Notes for Topics in Random Matrices Theory

May 29, 2020

Shenduo Zhang

E-mail: zhangshenduo@gmail.com

Contents

1 Problems before 5.3.2020

1.1 Spherical width of the unit ball is 1 or 2?

In Vershynin's book, *Exercise 7.5.7, Page 175*, says the unit sphere has spherical width 2.

But I think from his definition for spherical width $w_s(T)$,

$$w_s(T) = \mathbb{E} \sup_{x \in T} < \theta, x >, \theta \sim \text{Unif}(S^{n-1})$$

Under such definition, the spherical width shall be 1 instead of 2 isn't it? Although it's not a huge difference. It will be 2 if he define it to be the magnitude of absolute value $| < \cdot, \cdot > |$ or $\sup_{x \in T-T}$. And all of them are kind of equivalent.

2 Proof for Talagrand's concentration inequality.

You are not expected to answer this one because this can be viewed as part of my personal feeling instead of a real question. I should try to fix it on my own before I ask you. So I post it here for a shorter email for you to read and you only need to read this if you are interested in my experience.

I don't understand what the "combinatorial distance" or "support" means in Dr. Tao's book. Since he said he will rely on this heavily in the later context, I will try to read other literiture especially you recommended to me to figure it out. However I'm able to understand[1] it using Log-Soboblev inequality, but I didn't orgnize it in this note by the time I write this.

[1] *Maybe I don't.*

Actually I am doing it and in half way now. I find those literiture, specifily the one by Dr. Logusi, feels much more accessible because it's more detailed in proof, not involves that much asymptotic notation and usually works in $\mathbb{R}$ instead of $\mathbb{C}$. But I guess I have to get used to those things somehow in the future, especially asymptotic notation. It's hard, but it's getting more interesting, even before I see any new results involves random matrix. This part of contents is like an extended version of Vershynin's book but from a pure analytical perspective.

3 Problems before 4.26.2020

3.1 Dimension counts

Exercise 1.3.10, Page 47

When Dr. Tao was introducing the eigenvalue replusion phenomenon, he first try to show that it is a generic behaviour that a Hermitian matrix has simple spectrum, for which he used this *Excercise 1.3.10*.

Proposition 3.1. Suppose $n \geq 2$, the space of Hermitian matrices with at least one repeated eigenvalue has codimension-3 in the space of all Hermitian matrices. And the space of real symmetric matrices with at least one repeated eigenvalue has codimension 2 in the space of all symmetric matrices.

But first thing we notice that the space of all the Hermitian matrix with at least one repeated eigenvalue is not a linear space anymore. So it is hard for me to work with such kind of space or dimension in proving. May I just for now interprete it as the least number it takes to describe such a Hermitian matrix? (Which I tried to prove it in 19). Or just forget about why and go back to it after I know rigorous definition about it?

And when he said this space has codimension-3 in the space of all Hermitian matrices, I wonder when the size of matrix n goes large, the dimension of Hermitian matrix grows in speed of n^2. Why can a "subspace" of codimension-3 show that this is somehow generic behaviour?

4 Problems before 4.19.2020

Just in case you don't have the access to the book while you are at home, I attached the two book in pdf file in the attachments too.

4.1 Max of linear functionals = convex functional?

When Dr. Tao was introducing Courant-Fisher minimax theorem(*Theorem 1.3.2*,Page 42), he said

The i^th eigenvalue functional $A \to \lambda_i(A)$ is not a linear funcional, nor convex or concave functionals. (except in dimension 1). But it's the next best thing, a minimax expression of linear functionals.

Then he added in the footnote,

A convex functional is the same thing as a max of linear functionals, while a concave functional is the same thing as a min of linear functionals.

I don't understand why they are the "same thing". Or could you please explain what are those linear functionals that after taking maximum can become a "same thing" to our convex functional.

PART

I

Preparatory

5 Asymptotic notation

Definition 5.1. O, o, Θ
We use $X = O(Y), X \ll Y$ to denote the estimate $|X| \leq CY$ for some C independent of n and all $n \geq C$. If we need C to depend on a parameter k, indicate it by subscripts like $X = O_k(Y)$.
We use $X = o(Y)$ if $|X| \leq c(n)Y$ for some c that goes to zero as $n \to \infty$.
We write $X \sim Y$ or $X = \Theta(Y)$ if $X \ll Y \ll X$.

Definition 5.2 (High-prob notation)**.** Give an event $E = E_n$, we have five in decreasing order of confidence notation that an event is likely to hold:

1. *Surely*: An event is equal to sure event $\emptyset$.
2. *with full probability*: it occurs with probability 1.
3. *with overwhelming probability*:for every fixed $A > 0$, it holds with probability $1 - O_A(n^{-A})$.
4. *with high probability*:it holds with probability $1 - O(n^{-c})$for some $c > 0$ independent of n.
5. *asymptotically almost surely*: it holds with probability $1 - o(1)$.

We allow the factor depends on A, but it holds for all A not just a specific number. Which basically means the high probability can be derived by just take any A.

Proposition 5.1 (Union bound)**.** Now E is our event that depends on n, we can have a family of events index by α, namely E_α.

1. To keep *surely* after union bound, the family can have any cardinality.
2. To keep *almost surely* after union bound, the family of need to be at most countable.
3. When E_α holds with *uniformly overwhelming probability* in α, the family need to be polynomial cardinality $O(n^{O(1)})$.
4. When E_α holds with *uniformly high probability* in α, the family need to have cardinality $O(n^{o(1)})$.(In particular polylogarithmic $O(\log^{O(1)} n)$)

This asymptotic not easy is still not easy to handle when performing more complicated operation and having them everywhere in the estimates. But I guess I have to live with it and try to get used to it.

6 Probability

6.1 Basic probability

Theorem 6.1 (Lebesgue's dominate convergence theorem). Let f_n be a sequence of complex-valued measurable function on a measure space (S, Σ, μ). Suppose the sequence pointwise converge to a function f that is donimated by some integrable function g for all numbers n in the index set and all points $x \in S$. [2] Then f is integrable and we can switch the order of integral and limit.

[2] *If the measure space is complete, one can relax the dominate to almost dominate.*

Remark. In probability space, where $\mu(S) < \infty$, the uniform integrability is enough for such kind of switching. Which is know as Vitali's theorem.

Theorem 6.2 (Jensen's inequality). Let $F : \mathbb{R} \to \mathbb{R}$ be a convex function, let X be a bounded real-valued random variable. Show that $\mathbb{E}F(X) \geq F(\mathbb{E}X)$.

Proof. [Jensen's] Let $L : \mathbb{R} \to \mathbb{R}$ be a linear function such that $L(\mathbb{E}X) = F(\mathbb{E}X)$. Then F donimates L on $\mathbb{R}$. With linearty and monotoncity we have $\mathbb{E}F(X) \geq L(\mathbb{E}X) = F(\mathbb{E}X)$. □

Definition 6.1 (Usually bounded). In decreasing order of confidence,

1. *surely bounded*
2. *almost surely bounded*
3. *sub-Gaussian*
4. *sub-exponential* if there exist $C, c, a > 0$ such that $\mathbb{P}\{|X| \geq \lambda\} \leq C \exp\{-c\lambda^a\}\}$.[3]
5. *finite k^{th} moment*
6. *absolutely integrable*
7. *almost surely finite*

[3] *This is more general from the one in Vershynin's book's definition.*

Remark (Exponential and Fourier moments with moments). If X is sub-Gaussian(or has sub-exponential tails with exponent $a > 1$), then from dominate convergence we have the Taylor expansion

$$\mathbb{E}e^{tX} = 1 + \sum_{k=1}^{\infty} \frac{t^k}{k!} \mathbb{E}X^k \tag{6.1}$$

for any real or complex t.

Remind from Vershynin's book that the L^p norm is actually bounded by orcliz norm up to a factor depend on the orcliz space of $p \geq 1$, i.e.

$$||X||_{L^p} = O(\sqrt{p}||X||_{\psi_2}), = O(p||X||_{\psi_1}) \tag{6.2}$$

And it seems with a little more effort one can derive the high prob version of the magnitude of X under L^p norm.

6.2 Moments

6.2.1 Zeroth-moment method

By consider the indicator function of an event, the union bound can be rewrite(in the finitely many case) as follow zeroth-momen method

$$\mathbb{E}|\sum_i c_i X_i|^0 \leq \sum_i |c_i|^0 \mathbb{E}|X_i|^0 \tag{6.3}$$

for any scalar random variables X_i and scalars c_i.

Lemma 6.3 (Borel-Cantelli lemma)**.** Let $E_1, E_2, \ldots$ be a sequence of events such that $\sum_i \mathbb{P}\{E_i\} < \infty$. Show that, at most finitely many of the events E_i occur at once.

Proof. Let $I(E_n)$ denote the indicator function of the event E_n. Take expectation of $\sum_{i=1}^n I(E_n)$ and apply markov inequality we have

$$\mathbb{P}\{\sum_{n=1}^{\infty} I(E_n) \geq \lambda\} \leq \frac{1}{\lambda} \sum_{n=1}^{\infty} \mathbb{P}\{E_n\}$$

Let $\lambda \to \infty$ we obtain the claim. □

6.2.2 First-moment method

The applying of Markov inequality and computation of expectation is called first-moment method I guess.[4]

Fix X has finite k^{th} moment, then from Markov's theorem, the higher moments we control, the faster the tail decay.

$$\mathbb{P}\{|X| \geq \lambda\} \leq C\lambda^{-k}. \tag{6.4}$$

Noting that $\lambda^k \mathbf{1}_{|X|\geq\lambda}$ is donimated by $|X|^k$. And the previous one converges a.s. to zero function when $\lambda \to \infty$. So by dominate convergence theorem, take the expectation and take the limit inside the expectation, one can have a variant

$$\lim_{\lambda\to\infty} \lambda^k \mathbb{P}(|X| \geq \lambda) = 0. \tag{6.5}$$

But this result is qualitative since it provide no *rate of convergence* of $\lambda^k \mathbb{P}(|X| \geq \lambda)$ to zero. Which means the convergence is not uniform to the ambient n or other parameter. And we often have to strengthen the property of merely having uniformly finite bounded moments to abtain a uniformly quantitative convergence rate with respect to ambient n or other parameters.

But if the X_α are just identically distributed or it's just itself, then this will be just fine.

Lemma 6.4 (Magnitude from Moments)**.** Let $X = X_n$ be a scalar random variables depending on a parameter n.

1. $|X_n|$ has uniformly bounded expectation, for any $\epsilon > 0$ independent of n, we have $|X_n| = O(n^\epsilon)$ with high probability.
2. If X_n has uniformly bounded k^{th} moment, then for any $A > 0$, we have $|X_n| = O(n^{A/k})$ with probability $1 - O(n^{-A})$.
3. If X_n has uniform sub-exponential tails, then we have $|X_n| = O(\log^{O(1)} n)$ with overwhelming probability. [5]

[4] *He did not mentioned this in the book, maybe first moment method is usually refered to something else.*

[5] *We need to control $\mathbb{P}\{\forall n |X_n| \geq C \log^{O(1)} n\}$, this is what overwhelming probability means. We want a bound holds for any n so that it becomes independent of n.*

Is this overwhelming probability we have here same as taking the maximum over several independent random variables? Which will bring us back to the maximum of sub-Gaussian? The maximum of sub-Gaussian is given in expectation form in Vershynin's book, but here it seems one can easily derive a high prob version.

6.2.3 Second-moment method

The second moment method is usually refered to using Chebyshev's inequality with coputation of variance given the second order moment is finite.

Example. Denote the median of X as $\mathbf{M}(X)$. Show that if X has finite second moment, then

$$\mathbf{M}(X) = \mathbb{E}(X) + O(\mathbf{Var}(X)^{1/2}) \tag{6.6}$$

Straight forward from Chebyshev's inequality.

Theorem 6.5 (Pairwise independence $\to$ linearity of variance)**.** If $X_1, \cdots, X_k$ are pairwise independent scalar random variables of finite ean and variance, show that

$$\mathbf{Var}\left(\sum_{i=1}^{k} X_i\right) = \sum_{i=1}^{k} \mathbf{Var}(X_i) \tag{6.7}$$

and more generally,

$$\mathbf{Var}\left(\sum_{i=1}^{k} c_i X_i\right) = \sum_{i=1}^{k} |c_i|^2 \mathbf{Var}(X_i) \tag{6.8}$$

for any scalars c_i

6.2.4 Exponential-moment method

If $X_1, \cdots, X_k$ are jointly independent sub-Gaussian variables, then

$$\mathbb{E}\prod_{i=1}^{k} e^{tX_i} = \prod \mathbb{E}e^{tX_i} \tag{6.9}$$

for any complex t.

6.3 Conditioning/Freezing

Let E, F be events in some probability space. Conditioning magnify probability by a factor of at most $1/\mathbb{P}(E)$. In particular,

Example (How conditioning magnify probability)**.**

1. If F occurs unconditionally surely or almost surely, it occurs surely or almost surely after conditioning on E.
2. If F occurs unconditionally with overwhelming probability, it occurs with overwhelming probability conditioning on E, provided that $\mathbb{P}(E) \geq cn^{-C}$ for some $c, C > 0$ independent of n.
3. If F occurs unconditionally with high probability, it occurs with high probability conditioning on E, provided that $\mathbb{P}(E) \geq cn^{-a}$ for some $c > 0$ and sufficiently small $a > 0$ independent of n.
4. If F occurs unconditionally asymptotically almost surely, it occurs asymptotically almost surely conditioning on E also, provided that $\mathbb{P}(E) > c$ for some

$c > 0$ independent of n.

We define the conditional probability of an (unconditionally) event F conditioning on Y to be the (unconditionally) random variable that is define to equal $\mathbb{P}(F|Y = y)$ whenever $Y = y$. Same for the conditional expectation. Then if X, Y are independent, the equality holds always for any y almost surely which makes the conditional expectation a deterministic quantity.

And in the discret cases[6], we have the following identity

$$\mathbb{P}(F) = \mathbb{E}(\mathbb{P}(E|Y)) \tag{6.10}$$

$$\mathbb{E}(X) = \mathbb{E}(\mathbb{E}(X|Y)) \tag{6.11}$$

For conditioning on continuous random variables[7], One need disintegration.

Definition 6.2 (Disintegration)**.** Let Y be a random variable with range R. A *disintegration* of the underlying sample space ω with respect to Y is a subset R' of R of full measure in μ_Y, together with assignment of a probability measure $\mathbb{P}(|Y = y)$ on the subspace $\Omega_y := \{\omega \in \Omega : Y(\omega) = y\}$ of Ω for each $y \in R$. The probability measure we are assigning is measurable to y in the sense that the map $y \to \mathbb{P}(F|Y = y)$ is measureble for every event F, and such that

$$\mathbb{P}(F) = \mathbb{E}\mathbb{P}(F|Y) \tag{6.12}$$

for all such events, where $\mathbb{P}(F|Y)$ is the (almost surely defined) random variable to equal $\mathbb{P}(F|Y = y)$ whenever $Y = y$.

[6] *Conditioning on events or conditioning on discret variables is identically.*

[7] *which is analogue of dividing into a potentially uncountable number of cases.*

7 Stirling's formula

In short

$$n! = (1 + o(1))\sqrt{2\pi n} n^n e^{-n}.$$

7.1 Bound Using Taylor expansion

By just Taylor expansion of $e^x, x \geq 0$, let $x = n$ we have

$$n^n \leq n! \leq n^n e^{-n}. \tag{7.1}$$

So $n!$ is within an exponential factor of n^n.

7.2 Bound using Riemann sum

Start with the identity

$$\log n! = \sum_{m=1}^{n} \log m \tag{7.2}$$

viewing the right-hand side as a Riemann integral approximation to $\int_1^n \log \mathrm{d}x$. Compare the Upper Riemann sum and Lower Riemann sum, one have

$$\int_1^n \log x \mathrm{d}x \leq \sum_{m=1}^{n} \log m \leq \log n + \int_1^n \log x \mathrm{d}x \tag{7.3}$$

Which leads to another bounded

$$en^n e^{-n} \leq n! \leq en \times n^n e^{-n} \tag{7.4}$$

So the factor was not exponential but a factor of n.

7.3 Trapezoid rule

Lemma 7.1. The error of composite trapezoidal rule is

$$\text{error} = \int_a^b f(x)\mathrm{d}x - \frac{b-a}{N}\left[\frac{f(a)+f(b)}{2} + \sum_{k=1}^{N-1} f\left(a + k\frac{b-a}{N}\right)\right]$$

And there exist a number ξ between a and b such that

$$error = -\frac{(b-a)^3}{12N^2} f''(\xi)$$

On any interval $[m, m+1]$, $\log x$ has a second derivative of $O(1/m^2)$, which by Taylor expansion[8] leads to approxiamation

$$\int_m^{m+1} \log x \mathrm{d}x = \frac{1}{2}\log m + \frac{1}{2}\log(m+1) + \epsilon_m \tag{7.5}$$

for some error $\epsilon_m = O(1/m^2)$.

The error is absolutely convergent. And by integral test, we have $\sum_{m=1}^{n} \epsilon_m = C + O(1/n)$, where the constant $C := \sum_{m=1}^{\infty} \epsilon_m$. Perform this sum, rearrange and take exponent, we obtain

$$n! = (1 + O(1/n))e^{1-C}\sqrt{n}n^n e^{-n}. \tag{7.6}$$

Which indicates $n!$ lies roughly at the geometric means of the previous two bounds.

[8] *I didn't see why he need Taylor expansion here.*

7.4 Laplace method

Estimation of exponential integral $\int e^{\phi(x)}\mathrm{d}x$ when ϕ is large can be done through this principle.

First the integeral is donimated by the local maxima of ϕ. Then, near these maxima, $e^{\phi(x)}$ usually behaves like a rescaled Gaussia. So one can often understand the asymptotics of such integrals by a change of variables designed to reveal the Gaussian behaviour.

Start by interpreting the factorial via Gamma function,

$$n! = \int_0^\infty t^n e^{-t}\mathrm{d}t \tag{7.7}$$

$t^n e^{-n}$ achieves its maximum at n, so make the substitution $t = n + s$,

$$n! = \int_{-n}^\infty (n+s)^n e^{-n-s}\mathrm{d}s \tag{7.8}$$

Rearrange and simplify it impies

$$n! = n^n e^{-n}\int_{-n}^\infty \exp\Big(n\log\Big(1+\frac{s}{n}\Big) - s\Big)\mathrm{d}s.$$

By Taylor expansion can heuristically have

$$\exp\Big(n\log\Big(1+\frac{s}{n}\Big) - s\Big) \approx \exp\big(-s^2/2n\big).$$

To do this, we need toscale s by $\sqrt{n}$ to remove n in the denominator. Then

$$n! = \sqrt{n} n^n e^{-n} \int_{-\sqrt{n}}^{\infty} \exp\left(n \log\left(1 + \frac{x}{\sqrt{n}}\right) - \sqrt{n} x \right) \mathrm{d}x.$$

And Taylor expansion gives us for fixed x, we have the pointwise convergence

$$\exp\left(n \log\left(1 + \frac{x}{\sqrt{n}}\right) - \sqrt{n} x \right) \to \exp\left(-x^2/2\right) \tag{7.9}$$

as $n \to \infty$. And Lebesgue dominate convergence theorem tells us

$$\int_{-\sqrt{n}}^{\infty} \exp\left(n \log\left(1 + \frac{x}{\sqrt{n}}\right) - \sqrt{n} x \right) \mathrm{d}x \to \int_{-\infty}^{\infty} \exp\left(-x^2/2\right) \mathrm{d}x$$

Then we will conclude the *Stirling formula*

Theorem 7.2 (Stirling formula)**.**

$$n! = (1 + o(1)) \sqrt{2\pi n} n^n e^{-n}. \tag{7.10}$$

Example (Entropy formula)**.** Let n be large, let $0 < \gamma < 1$ be fixed, and let $1 \le m \le n$ be an integer of the form $m = (\gamma + o(1))n$. Show that $\binom{n}{m} = \exp((h(\gamma) + o(1))n)$, where h is the entropy function,

$$h(\gamma) := \gamma \log \frac{1}{\gamma} + (1 - \gamma) \log \frac{1}{1 - \gamma}.$$

Proof. [Entropy formula] Write binom in factorial form and plug in stirling formula. The exponential part will be just $\exp((h(\gamma) + o(1))n)$ itself. Luckily the term in the square root and the constant a head will be merged into $o(n)$ term after taking exponential. □

Example (Refined entropy formula)**.** Suppose $m = n/2 + k$ for some $k = o(n^{2/3})$. Show that

$$\binom{n}{m} = (\sqrt{\frac{2}{\pi}} + o(1)) \frac{2^n}{\sqrt{n}} \exp\left(-2k^2/n\right). \tag{7.11}$$

I was not able to prove this version of entropy formula.

Remark (Central limit theorem)**.** The Gaussian-type behaviour in k. This can be viewed as an illustration of the central limit theorem. When summing iid Berboulli variables. Indeed,

$$\mathbb{P}(X_1 + \cdots + X_n = n/2 + k) = (\sqrt{\frac{2}{\pi}} + o(1)) \frac{1}{\sqrt{n}} \exp\left(-2k^2/n\right).$$

When $k = o(n^{2/3})$. Which suggests the sum of X_i, S_n, is distributed roughly like Gaussian $N(n/2, n/4)$

8 Eigenvalues and sums of hermitian matrices

8.1 Spectral Theorem

For Hermitian $n \times n$ matrix A, B[9]. A basic question is to ask the extent to which the eigenvalues $\lambda_1(A), \ldots, \lambda_n(A)$ and $\lambda_1(B), \ldots, \lambda_n(B)$ of two Hermitian matrices A, B con-

[9] *Ajoint linear operator in finite vector space*

strain the eigenvalues $\lambda_1(A+B), \ldots, \lambda(A+B)$ of the sum. And in typical application to random matrices, one of the matrices (say, B) is "small" in some sense, so that $A+B$ is a perturbation of A.

The complete answer to this problem is in a survey(see the book). But in random matrices, we don't need the above theory but rather a simple aspect of it, which generates several *eigenvalue inequality*.

Proposition 8.1 (Weyl inequality)**.**

$$\lambda_{i+j-1}(A+B) \leq \lambda_i(A) + \lambda_j(B). \tag{8.1}$$

valid whenever $i, j \geq 1$ and $i+j-1 \leq n$.

Proposition 8.2 (Ky Fan inequality)**.**

$$\lambda_1(A+B) + \cdots + \lambda_k(A+B) \leq \lambda_1(A) + \cdots + \lambda_k(A) + \lambda_1(B) + \cdots + \lambda_k(B) \tag{8.2}$$

One consequence of these inequalities is that the spectrum of a Hermitian matrix is *stable* with respect to small perturbations.[10]

However the eigenvalues of non-Hermitian matrices can be far more unstable if *pseudospectrum* is present.[11]

Theorem 8.1 (Spectral theorem)**.** Let V be a finite-dimensional complex Hilbert space of some dimension n, and let $T : V \to V$ be a self-adjoint operator. Then there exist an orthonormal basis $v_1, \cdots, v_n \in V$ of V and eigenvalues $\lambda_1, \ldots, \lambda_n$ in $\mathbb{R}$ such that $Tv_i = \lambda_i v_i forall 1 \leq i \leq n$.

With spectrum theorem and finding a set of new basis[12] one can diagonalize matrix A which makes computation of matrix function easier. But when dealing with multiple matrices, it's hard to diagonalize matrices simultaneously, unless all the matrices are commute. However one can still normalize one of the eigenbases to be the standard basis.

8.2 Minimax formulae

Theorem 8.2 (Courant-Fischer minimax theorem)**.** Let A be an $n \times n$ Hermitian matrix. Then we have

$$\lambda_i(A) = \sup_{\dim(V)=i} \inf_{v \in V : |v|=1} v^* A v \tag{8.3}$$

and

$$\lambda_i(A) = \inf_{dim(V)=n-i+1} \sup_{v \in V : |v|=1} v^* A v \tag{8.4}$$

for all $1 \leq i \leq n$, where V ranges over all subspaces of C^n with the indicated dimension.

The i^{th} eigenvalue functional is not linear(expect in dimension one) nor convex nor concave. It is the next best thing, minimax expression of linear funcionals.[13]

Remark. By homogeneity, one can replace the constrain $|v| = 1$ with $v \neq 0$ provided that one can replaces the quadratic form v^*Av with the *Rayleigh quotient* v^*Av/v^*v.

Proof. One only need to prove the first equality and then replace A by $-A$, since the $-\lambda_i(A) = \lambda_{n-i+1}(-A)$.

[10] *In the sense that the Weyl's gives an upper bound of the deviation after perturbations.*

[11] *This reminds me of the stable dimension v.s. algebric dimension.*

[12] *orthonormal basis is unique up to a complex phase rotation $e^{i\theta_j}$ if there's no multiplicity of eigenvalues. If there is, I believe the uniqueness if up to a real rotation plus a complex rotation.*

[13] *convex functional is the same thing as a max of linear functionals, a concave functional is the same thing as a min of linear functionals.*

When $i = 1$. Then by Spectral theorem, we can assume A has the standard eigenbasis$e_1, \dots, e_n$, in which case we have

$$v^* A v = \sum_{i=1}^{m} \lambda_i |v_i|^2 \tag{8.5}$$

where $v = (v_1, v_2, \dots, v_n)$.. The supremum is taken in the space spanned by the first eigenbases.
When $i \geq 1$, consider the space spanned by $e_1, \dots, e_i$. Then we have the inequality

$$\lambda_i(A) \leq \sup_{\dim V = i} \inf_{v \in V : |v| = 1} v^* A v.$$

Then what left is to prove the reverse inequality. For every i-dimensional subspace V of $\mathbf{C}^n$, we have to show that V contains a unit vector v, such that

$$v^* A v \leq \lambda_i(A).$$

Let W be the space spanned by $e_i, \dots, e_n$. This space has codimension $i - 1$, so it must have non-trivial intersection with V. If we let v be a unit vector in $V \cap W$, the claim is proved.

□

Definition 8.1 (Partial Trace)**.** Given an $n \times n$ Hermitian matrix A and an m-dimensional subspace V of $\mathbb{C}^n$, we define the partial trace $\operatorname{tr}(A \restriction_V)$ be the expression

$$\operatorname{tr}(A \restriction_V) := \sum_{i=1}^{m} v_i^* A v_i$$

where $v_1, \dots, v_m$ is any orthonormal basis of V.

Lemma 8.3 (Commute hermitian matrices)**.** Two hermitian matrices are commute if they have the same eigenspace, which means can be diagonalized at the same time.

This expression is independent of the choices of orthonormal basis, since the tranformation between orthonormal basis are just rotation, which means they are Hermitian matrix. Then consider the eigenbases of A with restriction on V and any orthonormal basis chosen will be equal to this expression in eigenbases.

Proposition 8.3 (Extremal partial trace)**.** Let A be an $n \times n$ Hermitian matrix. Then for any $1 \leq k \leq n$, one has

$$\lambda_1(A) + \cdots + \lambda_k(A) = \sup_{\dim(V) = k} \operatorname{tr}(A \restriction_V)$$

and

$$\lambda_{n-k+1}(A) + \cdots + \lambda_n(A) = \inf_{\dim(V) = k} \operatorname{tr}(A \restriction_V)$$

As a corollary, we see that $A \to \lambda_1(A) + \cdots + \lambda_k(A)$ is a convex function, and $A \to \lambda_{n-k+1}(A) + \cdots + \lambda_n(A)$ is a concave function.

Since it's a supremum of a linear functional, it convex by default, vice versa.

Remark. Specialising 8.3 to the case when V is a coordinate sub-space(i.e. the span of k of the basis vectors $e_1, \dots, e_n$), we conclude the *Schur-Horn inequality*

$$\begin{aligned} \lambda_{n-k+1}(A) + \cdots + \lambda_n(A) &\leq a_{i_1 i_1} + a_{i_k i_k} \\ &\leq \lambda_1(A) + \cdots + \lambda_k(A) \end{aligned} \tag{8.6}$$

for any $1 \leq i_1 < \cdots < i_k \leq n$, where $a_{11}, a_{22}, \dots, a_{nn}$ are the diagonal entries of A.

I'm poor in Geometry too. From remark 1.3.5 in the book.

8.3 Eigenvalue inequalities

The basic idea is to exploid the linearity relationship

$$v^*(A+B)v = v^*Av + v^*Bv \tag{8.7}$$

for any unit vector v, and more generally,

$$\operatorname{tr}(A + B \restriction_V) = \operatorname{tr}(A \restriction_V) + \operatorname{tr}(B \restriction_V) \tag{8.8}$$

for any subspace V.

Proposition 8.4.

$$\lambda_1(A+B) \leq \lambda_1(A) + \lambda_1(B) \tag{8.9}$$

The result of supremum of sum is controlled by the sum of supremum.

To prove Ky Fan inequality 8.2, one need to observe plug in 8.3 to 8.8. By plugging in 8.6 instead, one get the following *Lidskii inequality*,

Proposition 8.5 (Lidskii inequality)**.**

$$\lambda_{i_1}(A+B) + \cdots + \lambda_{i_k}(A+B) \leq \lambda_{i_1}(A) + \cdots + \lambda_{i_k}(A) + \lambda_1(B) + \cdots + \lambda_k(B)$$

for any $1 \leq i_1 < \cdots < i_k \leq n$.

Using the inequality

$$|v^*Bv| \leq ||B||_{\text{op}} = \max(|\lambda_1(B)|, |\lambda_n(B)|)$$

for unit vector v, with 8.7, one can derive

Proposition 8.6 (Eigenvalue stability inequality)**.**

$$|\lambda_i(A+B) - \lambda_i(A)| \leq ||B||_{\text{op}}$$

Hence the spectrum of $A+B$ will be close to that of A if B is small in operator norm.

Hence the map $A \mapsto \lambda_i(A)$ is Lipschitz continuous on the space of Hermitian matrices, for fixed $1 \leq i \leq n$.

More generally, if one wants to establish Weyl inequality 8.1.

Proof. [Weyl inequality] From Courant-Fischer minimax theorem 8.2, it suffice to show that every $i+j-1$ dimensional subspace V contains a unit vector v such that

$$v^*(A+B)v \leq \lambda_i(A) + \lambda_j(B).$$

But one can find a subspace U of codimension $i-1$ such that $v^*Av \leq \lambda_i(A)$ for all unit vectors v in U[14], and a subspace W of codimension $j-1$ such that $v^*Bv \leq \lambda_j(B)$ for all unit vectors v in W.
The intersection $U \cap W$ has codimension at most $i+j-2$ and so has a non-trivial intersection with V. The claim follows □

[14] *From the inf sup version.*

One can establish the dual inequality for Lidskii and Weyl by consider $-A$.

8.4 p-Schatten norm

One can use the Lidskii inequality to establish the more general inequality

$$\sum_{i=1}^n c_i\lambda_i(A+B) \leq \sum_{i=1}^n c_i\lambda_i(A) + \sum_{i=1}^n c_i^*\lambda_i(B)$$

whenever $c_1, \ldots, c_n \geq 0$, and $c_1^* \leq \cdots \leq c_n^* \leq 0$ is the decreasing rearrangement of $c_1, \ldots, c_n$.

Proof.
By writing

$$\sum_{i=1}^n c_i\lambda_i(A+B) = \sum_{i=1}^n \int_0^\infty \mathbf{I}(c_i \geq \lambda)\lambda_i(A+B)\mathrm{d}\lambda = \int_0^\infty \sum_{i=1}^n \mathbf{I}(c_i \geq \lambda)\lambda_i(A+B)\mathrm{d}\lambda$$

Then notice that inside the integral the indicator function is merely 1 or 0, so we can apply Lidskii inequality then obtain the desired bounds.

□

Conbined this with Holder's inequality[15] to conclude the *p-Weilandt-Hoffman inequality*

[15] *by using it for* $\sum_{i=1}^n c_i^*\lambda_i(B)$.

Proposition 8.7 (p-Weilandt-Hoffman inequality)**.**

$$||\{\lambda_i(A+B) - \lambda_i(A)\}_{i=1}^n||_{l_n^p} \leq ||B||_{S^p} \tag{8.10}$$

for any $1 \leq p \leq \infty$. where

$$||B||_{S^p} := ||\{\lambda_i(B)\}_{i=1}^n||_{l_n^p} \tag{8.11}$$

is the *p-Schatten norm* of B.

For any $1 \leq p \leq \infty$ and any Hermitian matrix A, one has[16]

$$||\{a_{ii}\}_{i=1}^n||_{l_n^p} \leq ||A||_{S^p}. \tag{8.12}$$

[16] *By noticing that* $||A||_{S^p} = [tr\{(A^*A)^{p/2}\}]^{1/p}$, *and find that it automatically includes the left hand side with some other non-negative term.*

Proposition 8.8 (Non-commutative Holder inequality)**.**

$$|\mathrm{tr}(AB)| \leq ||A||_{S^p}||B||_{S^{p'}} \tag{8.13}$$

whenever $1 \leq p, p' \leq \infty$ with $1/p + 1/p' = 1$, and A, B are $n \times n$ Hermitian matrices.

Remark. The most important p-Schatten norm are ∞-schatten norm, which is operator norm and the 2-schatten norm, which is Frobenius norm(or Hilbert-Schmidt norm).[17]

[17] *1-schatten norm is nuclear norm.*

In the case of $p = 2$, we have

$$\sum_{i=1}^{n} |\lambda_i(A+B) - \lambda_i(A)|^2 \leq ||B||_F^2. \tag{8.14}$$

8.5 Eigenvalue deformation

Repeated eigenvalues are rare.

Proposition 8.9. Suppose $n \geq 2$, the space of Hermitian matrices with at least one repeated eigenvalue has codimension-3 in the space of all Hermitian matrices. And the space of real symmetric matrices with at least one repeated eigenvalue has codimension 2 in the space of all symmetric matrices.

Here the dimension is not the dimension of linear space, but rather the numbers needed to specify something.[18]

[18] *Maybe it's true, maybe it's not. Hermitian matrices obviously form a linear space, but in the proof we are not working with linear space, because we work with Unitary matrices.*

Proof. Consider a $n \times n$ Hermitian matrix H, then $H = U\Lambda U^*$, where $U \in U(n)$, and Λ is diagonal real matrix. Which will take us $n^2 + r$ parameters to describe it, n^2 for U, r for Λ. But then we find it might not take as that many parameters, because when we describe it with that much parameters, the representation is not unique, in other words, $\exists W \in U(n)$ such that $H = W\Lambda W^*$.
We always need r parameter to describe Λ. But for U, it suffices with less parameters. So we ask how many parameters in n^2 are redundant.
By thinking starting from U, how many parameters we need to reach all the other W that represents a same H as U does. And the answer is that we can write $W = UV$, where $V \in U(n)$ commutes with Λ. Denote the numbers of parameters needed for describe V as v. This implies there are v numbers of parameters redundant in our representation for H.
So totally we need $n^2 + r - v$ parameters for H.*(at least we can expect or it make sense.)*
In the case when H has at least one repeated eigenvalues, the dimension of this "space"[19] is determined by the space in which H only has one repeated eigenvalues.
We have $f = n - 1$. We can denote Λ as $\mathrm{diag}(\lambda_1 I_2, \lambda_3, \ldots, \lambda_n)$. What left is to see how many numbers it take to describe all $V \in U(n)$ and commutes with Λ.
We can denote the multiplicity of each eigenvalue as g_i, then $\sum_{i=1}^{k} g_i = n$. And we need V to commute with a diagonal matrix Λ. Noticing that Λ is formed by scaled identity matrix of size g_i. So V is a matrix whose diagonal are formed by the block of same size as Λ, we can expect $\sum_{i=1}^{k} g_i^2$ numbers to describe V.
Then in our case $f = n - 1, v = n + 2$, so it makes sense.(not proved) □

[19] *not linear space anymore.*

Definition 8.2 (Simple spectrum)**.** A Hermitian matrix has simple spectrum if it has no repeated eigenvalues.

We can see from the codimension of the space of hermitian matrix with simple spectrum and 8.6 that the set of them form a *open dense set* in the space of all Hermitian matrices.[20]

[20] *I don't know why.*

Thus simple spectrum is the generic behaviour of such matrices.

The unexpectedly high codimension of the non-simple matrices suggests a repulsion phenomenon: because it is unexpectedly rare for eigenvalues to be equal, there must be some "force" that "repels" eigenvalues of Herimitian matrices from getting too close to each other.

We first observe that when A has simple spectrum, the zeroes of characteristic polynomial $\lambda \mapsto \det(A - \lambda I)$ has non-zero derivative at those zeroes(which is called simple). From this and the inverse function theorem, we see that each of the eigenvalue maps $A \mapsto \lambda_i(A)$ are smooth on the region where A has simple spectrum. Because the eigenvectors $\mu_i(A)$ are determined(up to a phase[21]) by the equation $(A - \lambda_i(A)I)\mu_i(A) = 0$ and $\mu_i(A)^*\mu_i(A) = 1$, another application of the inverse function theorem tells us we can locally select the maps $A \mapsto \mu_i(A)$ to also be smooth.

[21] *spectral projection does not have such ambiguity.*

Differentiate the equations

$$A\mu_i = \lambda_i \mu_i \tag{8.15}$$
$$\mu_i^* \mu_i = 1 \tag{8.16}$$

to obtain

$$\dot{A}\mu_i + A\dot{\mu_i} = \dot{\lambda_i} + \lambda_i \dot{\mu_i} \tag{8.17}$$
$$\dot{\mu_i^*}\mu_i + \mu_i^*\dot{\mu_i} = 0 \tag{8.18}$$

8.18 simplifies to $\dot{\mu_i^*}\mu_i = 0$, which means $\dot{\mu_i}$ is orthogonal to μ_i. Taking the inner product of 8.17 with μ_i, we conclude the *Hadamard first variation formula*:

Proposition 8.10 (Hadamard first variatoin formula)**.**

$$\dot{\lambda_i} = \mu_i^* \dot{A} \mu_i. \tag{8.19}$$

If we apply this to $A(t) := A + tB$, we see that whenver $A + tB$ has simple spectrum,

$$\frac{\mathrm{d}}{\mathrm{d}t}\lambda_i(A + tB) = \mu_i(A + tB)^* B \mu_i(A + tB)$$

the right-hand side is bounded in magnitude by $||B||_{op}$. So the map $t \mapsto \lambda_i(A + tB)$ is Lipschitz continuous with Lipschitz constant $||B||_{op}$ whenever $A + tB$ has simple spectrum. One result of it is 8.6.

By differentiate twice on 8.17 and 8.18, we have the *Hadamard second variation formula*

Proposition 8.11 (Hadamard second variation formula)**.**

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\lambda_k = \mu_k^* \ddot{A} \mu_k + 2 \sum i \neq k \frac{|\mu_j^* \dot{A} \mu_k|^2}{\lambda_k - \lambda_j}$$

whenever A has simple spectrum and $1 \leq k \leq n$.

8.6 Minors

Proposition 8.12 (Cauchy interlacing law)**.** For any $n \times n$ Hermitian matrix A_n with top left $n - 1 \times n - 1$ minor A_{n-1}, then

$$\lambda_{i+1}(A_n) \leq \lambda_i(A_{n-1}) \leq \lambda_i(A_n) \tag{8.20}$$

for all $1 \leq i < n$.
The space of A_n for which equality holds in one of inequality 8.20 has codimension 2 (for Hermitian matrices) or 1 (for real symmetric matrices).

Remark. If one takes successive minors $A_{n-1}, A_{n-2}, \dots, A_1$ of an $n \times n$ Hermitian matrix A_n, and computes their spectra, then 8.20 shows that this triangulary array of numbers froms a pattern know as *Gelfand-Tsetlin pattern*.

Proof. The right side is obvious from Courant-min-max theorem. The left hand side need to argue about the nontrivial intersection as Weyl inequality. □

Proposition 8.13 (Eigenvalue equation)**.** Let A_n be an $n \times n$ Hermitian matrix with top left $n-1 \times n-1$minor A_{n-1}. Suppose that λ is an eigenvalue of A_n distinct from all the eigenvalues of A_{n-1}. One have

$$\sum_{j=1}^{n-1} \frac{|\mu_j(A_{n-1})^* X|^2}{\lambda_j(A_{n-1}) - \lambda} = a_{nn} - \lambda \tag{8.21}$$

where a_{nn} is the bottom right entry of A, and $X = (a_{nj})_{j=1}^{n-1} \in \mathbb{C}^{n-1}$ is the right column of A.

Proof. Expand the eigenvalue function and notice that since λ is distinct from all the eigenvalues of minor A_{n-1}, the matrix $A_{n-1} - \lambda I$ is invertible and has eigenvalue $\lambda_i(A_{n-1}) - \lambda$ with eigenvectors $\mu_j(A_{n-1})$. □

The function $\lambda \mapsto \sum_{j=1}^{n-1} \frac{|\mu_j(A_{n-1})^* X|^2}{\lambda_j(A_{n-1}) - \lambda}$ is a rational function of λ which is increasing away from the eigenvalues of A_{n-1}, where it has a pole.[22] By graphing this function one can see that the interlacing formula can be interpreted as a manifestation of the intermediate value theorem.

[22] *In rare case when the inner product vanishes, it has a removable singularity.*

It also suggests that under typical circumstances, an eigenvalue of A_n can only get close to an eigenvalue $\lambda_j(A_{n-1})$ if the associated inner product $\lambda_j(A_{n-1})^* X$ is small. This is useful to achieve eigenvalue repulsion.

8.7 Singular values

Theorem 8.4 (Singular value decomposition)**.** Let $0 \leq p \leq n$, and let A be a linear transformation from n-dimensional complex Hilbert space U to a p-dimensional complex Hilbert space V. Then there exist non-negative real numbers

$$\sigma_1(A) \geq \cdots \geq \sigma_p(A) \geq 0$$

and orthonormal set $u_1(A), ...u_p(A) \in U$ and $v_1(A), \dots, v_p(A) \in V$ known as singular vectors of A, such that

$$Au_j = \sigma_j v_j; \quad A^* v_j = \sigma_j u_j$$

for all $1 \leq j \leq p$, where we abbreviate $u_j = u_j(A)$. $Au = 0$ whenever u is orthogonal to all $u_1(A), \dots, u_p(A)$.

By noticing $\sigma_i(A^*) := \sigma_i(A)$, there won't be any problem for us to work with a tall matrix.

Remark. The singular value of a matrix A are unique. And if we have $\sigma_1(A) > \cdots > \sigma_p(A) > 0$, show that the singular vectors are unique up to rotation by a complex phase.
Singular value is invariant under left or right unitary transformation.[23]
Singular value of square Herimitian matrix are simply the absolute values of its eigenvalues.

[23] $\sigma_i(UAV) = \sigma_I(A)$ *whenever A is a unitary $p \times n$ matrix, U is a unitary $p \times p$ matrix, and V is a unitary $n \times n$ matrix.*

If A is a $p \times n$ complex matrix for some $1 \leq p \leq n$, show that the *augmented matrix*

$$\tilde{A} := \begin{pmatrix} 0 & A \\ A^* & 0 \end{pmatrix}$$

is a $p+n \times p+n$ Hermitian matrix whose eigenvalues consist of $\pm\sigma_1(A), \ldots, \pm\sigma_p(A)$ together with $n-p$ copies of the eigenvalue zero. The singular vectors of is the union of left and right eigenvectors.

Proposition 8.14 (Courant-Fischer minimax formula)**.** Let A be a $p \times n$ complex matrix for some $1 \leq p \leq n$.

$$\sigma_i(A) = \sup_{\dim V = i} \inf_{v \in V; |v|=1} |Av| \tag{8.22}$$

for all $1 \leq i \leq p$.

The analogous inequality about singular value we had before for Hermitian matrix can be deduced from here.

PART

II

Concentration of measure

The basic intuition here is that it is difficult for a large number of independent variables $X_1, \ldots, X_n$ to "work together" to simultaneously pull a combination $F(X_1, \ldots, X_n)$ too far away from its mean.

We will focus on a specific application which is useful on controlling the behaviour of random n-dimensional vectors with independent components, and in particular on the distance between such random vectors and a given subspace.

9 Linear combinations, and the moment method

9.1 First and second moment method

Zero moment method a bound is usually useless

$$\mathbb{P}\{S_n \neq 0\} \leq \sum_{i=1}^{n} \mathbb{P}\{X_i \neq 0\} \tag{9.1}$$

First moment method gives

$$\mathbb{P}(|S_n| \geq \lambda) \leq \frac{1}{\lambda} \sum_{i=1}^{n} \mathbb{E}|X_i|. \tag{9.2}$$

S_n typically has size $S_n = O(\sum_{i=1}^n |X_i|)$. The decay is linear in λ.

Second moment method suggests that if we have pairwise independence[24].

[24] *It suffices to have covariance vanish for all distinct pair*

$$\mathbb{P}(|S_n - \mathbb{E}S_n| \geq \lambda) \leq \frac{1}{\lambda^2} \sum_{i=1}^{n} \mathbf{Var}(X_i) \tag{9.3}$$

S_n has size $S_n = \mathbb{E}S_n + O((\sum_{i=1}^n \mathbf{Var}(X_i))^{1/2})$. The decay is quadratic in λ.

The 9.3 is sharp. We cannot expect any significant concentration in any range narrower than the standard deviation, because this would contradict with the size of Variance of S_n. Second the quadratic-type decay in λ is sharp given the pairwise independence hypothesis.

Example. Suppose that $n = 2^m - 1$, and that $X_j := (-1)^{a_j \cdot Y}$ where Y is drawn uniformly at random from the cube $\{0,1\}^m$, and $a_1, \ldots, a_n$ are an anumeration of the non-zero elements of $\{0,1\}^m$. Each $X_j, 1 \leq j \leq n$ has mean zero, variance 1, and pairwise independence in j: but S is equal to $(n+1)\mathbf{I}(Y = 0-1)$ which is equal to n with probability $1/(n+1)$. And the standard deviation of S is just $\sqrt{n}$. Which mean 9.3 is sharp up to constants when $\lambda = n$.

9.2 Higher moment method

Remark. In short, control of each individual moment $\mathbb{E}|S_n|^k$ only gives polynomial decay in λ, the tail prob, by using all the moments simultaneously one can obtain sub-Gaussian decay.

For higher moments. Assume X_i are normalised to be mean zero and variance at most 1, and are almost surely bounded in magnitude by some $K : |X_i| \leq K, K \geq 1$. And X_i are real-valued.

Assume there are k-wise independence for some positive integer k, then we can estimate k^{th} moment of S_n

$$\mathbb{E}|S_n|^k = \sum_{1 \leq i_1, \ldots, i_k \leq n} \mathbb{E} X_{i_1} \ldots X_{i_k}.$$

To compute the expectation of right hand side, we need to divide into cases depending on how various indices are repeated.

If one of the X_{i_j} only appear once, the expectation is zero. So we may assume that each of the X_{i_j} appear at least twice.

There are at most $k/2$ distinct X_j which appear. If it happens exactly, then the expectation has magnitude 1. If $l/2-r$ terms appear, since the upper bound K is assumed, we see the expectation has magnitude at most K^{2r}. This leads to the upper bound

$$\mathbb{E}|S_n|^k = \sum_{r=0}^{k/2} K^{2r} N_r$$

where N_r is the number of ways one can select integers $i_1, \ldots, i_k$ in $\{1, \ldots, n\}$ such that each i_j appears at least twice, and such that exactly $k/2-r$ integers appear. And the problem become combinatorial problem of estimating N_r.

There are $\binom{n}{k/2-r} \leq n^{k/2-r}/(k/2-r)!$ ways[25] to choose $k/2-r$ integers from $\{1, \ldots, n\}$. Each of the integers i_j has to come from one of these $k/2-r$ integers, leading to the crude bound

$$N_r \leq \frac{n^{k/2-r}}{(k/2-r)!}(k/2-r)^k$$

which after using a crude form of Stirling formula $n! \geq n^n e^{-n}$ gives

$$N_r \leq (en)^{k/2-r}(k/2)^{k/2+r},$$

and so

$$\mathbb{E}|S_n|^k \leq (enk/2)^{k/2} \sum_{r=0}^{k/2} (\frac{K^2 k}{en})^r.$$

Then if we make the mild assumption

$$K^2 \leq n/k \tag{9.4}$$

then from the geometric series formula we conclude that

$$\mathbb{E}|S_n|^k \leq 2(enk/2)^{k/2}$$

which leads to the large deviation inequality

$$\mathbb{P}\{|S_n| \geq \lambda\sqrt{n}\} \leq 2\left(\frac{\sqrt{ek/2}}{\lambda}\right)^k. \tag{9.5}$$

Compare this with 9.2 9.3, we find higher moment we control, the rate of decay in tail λ improves. But the range S_n concentrates in grows slowly, to $O(\sqrt{nk})$ rather than $O(\sqrt{n})$.[26]

[25] *With more care on this in to improve a explicit constant in the final bounds.*

[26] *The range means the bound for expectation, the decay means the tail probability.*

Now if we assume that $X_1, \dots, X_n$ are not just k-wise independent for any fixed k, but jointly independent. This bound leads to sub-Gaussian bounds by setting[27] $\sqrt{nk}$ to a small multiple of λ.

$$\mathbb{P}\{|S_n| \geq \lambda\sqrt{n}\} \leq C\exp\{-c\lambda^2\} \tag{9.6}$$

for some absolute constant $C, c \geq 0$, provided that $|\lambda| \leq c\sqrt{n}/\sqrt{K}$.

[27] *Because 9.5 holds for any k such that $K^2 \leq n/k$*

9.3 Exponential moments

Remark. Exponential are exploits independence much better than power moment methods. But it relie heavily on commutativity of the underlying variables, $e^{X+Y} = e^X e^Y$. So in random matrices it will be less powerful. But we can still expect good effects when we apply them to various components of random matrices.

Lemma 9.1 (Hoeffding's lemma)**.** Let X be a scalar variable taking values in an interval $[a, b]$. then for any $t > 0$,

$$\mathbb{E}e^{tX} \leq e^{t\mathbb{E}X}(1 + O(t^2\mathbf{Var}(X\exp\{O(t(b-a))\}))) \tag{9.7}$$

In particular,

$$\mathbb{E}e^{tX} \leq e^{t\mathbb{E}X}\exp\{O(t^2(b-a)^2)\} \tag{9.8}$$

Sharply

$$\mathbb{E}e^{tX} \leq e^{t\mathbb{E}X}\exp\{t^2(b-a)^2/8\} \tag{9.9}$$

Proof. It suffices to prove the first inequality, as the second then follows using the bound $\mathbf{Var} \leq (b-a)^2$ and from variaous elementary esimates[28].
And by normalising X, we may assume $\mathbb{E}(X) = 0$ and $b - a = 1$, which implies that $X = O(1)$. We then have the Taylor expansion

$$e^{tX} = 1 + tX + O(t^2X^2\exp\{O(t)\})$$

Which after taking the expectation gives the claim.
And the version with sharp constant can be derived as follow. First notice that when we assume X has mean zero it means that $a \leq 0$ a.s.. And convexity of e^x gives

$$e^{TX} \leq \frac{b-x}{b-a}e^{\lambda a} + \frac{x-a}{b-a}e^{\lambda b}, \quad \forall x \in [a, b]$$

After taking expectation, and denote $h = \lambda(b-a), p = -a/(b-a) \geq 0$ and $L(h) = -hp + \ln\big(1 - p + pe^h\big)$, we have

$$\frac{b}{b-a}e^{\lambda a} + \frac{a}{b-a}e^{\lambda b} = e^{L(h)}$$

Then by Taylor expansion and $L(0) = L'(0) = 0$,

$$L(h) \leq \frac{1}{8}h^2$$

Then the claim is proved. □

[28] *$O(t^2(b-a)^2\exp\{O(t(b-a))\}) = O(t^2(b-a)^2) + \exp\{O(t(b-a))\}$, then use $1 + x \leq e^x$ and combine the exponential term.*

Theorem 9.2 (Chernoff inequality)**.** Let $X_1, \dots, X_n$ be independent scalar random variables with $|X_i| \leq K$ almost surely, with mean μ_i and variance σ^2. Then for any

$\lambda \geq 0$, one has

$$\mathbb{P}(|S_n - \mu| \geq \lambda\sigma) \leq C \max(\exp\bigl(-c\lambda^2\bigr), \exp(-c\lambda\sigma/K))$$

for some absolute constants $C, c > 0$, where $\mu := \sum_{i=1}^n \mu_i$ and $\sigma^2 := \sum_{i=1}^n \sigma_i^2$.

Here the term $\exp(-c\lambda\sigma/K)$ can be replaced with $(\lambda K/\sigma)^{-c\lambda\sigma/K}$, which is superior when $\lambda K \gg \sigma$.

The Chernoff inequality asserts that S_n is sharply concentrated in the range $\mu + O(\sigma)$. The bound is sharp when λ is not too large, one can see that from the next example.

Example. Let $0 \leq p \leq \frac{1}{2}$ be fixed independently of n, and let $X_1, \ldots, X_n$ be idd copies of a Bernoulli random variable that equals 1 with probability p, thus $\mu_i = p$ and $\sigma_i^2 = p(1-p)$, so $@mu = np$ and $\sigma^2 = np(1-p)$. With Stirling formula we have

$$\mathbb{P}\{|S_n - \mu| \geq \lambda\sigma\} \geq c\exp\{-C\lambda^2\}$$

for some absolute constants $C, c > 0$ and all $\lambda \leq c\sigma$.

Proposition 9.1 (Hoeffding's inequality)**.** Let $X_1, \ldots, X_n$ be independent real variables, with X_i taking values in an interval $[a_i, b_i]$, and let $S_n := X_1 + \cdots + X_n$. One has

$$\mathbb{P}\{|S_n - \mathbb{E}S_n| \geq \lambda\sigma\} \leq C\exp\bigl(-c\lambda^2\bigr)$$

for some absolute constants $C, c > 0$, where $\sigma^2 := \sum_{i=1}^n |b_i - a_i|^2$

It means the independent random variable bounded in a same interval concentrates in sub-Gaussian way.

Theorem 9.3 (Azuma's inequality)**.** Let $X_1, \ldots, X_n$ be a sequence of scalar random variables with $|X_i| \leq 1$ almost surely. Assume also that we have the marginale difference property

$$\mathbb{E}(X_i | X_1, ..., X_{i-1}) = 0 \tag{9.10}$$

almost surely for all $i = 1, \ldots, n$. Then for any $\lambda > 0$, the sum $S_n := X_1 + \cdots + X_n$ obeys the large deviation inequality

$$\mathbb{P}\{|S_n| \geq \lambda\sqrt{n}\} \leq C\exp\{-c\lambda^2\} \tag{9.11}$$

for some absolute constants $C, c > 0$.

This means the full assumption of joint independence is not completely necessary for Chernoff-type bounds to be present. It suffices to have a weaker condition called *martingale difference sequence*, where dependence is allowed.

Proof. We assume X_i is real because we can take real and imaginary part. So it suffices to establish the upper tail estimate

$$\mathbb{P}\{S_n \geq \lambda\sqrt{n}\} \leq C\exp\{-c\lambda\}$$

Note that $|S_n| \leq n$ almost surely, so we may assume, WLOG that $\lambda \leq \sqrt{n}$.
We consider the exponential moment $\mathbb{E}\exp(tS_n)$ for some parameter $t > 0$. We write $S_n = S_{n-1} + X_n$. Then though we cannot split the expectation but we can condition

on the $X_1, \dots, X_{n-1}$ instead. After conditioning we can pull out $\exp(tS_{n-1})$.
Then apply Hoeffding's 9.8 to prove that

$$\mathbb{P}\{S_n \geq \lambda\sqrt{n}\} \leq \exp\big(O(nt^2) - t\lambda\sqrt{n}\big)$$

Then optimising in t.

□

9.4 Summary

Summary of inequality to control a sum S_n.

1. The zeroth moment method 9.1 requires no moment assumptions but is only useful when X_i is usually zero and has no decay in λ.
2. The first moment method 9.2 only requires abosolute integrability on X_i but has only a linear decay in λ.
3. The second moment method 9.3 requires second moment and pairwise independence bounds on X_i and gives a quadratic decay in λ.
4. Higher moment bounds 9.5 requires boundedness and k-wise independence, and give a k^{th} power decay in λ(Or quadratic exponential decay after optimising in k).
5. Exponential moment bounds 9.2 and 9.3 require boundedness and joint independence (or martingale behaviour) and gives quadratic-exponential decay in λ.

10 The truncation method

As we see from above, the strongest decay in λ require strong boundedness and independence hypothese.

But one can often partially extend these strong results from the case of bounded random variables to that of unbounded random variables (provided one still has strong decay of these variables) by *truncation method*.

The basic idea here is to take each random variable X_i and split it as $X_i = X_{i,leqN} + X_{i,>N}$, where N is a truncation parameter to be optimisied later(possibly in a manner depending on n).

The idea is then to estimate the tail of $S_{n,\leq N}$ and $S_{n,>N}$ by two different means. With $S_{n\leq N}$, the large deviation inequality is in role. With $S_{n>N}$, the underlying variables are not bounded, but they tend to have small zeroth and first moments, and so the bounds based on those moment methods tend to be powerful.

Theorem 10.1 (Weak Law of large numbers)**.** Let $X_1, X_2, \dots$ be iid scalar random variables with $X_i \equiv X$ for all i, where X is absolutely integrable. Then S_n/n converges in probability to $\mathbb{E}X$.

Proof. WLOG assume X has mean zero. Our task is then to show that $\mathbb{P}(|S_n| \geq \epsilon n) = o(1)$ for all fixed $\epsilon > 0$.
If X has finite variance, the large deviation inequality 9.3 derive the claim.
If X do not have finite variance, we perform a truncation method.
Let N be a large parameter to be chosen later, then split $X_i = X_{i,\leq N} + X_{i,>N}$. The variable $X_{i,\leq N}$ is bounded so has finite variance. From the dominated convergence theorem 6.1 we see that $|\mathbb{E}X_{\leq N}| \leq \epsilon/4$ if N is large enough. And apply the large

deviation inequality to S_n we have

$$\mathbb{P}\{|S_{n,\leq N}| \geq \epsilon n/2\} = o(1)$$

where the rate of decay here depends on N and ϵ.
For tail $X_{>N}$ we use the first moment method 9.2

$$\mathbb{P}\{|S_{n,>N} \geq \epsilon n/2\} \leq \frac{2}{\epsilon}\mathbb{E}|X_{>N}|.$$

But by dominated convergence theorem we may make $\mathbb{E}|X_{>N}$ as small as we please, smaller that $\delta > 0$ by taking N large enough.
Conbine we have

$$\mathbb{P}\{|S_n| \geq \epsilon n\} = \frac{2}{\epsilon}\delta + o(1)$$

since δ is arbitrary small, we obtain the claim. □

We can use the trick of *sparsification trick* to prove the strong law of large numbers.

Theorem 10.2 (Strong law of large numbers)**.** Let $X_1, X_2, \ldots$ be iid scalar random variables with $X_i \equiv X$ for all i, where X is absolutely integrable. Then S_n/n converges almost surely to $\mathbb{E}X$.

Proof. WLOG we assume X is real. By splitting X into positive and negative parts, we futhermore assume X is non-negative. In such case we won't assume it's mean zero, non-negative bring S_n to be non-decreasing.
Then once X is non-negative, we see that the empirical averages S_n/n cannot decrease too quickly in n. In particular we observe that [29]

$$S_m/m \leq (1 + O(\epsilon))\ S_n/n \text{ whenever } (1-\epsilon)n \leq m \leq n.$$

Next, we apply a sparsification trick. Let $0 < \epsilon < 1$. Suppose we knew that, almost surely, S_{n_m}/n_m converged to $\mathbb{E}X$ for $n = n_m$ of the form $n_m := \lfloor (1+\epsilon)^m \rfloor$ for some integer m. Then, for all other values of n, we see that asymptotically, S_n/n can only fluctuate by a multiplicative factor of $1 + O(\epsilon)$ because of the monotone nature of S_n [30]. Which is also illustrated above.
And because above and countable additivity, it suffices to show that S_{n_m}/n_m converges to $\mathbb{E}X$ [31]. Actually it will be enough to show that almost surely, one has $|S_{n_m}/n_m - \mathbb{E}X| \leq \epsilon$ for all but finitely many m. Because we can make m large enough for S_{n_m}/n_m to first get close to $\mathbb{E}X$.
Fix ϵ. Split $X = X_{\leq N_m} + X_{>N_m}$ and $S_{n_m} = S_{n_m,>N_m} + S_{n_m,\leq N_m}$ but we now allow $N = N_m$ to depend on m.
By dominate convergence theorem 6.1, for N_m large enough we have $|EX_{\leq N_m} - \mathbb{E}X| \leq \epsilon/2$.
Apply second moment method 9.3, we see that

$$\mathbb{P}\{|S_{n_m,\leq N_m/n_m} - \mathbb{E}X| > \epsilon\} \leq \frac{C_\epsilon}{n_m}\mathbb{E}|X_{\leq N_m}|^2$$

for some C_ϵ depending only on ϵ. Then apply the first moment method to handle the tail, we see that

$$\mathbb{P}\{|S_{n_m}/n_m - \mathbb{E}X| > \epsilon\} \leq \frac{C_\epsilon}{n_m}\mathbb{E}|X_{\leq N_m}|^2 + n_m\mathbb{P}\{|X| > N_m\}.$$

[29] *By expansion of geometric series.*

[30] *Check the n between n_m, n_{m+1}, we see the where the multiplicity comes from.*

[31] *Then take $\epsilon \to 0$ ny setting $\epsilon = 1/j$ to obtain the claim.*

By Borel Cantelli lemma 6.3, if suffices to prove that we can choose N_m such that

$$\sum_{m=1}^{\infty} \frac{C_\epsilon}{n_m} \mathbb{E}|X_{\leq N_m}|^2, \quad \sum_{m=1}^{\infty} n_m \mathbb{P}\{|X| > N_m\}$$

are both finite. [32] □

[32] *Then, almost surely, there are at most finitely many m for which the inequality holds*

Proposition 10.1. Let $X_1, \ldots, X_n = X$ be iid copies of sub-Gaussian random variable X, thus X obeys a bound of the form

$$\mathbb{P}\{|X| \geq t\} \leq C \exp\left(-ct^2\right) \tag{10.1}$$

for all $t > 0$ and some $C, c > 0$. Let $S_n := X_1 + \cdots + X_n$. Then for any suffiently large A independent of n, we have

$$\mathbb{P}\{|S_n - n\mathbb{E}X| \geq An\} \leq C_A \exp\{-c_A n\}$$

for some constants C_A, c_A depending on A, C, c. Furthermore, c_A grows linearly in A as $A \to \infty$.

Proof. We assume X is normalised to have 0 mean.
We perform a dyadic decomposition

$$X_i = X_{i,0} + \sum_{m=1}^{\infty} X_{i,m}$$

where $X_{i,0} := X_i \mathbf{I}(X_i \leq 1)$ and $X_{i,m} := X_i \mathbf{I}(2^{m-1} < X_i \leq 2^m)$. We similarly split S_n

$$S_n = S_{n,0} + \sum_{m=1}^{\infty} S_{n,m}$$

where $S_{n,m} = \sum_{i=1}^n X_{i,m}$. Then by the union bound and the pigeonhole principle[33] we have

$$\mathbb{P}(|S_n| \geq An) \leq \sum_{m=0}^{\infty} \mathbb{P}\left(|S_{n,m}| \geq \frac{A}{100(m+1)^2} n\right).$$

(say).
Which basically means we need to prove any upper bound of LHS.
Each $X_{i,m}$ is clearly bounded in magnitude by 2^m; from the sub-Gaussian hypothesis one can verify that the mean and variance of $X_{i,m}$ are at most $C' \exp\{-c' 2^{2m}\}$ for some $C', c' > 0$. If A is large enough, an application of the Chernoff bound gives By choosing $\lambda = c'' A \exp\{c' 2^{2m}/2\}$ for some small $c'' > 0$.[34]

$$\mathbb{P}\{|S_{n,m}| \geq 2^{-m-1} An\} \leq C' 2^{-m} \exp\{-c' An\}$$

(say) for some $C'c' > 0$, and the claim follows. □

[33] *By splitting An to different hole, there must be at least one hole that $S_{n,i}$ is larger or in the whole.*

[34] *I have no idea about how to make this calculation.*

Sub-Gaussian assumption can be generalised to a sub-exponential tail hypothesis

$$\mathbb{P}\{|X| \geq t\} \leq C \exp\{-ct^p\}$$

provided $p > 1$.

11 Lipschitz combinations

Theorem 11.1 (McDiarmid's inequality)**.** Let $X_1, \dots, X_n$ be independent random variables taking values in range $R_1, \dots, R_n$ and let $F : R_1 \times \cdots \times R_n \to \mathbf{C}$ be a function with the property that if one freezes all but the i^{th} coordinate of $F(x_1, \dots, x_n)$ for some $1 \le i \le n$, then F only fluctuates by at most $c_i > 0$, thus

$$|F(x_1, \dots, x_{i-1}, x_i, x_{i+1}, \dots, x_n) - F(x_1, \dots, x_{i-1}, x_i', x_{i+1}, \dots, x_n)| \le c_i$$

for all $x_j \in X_j, x_i' \in X_i$ for all $i \le j \le n$. Then for any $\lambda > 0$, one has

$$\mathbb{P}\{|F(X) - \mathbb{E}F(X)| \ge \lambda\sigma\} \le C \exp\{-c\lambda^2\}$$

for some absolute constants $C, c > 0$, where $\sigma^: = \sum_{i=1}^n c_i^2$.

Proof. We may assume that F is real. By symmetry, it suffices to show that one-sided estimate

$$\mathbb{P}\{F(X) - \mathbb{E}F(X) \ge \lambda\sigma^2\} \le C \exp\{-c\lambda^2\}.$$

Use exponential method using $Y := F(X) - \mathbb{E}(F(X)|X_1, .., X_{n-1})$ write

$$\mathbb{E}\{\exp(tF(X))|X_1, \dots, X_{n-1}\} = \mathbb{E}\{\exp(tY)|X_1, \dots, X_{n-1}\} \exp\{t\mathbb{E}(F(X)|X_1, \dots, X_{n-1})\}$$

Then apply Hoeffding's 9.8 with the property of F and integrate out conditioning to see

$$\le \exp\big(O(t^2 c_n^2)\big)\mathbb{E} \exp\{t\mathbb{E}(F(X)|X_1, \dots, X_{n-1})\}$$

Noticing $\mathbb{E}(F(X)|X_1, \dots, X_{n-1})$ is actually another function F_{n-1} have the property like F but hold for $n-1$ numebr of parameters. Then iterate the computation n times,

$$\le \exp\{\sum_{i=1}^n O(t^2 c_i^2)\} \exp\{t\mathbb{E}F(X)\}$$

Then use Markov's inequality and rearrangements

$$\mathbb{P}(F(X) - \mathbb{E}F(X) \ge \lambda\sigma) \le \exp\big(O(t^2\sigma^2) - t\lambda\sigma\big)$$

and optimise on t to obtain the claim.

□

McDiarmid implies Hoeffding's inequality 9.1, and it is a tensorisation of Hoeffding's lemma 9.1. By applying this trick to random variables taking value in more sophisticated metric spaces than an interval leading to a class of concentration of measure know as *transportation cost-information inequality.*

The more powerful concentration of measure results do not just exploit Lipschitz type behaviour in each individual variable, but joint Lipschitz behaviour.

Proposition 11.1 (Gaussian random variable)**.** Linear combination of Gaussian variables is a still Gaussian.
Gaussian random variables are invariant under rotation.

Lemma 11.2 (Measure-theoretic Jensen's inequality)**.** Let $a, b \in \mathbb{R}, f : [a, b] \to \mathbb{R}$ is non-negative Lebesgue integrable function, ϕ is a convex function on the real line.

The Lebesgue measure of $[a,b]$ need not to be unity. We can rescale to make the measure unity.

$$\phi(\frac{1}{b-a}\int_a^b f(x)\mathrm{d}x) \leq \frac{1}{b-a}\int_a^b \phi(f(x))\mathrm{d}x$$

Theorem 11.3 (Gaussian concentration inequality for Lipschitz function)**.** Let $X_1,\dots,X_n \equiv N(0,1)_{\mathbb{R}}$ be iid real Gaussian variables, and let $F:\mathbb{R}^n \to \mathbb{R}$ be a 1-Lipschitz function. Then for any λ one has

$$\mathbb{P}\{|F(X)-\mathbb{E}F(X)| \geq \lambda\} \leq C\exp\bigl(-c\lambda^2\bigr)$$

for some absolute constants $C,c>0$.

Proof. Centering does not harm, mean zero assumed. By symmetry it suffices to show the upper tail estimate

$$\mathbb{P}(F(X) \geq \lambda) \leq C\exp\bigl(-c\lambda^2\bigr).$$

The Lipschitz bound on F implies the gradient estimate

$$|\nabla F(x)| \leq 1$$

for all $x \in \mathbb{R}^n$.
Then use the exponential moment method, it suffices to show that

$$\mathbb{E}\exp\{tF(X)\} \leq \exp\{Ct^2\}$$

for some constant $C>$ and all $t>0$, then use Markov's inequality and optimisation in t gaves us the claim.
To exploit Lipschitz nature of F, we will need to introduce a second copy of $F(X)$. *Since we can expect the exponential moment is bounded by the exponential moment of their difference by independence, convexity and mean zero.(together with Jensen).* Then the difference will be bounded by the integral along a segament between $F(X),F(Y)$. [35]
Here goes the next technique. We do not integrate along the segament, instead we integrate along an arc

$$F(X)-F(Y) = \int_0^{\pi/2} \frac{\mathrm{d}}{\mathrm{d}\theta}F(Y\cos\theta + X\sin\theta)\mathrm{d}\theta.$$

The reason is that after all we can exploit Lipschitz property once we have the difference using integral on any curve connecting two points. Why don't we choose an orbit exploit the nature of the random variable better?
$X_\theta := Y\cos\theta + X\sin\theta$ is another Gaussian random variable equivalent to X[36]. And it's derivative $X'_\theta := -Y\sin\theta + X\cos\theta$ is the same. And most importantly, these two random variables are independent.[37]
Then apply general Jensen's 11.2. and chain rule and expectation obtain

$$\mathbb{E}\exp\{t(F(X)-F(Y))\} \leq \frac{2}{\pi}\int_0^{\pi/2} \mathbb{E}\exp\left(\frac{\pi t}{2}\nabla F(X_\theta)\cdot X'_\theta\right)$$

Conditioning on X_θ then estimate[38], and then integrate out the conditioning. □

[35] *Having the difference is the key to exploit the Lipschitz property.*

[36] *Same mean, variance*

[37] *Part of the reason is because they are Gaussian. This really exploit the nature of Gaussian.*

[38] *By independence.*

Gaussian concentration is the same thing as the blow-up phenomenoen.
11.3 is equivalent to the inequality

$$\mathbb{P}(X \in A)\mathbb{P}(X \notin A_\lambda) \leq C\exp\bigl(-c\lambda^2\bigr)$$

holding for all $\lambda > 0$ and all measurable sets A, where $X = (X_1, \ldots, X_n)$ is an $\mathbb{R}^n$-valued random variable with iid Gaussian components $X_1, \ldots, X_n \equiv N(0,1)_{\mathbb{R}}$, and A_λ is the λ-neighbourhood of A.

To prove this one need to first replace the expectation in to median. On oneside (from the theorem to this), first assuming A has probability measure larger than $1/2$ (if not just take the completion as A). Then set $F(x) := d(x, A)$ to obtain the bound. On the other side (from this to the theorem), Let $A := F(X) \leq \mathbf{M}F(X)$ to conclude the bound.

Theorem 11.4 (Talagrand concentration inequality)**.** Let $K > 0$, and let $X_1, \ldots, X_n$ be independent complex variables with $|X_i| \leq K$ for all $1 \leq i \leq n$. Let $F : \mathbb{C} \to \mathbb{R}$ be a 1-Lipschitz convex function (where we identify C^n with $\mathbb{R}^{2n}$ for the purpose of defining "Lipschitz" and "convex"). Then for any λ one has

$$\mathbb{P}\{|F(X) - \mathbf{M}F(X)| \geq \lambda K\} \leq C\exp\bigl(-c\lambda^2\bigr)$$

and

$$\mathbb{P}\{|F(X) - \mathbb{E}F(X)| \geq \lambda K\} \leq C\exp\bigl(-c\lambda^2\bigr)$$

for some absolute constants $C, c > 0$, where $\mathbf{M}F(X)$ is a median of $F(X)$.

The proof of Talagrand was based on a technique called combinatorial distance which I really didn't get it. I will try to read other material for the proof since he said this theorem will be relied based on.

I will first go for the proof using the Log-Soblev from Ledoux. This will give us the upper bound but not the lower bound since the symmetry was gone because of assumption that F is convex.

Theorem 11.5 (Log-Sobolev inequality)**.** Let $F : \mathbb{C}^n \to \mathbb{R}$ be a smooth convex function. Then

$$\mathbb{E}F(X)e^{F(X)} \leq (\mathbb{E}e^{F(X)})(\log \mathbb{E}e^{F(X)}) + C\mathbb{E}e^{F(X)}|\nabla F(X)|^2$$

for some absolute constant C (independent of n).

Remark. If one set $f := e^{F/2}$ and normalises $\mathbb{E}f(X)^2 = 1$, this inequality becomes

$$\mathbb{E}|f(X)|^2 \log|f(X)|^2 \leq 4C\mathbb{E}|\nabla f(X)|^2$$

which more closely resembles the classical log-Sobolev inequality. The constant C here can be taked to be 2.

Proof. We first establish the 1-dimensional case. If we let Y be an independent copy of X, observe that left-hand side can be rewritten as

$$\frac{1}{2}\mathbb{E}(F(X) - F(Y))(e^{F(X)} - e^{F(Y)}) + (\mathbb{E}F(X))(\mathbb{E}e^{F(X)}).$$

From Jensen's inequality we have

$$\mathbb{E}F(X) \leq \log \mathbb{E}e^{F(X)}$$

So it suffices to prove the estimation of the change term,

$$\frac{1}{2}\mathbb{E}(F(X)-F(Y))(e^{F(X)}-e^{F(Y)}) \leq C\mathbb{E}e^{F(X)}|\nabla F(X)|^2$$

X, Y is bounded, F is smooth and convex [39] we have the following when $F(X) \geq F(Y)$

$$\begin{aligned} F(X)-F(Y) &= O(|\nabla F(X)|) \\ e^{F(X)}-e^{F(Y)} &= O(|\nabla F(X)|e^{F(X)}) \end{aligned}$$

Hence by taking the expectation we obtain the claim. Vice versa.
Induct on n to show the general case,[40]
Write $X = (X', X_n)$, where $X' := (X_1, \ldots, X_{n-1})$. Denote the $n-1$ dimensional gradient and $f(X_n) := \log \mathbb{E}(e^F(X)|X_n)$.

$$\mathbb{E}(F(X)e^{F(X)}|X_n) \leq f(X_n)e^{f(X_n)} + C\mathbb{E}(e^{F(X)}|\nabla' F(X)|^2\big|X_n)$$

where the second expectation is taken over X_n.
Conditioning leaves $\log \mathbb{E}(e^{F(X)}|X_n)$ unsetteled. Integrate the conditioning out and we find we need to further estimate the term $\mathbb{E}f(X_n)e^{F(X_N)}$.

$$\mathbb{E}(F(X)e^{F(X)}) \leq \mathbb{E}f(X_n)e^{f(X_n)} + C\mathbb{E}(e^{F(X)}|\nabla' F(X)|^2)$$

First f is still convex. (By Holder's and Convexity of F)

$$\begin{aligned} (1-t)f(X_n)+tf(X_n) &= \log \mathbb{E}(e^{F(X)}|X_n)^{1-t} \cdot \mathbb{E}(e^{F(X)}|Y_n) \\ &\geq \log \mathbb{E}(e^{(1-t)F(X,X_n)+tF(X,Y_n)}|X_n, Y_n) \\ &\geq f((1-t)X_n + tY_n) \end{aligned}$$

By the result from $n = 1$

$$\mathbb{E}f(X_n)e^{f(X_n)} \leq (\mathbb{E}e^{F(X)})(\log \mathbb{E}e^{F(X)}) + C\mathbb{E}e^{f(X_n)}|f'(X_n)|^2$$

We see that the the only thing left here is to estimate $e^{f(X_n)}|f'(X_n)|^2$. By the chain rule,

$$\begin{aligned} e^{f(X_n)}|f'(X_n)|^2 &= e^{-f(X_n)}|\mathbb{E}_{X'}e^{F(X)}F_{X_n}(X)|^2 \\ &\leq e^{-f(X_n)}|\mathbb{E}_{X'}e^{F(X)}|^2|F_{X_n}(X)|^2 \\ &\leq \mathbb{E}e^{F(X)}|F_{x_n}(X)|^2 \end{aligned}$$

Taking the expectation then the claim follows. □

Now the proof for onesided Talagrand's inequality follows

Proof. To prove the Talagrand's concentration inequality 11.4, we let F to be convex and 1-Lipschitz. Applying the Log-Soblev inequality 11.5 to tF for any $t > 0$, we conclude

$$\mathbb{E}tF(X)e^{tF(X)} \leq (Ee^{tF(X)})(\log \mathbb{E}e^{tF(X)}) + Ct^2\mathbb{E}e^{tF(X)}$$

Setting $H(t) := \mathbb{E}e^{tF(X)}$, we can rewrite this as a differential inequality [41]

$$tH'(t) \leq H(t)\log H(t) + Ct^2 H(t)$$

[39] *bounded from below by its tangent line.*

[40] *Keeping care to ensure that the constant C does not change in this induction process.*

[41] *Log-Sobolev seems to make sense in this typical differential equation arguments.*

which we can rewrite as

$$\frac{\mathrm{d}}{\mathrm{d}t}(\frac{1}{t}\log H(t)) \leq C.$$

From Taylor expansion

$$\frac{1}{t}(\log \mathbb{E}e^{tF(X)}) = \frac{1}{t}\left(0 + t\frac{\mathbb{E}F(X)e^{tF(X)}}{\mathbb{E}e^{tFX}} + o(t^2)\right)$$

we see that

$$\frac{1}{t}\log H(t) \to \mathbb{E}F(X)$$

as $t \to \infty$, and thus

$$\frac{1}{t}\log H(t) \leq \mathbb{E}F(X) + Ct$$

for any $t > 0$.

$$\mathbb{E}e^{tF(X)} \leq \exp\big(t\mathbb{E}F(X) + Ct^2\big)$$

Then by Markov 9.2, we conclude that

$$\mathbb{P}\{F(X) - \mathbb{E}F(X) > \lambda\} \leq \exp\big(Ct^2 - t\lambda\big).$$

Optimising in t gives one-side Talagrand's for convex function. □

The same argument, starting with Gross's log-Sobolev inequality for the Gaussian measure gives the upper tail component of Talagrand's concentration with no convexity hypothesis on F. The situation is now symmetry with respect to reflection $F \mapsto -F$, so we can have upper and lower bound at the same time.

This method of obtaining concentration inequalities from log-Sobolev inequality (Poincare-type inequality) by combining the latter with the exponential moment method is know as *Herbst's argument*.

Theorem 11.6 (Distance between random vector and a subspace)**.** Let $X_1, \ldots, X_n$ be independent complex-valued random variables with mean zero and variance 1, and bounded almost surely in magnitude by K. Let V be a subspace of $\mathbb{C}^n$ of dimension d. Then for any $\lambda > 0$, one has

$$\mathbb{P}\{|d(X, V) - \sqrt{n-d}| \geq \lambda K\} \leq C\exp\big(-c\lambda^2\big)$$

for some absolute constants $C, c > 0$.

This can be interpreted as an assertation which claim the distance between a random vector X and an arbitraty subspace V is typically equal to $\sqrt{n - \dim V} + O(1)$.

Proof. The function $z \mapsto d(z, V)$ is convex and 1-Lipschitz. From Talagrand's concentration 11.4, one has

$$\mathbb{P}\{|d(X, V) - \mathbf{M}d(X, V) \geq \lambda K\} \leq C\exp\big(-c\lambda^2\big).$$

So it suffices to show that

$$\mathbf{M}d(X, V) = \sqrt{n-d} + O(K)$$

This can be done with a second moment calculation,

$$d(X,V)^2 = ||\pi(X)||^2$$

where π is the orthogonal projection to the $V^\perp$. By spectrum theorem[42], we can see that $\mathbb{E}d(X,V)^2 = \mathrm{tr}(\pi) = n-d$. For some concentration of around the mean of $d(X,V)^2$, we compute the variance of it.

$$d(X,V)^2 - \mathbb{E}d(X,V)^2 = \sum_{1\le i,j\le n} p_{ij}(X_i\bar{X}_j) - \delta_{ij}$$

The summand here are pairwise uncorrelated for $1 \le i < j \le n$. And they are independent hence uncorrelated for $1 \le i = j \le n$. Each summand also has a variance of $O(K^2)$. So we have the variance bound

$$\mathbf{Var}(d(X,V)^2) = O(K^2 \sum_{1\le i,j\le n} |p_{ij}|^2) + O(K^2 \sum_{1\le i\le n} |p_{ij}|^2) = O(K^2(n-d))$$

where p_{ij} is the components of π. The first term has bound $O(K^2 n)$ while the second has $O(K^2(n-k))$.

Then from Chebyshev's inequality 9.3, the median is equal to $n-d+O(\sqrt{K^2(n-d)})$, which implies taking square roots[43] that the median of $d(X,V)$ is $\sqrt{n-d}+O(K)$ as desired. □

[42] *One can write orthogonal projection as $\pi = EE^T$, where E has column of orthonormal basis e_i. Or $E(E^TE)^{-1}E^T$.*

[43] *Try to view $n-d = \sqrt{n-d}^2$.*

PART

III

Central limit theorem

Consider the sum $S_n := X_1 + \cdots + X_n$ of iid random variable $X_1, \ldots, X_n \equiv X$ of finite mean μ and variance σ^2. Then we can expect S_n has size $n\mu + O(\sqrt{n}\sigma)$. The *normalized sum*

$$Z_n := \frac{S_n - n\mu}{\sqrt{n}\sigma}, \tag{11.1}$$

By Chebyshev's inequality 9.3 we have

$$\mathbb{P}\{|Z_n| > \lambda\} \leq \lambda^{-2} \tag{11.2}$$

The Z_n has quadratic decay in tail or exponential if X is sub-Gaussian.

Theorem 11.7 (Central limit theorem)**.** Let $X_1, \ldots, X_n \equiv X$ be iid real random variables of finite mean μ and variance σ^2 for some $\sigma > 0$, and let Z_n be the normalized sum. Then as $n \to \infty$, Z_n converges in distribution to the standard normal distribution $N(0,1)_{\mathbb{R}}$.

Z_n does not converge in probability or a.s.. Because two very different values $n_1 \ll n_2$, the quantities Z_{n_1} and Z_{n_2} are almost independent of each other.[44]

Central limit theorem gives control on random walks, and can be veiwewd as a "commutative" analogue of various spectral results in random matrix theory. Wigner semicircle law can be viewed as a non-commutative or free version of the central limit theorem.

[44] *Let $n_2 = 2n_1$ shall obtain the claim.*

12 The Fourier method for CLT

12.1 Proof reductions for CLT

We can normalize X to have 0 mean and unit variance, in which case Z_n simplifies to

$$Z_n = \frac{X_1 + \cdots + X_n}{\sqrt{n}}. \tag{12.1}$$

And it suffices to prove the central limit theorem for bounded random variables X.

Lemma 12.1 (Linearty of convergence)**.** Let V be a finite-dimensional real or complex vector space, X_n, Y_n be sequences of V-valued random variables(not necessarily independent), and let X, Y be another pair of V-valued random variables. Let c_n, d_n be scalars converging to c, d respectively.

1. If X_n converges in distribution to X, and Y_n converges in distribution to Y, and at least one of X, Y is deterministic. Then

$$c_n X_n + d_n Y_n \to^d cX + dY.$$

2. If X_n converges in probability to X, Y_n converges in probability to Y,

$$c_n X_n + d_n Y_n \to^{\mathbb{P}} cX + dY.$$

3. If X_n converges almost surely to X, and Y_n converges almost surely to Y,

$$c_n X_n + d_n Y_n \to^{a.s.} cX + dY.$$

Now if we have the central limit theorem for bounded random variables. Let X be an normalized unbounded random variable, we apply truncation technique to it.

Let $N = N_n > 0$ be a truncation parameter depending on n to be optimised later. Split X in the usual fashion $X_{\leq N} + X_{>N}$, $S_n = S_{n,\leq N} + S_{n,>N}$.

Let $\mu_{\leq N}, \sigma^2_{\leq N}$ be the mean and variace of $X_{\leq N}$. As our assumption,

$$Z_{n,\leq} := \frac{S_{n,\leq N} - n\mu_{\leq N}}{\sqrt{n}\sigma_{\leq N}}$$

converges in distribution to $N(0,1)_{\mathbb{R}}$.

Lemma 12.2 (Diagonalization principle). $\forall x^i \in \mathbb{R}^\infty, i = 1.2, \dots$ if $|x^j_j| \leq C$ then there exist a subsequence x^{i_k} s.t. $\lim_{k\to\infty} x^{i_k}_j = x_j$.

Then there exist a sequence going (slowly) to infinity with n, such that $Z_{n,\leq N_n}$ still converges in distribution to $N(0,1)_{\mathbb{R}}$.

For such a sequence, we see from Lebesgue dominated convergence 6.1 that $\sigma_{\leq N_n}$ converges to $\sigma = 1$. So

$$\frac{S_{n,\leq N_n} - n\mu_{\leq N_n}}{\sqrt{n}} \to^d N(0,1)_{\mathbb{R}}.$$

Meanwhile from Lebesgue dominated convergence theorem 6.1, $\sigma_{\leq N_n} \to 0$. And since

$$\mathbb{P}\{|Z_n| > \lambda\} \leq \lambda^{-2}$$

By replacing λ to N_n, we see that

$$\frac{S_{n,>N_n} - n\mu_{>N_n}}{\sqrt{n}} \to^d 0.$$

And from the linearity of expectation we have $\mu_{\leq N_n} + \mu_{>N_n} = \mu = 0$. Summing up, by the linearity of convergence, we obtain the claim.

12.2 Proof With Fourier method

Definition 12.1 (characteristic function). Given any real random variable X, we introduce the characteristic function $F_X : \mathbb{R} \to \mathbb{C}$, defined by

$$F_X(t) := \mathbb{E}e^{itX} \tag{12.2}$$

Equivalently, F_X is the Fourier transform of the probability measure μ_X.
More generally, for a random variable X taking values in a real vector space $\mathbb{R}^d$, we define the characteristic function $F_X : \mathbb{R}^d \to \mathbb{C}$ by

$$F_X(t) := \mathbb{E}e^{it\cdot X} \tag{12.3}$$

Where $\cdot$ denotes the Euclidean inner product on $\mathbb{R}^d$.
One can define the characteristic function on complex vector space $\mathbb{C}^d$ by using the complex inner product

$$(z_1, \dots, z_d) \cdot (w_1, \dots, w_d) := \mathrm{Re}(z_1\bar{w}_1 + \cdots + z_d\bar{w}_d).$$

The characteristic function is bounded in magnitude by 1 and equals 1 at the origin. By the Lebesgue dominate convergence theorem, F_X is continuous in t. And it's uniformly continuous.

Lemma 12.3 (Riemann-Lebesgue lemma)**.** If X is an absolute continuous random variable taking values in $\mathbb{R}^d$ or $\mathbb{C}^d$, then $F_X(t) \to 0$ as $t \to \infty$.

The term absolute continuous cannot be dropped by thinking about Bernoulli random variable has characteristic function $\cos(x)$. To prove it, taking the real and imginary part and then apply the Riemann Lebesgue lemma[45].

[45] $\lim_{\lambda\to\infty} \int_a^b \sin(\lambda x) f(x) = 0$, *if* f *is integratable and absolute integratable.*

Theorem 12.4 (Taylor expansion of characteristic function)**.** Let X be a real random variable with finite k^{th} moment for some $k \geq 1$. F_X is k times continuously differentiable, and one has the partial Taylor expansion

$$F_X(t) = \sum_{j=0}^{k} \frac{(it)^j}{j!} \mathbb{E}X^j + o(|t|^k)$$

where $o(|t|^k)$ is a quantity goes to zero as $t \to \infty$, times $|t|^k$. In particular we have

$$\frac{\mathrm{d}^j}{\mathrm{d}t^j} F_X(t) = i^j \mathbb{E}X^j$$

for all $\leq j \leq k$.

When X is sub-Gaussian, we have

$$F_X = \sum_{k=0}^{\infty} \frac{(it)^k}{k!} \mathbb{E}X^k \tag{12.4}$$

converges locally uniformly in t.

Theorem 12.5 (Levy continuity theorem)**.** Let V be a finite dimensional real or complex vector space, and let X_n be a sequence of V-valued random variables, let X be an additional V-valued random variable. Then the following statements are equivalent:

1. F_X converges pointwise to F_X.
2. X_n converges in distribution to X.

This tell us the characteristic function depends only on the distribution and determin distribution somehow.

Distribution theory reinterprets functions as linear functionals acting on a space of test functions. Standard functions act by integration against a test function, but many other linear functionals do not arise in this way, and these are the "generalized functions". There are different possible choices for the space of test functions, leading to different spaces of distributions. The basic space of test function consists of smooth functions with compact support, leading to standard distributions. Use of the space of smooth, rapidly (faster than any polynomial increases) decreasing test functions (these functions are called Schwartz functions) gives instead the tempered distributions, which are important because they have a well-defined distributional Fourier transform.

Definition 12.2 (Schwartz function)**.** Let $\mathbb{N}_0$ be the set of non-negative integers, and for any $n \in \mathbb{N}_0$, we denote $\mathbb{N}_0^n := \mathbb{N}_0 \times \cdots \times \mathbb{N}_0$ to be the n-fold Cartesian product. The Schwartz space or the space of rapidly decreasing functions on $\mathbb{R}^n$ is the function space

$$\mathbf{S}\left(\mathbb{R}^{\mathbf{n}}, \mathbb{C}\right) := \left\{\mathbf{f} \in \mathbf{C}^{\infty}(\mathbb{R}^{\mathbf{n}}, \mathbb{C}) |\ \forall \alpha, \beta \in \mathbb{N}_{\mathbf{0}}^{\mathbf{n}}, \quad \|\mathbf{f}\|_{\alpha,\beta} < \infty\right\},$$

where $C^{\infty}(\mathbb{R}^n, \mathbb{C})$ is the set of smooth functions from $\mathbb{R}^n$ into $\mathbb{C}$, and

$$\|f\|_{\alpha,\beta} := \sup_{x \in \mathbb{R}^n} \left|x^{\alpha}(D^{\beta} f)(x)\right|.$$

Here we use multi-index.

In human langrage, a Schwartz function is a function such that any order of it derivitive exist on $\mathbb{R}$ and goes to zero as $x \to \pm\infty$ faster than any inverse power of x. In particular $\mathbf{S}(\mathbb{R}^n)$ is a subspace of the function space $C^{\infty}(\mathbb{R}^n)$.

Proposition 12.1 (Levy's continuity theorem,full version)**.** Let V be a finite dimensional real or complex vector space, and let X_n be a sequence of V-valued random variables, let X be an additional V-valued random variable. Suppose that F_{X_n} converges pointwise to a limit F. The following are equivalent,

1. F is continuous at 0.
2. X_n is a tight sequence.
3. F is the characteristic function of a V-valued random variable X.
4. X_n converge in distribution to some V-valued random variable X.

Proposition 12.2 (Esseen concentration inequality)**.** Let X be a random variable taking values in $\mathbb{R}^d$. Then for any $r > 0, \epsilon > 0$, show that

$$\sup_{x_0 \in \mathbb{R}^d} \mathbb{P}\{|X - x_0| \leq r\} \leq C_{d,\epsilon} r^d \int_{t \in \mathbb{R}^d : |t| \leq \epsilon/r} |F_X(t)| \mathrm{d}t \tag{12.5}$$

for some constant $C_{d,\epsilon}$ depending only on d and ϵ. The left-hand side is known as the small ball probability of X at radius r.

In Fourier analysis, we learn that the Fourier transform is particular well-suited tool for studying convolutions. The probability theory analogue of this fact is that characteristic functions are a particularly well-suited tool for studying sums of independent random variables.

Proposition 12.3 (Fourier identities)**.** Let V be a finite-dimensional real or complex vector space, and let X, Y be independent random variables taking values in V. Then

$$F_{X+Y}(t) = F_X(t) F_Y(t) \tag{12.6}$$

for all $t \in V$. Also, for any scalar c, one has

$$F_{cX}(t) = F_X(\bar{c}t)$$

and more generally, for any linear transformation $T : V \to V$, one has

$$F_{TX}(t) = F_X(T^*t)$$

The proof is based on the expansion. So it requires commutative assumption on X.

In particular in the normalized senerior, we have simple relationship

$$F_{Z_n}(t) = F_X(t/\sqrt{n})^n$$

that discribes the characteristic function of Z_n in terms of that of X.

Proof. Proof of central limit theorem. 11.7
We may normalise X to have mean zero and variance 1. By 12.4, we have

$$F_X(t) = 1 - t^2/2 + o(|t|^2)$$

for sufficiently small t, or equivalently

$$F_X(t) = \exp\big(-t^2/2 + o(|t|^2)\big)$$

for sufficiently small t.
Applying

$$F_{Z_n}(t) = F_X(t/\sqrt{n})^n$$

we conclude

$$F_{Z_n}(t) \to \exp\big(t^2/2\big)$$

as $n \to \infty$ for any fixed t. And this is the characteristic function for the normal distribution $N(0,1)_{\mathbb{R}}$. The claim follows from the Levy continuity theorem. □

Theorem 12.6 (Vector-valued central limit theorem)**.** Let $\overrightarrow{X} = (X_1, \dots, X_d)$ be a random variable taking values in $\mathbb{R}^d$ with finite second moment. Define the covariance matrix $\Sigma(\overrightarrow{X})$ to be the $d \times d$ matrix Σ whose ij^{th} entry is the covariance $\mathbb{E}(X_i - \mathbb{E}X_i)(X_j - \mathbb{E}X_j)$. The following is true

1. Covariance matrix is positive semi-definite real symmetric.
2. $\overrightarrow{S_n} := \overrightarrow{X_1} + \cdots + \overrightarrow{X_n}$ is the sum of n iid copies of $\overrightarrow{X}$, $\frac{\overrightarrow{S_n} - n\mu}{\sqrt{n}}$ converges in distribution to $N(0, \Sigma(X))_{\mathbb{R}^d}$.

Theorem 12.7 (Lindeberg central limit theorem)**.** Let $X_1, X_2, \dots$ be a sequence of independent (not necessarily identiclly ditributed) real random variables, normalized to have mean zero and variance one. Assume the *strong Lindeberg condition*

$$\lim_{N\to\infty} \limsup_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} \mathbb{E}|X_{j,>N}|^2 = 0$$

where $X_{j,>N} := X_j \mathbf{I}(|X_j| > N)$ is the truncation of X_j to large values. Then we have the normalized sum converges to $N(0,1)_{\mathbb{R}}$.

A more sophiscated version of the Fourier-analytic method gives a more quantitave form of the central limit theorem, the *Berry-Esseen theorem*.

Theorem 12.8 (Berry-Esseen theorem). Let X have mean zero and unit variance and finite third moment. Let $Z_n := (X_1 + \cdots + X_n)/\sqrt{n}$, where $X_1, \ldots, X_n$ are iid copies of X. Then we have

$$\mathbb{P}(Z_n < a) = \mathbb{P}(G < a) + O(\frac{1}{\sqrt{n}}(\mathbb{E}|X|^3)) \tag{12.7}$$

uniformly for all $a \in \mathbb{R}$, where $G \equiv N(0,1)_{\mathbb{R}}$, and the implied constant is absolute.

13 The moment method for CLT

The Fourier proof relies heavily on the Fourier-analytic identities 12.6. It uses the identity $e^{A+B} = e^A e^B$ and on the independent situation $\mathbb{E}(e^A e^B) = \mathbb{E}e^A \mathbb{E}e^B$. When we turn to random matrix theory, we will often lose some of these properties, which makes it hard to apply Fourier analytic method.

The moment method is equivalent to the Fourier method in principle, but in practice it looks somewhat different. And it is often more apparant how to modify them to non-independent or non-commutative settings.

First we need an analogue of Levy's continuity theorem. Here we encounter a technical issue, whereas the Fourier phase $x \mapsto e^{itx}$ were bounded, the moment function $x \mapsto x^k$ become unbounded at infinity.

One can deal with this issue as long as one has sufficient decay:

Theorem 13.1 (Carleman continuity theorem). Let X_n be a sequence of uniformly sub-Gaussian real random variable, and let X_4 be another sub-Gaussian random variable. Then the following statements are equivalent:

1. For every $k = 0, 1, \ldots$, $\mathbb{E}X_n^k$ converges pointwise to $\mathbb{E}X^k$.
2. X_n converges in distribution to X.

Proof. From 2. to 1..:
Let $N > 0$ be a truncation parameter, let $\phi : \mathbb{R} \to \mathbb{R}$ be a smooth function that equals 1 on $[-1, 1]$ and vanish outside of $[-2, 2]$.[46] Then convergence in distribution implies that

$$\mathbb{E}X_n^k \phi(X_n/N) \to^d \mathbb{E}X^k \phi(X/N).$$

On the other hand, from the uniform sub-Gaussian hypothesis, one can make $\mathbb{E}X_n^k(1 - \phi(X_n/N))$ and $\mathbb{E}X^k(1 - \phi(X/N))$ arbitrarily small for fixed k by making N large enough. Summing and letting N go to infinity we obtain the claim.
From 1. to 2.. sub-Gaussian implies $(k+1)^{th}$ moment is bounded by $(Ck)^{k/2}$ for all $k \geq 1$ and some C independent of k.[47]
From Taylor's theorem with remainder we conclude

$$F_{X_n}(t) = \sum_{j=0}^{k} \frac{(it)^j}{j!} \mathbb{E}X_n^j + O((Ck)^{-k/2}|t|^{k+1})$$

uniformly in t and n. Substracting the expansion of X_n with X and taking the limit (with the help of 1.), we conclude

$$\limsup_{n \to \infty} |F_{X_n}(t) - F_X(t)| = O((Ck)^{-k/2}|t|^{k+1}).$$

[46] *For requirements of boundedness. Because the moment function is not bounded, we need this to exploit sub-Gaussian properties.*

[47] *While sub-exponential implies bounded by $(Ck)^k$.*

Then letting $k \to \infty$ and keeping t fixed, we have the pointwise convergence of characteristic function. Then apply the Levy's continuity theorem 12.5 to obtain the claim. □

We can see from this theorem, a sub-Gaussian random variable is uniquely determined by its moment.

If the tail is heavier, this cound fail. Like a smooth function is not determined by its derivatives at one point if that function is not analytic.

13.1 Proof of central limit theorem

When it turns to the proof of central limit theorem, WLOG we assume X is bounded, notice that then X becomes sub-Gaussian automatically. So it suffices to show that

$$\mathbb{E}Z_n^k \to \mathbb{E}G^k$$

for all $k = 0, 1, 2, \ldots$, where $G \equiv N(0,1)_{\mathbb{R}}$ is a standard Gaussian variable.

Proposition 13.1 (Moment of standard Gaussian random variable)**.** Let k be a natural numbner, and let $G \equiv N(0,1)_{\mathbb{R}}$. Then $\mathbb{E}G^k$ vanish when k is odd, and equal to $\frac{k!}{2^{k/2}(k/2)!}$ when k is even.

We can do it by checking the characteristic function of Gaussian distribution

$$F(t) = e^{it\mu}e^{-\sigma^2 t^2/2}.$$

taking the k times of the derivitive to generate kth moments.

Using the definition of Z_n and linearity of expectation, we can expand the kth moment as

$$\mathbb{E}Z_n^k = n^{-k/2} \sum_{1 \le i_1, \ldots, i_k \le n} \mathbb{E}X_{i_1} \ldots X_{i_k}$$

First we look at some small values of k.

1. For $k = 0$, the expression is trivially 1.
2. For $k = 1$, this expression is 0 thanks to the mean zero hypothesis.
3. For $k = 2$, we can split the expression into the diagonal and off diagonal components:

$$n^{-1} \sum_{1 \le i \le n} \mathbb{E}X^2 + n^{-1} \sum_{1 \le i \le j} \mathbb{E}2X_i, X_j.$$

   Each summand in the first sum is 1 because of the unit variance. Each summand in the second sum is 0 since the mean zero and independent hypothesis. So the second moment $\mathbb{E}Z_n^2$ is 1.

4. For $k = 3$, we do the similar expansion

$$n^{-3/2} \sum_{1 \le i \le n} \mathbb{E}X_i^3 + n^{-3/2} \sum_{1 \le i < j \le n} \mathbb{E}3X_i^2 X_j + 3X_i X_j^2 + n^{-3/2} \sum_{1 \le i < j < k \le n} \mathbb{E}6X_i X_j X_k.$$

   The summands in the latter two sums vanish because of the joint independence and mean zero hypothesis. The summand in the first sum does not vanish, but are $O(1)$.[48] Hence the first term is $O(n^{-1/2})$, which is asymptotically negligible. The third moment $\mathbb{E}Z_n^3$ goes to zero.

[48] $O(1)$ *in the sense that bounded by* n.

5. For $k = 4$, the expansion goes as

$$n^{-2}\sum_{1\leq i\leq n}\mathbf{E}X_i^4 + n^{-2}\sum_{1\leq i<j\leq n}\mathbf{E}4X_i^3X_j + 6X_i^2X_j^2 + 4X_iX_j^3$$
$$+n^{-2}\sum_{1\leq i<j<k\leq n}\mathbf{E}12X_i^2X_jX_k + 12X_iX_j^2X_k + 12X_iX_jX_k^2$$
$$+n^{-2}\sum_{1\leq i<j<k<l\leq n}\mathbf{E}24X_iX_jX_kX_l$$

Most of the term vanish expect the first sum, which is $O(n^{-1})$, and is again asymptotically negligible, and the sum $n^{-2}\sum_{1\leq i<j\leq n}\mathbb{E}6X_i^2X_j^2$, which by the independence and unit variance assumption works out to $n^{-2}6\binom{n}{2} = 3 + o(1)$. Thus the fourth moment $\mathbb{E}Z_n^4$ goes to 3.

Now we tackle the general case. Ordering the indices $i_1, \ldots, i_k$ as $j_1 < \cdots < j_m$ for some $1 \leq m \leq k$, with each j_r occuring with multiplicity $a_r \geq 1$. [49]

[49] *m:numbers of different X_i, a_i: degree of each X_i, j: the index of X_i.*

We see that $\mathbb{E}Z_n^k$ is the sum of all terms of the form

$$n^{-k/2}\sum_{1\leq j_1<\cdots<j_m\leq n} c_{k,a_1,\ldots,a_m}\mathbb{E}X_{j_1}^{a_1}\ldots X_{j_m}^{a_m} \tag{13.1}$$

where $1 \leq m \leq k$, $a_1, \ldots, a_m$ are positive integers adding up to k, and $c_{k,a_1,\ldots,a_m}$ is the multinomial coefficient

$$c_{k,a_1,\ldots,a_m} := \frac{k!}{a_1!\ldots a_m!}$$

The total number of such terms is 2^{k-1}.[50]

[50] *One must*

As we obseved, if any of the a_r are equal to 1, then every summand in 13.1 vanishes by joint independence and the mean zero hypothesis. Thus we restrict attention to those expressions for which all the a_r are at least 2. Since all the a_r sum up to k, we conclude that m is at most $k/2$.

On the other hand, the total number of summands in 13.1 is at most n^m[51]and the summands are bounded for fixed k since X is bounded. Thus if m is strictly less than $k/2$, then the expression

[51] *Actually it's* $\binom{n}{m}$