is sub-exponential.

Remark 1.19. Ledoux and Talagrand work with heavy tails a lot. If you have ten moments and nothing else. They don't get sub-Gaussian behavior though.

2 Concentration of norm of a random variable

Now we will start to see some high-dimensional phenomena. Suppose you want to do a search of high-dimensional space. Suppose you have a cube of unit size, and also suppose you have your procedure that works in n dimensions. How does your procedure scale, if you have a cube of size 2? The volume of the original cube is 1. As we scale by 2, the volume of the doubled cube is 2^n!! Much much larger than 1 — it scales exponentially. That is the basis of many difficulties in statistical inference. You have to do something smart to overcome this curse. First thing we will consider is take a random vector in high-dimensional space. We would like to say with exponentially high probability, that vector will be on the sphere, which is kind of non-intuitive. But it will mostly be on the sphere – e.g., concentration of the norm.

Theorem 2.1. *Concentration of the norm.*
Let X be a random vector with independent, sub-gaussian coordinates in $\mathbb{R}^n$. We want to scale things properly, so the variance of each coordinate is 1. Then:

$$\mathbb{E}\left[\|X\|_2^2\right] = \mathbb{E}\left[\sum_{i=1}^n X_i^2\right] = n \tag{17}$$

So we should expect the norm to be approximately $\sqrt{n}$. We will show that the random vector will be with very high probability be in the thin shell. We are claiming this random variable is sub-gaussian:

$$\|\|X\|_2 - \sqrt{n}\|_{\psi_2} \leq C \max_i \|X_i\|_{\psi_2}^2 \tag{18}$$

Proof. This is basically a re-formulation of Bernstein's inequality. It's much easier to work with norm squared: Let's look at $\|X\|_2^2 - n$, and take square roots in the end. Note that X_i^2 is sub-exponential — in particular, $\|X_i^2 - 1\|_{\psi_1} \leq \|X_i\|_{\psi_2}^2$. We have

$$\begin{aligned} \|X\|_2^2 - n &= \sum_{i=1}^n (X_i^2 - 1) \\ \mathbb{P}\left\{|\|X\|_2^2 - n| > t\right\} &\leq 2\text{exp}\left(-C * \min\{t^2/n, t\}\right) \end{aligned} \tag{19}$$

applying Bernstein's inequality. So now we have a concentration inequality for the squares. Now we have to "take square roots", since we really want deviation for the norm. This is more like a delta-method. This will be an exercise for you guys, and you will end up with something like

$$\mathbb{P}\left\{|\|X\|_2 - \sqrt{n}| \geq u\right\} \leq 2\text{exp}\left(-cu^2\right) \tag{20}$$

for all u. And that is it! This is exactly the definition of sub-gaussian tail; if you do it carefully, the constant will be the maximum (e.g., use Bernstein tail properly — the part where it gets cut off). □

Remark 2.2. Now how do we reason that everything is on the boundary? If $u = 1000$, we get

$$\mathbb{P}\left\{|\|X\|_2 - \sqrt{n}| \geq 1000\right\} \leq 0.01$$

So with high probability, the random variable will show up in the radius of $\sqrt{n}$: This is a concentration of the norm.

Intuition: You can think of it in the following way: The volume of the inner part will essentially deflate — there will almost be no volume inside, and the random variable will avoid the inner part. You can think of χ^2 distribution. If X_i's are Gaussian, then X_i^2 is the norm – we are stating the χ^2 distribution concentrates like this. Another piece of intuition you can give is that $\|X\|_2^2$ is a sum of independent random variables. The mean is $n \pm \mathcal{O}(\sqrt{n})$ – if you take the square root, you get $\|X\|_2 = \sqrt{n \pm \mathcal{O}(\sqrt{n})} = \sqrt{n} \pm \mathcal{O}(1)$, actually. You can therefore intuitively see what concentration of norm says. But it gives much more! Exponential, sharp decay. For that, there is no good intuition.

2.1 Grothendieck's Inequality and SDPs

Today, we will use high dimensional probability to prove Grothendieck's inequality and look at applications to semidefinite programming. This is a tool that's starting to be appreciated by machine learning and computer science communities, and it's slowly coming to statistics. This is a beautiful theorem of Grothendieck, which on the surface has nothing to do with probability.

Theorem 2.3. *Grothendieck's inequality.*
Consider an $m \times n$ real matrix a_{ij}, and assume the following holds: For any signs x_i's, y_j's (.e.g, in $\{\pm 1\}$); we have the quadratic form

$$\left|\sum_{i,j} a_{i,j} x_i y_j\right| \leq 1 \tag{21}$$

Then, the same happens if we vectorize this inequality – that is, we replace by x_i's, y_j's by vectors, and where multiplication becomes a dot product instead. Specifically, for any ℓ_2 unit *vectors $u_i, v_j \in \mathbb{R}^n$, we have*

$$\left|\sum_{i,j} a_{i,j}\langle u_i, v_j\rangle\right| \leq K \tag{22}$$

where K is an absolute constant that does not depend on anything. The best known value is $K \approx 1.783$. We know that 1.4 is a lower bound (1 is a trivial lower bound). We also know that the best known upper bound is not the best possible — this problem is still open.

In the condition, we take maximal quadratic forms – but not quite, only over signs – this is related to the cut norm in computer science. Before we prove this statement, let's make one remark.

Remark 2.4. Instead of insisting that $x_i, y_j \in \{\pm 1\}$, we can require more: That the condition holds for any numbers which are bounded by 1. Since this is a convex function, and the maxima are attained on the vertices of the discrete cube (± 1), we can extend it as follows: If $\left|\sum_{i,j} a_{i,j}x_iy_j\right| \leq \max_i |x_i| \max_j |y_j|$ for any x_i, y_j, then

$$\left|\sum_{i,j} a_{i,j}\langle u_i, v_j\rangle\right| \&leqK \max_i \|u_i\|_2 \max_j \|v_j\|_2 \tag{23}$$

for all u_i, v_j.

Remark 2.5. Since the conclusion holds for finite dimension spaces, and the dimension doesn't play a role in the bound, it actually turns out to be true for vectors in Hilbert spaces – we will prove this.

Proof. First of all, the trivial reduction is that there exists $K := K(A)$, perhaps depending on the matrix, for which Grothendieck holds. This is trivial. Let K be the smallest such number. We will now turn the statement into a random one.

Let $g \sim \mathcal{N}(0, I_n)$. Let $U_i := \langle g, u_i\rangle, V_j := \langle g, v_j\rangle$. For simplicity assume $\|u_i\|_2 = \|v_i\|_2 = 1$. These are also normal random variables. We know that $\mathbb{E}\left[U_iV_j\right] = \langle u_i, v_j\rangle$. $K = \sum_{i,j} a_{i,j}\langle u_i, v_j\rangle = \sum_{i,j} a_{i,j}\mathbb{E}\left[U_iV_j\right]$. We want to bound this quantity. By the assumption of the Grothendieck inequality, we can convert the vector valued Grothendieck into the single-dimensional Grothendieck. We're going to do a recursive proof — we are going to give some originally bad bound for K, and then use this bound further down the line to get an equation for K which depends on K, which we will then optimize. Note that we show K is finite because we have $\left|\sum_{i,j} a_{ij}\langle u_i, v_j\rangle\right| \leq \sum |a_{ij}| =: K$.

Thus, look at

$$\mathbb{E}\left[\sum_{i,j} a_{i,j}U_iV_j\right] \leq R^2 \tag{24}$$

where we're assuming $U_i, V_j \leq R$. However, this doesn't quite work – normal vectors are not bounded almost surely – we would get it automatically at this point! So what do you do if you want to bound it – truncate it! The rest is just a truncation procedure.

Let's introduce some level R, which we will optimize later. Then $U_i = U_i^- + U_i^+$, where $U_i^- = U_i \mathbf{1}\left\{|U_i| \leq R\right\}$ and $U_i^+ = U_i \mathbf{1}\left\{|U_i| > R\right\}$, and similarly for V_i^-, V_i^+. Then,

$$\begin{aligned} K &= \mathbb{E}\left[\sum_{i,j} a_{ij}\left(U_i^- + U_i^+\right)\left(V_i^- + V_i^+\right)\right] \\ &= S_1 + S_2 + S_3 + S_4 \end{aligned} \tag{25}$$

We have $S_1 = \mathbb{E}\left[\sum_{i,j} a_{ij} U_i^- V_j^-\right] \leq R^2$, by assumption of Grothendieck. So this one is easy. Let's do the second sum now: $S_2 = \mathbb{E}\left[\sum_{i,j} a_{ij} U_i^+ V_j^-\right]$ – we can't just use Grothendieck here. What we can do is apply Grothendieck on the random variables, which are actually functionals and belong to Hilbert space. We have $U_i^+, V_j^- \in L^2$ of random variables. The inner product is $\langle U, V\rangle_{L^2} = \mathbb{E}\left[UV\right]$. So we introduce a Hilbert space by defining this inner product norm.

Now we have $\|U_i^+\|_{L^2} = \mathbb{E}\left[\left(U_i^+\right)^2\right] = \mathbb{E}\left[(g^+)^2\right] = \mathbb{E}\left[g^2 \mathbf{1}|g| > R\right]$. As an exercise, prove that this value is $\leq 4/R^2$ – this is a crude bound. For V_i^-, we don't need much. We just say $\|V_i^-\|_{L^2} \leq \|V_i\|_{L^2} = 1$. Then for Hilbert space L^2, we get

$$\mathbb{E}\left[\sum_{i,j} a_{ij} U_i^+ V_j^-\right] = \sum_{i,j} a_{ij} \langle U_i^+, V_j^-\rangle_{L^2} \leq K \cdot \sqrt{\frac{4}{R^2}} \cdot 1 = \frac{2K}{R} \tag{26}$$

We bound the other cross-term the same way, and the fourth term is even simpler – since we truncate them a lot, they will be even smaller in L_2 norm. So S_3, S_4 are similar. As a result, putting everything together, we get

$$K \leq S_1 + S_2 + S_3 + S_4 \leq R^2 + \frac{6K}{R} \tag{27}$$

This holds for any R, so if we choose $R = 12$, for example, then we have $K \leq 144 + \frac{K}{2}$, and we get a constant bound for K. This is a nice probabilistic / functional analytic proof. □

Remark 2.6. To get the best bound, we can use the kernel trick. See the textbook, where I do it in full.

Remark 2.7. Grothendieck was one of the best known mathematicians in the 50s, 60s, 70s. Before algebraic geometry, he did functional analysis. He wrote a paper called tensor products in function spaces around 1954 in some obscure journal – he had a purely functional theoretic motivation. In 1968, someone noticed it and it became headlines of research in functional analysis, until it came to computer science in the 1980s.

2.2 The relation to semidefinite programming

The Grothendieck inequality can be used as a black-box in relaxing computationally hard problems to semidefinite programs. This is actually the link between NP-hard and not NP-hard problems. A linear program where you maximize $\max_x \langle a, x\rangle$ over $x \in$ polyhedron, or generally, a convex set. You maximize until you touch the tip, and that's a solution.

A semidefinite program is a subclass of convex programs – instead of x a vector, you have a positive-semi-definite matrix X: You maximize $\max_X \langle A, X\rangle$, $X \succeq 0$, $\langle B_i, x\rangle = b_i$. The inner product of two matrices is just $\langle A, X\rangle = \sum_{i,j} A_{ij}X_{ij} = \mathrm{Tr}(A^T X)$. This is convex. You have a linear functional over a convex set. It's not hard to check that PSD matrices form a convex set (in fact it is a cone). You intersect the cone with a linear subspace. This is called the *spectrahedron*. SDPs are kind of slow polynomial time, but are not NP-hard.

Suppose you want to solve a combinatorially problem which is NP-hard. In particular, say you want to maximize $\max \sum_{i,j} A_{ij}x_i x_j$, with $x_i = \pm 1$ for all i. This is hard and important. We will in a second reduce this to the SDP $\max \sum_{i,j} A_{ij}\langle X_i, X_j\rangle$ given that $\|X_i\|_2 = 1$ for all i. We will show that the solutions are approximately the same. The link will be Grothendieck's inequality. NP-hard corresponds to the assumption, and the vector-valued Grothendieck corresponds to the SDP relaxation.

Let G denote the Gram matrix, so that $G_{ij} = \langle X_i, X_j\rangle$. Thus we want to maximize $\langle A, G\rangle$, subject to $G_{ii} = 1$ for all i. Also note that G is always PSD — this it can be solved. Now we get the following theorem:

Theorem 2.8. *SDP from Grothendieck.*
We can relax the integer optimization problem (NP-hard problem) to an SDP; the error is smaller than Grothendieck's constant K, a constant multiplicative factor, which is about 1.783. *Recall we want this constant to be as small as possible.*

Remark 2.9. There is a form of Grothendieck where $x = y; u = v$. We can look at $\langle Ax, y\rangle = \langle Au, u\rangle - \langle Av, v\rangle$ where $u = \frac{1}{2}(x+y), v = \frac{1}{2}(x-y)$ — this is some polar identity. So it doesn't really matter whether you want to talk about different x, y or the same. Just do polarization as pre-processing.

But how do we get from the vectors output by the SDP to the signs? It is not trivial, but it is possible to do. It is not obvious. It's still an open problem for doing this in general. But there is a nice special case — max-cut for networks (Goemans-Williamson). Let A be the adjacency matrix of the graph in question. We can describe the cut with a ± 1 vector – they will encode the cut ($\{\pm 1\}^n$). How many edges are there? We look at $\frac{1}{2}\sum_{x_i=-x_j} A_{ij} = \frac{1}{4}\sum_{i,j} A_{i,j}(1 - x_i x_j)$. This is the same kind of problem as we discovered before.

Goemans and Williamson proposed randomized rounding, a scheme to get signs from vectors. It's super simple: They say, take the output vectors, and just project them onto a random line. For instance, x_3, x_4, x_5 will fall on one side, and the rest will fall on the other: Call one side $+1$, and the other side -1. So just project onto a number line – will it work? Indeed, it will. Formally, choose $g \sim \mathcal{N}(0, I_n)$ and define $x_i = \mathrm{sgn}\langle X_i, g\rangle \in \{\pm 1\}^n$.

Theorem 2.10. *The SDP + randomized rounding gives us an expected* 0.878 *approximation lower bound.*
Note that this is better than Grothendieck constant, and that will not show up in our calculations. The expected max cut will be at least this good.

We need some theorems:

Lemma 2.11. *(Grothendieck identity). For all u, v unit vectors, $\mathbb{E}\left[\operatorname{sgn}(g, u)\operatorname{sgn}(g, v)\right] = f(\langle u, v\rangle)$ by rotational invariance. In particular, this function is* $(2/\pi)\arcsin$ *(exercise).*

Unfortunately, arcsin is not linear, so we have to linearize it. Otherwise, we'd be done directly. We want $1 - \frac{2}{\pi}\arcsin(t) = \frac{2}{\pi}\arccos(t)$. We can linearize it by choosing the line which lower bounds the curve between 0 and 1 — this is $0.878(1 - t)$. This is where the linearization comes from.

Proof. Now we prove the theorem.

$$\begin{aligned}
\mathbb{E}\left[\mathrm{CUT}(G, x)\right] &= \frac{1}{4}\sum_{i,j=1}^{n} A_{ij}(1 - \mathbb{E}\left[x_i x_j\right]) \\
&= \frac{1}{4}\sum_{i,j=1}^{n} A_{ij}(1 - \mathbb{E}\left[\operatorname{sgn}\langle X_i, g\rangle\langle X_j, g\rangle\right]) \\
&= \frac{1}{4}\sum_{i,j=1}^{n} A_{ij}\left(1 - \frac{2}{\pi}\arcsin\langle X_i, X_j\rangle\right) \\
&\geq 0.878\frac{1}{4}\sum_{i,j=1}^{n} A_{ij}(1 - \langle X_i, X_j\rangle) = 0.878 \cdot \text{SDP-relaxation}
\end{aligned} \tag{28}$$

Note in the second to last step we used A_{ij} is non-negative. □

Remark 2.12. Note that when you solve the SDP, the number you get in the SDP is a bigger number (e.g., see the Grothendieck constant factor larger). However, the actual *value* of the max-cut resulting from rounding is smaller.

Remark 2.13. Grothendieck relation to networks and SDPs has not been sufficiently explored. There are some papers which use it for min-cut (even though it's a linear program – we want an additional constraint on sizes of communities), for community detection. The first paper was by Guedon and myself in 2013, the second paper is from Andrea Montanari and co-authors in 2016.

3 Covariance Estimation and Random Matrices

3.1 Background on Covering and Packing Arguments

These notions are very helpful in discretizing hard problems.

Definition 3.1. Epsilon Net.
Let T be a metric space (for instance, $T \in \mathbb{R}^n, \|\cdot\|_2$). Consider subset $K \subset T$. We want to construct an epsilon net which will capture the geometry. A subset N in K is called an *epsilon net* if for all $x \in K$, there exists a point $y \in N$ which has $d(x-y) \le \epsilon$. That is, the balls of radius ϵ at each point in the net N will cover K completely.

Definition 3.2. Covering number.
The covering number $\mathcal{N}(K, d, \epsilon)$ is the minimum cardinality of an epsilon net for K with metric d and scale ϵ.

A dual notion to covering is packing:

Definition 3.3. Packing number.
Consider a subset $N \subset T$ is called ϵ-separated if the points in it are at least ϵ apart: $d(x, y) > \epsilon$ for all $x \neq y$, $x, y \in N$. We define $\mathcal{P}(K, d, \epsilon)$ as the maximal cardinality of an ϵ-separated set. If you draw balls of radius $\epsilon/2$, they will not intersect (triangle inequality).

Packing numbers and covering numbers are almost equivalent to each other.

Lemma 3.4. *$\mathcal{P}(K, d, 2\epsilon) \le \mathcal{N}(K, d, \epsilon) \le \mathcal{P}(K, d, \epsilon)$.*

Proof. We can easily prove the upper bound. Suppose we have a maximally ϵ-separated set $\mathcal{P}$. Then we want to conclude that we have a good covering. We want to show it's automatically an ϵ-net. What happens if it's not an ϵ-net? Then we can find a point such that there is no point in the net which is close to it. Thus we must have a point which is at least ϵ away from every point in the set. If that were true, we could add it to our packing, it'll be a larger packing, and it'll still be ϵ-separated — that's impossible by maximality of the packing. □

Now how to construct an ϵ-net algorithmically? Pick a point, exclude everything in the ϵ-neighborhood, then pick another point. Keep going — that way everything will be ϵ-separated. If you keep going, you'll construct a maximum ϵ-separated set, and by the previous lemma you'll get an ϵ-net. If the set is compact, there will be a finite cover, so you'll be fine. We are working with compact sets.

Now, what about specific volumes and fix an ϵ — how large is the ϵ-net for these objects? That is our next bound. These volumetric bounds are usually easy and efficient, though there can be better bounds.

Lemma 3.5. *Covering number and volume (volumetric bounds).*
Suppose we have a set $K \subset \mathbb{R}^n$. Then

$$\frac{\mathrm{vol}(K)}{\mathrm{vol}(\epsilon B)} \le \mathcal{N}(K, \epsilon) \le \mathcal{P}(K, \epsilon) \le \frac{\mathrm{vol}(K + \frac{\epsilon}{2}B)}{\mathrm{vol}(\frac{\epsilon}{2}B)} \tag{29}$$

where $A + B = \{a + b : a \in A, b \in B\}$, which is known as a Minkowski sum.

Proof. First we prove the lower bound. If we have a covering, then K is completely inside the union of the balls of the covering. So $\text{vol}(K) \leq \#$ balls $\cdot \text{vol}(\epsilon B)$. The upper bound is similar, except for slightly inflated balls. We are able to pack inside K a lot of disjoint balls, with one exception — they are protruding a bit from K. So we can say inside $K + \frac{\epsilon}{2}B$, there are at least $\mathcal{P}(K, \epsilon)$ balls. □

Corollary 3.6. *For the unit ball, you need*

$$\left(\frac{1}{\epsilon}\right)^n \leq \mathcal{N}(B, \epsilon) \leq \left(\frac{3}{\epsilon}\right)^n \tag{30}$$

for all $\epsilon < 1$.

Proof. For the lower bound, just apply previous lemma directly, and use the fact that volume scales with radius as r^n, and cancel out the volumes.

For the upper bound, you get a $(1 + \frac{\epsilon}{2})$-radius ball — this gives you $\frac{(1+\frac{\epsilon}{2})^n}{(\epsilon/2)^n} = (\frac{2}{\epsilon} + 1)^n \leq (3/\epsilon)^n$. □

The takeaway from these estimates is that usually covering and packing numbers are exponential in dimension — this is kind of bad news, the curse of dimensionality. But on the other side, the probability bounds are exponentially *good* in dimension —- so sometimes, these things cancel each other out and you can do good things. This is the blessing of dimensionality in probability.

Next time — chapters $5, 6$ of the book.

4 MISSED (Covariance and random matrices?: FILL IN WITH RISHABH NOTES

5 Matrix Bernstein

5.1 Proof

Last time, we talked about Matrix Bernstein inequality. I'll state it again:

Theorem 5.1. *Matrix-Bernstein Inequality.*
Let $X_1, \cdots, X_n$ be independent, mean 0, $n \times n$ symmetric random matrices $\|X_i\| \leq K$ a.s. where the norm is the operator norm. Then, we have

$$\mathbb{P}\left\{\left\|\sum_{i=1}^n X_i\right\| > t\right\} \leq 2n\text{exp}\left(-\frac{t^2/2}{\sigma^2 + kt/3}\right) \tag{31}$$

Here $\sigma^2 ='' \text{Var}(\sum_{i=1}^n X_i)'' \leq \|\sum_{i=1}^n \mathbb{E}\left[X_i^2\right]\|$.

Again, note that we have the mixture of two tails: sub-gaussian and sub-exponential. The n in the statement is actually optimal — you need an n. Just choose a symmetric random matrix, and put a 1 somewhere on the diagonal — then you'll see that n is necessary. These examples are very discrete typically.

Remark 5.2. Tightness.
This inequality is tight for the Gaussian tail, by central limit theorem. For the exponential tail, it is not tight, even for numbers instead of matrices. There are two regimes: You either have Gaussian distribution limit (ClT) or Poisson in the limit. But Poisson is not exactly e^-t: Instead, optimal would be the Poisson tail, t^{-t}. For random variables, we have a strengthening — Bennet's inequality. For matrices I don't know, but probably we don't have it.

We noted we will prove this result using matrix calculus. We write down the spectral decomposition $X = \sum_i \lambda_i u_i u_i^T$, and extend functions $f : \mathbb{R} \to \mathbb{R}$ by applying them to matrices as $f(X) = \sum_i f(\lambda_i) u_i u_i^T$. After defining an ordering on matrices with $\succ$, it turns out a lot of properties of functions on numbers pass up to matrices as well. Note that we don't always get inequality though – when matrices don't commute, some things don't extend, for instance $e^{X+Y} \neq e^X e^Y$ for noncommuting matrices (when things do commute, they are exactly the same). We have the following great inequality:

Theorem 5.3. *Golden-Thompson Inequality.*

$$\mathrm{Tr}(e^{X+Y}) \leq \mathrm{Tr}(e^X e^Y) \tag{32}$$

We can use this and apply same approach, apply MGF etc. There's another approach, which is morally stronger than Golden-Thompson, which is Lieb's inequality. In our proof, we will use Lieb's inequality:

Theorem 5.4. *Lieb's inequality.*
Let H be an $n \times n$ symmetric matrix. Consider $f(X) = \mathrm{Tr}(\exp(H + \log X))$. Then f is ***concave*** *on the space of positive-definite matrices. In the scalar case, we have linearity, e.g. $f(x) = e^{H+logx} = xe^H$.*

Note that this is a deterministic statement. But let's move it to a probabilistic one for matrices. Jensen's inequality says that if X is a random matrix, then $\mathbb{E}[f(X)] \leq f(\mathbb{E}[X])$ if f is concave. Let's apply this for $X = e^Z$. Therefore,

Corollary 5.5. *Let H be an $n \times n$ symmetric matrix and Z be a random $n \times n$ symmetric matrix. Then applying Lieb's inequality,*

$$\mathbb{E}[\mathrm{Tr}(\exp(H + Z))] \leq \mathrm{Tr}\left(\exp(H + \log(\mathbb{E}\left[e^Z\right]))\right) \tag{33}$$

This corollary will be our primary method in proving Matrix Bernstein. We'll look at the MGF on Z, and then chop off terms one at a time and iterate this inequality. Now let's do the proof of Matrix Bernstein.

Proof. Let $S = \sum_{i=1}^n X_i$. We want to bound the operator norm. The operator norm of S is the maximal eigenvalue: $\|S\| = \max_i |\lambda_i(S)| = \max\{\lambda_{\max}(S), \lambda_{\max}(-S)\}$. So we will look at the maximum eigenvalues. How do we do that? We want

$$\begin{aligned}\mathbb{P}\left\{\lambda_{\max}(S) > t\right\} &= \mathbb{P}\left\{e^{\lambda\lambda_{\max}(S)} > e^{\lambda t}\right\} \\ &\leq e^{-\lambda t}\mathbb{E}\left[e^{\lambda\lambda_{\max}(S)}\right]\end{aligned} \tag{34}$$

It's still a mystery how to handle the maximum eigenvalue, we will replace it by the trace, which is a sum of all eigenvalues. We have $e^{\lambda\lambda_{\max}(S)} = \lambda_{\max}(e^{\lambda S})$ by monotonicity. We don't know how to bound this, but we can bound $\leq \mathrm{Tr}(e^{\lambda S})$ since the trace is the sum of the eigenvalues and the eigenvalues are non-negative, which is the case since we exponentiate. We can then take expectations of both sides. We are almost ready to apply Lieb's inequality.

$$\begin{aligned}\mathbb{E}\left[\mathrm{Tr}(e^{\lambda S})\right] &= \mathbb{E}_{X_1,\cdots,X_{n-1}}\left[\mathrm{Tr}(\exp(\sum_{i=1}^{n-1}\lambda X_i + \lambda X_n))\right] \\ &\leq^{\text{Lieb's inequality}} \mathbb{E}\left[\mathrm{Tr}(\exp(\sum_{i=1}^{n-1}\lambda X_i + \log(\mathbb{E}_{X_n}\left[e^{\lambda X_n}\right])))\right]\end{aligned} \tag{35}$$

What we really do is condition on all random variables except the last one, and take Lieb's with respect to X_n, and then uncondition afterwards. Now we can again apply Lieb's inequality in the same fashion to get

$$\mathbb{E}\left[\mathrm{Tr}(\exp(\sum_{i=1}^{n-2}\lambda X_i + \log\mathbb{E}\left[e^{\lambda X_{n-1}}\right] + \log\mathbb{E}\left[e^{\lambda X_n}\right]))\right] \leq \mathrm{Tr}\left(\exp(\sum_{i=1}^{n}\log\mathbb{E}\left[e^{\lambda X_i}\right])\right) \tag{36}$$

Now this is a problem about the MGF of one random variable — in particular, it's called the cumulant, and this is not difficult since it does not involve a sum of random variables. Let's try to find a good bound on the cumulant. We have $\log\mathbb{E}\left[e^{\lambda X}\right]$ for X mean zero and $\|X\| \leq K$. The best way is to just do Taylor expansion of exponential and look at how the terms behave. We get $e^z = 1 + z + z^2/2 + \cdots$. Then we can bound for $|z| \leq 1$, $e^z \leq 1 + z + z^2$. Let's apply this to the matrix: $e^{\lambda x} \leq 1 + \lambda x + \lambda^2 x^2$ if $|x| \leq K$ and $\lambda \leq 1/K$. Now it's an exercise to see that this inequality lifts to the matrix case: $e^{\lambda X} \leq I + \lambda X + \lambda^2 X^2$ for $\mathbb{E}[X] = 0, \|X\| \leq K$ if $\lambda \leq 1/K$. Now we need to take expected value: $\mathbb{E}\left[e^{\lambda X}\right] \leq I + 0 + \lambda^2\mathbb{E}[X^2]$ if $\lambda \leq 1/K$. Then

$$\log(\mathbb{E}\left[e^{\lambda X}\right]) \leq \lambda^2\mathbb{E}\left[X^2\right] \tag{37}$$

Then, we plug it into the proof from before: Letting $Z = \sum_{i=1}^n \mathbb{E}\left[X_i^2\right]$,

$$\begin{aligned}\mathbb{E}\left[e^{\lambda\lambda_{\max}(S)}\right] &\leq \mathrm{Tr}(\exp(\sum_{i=1}^n \lambda^2 \mathbb{E}\left[X_i^2\right])) \\ &= \mathrm{Tr}(\exp(\lambda^2 Z)) \\ &\leq n\lambda_{\max}(\exp(\lambda^2 Z)) \\ &\leq^{\text{monotonicity}} n\exp\left(\lambda^2 \cdot \lambda_{\max}(Z)\right) \\ &\leq n\exp(\lambda^2 \|Z\|) \\ &= n\exp(\lambda^2\sigma^2)\end{aligned} \tag{38}$$

noting that we lose a factor in the fact that we're bounding the trace again by n times the maximum eigenvalue, and σ^2 is the matrix variance. So we finally have the MGF of the sum. Now we can complete the proof:

$$\mathbb{P}\left\{\lambda_{\max}(S) > t\right\} \leq e^{-\lambda t} n\exp(\lambda^2\sigma^2) = n\exp(-\lambda t + \lambda^2\sigma^2) \tag{39}$$

where $|\lambda| \leq 1/K$: Then optimize to get the result. □

Now suppose you're not interested in tail bounds, just the expected value of the norm of the sum — even this is not trivial. Currently, the best way to get that is to go through matrix Bernstein. So what do we have? It's trivial in the scalar case, it's just the variance, but not in the matrix case.

Corollary 5.6. *Expectation.*
We can expect the following dependencies on K and σ by examining the sub-Gaussian case and the sub-exponential case.

$$\mathbb{E}\left[\|\sum_{i=1}^n X_i\|\right] \lesssim \sigma + K \tag{40}$$

If we are more careful, we remember there's an n on the outside — that will need to be pulled inside, and that creates a logarithmic dependence on n:

$$\mathbb{E}\left[\|\sum_{i=1}^n X_i\|\right] \lesssim \sigma\sqrt{\log n} + K\log n \tag{41}$$

It's an exercise to check this precisely. The $\sigma\sqrt{\log n}$ is tight for sure (Gaussian maxima), but the other term may not be.

5.2 Application to covariance matrix estimation

Now we will apply Matrix Bernstein in a few contexts: We'll go back to general covariance matrix estimation, trying to assume as little as possible about the distribution. What happened before is that we assumed that if X is distributed as a sub-gaussian in our data, then

$$\|\Sigma_N - \Sigma\| \leq \sqrt{\frac{n}{N}} + \frac{n}{N}\|\Sigma\|$$

where $\Sigma = \mathbb{E}\left[XX^T\right]$ is covariance of X and $\Sigma_N = \frac{1}{N}\sum_{i=1}^N X_iX_i^T$ is the sample covariance matrix. Here n is dimension. The weak point in this is that we're assuming the distribution is sub-gaussian.

Now we will be able to do this for any distribution. Let's try to imagine what we can do. Maybe it's too difficult to have no assumptions because of adversaries. So you can't really estimate anything. What we *can* assume is that if X is a random vector in $\mathbb{R}^n$, we have $\|X\|_2^2 \lesssim \mathbb{E}\left[\|X\|_2^2\right]$ almost surely (suppose the constant is something like 10).

Theorem 5.7. *Let X be a random vector in $\mathbb{R}^n$. If $\|X\|_2^2 \lesssim \mathbb{E}\left[\|X\|_2^2\right]$ almost surely, then*

$$\mathbb{E}\left[\|\Sigma_N - \Sigma\|\right] \leq \left(\sqrt{\frac{n\log n}{N}} + \frac{n\log n}{N}\right)\|\Sigma\| \tag{42}$$

Remark 5.8. Let us clarify the boundedness assumption: It's an exercise to check that $\mathbb{E}\left[\|X\|_2^2\right] = \mathrm{Tr}(\Sigma)$. If the covariance matrix is identity, this value is n. So our assumption is really that $\|X\|_2^2 \lesssim \mathrm{Tr}(\Sigma)$ almost surely.

Proof. We have $\Sigma_N - \Sigma = \frac{1}{N}\sum_{i=1}^N(X_iX_i^T - \Sigma)$. Then apply the corollary to Matrix-Bernstein to get

$$\mathbb{E}\left[\|\Sigma_N - \Sigma\|\right] \leq \frac{1}{N}(\sigma\sqrt{\log n} + K\log n) \tag{43}$$

where $\sigma^2 = \|\sum_{i=1}^N \mathbb{E}\left[(X_iX_i^T - \Sigma)^2\right]$ and $\|XX^T - \Sigma\| \leq K$. The rest is a computation — we need to bound σ^2 and K. By triangle inequality,

$$\begin{aligned}\sigma^2 = \|\sum_{i=1}^N \mathbb{E}\left[(X_iX_i^T - \Sigma)^2\right] &\leq N\|\mathbb{E}\left[(XX^T - \Sigma)^2\right]\| \\ &= N \cdot \|\mathbb{E}\left[(XX^T)^2\right] - \Sigma^2\| \preceq \mathbb{E}\left[(XX^T)^2\right] = \mathbb{E}\left[XX^TXX^T\right] \preceq \mathbb{E}\left[\|X\|_2^2XX^T\right] \\ &\preceq \mathrm{Tr}(\Sigma)\mathbb{E}\left[XX^T\right] = \mathrm{Tr}(\Sigma)\Sigma \\ \sigma^2 &\leq N \cdot \mathrm{Tr}(\Sigma)\|\Sigma\|\end{aligned} \tag{44}$$

Now we need a bound on K: As an exercise, show

$$\|XX^T - \Sigma\| \lesssim \mathrm{Tr}(\Sigma) \tag{45}$$

Again, use triangle inequality on the norm and proceed. So we have everything we need: A good bound on σ and K. Now we substitute back in to Matrix Bernstein to get

$$\mathbb{E}\left[\|\Sigma_N - \Sigma\|\right] \lesssim \frac{1}{N}\left(\sqrt{N \cdot \mathrm{Tr}(\Sigma) \cdot \|\Sigma\|}\sqrt{\log n} + \mathrm{Tr}(\Sigma)\log n\right) \tag{46}$$

Then, noting that $\text{Tr}(\Sigma) \leq n\|\Sigma\|$ gives us

$$\lesssim \frac{1}{N}\left(\sqrt{Nn\|\Sigma\|^2}\sqrt{\log n} + n\|\Sigma\|\log n\right) \tag{47}$$

$$\lesssim \|\Sigma\|\left(\sqrt{\frac{n\log n}{N}} + \frac{n\log n}{N}\right) \tag{48}$$

□

So we were able to get $N \sim n\log n$ as enough for covariance matrix estimation — all we had to do in comparison to the subgaussian assumption is get an extra $\log n$ factor correction, which is somewhat surprising.

Remark 5.9. Low-dimensional distributions.
We would hope that for actual low-dimensional distributions that it would not depend on the true dimension. We usually care about covariance estimation for PCA, anyways, where we kind of are hoping there is low-dimensional structure. If X is distribution as something close to a pancake (e.g., low dimension), then $r = \frac{\text{Tr}(\Sigma)}{\|\Sigma\|} = \frac{\sum_{i=1}^n \sigma_i}{\max_i \lambda_i}$. If the distribution is supported on a d-dimensional subspace, then $r \leq \frac{\sum_{i=1}^d \lambda_i}{\max_i \lambda_i} \leq d$. This is natural, we call d the **intrinsic dimension**. If the distribution is isotropic ($\Sigma = I$), then all eigenvalues are 1 and $r = n$. The intrinsic dimension is **stable** compared to the usual linear algebraic dimension: Suppose we have a couple outliers — of course algebraic dimension will be thrown off by this. But the intrinsic dimension probably will not — because we are talking about the bulk of eigenvalues. You can always assume what d is, and there's also a straightforward approach to estimate d by just calculating for different choices of d, and seeing how it changes (e.g., look for kinks in the spectrum curve). We can essentially replace all the n's in the proof above with r instead, whenever we make the upper bound of $n \cdot \lambda_{\max}$ for the trace, replace n with r. Thus $N \sim r\log n$ samples suffice, if you know intrinsic dimension is small.

Remark 5.10. What about $\log r$ dependence here? Intuitively, if $n = r$, then we could say $r\log r$ (just represent everything in that space) — if we are more careful, can we get the r if it's intrinsic dimension? But it's kind of open — don't know how to do this in the approximate case. But perhaps we have to pay for adaptation in some sense — we have to estimate what r is. Not sure whether the true result should have any dependence on n.

Remark 5.11. You have to center the data matrix because throughout, we've been assuming $\mathbb{E}[X] = 0$. This does not affect the rate, because it is easier to estimate the matrix: The rate is $\|\mathbb{E}[X] - \frac{1}{N}\sum_{i=1}^N X_i\| \lesssim \sqrt{\frac{n}{N}}$, which is at least as good as the rate for covariance estimation.

Remark 5.12. The 4^{th} moment assumption is enough to replace the strong assumption $\|X\|_2^2 \lesssim \mathbb{E}[\|X\|_2^2]$ a.s. assumption. Under $2+\epsilon$ moments, you have this bound (Vershynin-Srivastava). But the original assumption is not too strong. Our truncation level here is at the level n. If you talk about 4^{th} moments, you're proposing to truncate at the $O(1)$ level. Note: truncation in referring to handling the split for what datapoints we can "throw away" — analagous to the sub-exponential truncation approach to the proof.

5.3 Application to community detection in networks

This is an especially active research area right now. We have two communities and the network is a graph. More edges run inside each community (more tightly connected) than across the communities. We'll suppose there are $n/2$ nodes in each community.

One model for this setup is the Erdos-Renyi random graph model $G(n, p)$ where you have n edges and connect any vertices i, j by an edge independently with probability p. This graph has no communities — it is homogeneous.

Second is a slight generalization of this, the stochastic block model (SBM), where you have two communities: $G(n, p, q)$. You connect nodes in the same community with probability p, and across communities with probability q, with $p < q$. Each community has $n/2$ nodes. There are many other models (e.g., preferential attachment). There are many algorithms that input the network and output the communities. We will consider the basic spectral algorithm.

Definition 5.13. Spectral algorithm for community detection.

(a) Input adjacency matrix $A \in \mathbb{R}^{n\times n}$ symmetric.

(b) Look at eigenvalues $\lambda_1(A), \lambda_2(A)$ and the corresponding eigenvectors.

(c) Look at coordinates of second eigenvector in decreasing order. Hopefully there is a sharp threshold: Say one is community one, the other is community two.

This can be extended to more than two communities: You look at second and third eigenvectors together, and you'll see clusters in the plane. Just run K-means.

It's easy to see the second eigenvector is related to community detection. We want to show first that adjacency matrix A is close to its expectation $\mathbb{E}[A]$. For that we will use Matrix-Bernstein. Note that $\mathbb{E}[A]$ will have all p for the first community and the second community edges, and q otherwise. The first eigenvector is all 1s — this is useless (true if the graph is connected). The second eigenvector will have $+1$ for one community and -1 for the other (recall your spectral graph theory). We have $\lambda_1 = \frac{p+q}{2}n$, $\lambda_2 = \frac{p-q}{2}n$. We remark that λ_1 is the expected average degree of the network. So as long as the realized adjacency matrix is close enough to $\mathbb{E}[A]$, our procedure will work. We care about **how close the adjacency matrix is to the expected adjacency matrix**.

When can we be sure that this procedure works? e.g., when do we know $v_s(A) \approx v_w(\mathbb{E}[A])$? This is true if $A \approx \mathbb{E}[A]$, and this is resolved by Davis-Kahan: The eigenvalues and eigenvectors resulting from perturbation are good. We write a noisy version of the signal:

$$A = \mathbb{E}[A] + (A - \mathbb{E}[A]) \tag{49}$$

We want to prove that the signal-to-noise ratio is large. In this case, we know that perturbation works. We have $\|\mathbb{E}[A]\| = \lambda_1 = \frac{p+q}{2}n = d$, the expected average degree. Now we need the magnitude of the noise, hoping that $\|A - \mathbb{E}[A]\| \leq d$. All we do is apply Matrix-Bernstein for matrix A to compute the noise. We need to write A as a sum of random matrices. We can decompose it into matrices, each of which is everywhere 0 with one entry as 1 where there is one edge, everything else 0 except for the symmetric position where there is also a 1.

$$\begin{aligned} A &= \sum_{i \leq j} X_{ij} \\ A - \mathbb{E}[A] &= \sum_{i \leq j} (X_{ij} - \mathbb{E}[X_{ij}]) \qquad (50) \\ \mathbb{E}[\|A - \mathbb{E}[A]\|] &\lesssim \sigma\sqrt{\log n} + K \log n \end{aligned}$$

where $\sigma^2 = \left\|\sum_{i \leq j} \mathbb{E}[(X_{ij} - \mathbb{E}[X_{ij}])^2]\right\|$ and $K = \max_{ij} \|X_{ij} - \mathbb{E}[X_{ij}]\| \leq 4$ (look at the amount of non-zero entries – should be 2, then square it). For σ^2, you can get (exercise) that $\sigma^2 \leq \frac{p+q}{2} n = d$. Then we conclude

$$\mathbb{E}[\|A - \mathbb{E}[A]\|] \lesssim \sqrt{d \log n} + \log n \qquad (51)$$

So what's our hope now? We want $\|A - \mathbb{E}[A]\| \leq \frac{1}{10} d$. So we want $\sqrt{d \log n} + \log n \leq \frac{d}{10}$. So this is true if $d \geq c \log n$, and we're done. So, non-rigorously,

Theorem 5.14. *Spectral clustering works if the expected average degree of the network is* $\geq c \log n$.

It's a result that this is actually sharp — cannot go below $c \log n$ with spectral clustering. If the average degree is below $\log n$, the network becomes very sparse and the distribution of degree becomes wild. The degrees of the very popular people is over $\log n$. So the algorithm will think they are the communities, and the second eigenvector will pick super popular people or those who do not have friends at all. These outliers will prevent spectral method from working.

However, there are other algorithms which go below $\log n$ for sparser networks, all the way down to $d = O(1)$ — this is a very popular area right now. Basically do pre-processing, and remove very popular people and so on.