Random Matrix Theory

Manjunath Krishnapur
Indian Institute of Science, Bangalore

Contents

Acknowledgments: Lecture notes from a course on random matrix theory in the spring of 2011 at IISc, Bangalore. Thanks to those who attended the course (Rajesh Sundaresan, Tulasi Ram Reddy, Kartick Adhikari, Indrajit Jana and Subhamay Saha). Thanks to Anirban Dasgupta for pointing out some errors in the notes.

CHAPTER 1

Introduction

1. What is random matrix theory?

A random matrix is a matrix whose entries are random variables. The eigenvalues and eigenvectors are then random too, and the main objective of the subject is to understand their distributions. This statement omits many other interesting aspects of random matrices, but is operationally useful to keep in mind. We start with examples.

(1) Let $X_1, \ldots X_n$ be i.i.d $p \times 1$ random vectors having $N_p(0, \Sigma)$ distribution. Assume that Σ is unknown. Based on the data a natural estimate for Σ is the sample covariance matrix

$$S_n := \frac{1}{n} \sum_{k=1}^{n} X_k X_k^t.$$

Historically, this was the first random matrix to be studied,and goes by the name of *Wishart matrix*.

(2) Let $X = (X_{i,j})_{i,j \leq n}$ where $X_{i,j}$, $i \leq j$ are i.i.d real or complex valued random variables and $X_{i,j} = \overline{X}_{j,i}$. Then X is a Hermitian random matrix and hence has real eigenvalues. If we assume that $X_{i,j}$ have finite second moment, this matrix is called *Wigner matrix*.

Its origin lies in the study of heavy nucleii in Physics. Essentially, the behaviour of a nucleus is determined by a Hermitian operator (the Hamiltonian that appears in Schrodinger's equation). This opearator is a second order differential operator in about as many variables as the number of protons and neutrons and hence is beyond exact determination except in the simplest atoms. Eugene Wigner approached this problem by assuming that the exact details did not matter and replaced the Hermitian operator by a *random Hermitian matrix* of high dimensions. The eigenvalues of the original operator denote the energy levels and are of physical interest. By considering the eigenvalues of the random matrix, Wigner observed that statistically speaking, the

(3) Consider the matrix $A = (a_{i,j})_{i,j \leq n}$ with i.i.d entries. There is less physical motivation for this model but probabilistically appears even simpler than the previous model as there is more independence. This is a false appearance, but we will come to that later!

(4) Patterned random matrices have come into fashion lately. For example, let X_i be i.i.d random variables and define the random *Toeplitz matrix* $T = \left(X_{|i-j|}\right)_{i,j \leq n}$. One can also consider the asymmetric Toeplitz matrix. Many questions about the eigenvalues of these matrices are still open.

(5) Random unitary matrices.

(6) Random Schrodinger operators or random tridiagonal matrices.

2. Principal component analysis - a case for studying eigenvalues

We saw some situations in which random matrices arise naturally. But why study their eigenvalues. For Wigner matrices, we made the case that eigenvalues of the Hamiltonian are important in physics, and hence one must study eigenvalues of Wigner matrices which are supposed to model the Hamiltonian.

Here we make a case for studying the spectrum of the Wishart matrix which is more easy to understand for those of us physicsly challenged. Suppose $X_1, \ldots, X_n$ are $p \times 1$ vectors. For example, they could be vectors obtained by digitizing the photographs of employees in an office, in which case $n = 100$ and $p = 10000$ are not unreasonable values. Now presented with another vector Y which is one of the employees, we want a procedure to determine which of the X_is it is (for example, there is a door to a secure room where a photo is taken of anyone who enters the room, and the person is identified automatically). The obvious way to do it is to find the L^2 norm $\|Y - X_i\|_2$ for all $i \leq n$ and pick the value of i which minimizes the distance. As p is large, this involves a substantial amount of computation. Is there a more efficient way to do it?

There are many redundancies in the photograph. For example, if all employees have black hair, some of the co-ordinates have the same value in each of the X_is and hence is not helpful in distinguishing between individuals. Further, there are correlations. That is, if a few pixels (indicating the skin colour) are seen to be white, there is no need to check several other pixels which will probably be the same. How to use this redundancy in a systematic way to reduce computations?

We look for the unit vector $\alpha \in \mathbb{R}^p$ such that $\alpha^t X_1, \ldots, \alpha^t X_n$ have maximum variability. For simplicity assume that $X_1 + \ldots + X_n = 0$. Then, the variance of the set $\alpha^t X_j$ is

$$\frac{1}{n}\sum_{j=1}^{n}(\alpha^t X_j)^2 = \alpha^t \left(\sum_{j=1}^{n} X_j X_j^t\right)\alpha = \alpha^t S_n \alpha$$

where S_n is the sample covariance matrix of X_js. But we know from linear algebra that the maximum of $\alpha^t S_n \alpha$ is the maximum eigenvalue of S_n and the maximizing α is the corresponding eigenvector. Thus we are led to eigenvalues and eigenvectors of S_n. In this problem, X_j are random, but it may be reasonable to suppose that X_js themselves (the employees) are samples from a larger population, say $N_p(0, \Sigma)$. If we knew Σ, we could use the first eigenvector of Σ, but if we do not know Σ, we would have to use the first eigenvector of S_n. The leads to the question of whether the first eigenvalue of S_n and of Σ are close to each other? If p is not small compared to n, one cannot expect such luck. More generally, by taking the top d eigenvectors, α_i, $i \leq d$, we reduce the dimension of vectors from p to d by replacing X_j by the vector $Y_j := (\alpha_1^t X_j, \ldots, \alpha_d^t X_j)$.

In any case, for now, this was just a motivation for looking into eigenvalues and eigenvectors of random matrices. In the remaining part of this chapter we introduce the language and terminology needed and also give some of the background knowledge needed later.

3. Gaussian random variables

A standard normal random variable X is one that has density $(2\pi)^{-1/2}\exp\{-x^2/2\}$. We write $X \sim N(0,1)$. If X, Y are i.i.d $N(0,1)$, then the complex random variable $a := (X+iY)/\sqrt{2}$ is said to have standard complex Gaussian distribution. We write $a \sim \mathbb{C}N(0,1)$. a has density $\pi^{-1}\exp\{-|z|^2\}$ on the complex plane.

We assume that you know all about multivariate normal distributions. Here is a quick recap of some facts, but stated for complex Gaussians which may be a tad unfamiliar. Let $a = (a_1, \ldots, a_n)^t$ where a_k are i.i.d $\mathbb{C}N(0,1)$. If $Q_{m\times n}$ is a complex matrix and $u_{m\times 1}$ a complex vector, we say that $b = u + Qa$ has $\mathbb{C}N_m(u,\Sigma)$ distribution, where $\Sigma = QQ^*$.

Exercise 1. Let a, b be as above.

(1) Show that that distribution of b depends only on v and $\Sigma = QQ^*$.
(2) Show that $\mathbf{E}[b_k] = u_k$ and $\mathbf{E}[(b_k - u_k)(b_\ell - u_\ell)] = 0$ while $\mathbf{E}[(b_k - u_k)\overline{(b_\ell - u_\ell)}] = \Sigma_{k,\ell}$.
(3) If Q is nonsingular, show that b has density $\frac{1}{\pi^n \det(\Sigma)}\exp\{-(z-u)^*\Sigma^{-1}(z-u)\}$ on $\mathbb{C}^n$.
(4) If $b \sim \mathbb{C}N_m(u,\Sigma)$, find the distribution of $c := w + Rb$ where $w_{p\times 1}$ and $R_{p\times m}$.
(5) The characteristic function of a $\mathbb{C}^m$-valued random vector c is the function $\phi : \mathbb{C}^m \to \mathbb{C}$ defined as $\phi(w) := \mathbf{E}[\exp\{i\Im\{w^*c\}\}]$. Show that if $u = 0$, then the characteristic function of b is $\phi(w) = \exp\{-\frac{1}{4}w^*\Sigma w\}$.
(6) If $b_{m\times 1}$ and $c_{n\times 1}$ are such that $(b^t, c^t)^t$ has $\mathbb{C}N(u,\Sigma)$ we say that (b,c) has joint complex Gaussian distribution. Write

$$u = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}, \qquad \Sigma = \begin{bmatrix} A & B \\ B^* & C \end{bmatrix} \tag{1}$$

where the dimensions of u_i and A, B, C are self-evident. Then, show that $b \sim \mathbb{C}N_m(u_1, A)$ and the conditional distribution of c given b is $\mathbb{C}N(u_2 - B^*A^{-1}(b-u_1), C - B^*A^{-1}B)$.
(7) Suppose $X_{m\times 1}$ and $Y_{m\times 1}$ are real Gaussian vectors. Under what conditions is $X + iY$ have a complex Gaussian distribution?

Wick formula/Feynman diagram formula: Since the distribution of a real or complex Gaussian vector depends only on the mean vector and and covariance matrix, answers to all questions about the distribution must be presentable as a function of these parameters. Of course, in practice this may be impossible. One instance is the expectation of a product of Gaussians, and we show now that it can be written as a weighted sum over certain combinatorial objects. We first define two multilinear functions on matrices (the functions are linear in each column or each row). Let $\mathcal{S}_n$ denote the symmetric group on n letters. A *matching* of the set $[n]$ is a partition of $[n]$ into disjoint subsets of size two each. Let $\mathcal{M}_n$ denote the set of *matchings* of the set $[n]$ (it is nonempty if and only if n is even).

Definition 2. Let A be an $n \times n$ matrix. The *permanent of A* is defined as $\mathrm{per}(A) := \sum_{\pi\in\mathcal{S}_n}\prod_{i=1}^n a_{i,\pi_i}$. If A is symmetric, the *hafnian of A* is defined as $\mathrm{haf}(A) := \sum_{M\in\mathcal{M}_n}\prod a_{i,M_i}$. Here for each matching M, we take the product over all pairs in M, and each pair is taken only once.

Lemma 3. *Let $(b^t, c^t)^t$ be a complex Gaussian vector as in* (1)*. Then*

$$\mathbf{E}\left[\prod_{i=1}^{k} b_i \prod_{j=1}^{\ell} \overline{c}_\ell\right] = per(B).$$

In particular, if $b \sim \mathbb{C}N(0,\Sigma)$ then $\mathbf{E}[|b_1|^2 \dots |b_m|^2] = per(\Sigma)$.

PROOF. It suffices to prove the second statement (why?). Thus, let $b \sim \mathbb{C}N_m(0,\Sigma)$. Then, by exercise 1 we have its characteristic function

$$\mathbf{E}\left[\exp\left\{\frac{1}{2}w^*b - \frac{1}{2}b^*w\right\}\right] = \exp\left\{-\frac{1}{4}w^*\Sigma w\right\}.$$

Differentiate once with respect to $w_1, \dots, w_m$ and once with respect $\overline{w}_1, \dots, \overline{w}_m$ and then set $w = 0$. Differentiating under the expectation, on the left side we get $\frac{(-i)^m i^m}{2^{2m}}\mathbf{E}[|b_1 \dots b_k|^2]$. On the right side, expanding the exponential in series we get $\sum (k!)^{-1} 4^{-k} (w^*\Sigma w)^k$. Terms with $k < m$ vanish upon differentiation, while those with $k > m$ vanish when we set $w = 0$ (since at least one w_j factor remains after differentiating). Thus we only need to differentiate

$$(w^*\Sigma w)^m = \sum_{\substack{i_1,\dots,i_m \\ j_1,\dots,j_m}} \overline{w}_{i_1} w_{j_1} \dots \overline{w}_{i_m} w_{j_m} \sigma_{i_1,j_1} \dots \sigma_{i_m,j_m}.$$

Only those summands in which $\{i_1, \dots, i_m\}$ and $\{j_1, \dots, j_m\}$ are both permutations of $\{1, \dots, m\}$ survive the differentiation, and such a term contributes $\prod_k \sigma_{i_k,j_k}$. Thus, the right hand side finally reduces to

$$(m!)^{-1} 4^{-m} \sum_{\pi,\tau \in \mathcal{S}_m} \prod_{k=1}^{m} \sigma_{\pi_k,\tau_k} = m! 4^{-m} \sum_{\pi,\tau \in \mathcal{S}_m} \prod_{k=1}^{m} \sigma_{k,\tau\pi^{-1}(k)} = 4^{-m} \mathrm{per}(\Sigma)$$

since each permutation in $\mathcal{S}_m$ occurs $m!$ times as $\tau\pi^{-1}$. ■

On similar lines (or can you think of another way without using characteristic functions?), prove the following Feynman diagram formula for real Gaussians.

Exercise 4.
(1) Let $X \sim N_m(0,\Sigma)$. Then $\mathbf{E}[X_1 X_2 \dots X_m] = \mathrm{haf}(\Sigma)$. In particular, the expectation is zero if m is odd.
(2) For $X \sim N(0,1)$, we have $\mathbf{E}[X^{2m}] = (2m-1)(2m-3)\dots(3)(1)$, the number of matchings of the set $[2m]$.

The semicircle law: A probability distribution that arises frequently in random matrix theory and related subjects, but was never seen elsewhere in probability theory (as far as I know) is the *semicircular distribution* $\mu_{s.c}$ with density $\frac{1}{2\pi}\sqrt{4-x^2}$ on $[-2,2]$.

Exercise 5. Show that the odd moments of $\mu_{s.c}$ are zero and that the even moments are given by

$$\int x^{2n} \mu_{s.c}(dx) = \frac{1}{n+1}\binom{2n}{n}. \tag{2}$$

Catalan numbers: The number $C_n = \frac{1}{n+1}\binom{2n}{n}$ is called the n^{th} *Catalan number*. It has many combinatorial interpretations and arises frequently in mathematics. Here are some basic properties of Catalan numbers.

Exercise 6. (1) Show the recursion $C_{n+1} = \sum_{i=1}^{n} C_{i-1}C_{n-i}$ where the convention is that $C_0 = 1$.

(2) Show that the generating function of the Catalan numbers, $C(t) := \sum_{n=0}^{\infty} C_n t^n$ is satisfies $tC(t)^2 = C(t) + 1$. Conclude that $C(t) = \frac{1}{2t}(1 + \sqrt{1-4t})$. [**Note:** By Stirling's formula, estimate C_n and thus observe that $C(t)$ is indeed convergent on some neighbourhood of 0. This justifies all the manipulations in this exercise].

We show that Catalan numbers count various interesting sets of objects. The first is the set of *Dyck paths*.

Definition 7. If $X_1, \ldots, X_n \in \{+1, -1\}$, let $S_k = X_1 + \ldots + X_k$. The sequence of lattice points $(0,0), (1,S_1), (2,S_2), \ldots, (n,S$ is called a "simple random walk path". A simple random walk path of length $2n$ is called a *bridge* if $S_{2n} = 0$. A simple random walk bridge is called a *Dyck path* of length $2n$ if $S_k \geq 0$ for all $k \leq 2n$.

Lemma 8. *The number of Dyck paths of length* $2n$ *is* C_n[1]

PROOF. Let A_q be the set of all sequences $X \in \{+1, -1\}^{2q+1}$ such that $\sum_i X_i = -1$ and such that $X_{2q+1} = -1$. Let B_q be the set of sequences X in A_q for which $S_j > -1$ for all $j \leq 2q$. Obviously, A_q is in one-to one correspondence with simple random walk bridges of length $2q$ (just pad a -1 at the end) and hence $|A_q| = \binom{2q}{q}$. Further, B_q is in bijection with the set of Dyck paths of length $2q$.

If $X, Y \in A_q$, define $X \sim Y$ if $(X_1, \ldots, X_{2q})$ can be got by a cyclic permutation of $(Y_1, \ldots, Y_q)$. This is an equivalence relationship and the equivalence classes all have size $q+1$ (since there are $q+1$ negative signs, and any of them can occur as the last one). We claim that exactly one path in each equivalence class belongs to B_q.

Indeed, fix $X \in A_q$, and consider the *first* index J such that $S_J = \min\{S_0, \ldots, S_{2q}\}$. Obviously we must have $X_J = -1$. Consider the cyclic permute $Y = (X_{J+1}, \ldots, X_J)$. We leave it as an exercise to check that $Y \in B_q$ and that $Y' \notin B_q$ for any other cyclic shift of X. This shows that exactly one path in each equivalence class belongs to B_q and hence $|B_q| = (q+1)^{-1}|A_q| = C_q$. ■

Exercise 9. In each of the following cases, show that the desired number is C_n by setting up a bijection with the set of Dyck paths. This is a small sample from Stanley's *Enumerative combinatorics*, where he gives sixty six such instances!

(1) The number of ways of writing n left braces "(" and n right braces ")" legitimately (so that when read from the left, the number of right braces never exceeds the number of left braces).

(2) A matching of the set $[2n]$ is a partition of this set into n pairwise disjoint two-element subsets. A matching is said to be non-crossing if there do not exist indices $i < j < k < \ell$ such that i is paired with k and j is paired with ℓ. The number of non-crossing matchings of $[2n]$ is C_n.

[1]The beautiful proof given here is due to Takács. An easy generalization is that if $X_i \geq -1$ are integers such that $S_n = -k$, then there are exactly k cyclic shifts of X for which $\min_{m<n} S_m > -k$. An interesting consequence is *Kemperman's formula*: If $X_i \geq -1$ are i.i.d integer valued random variables, then $\mathbf{P}(\tau_{-k} = n) = \frac{k}{n}\mathbf{P}(S_n = -k)$. Here τ_{-k} is the first hitting time of $-k$.

(3) $a_1, a_2, \ldots, a_n$ are elements in a group and they have no relationships among them. Consider all words of length $2n$ that use each a_k and a_k^{-1} exactly once (there are $(2n)!$ such words). The number of these words that reduce to identity is C_n.

Combine part (2) of exercise 5 with part (3) of exercise 9 to see that the $2n$ moment of the semicircle equals the number of non-crossing matchings of $[2n]$. Except for the phrase "non-crossing", this is identical to the combinatorial interpretation of Gaussian moments as given in part (2) of exercise 4. This analogy between the semicircle and Gaussian goes very deep as we shall see later.

4. The space of probability measures

Let $\mathcal{P}(\mathbb{R})$ denote the space of Borel probability measures on $\mathbb{R}$. On $\mathcal{P}(\mathbb{R})$, define he Lévy metric

$$\mathcal{D}(\mu,\nu) = \inf\{a > 0 \,:\, F_\mu(t-a) - a \le F_\nu(t) \le F_\mu(t+a) + a \;\forall t \in \mathbb{R}\}.$$

$\mathcal{P}(\mathbb{R})$ becomes a complete seperable metric space with the metric $\mathcal{D}$. An important but easy fact is that $\mathcal{D}(\mu_n,\mu) \to 0$ if and only if $\mu_n \to \mu$ in the sense of distribution (its importance is in that it shows weak convergence to be metrizable). Recall that *convergence in distribution* or *convergence weakly* means that in terms of distribution functions, $F_{\mu_n}(x) \to F_\mu(x)$ for all x that are continuity points of F_μ.

The following exercise shows how to bound Lévy ditance for measures with densities.

Exercise 10. If μ and ν are probability measures with densities f and g respectively, show that for any $A \le \infty$[2]

$$\mathcal{D}(\mu,\nu) \le \int_{-A}^{A} |f(x) - g(x)|dx + \mu([-A,A]^c) + \nu([-A,A]^c). \tag{3}$$

5. Empirical spectral distributions

Consider an $n \times n$ Hermitian matrix X with eigenvalues $\lambda_1, \ldots, \lambda_n$. The *empirical spectral distribution* (ESD) of X is the random measure $L_X := \sum_{k=1}^n \delta_{\lambda_k}$. If X is random, let $\overline{L}_X = \mathbf{E}[L_n]$ be the *expected ESD* of X. This means that $\overline{L}[a,b] = \mathbf{E}[L[a,b]] = \frac{1}{n}\mathbf{E}[\#\{k \,:\, \lambda_k \in [a,b]\}]$.

For a fixed matrix X, L_X is an element of $\mathcal{P}(\mathbb{R})$. If X is random, $\overline{L}_X$ is an element of $\mathcal{P}(\mathbb{R})$, while L_X is a random variable taking values in $\mathcal{P}(\mathbb{R})$ (that is, a measurable function with respect to the Borel sigma algebra on $\mathcal{P}(\mathbb{R})$).

Why do we talk about the empirical measures instead of eigenvalues directly? There are two advantages. Firstly, the eigenvalues of a matrix come without any special order, and L_X equally disregards the order and merely considers eigenvalues as a set (with appropriate multiplicities). Secondly, most often we study asymptotics of eigenvalues of a sequence of matrices X_n as the dimension n increases. If we think of eigenvalues as a vector $(\lambda_1, \ldots, \lambda_n)$, say by writing them in ascending order, then the space in which the vector takes values is $\mathbb{R}^n$ which changes with n. To

[2] the case $A = \infty$ gives the *total variation distance* $\|\mu - \nu\|_{TV} = \int |f - g|$.

talk of the limit of the vector becomes meaningless. But if we encode the eigenvalues by the ESD L_{X_n}, then they all take values in one space $\mathcal{P}(\mathbb{R})$ and we can talk about taking limits.

Exercise 11. Make sure you understand what the following statements mean.

(1) $L_{X_n} \to \mu$ where X_n is a sequence of non-random matrices and $\mu \in \mathcal{P}(\mathbb{R})$.
(2) $L_{X_n} \xrightarrow{P} \mu$ or $L_X \xrightarrow{a.s.} \mu$ where X_n is a sequence of random matrices and $\mu \in \mathcal{P}(\mathbb{R})$. Does this make sense if μ is itself a random probability measure?

For instance, is the first statement equivalent to saying $\mathcal{D}(L_{X_n}, \mu) \xrightarrow{P} 0$ in the usual sense for real-valued random variables? Is it equivalent to saying that $\int f dL_{X_n} \xrightarrow{P} \int f d\mu$ for all bounded continuous f?

CHAPTER 2

Wigner's semicircle law

1. Wigner matrices

Definition 12. A *Wigner matrix* is a random matrix $X = (X_{i,j})_{i,j\le n}$ where

(1) $X_{i,j}$, $i < j$ are i.i.d (real or complex valued).
(2) $X_{i,i}$, $i \le n$ are i.i.d real random variables (possibly a different distribution)
(3) $X_{i,j} = \overline{X}_{j,i}$ for all i, j.
(4) $\mathbf{E}[X_{1,2}] = 0$, $\mathbf{E}[|X_{1,2}|^2] = 1$. $\mathbf{E}[X_{1,1}] = 0$, $\mathbf{E}[X_{1,1}^2] < \infty$.

Definition 13. Let A have i.i.d $\mathbb{C}N(0,1)$ entries and let H have i.i.d $N(0,1)$ entries. Set $X = \frac{A+A^*}{\sqrt{2}}$ and $Y = \frac{H+H^*}{\sqrt{2}}$. X is called the *GUE matrix* and Y is called the *GOE matrix*. Equivalently, we could have defined X (or Y) as a Wigner matrix with $X_{1,2} \sim \mathbb{C}N(0,1)$ (resp. $Y_{1,2} \sim N(0,1)$) and $X_{1,1} \sim N(0,2)$ (resp. $Y_{1,1} \sim N(0,2)$). GUE and GOE stand for Gaussian unitary ensemble and Gaussian orthogonal ensemble, respectively.

The significance of GUE and GOE matrices is that their eigenvalue distributions can be computed *exactly*! We shall see that later in the course. However, for the current purpose of getting limits of ESDs, they offer dispensable, but helpful, simplifications in calculations. The following exercise explains the reson for the choice of names.

Exercise 14. Let X be a GOE (or GUE) matrix. Let P be a non-random orthognal (respectively, unitary) $n \times n$ matrix. Then $P^*XP \stackrel{d}{=} P$.

Let X be a Wigner matrix and let $\tilde{\lambda}_1, \dots, \tilde{\lambda}_n$ denote the eigenvalues of X (real numbers, since X is Hermitian). Observe that $\sum \tilde{\lambda}_k^2 = \mathrm{tr}(X^2) = \sum_{i,j} |X_{i,j}|^2$. By the law of large numbers, the latter converges in probability if we divide by n^2 and let $n \to \infty$. Hence, if we let $\lambda_k = \tilde{\lambda}_k/\sqrt{n}$ be the eigenvalues of $X/\sqrt{n}$, then $\mathbf{n}^{-1}\sum_{k=1}^n \lambda_k^2$ converges in probability to a constant. This indicates that we should scale X down by $\sqrt{n}$[1]. Let L_n and $\overline{L}_n$ denote the ESD and the expected ESD of $X/\sqrt{n}$ respectively. Note that we used the finiteness of variance of entries of X in arguing for the $1/\sqrt{n}$ scaling. For heavy tailed entries, the scaling will be different.

Theorem 15. *Let X_n be an $n \times n$ Wigner random matrix. Then $\overline{L}_n \to \mu_{s.c}$ and $L_n \xrightarrow{P} \mu_{s.c}$.*

In this chapter we shall see three approaches to proving this theorem.

[1]Recall that for a sequence of probability measures to converge, it must be *tight*. Often the simplest way to check tightness is to check that the variances or second moments are bounded. This is what we did here.

(a) The method of moments.
(b) The method of Stieltjes' transforms
(c) The method of invariance principle.

Roughly, these methods can be classified as combinatorial, analytic and probabilistic, respectively. The first two methods are capable of proving Theorem 15 fully. The last method is a general probabilistic technique which does not directly prove the theorem, but easily shows that the limit must be the same for all Wigner matrices.

Since part of the goal is to introduce these techniques themselves, we shall not carry out each proof to the end, particularly as the finer details get more technical than illuminating. For example, with the method of moments we show that expected ESD of GOE matrices converges to semi-circular law and only make broad remarks about general Wigner matrices. Similarly, in the Stieltjes transform proof, we shall assume the existence of fourth moments of $X_{i,j}$. However, putting everthing together, we shall havea complete proof of Theorem 15. These techniques can be applied with minimal modifications to several other models of random matrices, but these will be mostly left as exercises.

2. The method of moments for expected ESD of GOE and GUE matrix

The idea behind the method of moments is to show that $\mu_n \to \mu$, whene $\mu_n, \mu \in \mathcal{P}(\mathbb{R})$ by showing that the moments $\int x^p \mu_n(dx) \to \int x^p \mu(dx)$ for all non-negative integer p. Of course this does not always work. In fact one can find two probability measures μ and ν with the same moments of all orders. Taking $\mu_n = \nu$ gives a counterexample.

Result 16. Let $\mu_n, \mu \in \mathcal{P}(\mathbb{R})$ and ssume that $\int x^p \mu_n(dx) \to \int x^p \mu(dx)$ for all $p \geq 1$. If μ is determined by its moments, then $\mu_n \to \mu$.

Checking if a probability measure is determined by its moments is not easy. An often used sufficient condition is summability of $(\int x^{2p} \mu(dx))^{-1/2p}$, called *Carlemann's condition*. An even easier version which suffices for our purposes (for example when the limit is the semicircle distribution) is in the following exercise.

Exercise 17. Let $\mu_n, \mu \in \mathcal{P}(\mathbb{R})$. Suppose μ is compactly supported. If $\int x^p \mu_n(dx) \to \int x^p \mu(dx)$ for all $p \geq 1$, then $\mu_n \to \mu$.

The first technique we shall use to show Wigner's semicircle law is the method of moments as applied to L_n. Since $\mu_{s.c}$ is compactly supported, exercise 17 shows that it is sufficient to prove that $\int x^p L_n(dx) \to \int x^p \mu_{s.c}(dx)$ for all p. The key observation is the formula

$$\int x^p L_n(dx) = \frac{1}{n} \sum_{k=1}^{n} \lambda_k^p = \frac{1}{n} \mathrm{tr}(X/\sqrt{n})^p = \frac{1}{n^{1+\frac{p}{2}}} \sum_{i_1, \ldots, i_p = 1}^{n} X_{i_1, i_2} \ldots X_{i_p, i_1} \tag{4}$$

which links spectral quantities to sums over entries of the matrix X. By taking expectations, we also get

$$\int x^p \overline{L}_n(dx) = \frac{1}{n}\sum_{k=1}^{n} \lambda_k^p = \frac{1}{n}\mathrm{tr}(X/\sqrt{n})^p = \frac{1}{n^{1+\frac{p}{2}}} \sum_{i_1,\ldots,i_p=1}^{n} \mathbf{E}[X_{i_1,i_2}\ldots X_{i_p,i_1}] \tag{5}$$

which will help in showing that $\overline{L}_n \to \mu_{s.c.}$. We first carry out the method of moments for the expected ESD of a GOE matrix, and later go on to the more involved statement about the ESD of a general Wigner matrix. The first goal is to see how the semicircle distribution arises.

The idea is to use the formula (5) and evaluate the expectation on the right hand side with the help of the Wick formula of exercise 2. The rest of the work is in keeping track of the combinatorics to see how the semicircle moments emerge. To get the idea, we first do it by hand for a few small values of q in (5). We work with the GOE matrix X. Remember that $X_{i,i} \sim N(0,2)$ and $X_{i,j} \sim N(0,1)$ for $i < j$.

(i) Case, q=1. $\mathbf{E}[X_{i,j}X_{j,i}] = 1$ for $j \neq i$ and 2 for $j = i$. Hence $\mathbf{E}[\mathrm{tr}(X^2)] = 2n + 2\binom{n}{2} = n^2 + n$ and

$$\int x^2 \, \overline{L}_n(dx) = \frac{1}{n^2}\mathbf{E}[\mathrm{tr}X^2] = 1.$$

(ii) Case $q = 2$. From the Wick formula for real Gaussians, $\mathbf{E}[X_{i,j}X_{j,k}X_{k,\ell}X_{\ell,i}]$ becomes

$$\begin{aligned} &= \mathbf{E}[X_{i,j}X_{j,k}]\mathbf{E}[X_{k,\ell}X_{\ell,i}] + \mathbf{E}[X_{i,j}X_{k,\ell}]\mathbf{E}[X_{j,k}X_{\ell,i}] + \mathbf{E}[X_{i,j}X_{\ell,i}]\mathbf{E}[X_{j,k}X_{k,\ell}] \\ &= (\delta_{i,k}+\delta_{i,j,k}) + (\delta_{i,k}\delta_{j,\ell}+\delta_{i,\ell}\delta_{j,k})(\delta_{i,k}\delta_{j,\ell}+\delta_{i,j}\delta_{k,\ell}) + (\delta_{j,\ell}+\delta_{i,j,\ell})(\delta_{j,\ell}+\delta_{j,k,\ell}) \end{aligned}$$

corresponding to the three matchings $\{\{1,2\},\{3,4\}\}$, $\{\{1,3\},\{2,4\}\}$, $\{\{1,4\},\{2,3\}\}$ respectively. Observe that the diagonal entries are also taken care of, since their variance is 2. This looks messy, but look at the first few terms. When we sum over all i,j,k,ℓ, we get

$$\sum_{i,j,k,\ell} \delta_{i,k} = n^3, \qquad \sum_{i,j,k,\ell} \delta_{i,j,k} = n^2, \qquad \sum_{i,j,k,\ell} (\delta_{i,k}\delta_{j,\ell})^2 = n^2.$$

It is clear that what matters is how many of the indices i,j,k,ℓ are forced to be equal by the delta functions. The more the constraints, the smaller the contribution upon summing. Going back, we can see that only two terms ($\delta_{i,k}$ in the first summand and $\delta_{j,\ell}$ term in the third summand) contribute n^3, while the other give n^2 or n only.

$$\int x^4 \, \overline{L}_n(dx) = \frac{1}{n^3}\mathbf{E}[\mathrm{tr}X^4] = \frac{1}{n^3}\sum_{i,j,k,\ell} (\delta_{i,k}+\delta_{j,\ell}) \; + \; \frac{1}{n^3}O(n^2) = 2 + O(n^{-1}).$$

Observe that the two non-crossing matchings $\{\{1,2\},\{3,4\}\}$ and $\{\{1,4\},\{2,3\}\}$ contributed 1 each, while the crossing-matching $\{\{1,3\},\{2,4\}\}$ contributed zero in the limit. Thus, recalling exercise 2, we find that $\int x^4 \, \overline{L}_n(dx) \to \int x^4 \, \mu_{s.c.}(dx)$

(iii) Case $q = 3$. We need to evaluate $\mathbf{E}[X_{i_1,i_2}X_{i_2,i_3}\ldots X_{i_6,i_1}]$. By the wick formula, we get a sum over matching of [6]. Consider two of these matchings.

(a) $\{1,4\},\{2,3\},\{5,6\}$: This is a non-crossing matching. We get

$$\begin{aligned} & \mathbf{E}[X_{i_1,i_2}X_{i_4,i_5}]\mathbf{E}[X_{i_2,i_3}X_{i_3,i_4}]\mathbf{E}[X_{i_5,i_6}X_{i_6,i_1}] \\ = \; & (\delta_{i_1,i_4}\delta_{i_2,i_5}+\delta_{i_1,i_5}\delta_{i_2,i_4})(\delta_{i_2,i_4}+\delta_{i_2,i_3,i_4})(\delta_{i_5,i_1}+\delta_{i_5,i_1,i_6}) \\ = \; & \delta_{i_1,i_5}\delta_{i_2,i_4}+[\ldots]. \end{aligned}$$

When we sum over $i_1,\ldots,i_6$, the first summand gives n^4 while all the other terms (pushed under $[\ldots]$) give $O(n^3)$. Thus the contribution from this matching is $n^4+O(n^3)$.

(b) $\{1,5\},\{2,6\},\{3,4\}$: A crossing matching. We get which is equal to

$$\begin{aligned} & \mathbf{E}[X_{i_1,i_2}X_{i_5,i_6}]\mathbf{E}[X_{i_2,i_3}X_{i_6,i_1}]\mathbf{E}[X_{i_3,i_4}X_{i_4,i_5}] \\ = \; & (\delta_{i_1,i_5}\delta_{i_2,i_6}+\delta_{i_1,i_6}\delta_{i_2,i_5})(\delta_{i_2,i_6}\delta_{i_3,i_1}+\delta_{i_2,i_1}\delta_{i_3,i_6})(\delta_{i_3,i_5}+\delta_{i_3,i_4,i_5}) \end{aligned}$$

It is easy to see that all terms are $O(n^3)$. Thus the total contribution from this matching is $O(n^3)$.

We leave it as an exercise to check that all crossing matchings of [6] give $O(n^3)$ contribution while the non-crossing ones give $n^4+O(n^3)$. Thus,

$$\int x^6\,\overline{L}_n(dx) = \frac{1}{n^4}\mathbf{E}[\mathrm{tr}X^6] = \frac{1}{n^4}(C_6n^4+O(n^3)) \to C_6 = \int x^6\mu_{s.c}(dx).$$

3. Expected ESD of GOE or GUE matrix goes to semicircle

Proposition 18. *Let $X=(X_{i,j})_{i,j\le n}$ be the GOE matrix and let L_n be the ESD of $X_n/\sqrt{n}$. Then $\overline{L}_n \to \mu_{s.c}$.*

To carry out the convergence of moments $\int x^{2q}\overline{L}_n(dx) \to \int x^{2q}\mu(dx)$ for general q, we need some preparation in combinatorics.

Definition 19. Let P be a polygon with $2q$ vertices labeled $1,2,\ldots,2q$. A *gluing of P* is a matching of the edges into pairs along with an assignment of sign $\{+,-\}$ to each matched pair of edges. Let $\mathcal{M}_{2q}^{\dagger}$ denote the set of all gluings of P. Thus, there are $2^q(2q-1)!!$ gluings of a polygon with $2q$ sides.

Further, let us call a gluing $M\in\mathcal{M}_{2q}^{\dagger}$ to be *good* if the underlying matching of edges is non-crossing and the orientations are such that matched edges are oriented in opposite directions. That is, $[r,r+1]$ can be matched by $[s+1,s]$ but not with $[s,s+1]$. The number of good matchings is C_q, by part (3) of exercise 9.

Example 20. Let P be a quadrilateral with vertices $1,2,3,4$. Consider the gluing $M=\{\{[1,2],[4,3]\},\{[2,3],[1,4]\}\}$. It means that the edge $[1,2]$ is identified with $[4,3]$ and the edge $[2,3]$ is identified with $[1,4]$. If we actually glue the edges of the polygon according to these rules, we get a torus[2]. The gluing

[2]Informally, gluing means just that. Formally, gluing means that we fix homeomorphism $f:[1,2]\to[3,4]$ such that $f(1)=3$ and $f(2)=4$ and a homeomorphism $g:[2,3]\to[1,4]$ such that $g(2)=1$ and $g(3)=4$. Then define the equivalences $x\sim f(x)$, $y\sim g(y)$. The resulting quotient space is what we refer to as the glued surface. It is locally homeomorphic to $\mathbb{R}^2$ which justifies the word "surface". The quotient space does not depend on the choice of homeomorphisms f and g. In particular, if we reverse the orientations of all the edges, we get the same quotient space.

$M' = \{\{[1,2],[3,4]\},\{[2,3],[1,4]\}\}$ is different from M. What does the gluing give us? We identify the edges $[2,3]$ and $[1,4]$ as before, getting a cylinder. Then we glue the two circular ends in *reverse* orientation. Hence the resulting surface is Klein's bottle.

For a polygon P and a gluing M, let V_M denote the number of distinct vertices in P after gluing by M. In other words, the gluing M gives an equivalence relationship on the vertices of P, and V_M is the number of equivalence classes.

Lemma 21. *Let P be a polygon with $2q$ edges and let $M \in \mathcal{M}_{2q}^{\dagger}$. Then $V_M \leq q+1$ with equality if and only if M is good.*

Assuming the lemma we prove the convergence of $\overline{L}_n$ to semicircle.

PROOF OF PROPOSITION 18.

$$\begin{aligned}
\mathbf{E}[X_{i_1,i_2}\ldots X_{i_{2q},i_1}] &= \sum_{M\in\mathcal{M}_{2q}} \prod_{\{r,s\}\in M} \mathbf{E}[X_{i_r,i_{r+1}} X_{i_s,i_{s+1}}] \\
&= \sum_{M\in\mathcal{M}_{2q}} \prod_{\{r,s\}\in M} (\delta_{i_r,i_s}\delta_{i_{r+1},i_{s+1}} + \delta_{r,s+1}\delta_{r+1,s}) \\
&= \sum_{M\in\mathcal{M}_{2q}^{\dagger}} \prod_{\{e,f\}\in M} \delta_{i_e,i_f}. \qquad (6)
\end{aligned}$$

Here for two edges e, f, if $e = [r, r+1]$ and $s = [s, s+1]$ (or $f = [s+1, s]$), then δ_{i_e,i_f} is just $\delta_{i_r,i_s}\delta_{i_{r+1},i_{s+1}}$ (respectively $\delta_{i_r,i_{s+1}}\delta_{i_{r+1},i_s}$). Also observe that diagonal entries are automatically taken care of since they have have variance 2 (as opposed to variance 1 for off-diagonal entries).

Sum (6) over $i_1, \ldots, i_{2q}$ and compare with Recall (5) to get

$$\int x^{2q}\overline{L}_n(dx) = \frac{1}{n^{1+q}} \sum_{M\in\mathcal{M}_{2q}^{\dagger}} \sum_{i_1,\ldots,i_{2q}} \prod_{\{e,f\}\in M} \delta_{i_e,i_f} = \frac{1}{n^{1+q}} \sum_{M\in\mathcal{M}_{2q}^{\dagger}} n^{V_M}. \qquad (7)$$

We explain the last equality. Fix M, and suppose some two vertices r, s are identified by M. If we choose indices $i_1, \ldots, i_{2q}$ so that some $i_r \neq i_s$, then the δ-functions force the term to vanish. Thus, we can only choose one index for each equivalence class of vertices. This can be done in n^{V_M} ways.

Invoke Lemma 21, and let $n \to \infty$ in (7). Good matchings contribute 1 and others contribute zero in the limit. Hence, $\lim_{n\to\infty} \int x^{2q}\overline{L}_n(dx) = C_q$. The odd moments of $\overline{L}_n$ as well as $\mu_{s.c}$ are obviously zero. By exercise 5, and employing exercise 17 we conclude that $\overline{L}_n \to \mu_{s.c}$. ■

It remains to prove Lemma 21. If one knows a little algebraic topology, this is clear. First we describe this "high level picture". For the benefit of those not unfamiliar with Euler characteristic and genus of a surface, we give a self-contained proof later[3].

[3]However, the connection given here is at the edge of something deep. Note the exact formula for GOE $\int t^{2q} d\overline{L}_n(t) = \sum_{g=0}^{q} n^{-g} A_{q,g}$, where $A_{q,g}$ is the number of gluings of P_{2q} that lead to a surface with Euler characteristic $2-2g$. The number g is called the genus. The right hand side can be thought of as a generating function for the number $A_{q,g}$ in the variable n^{-1}. This, and other related formulas express generating functions for maps drawn on surfaces of varying genus in terms of Gaussian integrals over hermitian matrices, which is what the left side is. In particular, such

A detour into algebraic topology: Recall that a *surface* is a topological space in which each point has a neighbourhood that is homeomorphic to the open disk in the plane. For example, a polygon (where we mean the interior of the polygon as well as its boundary) is not a surface, since points on the boundary do not have disk-like neighbourhoods. A sphere, torus, Klein bottle, projective plane are all surfaces. In fact, these can be obtained from the square P_4 by the gluing edges appropriately.

(1) Let $P = P_{2q}$ and $M \in \mathcal{M}_{2q}^{\dagger}$. After gluing P according to M, we get a surface (means a topological space that is locally homeomorphic to an open disk in the plane) which we denote P/M. See examples 20.

(2) If we project the edges of P via the quotient map to P/M, we get a graph G_M drawn (or "embedded") on the surface P/M. A graph is a combinatorial object, defined by a set of vertices V and a set of edges E. An *embedding of a graph* on a surface is a collection of function $f : V \to S$ and $f_e : [0,1] \to S$ for each $e \in E$ such that f is one-one, for $e = (u,v)$ the function f_e is a homeomorphism such that $f_e(0) = f(u)$ and $f_e(1) = f(v)$, and such that $f_e((0,1))$ are pairwise disjoint. For an embedding, each connected component of $S \setminus \cup_{e \in E} f_e[0,1]$ is called a *face*. A *map* is an embedding of the graph such hat each face is homeomorphic to a disk.

(3) For any surface, there is a number χ called the *Euler characteristic* of the surface, such that for any map drawn on the surface, $V - E + F = \chi$, where V is the number of vertices, E is the number of edges and F is the number of faces of the graph. For example, the sphere has $\chi = 2$ and the torus has $\chi = 0$. The Klein bottle also has $\chi = 0$. The *genus* of the surface is related to the Euler characteristic by $\chi = 2 - 2g$.

(4) A general fact is that $\chi \leq 2$ for any surface, with equality if and only if the surface is simply connected (in which case it is homeomorphic to the sphere).

(5) The graph G_M has $F = 1$ face (the interior of the polygon is the one face, as it is homeomorphically mapped under the quotient map), $E = q$ edges (since we have merged $2q$ edges in pairs) and $V = V_M$ vertices. Thus, $V_M = \chi(G_M) - 1 + q$. By the previous remark, we get $V_M \leq q + 1$ with equality if and only if P/M is simply connected.

(6) Only good gluings lead to simply connected P/M.

From these statements, it is clear that Lemma 21 follows. However, for someone unfamiliar with algebraic topology, it may seem that we have restated the problem without solving it. Therefore we give a self-contained proof of the lemma now.

PROOF OF LEMMA 21. After gluing by M, certain vertices of P are identified. If $V_M > q$, there must be at least one vertex, say r, of P that was not identified with any other vertex. Clearly, then M must glue $[r-1, r]$ with $[r, r+1]$. Glue these two edges, and we are left with a polygon Q with $2q-2$ sides with an edge sticking out. For r to remain isolated, it must not enter the gluing at any

formulas have been used to study "random quadrangulations of the sphere", and other similar objects, using random matrix theory. Random planar maps are a fascinating and active research are in probability, motivated by the notion of "quantum gravity" in physics.

future stage. This means, the gluing will continue within the polygon Q. Inductively, we conclude that Q must be glued by a good gluing. Retracing this to P, we see that M must be a good gluing of P. Conversely, if M is a good gluing, it is easy to see that $V_M = q+1$[4]. ■

Exercise 22. Show that the expected ESD of the GUE matrix also converges to $\mu_{s.c.}$.

4. Wishart matrices

The methods that we are going to present, including the moment method, are applicable beyond the simplest model of Wigner matrices. Here we remark on what we get for Wishart matrices. Most of the steps are left as exercises.

Definition 23. Let $m < n$ and let $X_{m\times n}$ be a random matrix whose entries are i.i.d. If $\mathbf{E}[X_{i,j}] = 0$ and $\mathbf{E}[|X_{i,j}|^2] = 1$, we say that the $m \times m$ matrix $A = XX^*$ is a *Wishart matrix*. If in addition, $X_{i,j}$ are i.i.d $N(0,1)$ (or $\mathbb{C}N(0,1)$), then A is called a real (or complex, respectively) Gaussian Wishart matrix.

Note that X is not hermitian, but A is. The positive square roots of the eigenvalues of A are called the *singular values* of X. Then the following is true.

Theorem 24. *Let $X_{m,n}$ be a real or complex Gaussian Wishart matrix. Suppose m and n go to infinity in such a way that $m/n \to c$ for a finite positive constant c. Let L_n be the ESD of A_n/n. Then, the expected ESD $\overline{L}_n \to \mu^c_{m.p}$ which is the* Marcenko-Pastur distribution*, defined as the probability measure with density*

$$\frac{d\mu^c_{m.p}(t)}{dt} = \frac{1}{2\pi c}\frac{\sqrt{(b-t)(t-a)}}{t}, \qquad b = (1+\sqrt{c})^2,\ a = (1-\sqrt{c})^2,\ \text{for } t \in [a,b].$$

Exercise 25. Prove Theorem 24.
Hint: The following trick is not necessary, but often convenient. Given an $m \times n$ matrix X, define the $(m+n)\times(m+n)$ matrix

$$B = \begin{bmatrix} 0_{m\times m} & X_{m\times n} \\ X^t_{n\times m} & 0_{n\times n} \end{bmatrix}.$$

Assume $m \le n$. By exercise 26 below, to study the ESD of $A = XX^*$, one might as well study the ESD of B.

Exercise 26. For A and B as in the hint for the previous exercise, suppose $m < n$. If s_k^2, $k \le m$ are the eigenvalues of A, then the eigenvalues of B are $\pm s_k$, $k \le m$ together with $n-m$ zero eigenvalues.

[4]Thanks to R. Deepak for this neat proof. Another way to state it is as follows. Consider the polygon P (now a topological space homeomorphic to the closed disk). Glue it by M to get a quotient space P/M. Consider the graph G formed by the edges of P (so G is a cycle). Project to G to P/M. The resulting graph G_M is connected (since G was), and has q edges. Hence it can have at most $q+1$ vertices, and it has $q+1$ vertices if and only if the G_M is a tree. Work backwards to see that M must be good. The induction step is implicit in proving that a graph has $V \le E+1$ with equality for and only for trees.

5. Continuity of eigenvalues

Suppose we drop the mean zero condition in the definition of a Wigner matrix. Does the ESD converge to semicircle law again? Such questions can be addressed by changing the matrix so that it becomes a standard Wigner matrix (for example, subtract cJ_n where J_n is the matrix of all ones). The relevant question is how the ESD changes under such perturbations of the matrix. We prove some results that will be used many times later. We start with an example.

Example 27. Let A be an $n \times n$ matrix with $a_{i,i+1} = 1$ for all $i \leq n-1$, and $a_{i,j} = 0$ for all other i, j. Let $\varepsilon > 0$ and define $A_\varepsilon = A + \varepsilon \mathbf{e}_n \mathbf{e}_1^t$. In other words, we get A_ε from A by adding ε to the $(n,1)$ entry. The eigenvalues of A are all zero while the eigenvalues of A_ε are $\pm \sqrt[n]{\varepsilon} e^{2\pi i k/n}$, $0 \leq k \leq n-1$ (the sign depends on the parity of n). For fixed n, as $\varepsilon \to 0$, the eigenvalues of A_ε converge to those of A. However, the continuity is hardly uniform in n. Indeed, if we let $n \to \infty$ first, $L_A \to \delta_0$ while for any ε positive, L_{A_ε} converges to the uniform distribution on the unit circle in the complex plane. Thus, the LSD (limiting spectral distribution) is not continuous in the perturbation ε.

For Hermitian matrices (or for normal matrices), the eigenvalues are much better tied up with the entries of the matrix. The following lemma gives several statements to that effect.

Lemma 28. *Let A and B be $n \times n$ Hermitian matrices. Let F_A and F_B denote the distribution functions of the ESDs of A and B respectively.*

(a) Rank inequality: *Suppose* $\mathrm{rank}(A-B) = 1$. *Then* $\sup_{x \in \mathbb{R}} |F_A(x) - F_B(x)| \leq \frac{1}{n}$.

(b) Hoffman-Wielandt inequality: *Let* $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ *and* $\mu_1 \geq \mu_2 \geq \ldots \geq \mu_n$ *be the eigenvalues of A and B respectively. Then,* $\sum_{k=1}^n |\lambda_k - \mu_k|^2 \leq tr(A-B)^2$.

(c) Bound on Lévy distance: $\mathcal{D}(L_A, L_B) \leq \sqrt[3]{\frac{1}{n} tr(A-B)^2}$.

If we change a matrix by making the means to be non-zero but the same for all entries, then the overall change could be big, but is of rank one. In such situations, part (a) is useful. If we make a truncation of entries at some threshold, then the magnitudes of changes may be small, but the perturbation is generally of large rank. In such part (c) is useful.

PROOF. (a) Let E_λ^A denote the eigenspace of A corresponding to the eigenvalue λ. As A is Hermitian, by the spectral theorem E_λ^A are orthogonal to one another and $\oplus E_\lambda^A = \mathbb{C}^n$. Fix any $x \in \mathbb{R}$ and define the subspaces

$$V = \bigoplus_{\lambda \leq x} E_\lambda^A, \qquad W = \bigoplus_{\lambda > x} E_\lambda^B.$$

If the smallest eigenvalue of B greater than x is $x + \delta$, then for any $u \in V \cap W$ we have $\langle (B-A)u, u \rangle \geq \delta \|u\|^2$. As $\mathrm{rank}(B-A) = 1$, this shows that $\dim(V \cap W) \leq 1$. From the formula $\dim(V) + \dim(W) - \dim(V \cap W) = n$ we see therefore get $\dim(V) - (n - \dim(W)) \leq 1$. Observe that $nF_A(x) = \dim(V)$ and $n - nF_B(x) = \dim(W)$ and hence the previous inequality becomes $F_A(x) - F_B(x) \leq n^{-1}$. Interchanging the roles of A and B we get the first statement of the lemma.

(b) Expanding both sides and using $\mathrm{tr}A^2 = \sum \lambda_i^2$ and $\mathrm{tr}B^2 = \sum \mu_i^2$, the inequality we need is equivalent to $\mathrm{tr}(AB) \leq \sum_i \lambda_i \mu_i$. By the spectral theorem, write $A = UDU^*$ and $B = VCV^*$ where $D = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ and $C = \mathrm{diag}(\mu_1, \ldots, \mu_n)$ and U, V are unitary matrices. Let $Q = U^*V$. Then,

$$\mathrm{tr}(AB) = \mathrm{tr}(UDU^*VCV^*) = \mathrm{tr}(DQCQ^*) = \sum_{i,j} \lambda_i \mu_j |Q_{i,j}|^2.$$

The claim is that the choice of Q which maximizes this expression is the identity matrix in which case we get $\sum_i \lambda_i \mu_i$ as required. It remains to show the claim. xxx

(c) If $\mathcal{D}(F_A, F_B) > \delta$, then by definition of Lévy distance, there is some $x \in \mathbb{R}$ such that $F_B(x+\delta) + \delta < F_A(x)$ (or perhaps the same inequality with A and B reversed). Order the eigenvalues of A and B as in part (ii). Then there is some k such that

$$\lambda_{k+1} > x \geq \lambda_k, \qquad \mu_{k-n\delta} > x + \delta.$$

But then $\sum(\lambda_j - \mu_j)^2 \geq \sum_{j=k-n\delta}^{k} (\lambda_j - \mu_j)^2 \geq n\delta^3$. Thus, by part (ii) of the lemma, we see that $\mathrm{tr}(A-B)^2 \geq n\delta^3$. ■

These inequalities will be used many times later.

6. WSL for Wigner matrices by method of moments

In this section we make brief remarks on how the method of moments can give a full proof of Theorem 15. The method of moments will be based on equation (4).We shall need to address the following questions.

(1) To study the expected ESD, we shall have to look at (5). However, if $X_{i,j}$ do not have higher moments, exprssions such as $\mathbf{E}[X_{i_1,i_2} \ldots X_{i_{2q},i_1}]$ may not exist. For instance, if i_r are all equal, this becomes $\mathbf{E}[X_{1,1}^{2q}]$.

(2) Secondly, the evaluation of $\mathbf{E}[X_{i_1,i_2} \ldots X_{i_{2q},i_1}]$ used Wick formulas, which apply to joint Gaussians only.

(3) Lastly, we must prove the result for the ESD itself, not just the expected ESD.

We now indicate how these problems are to be addressed.

(a) The main idea is truncation. But to prevent the complication that diagonal entries are allowed to be different, let us assume that $X_{i,j}$, $i \leq j$ are all i.i.d. from Fix any $\delta > 0$ and find $A > 0$ large enough so that $\mathbf{E}[|X_{1,2}|^2 \mathbf{1}_{|X_{1,2}|>A}] \leq \delta$. Let $\alpha_A := \mathbf{E}[X_{1,2}\mathbf{1}_{|X_{1,2}|\leq A}] = -\mathbf{E}[X_{1,2}\mathbf{1}_{|X_{1,2}|>A}]$ and $\beta_A = \mathrm{Var}(|X_{1,2}|^2 \mathbf{1}_{|X_{1,2}|\leq A})$. By choice of A we get

$$|\alpha_A| \leq \sqrt{\mathbf{E}[|X_{1,2}|^2 \mathbf{1}_{|X_{1,2}|>A}]} \leq \sqrt{\delta}$$
$$\beta_A = \mathbf{E}[|X_{1,2}|^2 \mathbf{1}_{|X_{1,2}|\leq A}] - \alpha_A^2 \in [1-2\delta, 1]$$

Define,

$$Y_{i,j} = X_{i,j}\mathbf{1}_{|X_{i,j}|\le A}. \qquad |Y_{i,j}| \le A. \text{ Perhaps not centered or scaled right.}$$
$$Z_{i,j} = Y_{i,j} - \alpha_A. \qquad |Z_{i,j}| \le 2A. \text{ Centered, but entries have variance } \beta_A, \text{ not 1.}$$

Let L_n^X be the ESD of $X_n/\sqrt{n}$ and similarly for Y and Z. we want to show that $\mathcal{D}(L_n^X, \mu_{s.c}) \xrightarrow{P} 0$. We go from L_n^X to $\mu_{s.c}$ in steps.

(A) By part (c) of Lemma 28

$$\mathcal{D}(L_n^X, L_n^Y)^3 \le \frac{1}{n^2}\sum_{i,j}|X_{i,j}|^2\mathbf{1}_{|X_{i,j}|>A} \qquad \text{and hence} \qquad \mathbf{E}\left[\mathcal{D}(L_n^X, L_n^Y)^3\right] \le \delta.$$

(B) Since $Z = Y - \alpha_A J_n$, by part (a) of Lemma 28 we see that

$$\mathcal{D}(L_n^Z, L_n^Y) \le \sup_{x\in\mathbb{R}}|F_{L_n^Y}(x) - F_{L_n^Z}(x)| \le \frac{1}{n}.$$

(C) Z is a (scaled) Wigner matrix with entries having variance β_A and such that the entries are bounded random variables. For Z, the moment method can be tried, and in step (b) we show how this is done. Thus *assume* that we are able to show that

$$\overline{L}_n^Z \to \mu_{s.c}^{\beta_A}, \qquad \text{and} \qquad L_n^Z \xrightarrow{P} \mu_{s.c}^{\beta_A}.$$

(D) Lastly, we leave it as an exercise to show that $\varepsilon(\delta) := \mathcal{D}(\mu_{s.c}^{\beta_A}, \mu_{s.c}) \to 0$ as $\delta \to 0$.

Combining (A)-(D), we get

$$\begin{aligned}\mathcal{D}(L_n^X, \mu_{s.c}) &\le \mathcal{D}(L_n^X, L_n^Y) + \mathcal{D}(L_n^Y, L_n^Z) + \mathcal{D}(L_n^Z, \mu_{s.c}^{\beta_A}) + \mathcal{D}(\mu_{s.c}^{\beta_A}, \mu_{s.c}) \\ &\le \left(\frac{1}{n^2}\sum_{i,j}|X_{i,j}|^2\mathbf{1}_{|X_{i,j}|>A}\right)^{\frac{1}{3}} + \frac{1}{n} + \mathcal{D}(L_n^Z, \mu_{s.c}^{\beta_A}) + \varepsilon(\delta)\end{aligned}$$

The first term goes in probability to $(\mathbf{E}[|X_{1,2}|^2\mathbf{1}_{|X_{1,2}|>A}])^{1/3} \le \sqrt[3]{\delta}$. The second and third terms are not random and go to zero as $n \to \infty$. Hence

$$\mathbf{P}\left(\mathcal{D}(L_n^X, \mu_{s.c}) > 2\sqrt[3]{\delta} + 2\varepsilon(\delta)\right) \to 0$$

as $n \to \infty$. This implies that $\mathcal{D}(L_n^X, \mu_{s.c}) \xrightarrow{P} 0$. This is precisely the same as saying $L_n^X \xrightarrow{P} \mu_{s.c}$.

(b) Let us assume that $X_{i,j}$bounded random variables. Then we again come to the equation (5) for the moments of $\overline{L}_n$. We do not have Wick formula to evaluate the expectation, but because of independence of $X_{i,j}$ for $i < j$, the expectation factors easily. For simplicity let us assume that the entries are real valued

$$\mathbf{E}\left[X_{i_1,i_2}\dots X_{i_p,i_1}\right] = \prod_{j\le k}\mathbf{E}\left[X_{1,2}^{N_{j,k}(\mathbf{i})}\right] \tag{8}$$
$$\text{where } N_{j,k}(\mathbf{i}) = \#\{r \le p \,:\, (i_r, i_{r+1}) = (j,k) \text{ or } (k,j)\}.$$

As always, when we fix p, i_{p+1} is just i_1. As $X_{i,j}$ all have mean zero, each $N_{j,k}(\mathbf{i})$ and $N_j(\mathbf{i})$ should be either zero or at least two (to get a non-zero expectation). Hence, the number of distinct indices that occur in $\mathbf{i}$ can be atmost q.

From a vector of indices $\mathbf{i}$, we make a graph G as follows. Scan $\mathbf{i}$ from the left, and for each new index that occurs in $\mathbf{i}$, introduce a new vertex named $v_1, v_2, \ldots$. Say that $r \leq p$ is associated to v_k if i_r is equal to the k^{th} new index that appeared as we scanned $\mathbf{i}$ from the left. For each $r \leq p$, find v_j, v_k that are associated to r and $r+1$ respectively, and draw an edge from v_j to v_k. Denote the resulting (undirected) multi-graph by $G(\mathbf{i})$ (multi-graph means loops are allowed as well as more than one edge between the same pair of vertices).

Example 29. Let $p = 7$ and $\mathbf{i} = (3,8,1,8,3,3,1)$. Then $G(\mathbf{i})$ has vertices $v_1.v_2, v_3, v_4$ and edges $[v_1, v_2]$, $[v_2, v_3]$, $[v_3, v_2]$, $[v_2, v_1]$, $[v_1, v_1]$, $[v_1, v_3]$ and $[v_3, v_1]$. We can also write $G(\mathbf{i})$ as a graph (loops allowed) with edges $[v_1, v_1]$, $[v_1, v_2]$, $[v_2, v_3]$, $[v_1, v_3]$ with weights (multiplicities) $1, 2, 2, 2$ respectively.

Observe that if there is a permutation $\pi \in \mathcal{S}_n$ such that $\mathbf{j} = \pi(\mathbf{i})$ (that is $j_r = i_{\pi(r)}$), then $G(\mathbf{i}) = G(\mathbf{j})$. Conversely, if $G(\mathbf{i}) = G(\mathbf{j})$, then there is such a permutation π.

Example 30. Let $\mathbf{i} = (1,2,1,3,4,3)$ and $\mathbf{j} = (1,2,3,2,1,4)$. Then $G(\mathbf{i})$ and $G(\mathbf{j})$ are both trees with four vertices, v_1, v_2, v_3, v_4. However, in $G(\mathbf{i})$, the vertex v_2 is aleaf while in $G(\mathbf{j})$ v_2 is not. In our interpretation, these two trees are not isomorphic, although combinatorially these two trees are isomorphic. In other words, our graphs are labelled by $v_1, v_2, \ldots$, and an isomorphism is supposed to preserve the vertex labels also.

The weight of the graph, $w[G] := \prod_{j \leq k} \mathbf{E}\left[X_{1,2}^{N_{j,k}(\mathbf{i})}\right]$ can be read off from G. Let $N_n[G]$ denote the number of $\mathbf{i} \in [n]^p$ such that $G(\mathbf{i}) = G$.

Observe that $N_n[G] = 0$ unless G is connected and the number of edges (counted with multiplicities) in G is equal to p. Further, $w[G] = 0$ if some edge has multiplicity 1. We exclude such graphs in the discussion below. If the number of vertices of G is k, then $N_n[G] = n(n-1)\ldots(n-k+1)$.

There are atmost $\lfloor p/2 \rfloor$ distinct (counting without multiplicities) edges in G since each must be repeated at least twice. Hence, the number of vertices in G is atmost $\lfloor p/2 \rfloor + 1$. If p is even this is attained if and only if G is a tree and $\mathbf{i}$ is a depth-first search of G (hence each edge is traversed twice, once in each direction). If p is odd and there are $\lfloor p/2 \rfloor + 1$ vertices, then some edge will have to be traversed three times exactly, and that is not possible if G is a tree (since $\mathbf{i}$ starts and ends at the same vertex). Note that because of the way our isomorphism works, isomorphism of

trees is really isomorphism of plane trees[5].

$$\begin{aligned}\int x^p \overline{L}_n(dx) &= \frac{1}{n^{1+\frac{p}{2}}} \sum_{\mathbf{i}} w[G(\mathbf{i})] \\ &= \frac{1}{n^{1+\frac{p}{2}}} \sum_{G} N_n[G] w[G] \\ &\to \#\left\{\text{of plane trees with vertices } v_1, \ldots, v_{p/2}\right\} \\ &= C_{p/2}\end{aligned}$$

where $C_{p/2}$ is defined to be zero if p is odd.

(c) Fix p and consider $M_n = \int x^p dL_n(x)$. We know that $\mathbf{E}[M_n] \to C_{p/2}$ (defined to be zero if p is odd). It will follow that $M_n \xrightarrow{P} C_{p/2}$ if we can show that $\mathrm{Var}(M_n) \to 0$. Now,

$$\begin{aligned}\mathrm{Var}(M_n) &= \mathbf{E}\left[M_p^2\right] - \mathbf{E}[M_p]^2 \\ &= \frac{1}{n^{2+p}} \left\{ \sum_{\mathbf{i},\mathbf{j} \in [n]^p} \mathbf{E}\left[\prod_{r=1}^{p} X_{i_r, i_{r+1}} X_{j_r, j_{r+1}}\right] - \sum_{\mathbf{i},\mathbf{j} \in [n]^p} \mathbf{E}\left[\prod_{r=1}^{p} X_{i_r, i_{r+1}}\right] \mathbf{E}\left[\prod_{r=1}^{p} X_{j_r, j_{r+1}}\right] \right\}\end{aligned}$$

which can again be analyzed by some combinatorics. Basically, in the second summand, the leading order terms come from cases when both $G(\mathbf{i})$ and $G(\mathbf{j})$ are trees. But these two trees combined together will occur as $G(\mathbf{i},\mathbf{j})$ in the first summand. Thus all leading order terms cancel, and what are left are of a lower power of n then in the denominator. The calculations will lead to $\mathrm{Var}(M_n) \le C(p) n^{-2}$ for some constant $C(p)$. Hence M_n converges in probability. For more details we refer to the book **?**.

We have shown that $\int x^p L_n(dx)$ converges in probability to $\int x^p d\mu_{s.c}(x)$. Does that imply that $\int f dL_n$ converges in probability to $\int f d\mu_{s.c}$? Does it imply that $\mathcal{D}(L_n, \mu_{s.c})$ converges in probability to 0?

7. Stieltjes' transform of a probability measure

Definition 31. For $\mu \in \mathcal{P}(\mathbb{R})$, its *Stieltjes' transform* is defined as $G_\mu(z) = \int_{\mathbb{R}} \frac{1}{z-x} \mu(dx)$. It is well-defined on $\mathbb{C}\backslash\mathrm{support}(\mu)$, in particular for $z \in \mathbb{H} := \{u + iv \,:\, v > 0\}$, the upper half-plane. If $X \sim \mu$, we can write $G_\mu(z) = \mathbf{E}\left[\frac{1}{z-X}\right]$.

Some simple observations on Stieltjes' transforms.

(a) For any $\mu \in \mathcal{P}(\mathbb{R})$, $|G_\mu(z)| \le \frac{1}{\Im z}$ for $z \in \mathbb{H}$.
(b) G_μ is analytic in $\mathbb{C}\backslash\mathrm{support}(\mu)$, as can be seen by integrating over any contour (that does not enclose the support) and interchanging integrals (integrating $1/(z-x)$ gives zero by Cauchy's theorem).

[5] A plane tree is a tree with a marked root vertex, and such that the offsprings of every individual are ordered. A good way to think of them is as genealogical trees, where in each family the offsprings are distinguished by order of birth.

(c) Suppose μ is supported on a compact interval $[-a,a]$. Then, its moments $m_k := \int x^k \mu(dx)$ satisfy $|m_k| \le a^k$ and hence $\sum m_k z^{-k-1}$ converges for $|z| > a$ and uniformly for $|z| \ge a+\delta$ for any $\delta > 0$. Hence,

$$\sum_{k=0}^{\infty} \frac{m_k}{z^{k+1}} = \mathbf{E}\left[\sum_{k=0}^{\infty} \frac{X^k}{z^k}\right] = \mathbf{E}\left[\frac{1}{z-X}\right] = G_\mu(z) \tag{9}$$

where the first equality follows by DCT. One can legitimately define $G_\mu(\infty) = 0$ and then (9) just gives the power series expansion of $w \to G_\mu(1/w)$ around 0. Since the power series coefficients are determined by the analytic function in any neighbourhood of 0, we see that if $G_\mu(z) = G_\nu(z)$ for all z in some open subset of $\mathbb{H}$, then $\mu = \nu$.

(d) For compactly supported *mu*, $G_\mu(z) \sim \frac{1}{z}$ as $z \to \infty$. If μ is not cmpactly supported, the same is true for $z = iy$ as $y \uparrow \infty$.

Equation (9) also shows that the Stieltjes transform is some variant of the moment generating function or the Fourier transform. Its usefulness in random matrix theory is analogous to the use of characteristic functions in proving central limit theorems. The following lemma gives analogues of Fourier inversion and Lévy's continuity theorems.

Lemma 32. *Let μ, ν be probability measures on $\mathbb{R}$.*

(1) For any $a < b$

$$\lim_{y \downarrow 0} \int_a^b -\frac{1}{\pi} \Im\{G_\mu(x+iy)\} dx = \mu(a,b) + \frac{1}{2}\mu\{a\} + \frac{1}{2}\mu\{b\}.$$

(2) If $G_\mu(z) = G_\nu(z)$ for all z in an open subset of $\mathbb{H}$, then $\mu = \nu$.

(3) If $\mu_n \to \mu$, then $G_{\mu_n} \to G_\mu$ pointwise on $\mathbb{H}$.

(4) If $G_{\mu_n} \to G$ pointwise on $\mathbb{H}$ for some $G : \mathbb{H} \to \mathbb{C}$, then G is the Stieltjes' transform of a possibly defective measure. If further, $iyG(iy) \to 1$ as $y \uparrow \infty$, then, $G = G_\mu$ for a probability measure μ and $\mu_n \to \mu$.

Exercise 33. If μ has a continuous density f, then show that $f(x) = -\frac{1}{\pi} \lim_{y \downarrow 0} y \Im\{G_\mu(x+iy)\}$.

PROOF. (1) Observe that

$$\frac{-1}{\pi} \Im G_\mu(x+iy) = \frac{-1}{\pi} \int_{\mathbb{R}} \Im\left\{\frac{1}{x+iy-t}\right\} \mu(dt) = \int_{\mathbb{R}} \frac{1}{\pi} \frac{y}{(x-t)^2+y^2} \mu(dt).$$

The last quantity is the density of $\mu \star C_y$, where C_y is the Cauchy distribution with scale parameter y.

On some probability space, let X and Z be independent random variables such that $X \sim \mu$ and $Z \sim C_1$. Then by the above observation, we get

$$\int_a^b -\frac{1}{\pi} \Im\{G_\mu(x+iy)\} dx = \mathbf{P}(X + yZ \in [a,b]) = \mathbf{E}\left[\mathbf{1}_{X+yZ \in [a,b]}\right].$$

Observe that $\mathbf{1}_{X+yZ \in [a,b]} \to \mathbf{1}_{X \in (a,b)} + \mathbf{1}_{X=a, Z>0} + \mathbf{1}_{X=b, Z<0}$ as $y \downarrow 0$. Take expectations, apply DCT, and use independence of X and Z to get $\mu(a,b) + \frac{1}{2}\mu\{a\} + \frac{1}{2}\mu\{b\}$.

(2) Follows immediately from the first part.

(3) If $\mu_n \to \mu$, then $\int f d\mu_n \to \int f d\mu$ for all bounded continuous functions f. For fixed $z \in \mathbb{H}$, the function $x \to \frac{1}{z-x}$ is bounded and continuous on $\mathbb{R}$ and hence $G_{\mu_n}(z) \to G_\mu(z)$.

(4) Conversely suppose that $G_{\mu_n} \to G$ pointwise for some function G. By Helly's selection principle, some subsequence μ_{n_k} converges vaguely to a possibly defective measure μ. As $(z-x)^{-1}$ is continuous and *vanishes at infinity*, $G_{\mu_{n_k}}(z) \to G_\mu(z)$ for all $z \in \mathbb{H}$.

Hence $G_\mu = G$ which shows that all subsequential limits have the same Stieltjes transform G. Further $iyG(iy) \to 1$ which shows that μ is a probability measure. By uniqueness of Stieltjes transforms, all subsequential limits are the same and hence $\mu_n \to \mu$. ■

Our next lemma gives a sharper version of the uniqueness theorem, by getting a bound on the Lévy distance between two probability measures in terms of the difference between their Stieltjes transforms.

8. Bounding Lévy distance in terms of Stieltjes transform

The following lemma is a quantitative statement that implies parts (2) and (4) of Lemma 32 as easy corollaries (how do you get part (4) of Lemma 32?).

Lemma 34. *Let $\mu, \nu \in \mathcal{P}(\mathbb{R})$. Then, for any $y > 0$ and $\delta > 0$ we have*

$$\mathcal{D}(\mu,\nu) \leq \frac{2}{\pi}\delta^{-1}y + \frac{1}{\pi}\int_{\mathbb{R}} |\Im G_\mu(x+iy) - \Im G_\nu(x+iy)|dx.$$

PROOF. Let $\mu_y = \mu \star C_y$ and $\nu_y = \nu \star C_y$. We bound the Lévy distance between μ and ν in three stages.

$$\mathcal{D}(\mu,\nu) \leq \mathcal{D}(\mu_y,\mu) + \mathcal{D}(\nu_y,\nu) + \mathcal{D}(\mu_y,\nu_y).$$

By the proof of Lemma 32 we know that μ_y has density $-\pi^{-1}\Im G_\mu(x+iy)$ and similary for ν_y. Hence, by exercise 35

$$\mathcal{D}(\mu_y,\nu_y) \leq \frac{1}{\pi}\int_{\mathbb{R}} |\Im G_\mu(x+iy) - \Im G_\nu(x+iy)|dx. \tag{10}$$

Next we control $\mathcal{D}(\mu_y,\mu)$. Let $X \sim \mu$ and $Z \sim C_1$ so that $V = X + yZ \sim \mu_y$. For $t > 0$ observe that $\mathbf{P}(Z > t) = \int_t^\infty \pi^{-1}(1+u^2)^{-1}du \leq \int_t^\infty \pi^{-1}u^{-2}du = \pi^{-1}t^{-1}$. Thus, for any $\delta > 0$, we get

$$\mathbf{P}\left(X \leq t,\, V > t+\delta\right) \leq \mathbf{P}\left(Z > y^{-1}\delta\right) \leq \pi^{-1}\delta^{-1}y$$
$$\mathbf{P}\left(V \leq t,\, X > t+\delta\right) \leq \mathbf{P}\left(Z < -y^{-1}\delta\right) \leq \pi^{-1}\delta^{-1}y.$$

These immediately give $\mathcal{D}(\mu,\mu_y) \leq \delta + \frac{y}{\pi\delta}$. Similarly $\mathcal{D}(\nu,\nu_y) \leq \delta + \frac{y}{\pi\delta}$. Combine with (10) to get the inequality in the statement. ■

Exercise 35. Let μ and ν have densities f and g respectively. Then show that $\mathcal{D}(\mu,\nu) \leq \int |f-g|$ (the latter is called the total variation distance between μ and ν).

9. Heuristic idea of the Stieltjes' transform proof of WSL for GOE

Let X_n be a GOE matrix. Let $A_n = \frac{1}{\sqrt{n}} X_n$ have eigenvalues λ_k and ESD L_n. The Stieltjes' transform of L_n is

$$G_n(z) := \int \frac{1}{z-x} L_n(dx) = \frac{1}{n} \sum_{k=1}^{n} \frac{1}{z-\lambda_k} = \frac{1}{n} \mathrm{tr}(zI - X_n)^{-1}.$$

We show that $L_n \to \mu_{s.c}$ by showing that $G_n(z) \to G_{s.c}(z)$ for all $z \in \mathbb{H}$. By Lemma **??**, this proves the claim. We are being a little vague about the mode of convergence but that will come in a moment.[6]

We introduce the following notations. We fix n for now. Y_k will denote the matrix obtained from X by deleting the k^{th} row and the k^{th} column. And $\mathbf{u}_k \in \mathbb{C}^{n-1}$ will denote the column vector got by deleting the k^{th} entry in the k^{th} column of X.

From the formulas for the entries of the inverse matrix, we know that for any M,

$$(zI - A)^{k,k} = \frac{1}{z - \frac{1}{\sqrt{n}} X_{k,k} - \frac{1}{n} \mathbf{u}_k^* (zI - \frac{1}{\sqrt{n}} Y_k)^{-1} \mathbf{u}_k}$$

and hence letting V_k denote the denominator on the right side, we can write

$$G_n(z) = \frac{1}{n} \sum_{k=1}^{n} \frac{1}{V_k}. \tag{11}$$

The key observations are

(1) Y_k is just an $(n-1)$-dimensional GOE matrix.
(2) $\mathbf{u}_k$ is a standard Gaussian vector in $(n-1)$-dimensions.
(3) Y_k and $\mathbf{u}_k$ are independent.

Therefore,

$$\begin{aligned} \mathbf{E}[V_1] &= z - \frac{1}{n} \mathbf{E}\left[\mathbf{E}\left[\mathbf{u}_1^* (zI - Y_1)^{-1} \mathbf{u}_1 \,\middle|\, \frac{1}{\sqrt{n}} Y_1 \right] \right] \\ &= z - \frac{1}{n} \mathbf{E}\left[\mathrm{tr}(zI - \frac{1}{\sqrt{n}} Y_1)^{-1} \right] \\ &\approx z - \mathbf{E}[G_{n-1}(z)]. \end{aligned} \tag{12}$$

provided we ignore the difference between n and $n-1$. As V_k are identically distributed, $\mathbf{E}[V_k]$ is equal to the same quantity.

Let us assume that each V_k is very close to its expectation. This will be a consequence of high dimensionality and needs justification. Then return to (11) and write

$$G_n(z) \approx \frac{1}{n} \sum_{k=1}^{n} \frac{1}{\mathbf{E}[V_k]} = \frac{1}{z - \frac{n-1}{n} \mathbf{E}[G_{n-1}(z)]}.$$

[6]The method of Stieltjes' transform for the study of ESDs, as well as the idea for getting a recursive equation for G_n is originally due to the physicist Leonid Pastur **?**. The method was pioneered in many papers by Zhidong Bai.

There are two implications in this. Firstly, the random quantity on the left is close to the non-random quantity on the right, and hence if we assume that $\mathbf{E}[G_n(\cdot)]$ converges to some $G(\cdot)$, then so that $G_n(\cdot)$, and to the same limit. Secondly, for G we get the equation

$$G(z) = \frac{1}{z - G(z)}.$$

This reduces to the quadratic equation $G(z)^2 - zG(z) + 1 = 0$ with solutions $G(z) = (z \pm \sqrt{z^2-4})/2$. By virtue of being Stieltjes' transforms, $G_n(z) \sim z^{-1}$ as $z \to \infty$ and G must inherit this property. Thus we are forced to take $G(z) = (z - \sqrt{z^2-4})/2$ where the appropriate square root is to be chosen. By direct calculation, the Stieltjes transform of $\mu_{s.c}$ is identified to be the same. This completes the heuristic.

Exercise 36. Show that $G = G_{\mu_{s.c}}$ satisfies the equation $(G(z))^2 - zG(z) + 1 = 0$ for all $z \in \mathbb{H}$. Argue that no other Stieltjes' transform satisfies this equation. One can then write

$$G(z) = \frac{z - \sqrt{z^2-4}}{2}$$

where the branch of square root used is the one defined by $\sqrt{re^{i\theta}} = \sqrt{r}e^{i\theta/2}$ with $\theta \in (-\pi, \pi)$. Expand by Binomial theorem and verify that the even moments are given by Catalan numbers.

10. The Stieltjes' transform proof of WSL

Now for the rigorous proof. The crucial point in the heuristic that needs justification is that V_k is close to its expected value. The following two lemmas will come in handy.

Lemma 37. *Let V be a complex valued random variable and assume that almost surely, $\Im V \geq t$ for some constant $t > 0$. Then, for any $p > 0$*

$$\mathbf{E}\left[\left| \frac{1}{V} - \frac{1}{\mathbf{E}\, V} \right|^p \right] \leq t^{-2p} \mathbf{E}\left[|V - \mathbf{E}\, V|^p \right].$$

PROOF. Almost surely, $\Im V \geq t$ and hence $\Im\{\mathbf{E}\, V\} \geq t$ too. Hence, $|V| \geq t$ a.s., and $|\mathbf{E}\, V| \geq t$. Thus,

$$\left| \frac{1}{V} - \frac{1}{\mathbf{E}\, V} \right| = \frac{|V - \mathbf{E}\, V|}{|V|\, |\mathbf{E}\, V|} \leq t^{-2} |V - \mathbf{E}\, V|.$$

Raise to power p and take expectations. ■

Lemma 38. *Let $\mathbf{u}$ be an $n \times 1$ random vector where u_i are independent real or complex valued random variables with zero mean and unit variance. Let M be a non-random $n \times n$ complex matrix. Then,*

(a) $\mathbf{E}\left[\mathbf{u}^* M \mathbf{u}\right] = trM$.

(b) If in addition $m_4 := \mathbf{E}[|u_i|^4] < \infty$, *then* $Var\left(\mathbf{u}^* M \mathbf{u}\right) \leq (2 + m_4) tr(M^* M)$.

PROOF. Write $\mathbf{u}^*M\mathbf{u} = \sum_{i,j=1}^n M_{i,j}\overline{u}_i u_j$. When we take expectations, terms with $i \neq j$ vanish and those with $i = j$ give $M_{i,i}$. The first claim follows. To find the variance,[7] we compute the second moment $\mathbf{E}\left[|\mathbf{u}^*M\mathbf{u}|^2\right] = \sum_{i,j}\sum_{k,\ell} M_{i,j}\overline{M}_{k,\ell}\mathbf{E}[\overline{u}_i u_j \overline{u}_k u_\ell]$.

$\mathbf{E}[\overline{u}_i u_j \overline{u}_k u_\ell]$ vanishes unless each index appears at least twice. Thus, letting $m_2 = \mathbf{E}[u_1^2]$

$$\mathbf{E}[\overline{u}_i u_j \overline{u}_k u_\ell] = \delta_{i,j}\delta_{k,\ell} + \delta_{i,\ell}\delta_{j,k} + |m_2|^2\delta_{i,k}\delta_{j,\ell} + m_4\delta_{i,j,k,\ell}.$$

Thus

$$\begin{aligned}
\mathbf{E}\left[|\mathbf{u}^*M\mathbf{u}|^2\right] &= \sum_{i,k} M_{i,i}\overline{M}_{k,k} + \sum_{i,j} M_{i,j}\overline{M}_{j,i} + |m_2|^2\sum_{i,j} M_{i,j}\overline{M}_{i,j} + m_4\sum_i M_{i,i}\overline{M}_{i,i} \\
&= (\mathrm{tr}M)^2 + \mathrm{tr}(M^*M^t) + |m_2|^2\mathrm{tr}(M^*M) + m_4\sum_i |M_{i,i}|^2 \\
&\leq (\mathrm{tr}M)^2 + (1 + |m_2|^2 + m_4)\mathrm{tr}(M^*M).
\end{aligned}$$

Observe that $|m_2|^2 \leq \mathbf{E}[|u_1|^2] \leq 1$ where equality may not hold as u_1 is allowed to be complex valued. Subtract $\mathbf{E}[\mathbf{u}^*M\mathbf{u}]^2 = (\mathrm{tr}M)^2$ to get $\mathrm{Var}(\mathbf{u}^*M\mathbf{u}) \leq (2 + m_4)\mathrm{tr}(M^*M)$. ■

Now we are ready to prove Wigner's semicircle law under fourth moment assumption.

Theorem 39. *Let X_n be a Wigner matrix. Assume $m_4 = \max\{\mathbf{E}[|X_{1,2}|^4], \mathbf{E}[X_{1,1}^4]\}$ is finite. Then, $L_n \xrightarrow{P} \mu_{s.c}$ and $\overline{L}_n \to \mu_{s.c}$.*

PROOF. Let G_n and $\overline{G}_n$ denote the Stieltjes' transforms of L_n and $\overline{L}_n$ respectively. Of course, $\overline{G}_n(z) = \mathbf{E}[G_n(z)]$. Fix $z \in \mathbb{H}$. From (11) we have $G_n(z) = n^{-1}\sum_{k=1}^n 1/V_k$ where

$$V_k = (zI - X)^{k,k} = z - \frac{X_{k,k}}{\sqrt{n}} - \frac{1}{n}\mathbf{u}_k^*\left(zI - \frac{Y_k}{\sqrt{n}}\right)^{-1}\mathbf{u}_k. \tag{13}$$

Here Y_k is the $(n-1)\times(n-1)$ matrix obtained from X by deleting the k^{th} row and k^{th} column, and $\mathbf{u}_k$ is the $(n-1)\times 1$ vector obtained by deleting the k^{th} element of the k^{th} column of X. Clearly Y_k is a Wigner matrix of dimension $(n-1)$ and $\mathbf{u}_k$ is a vector of iid copies of $X_{1,2}$, and $\mathbf{u}_k$ is independent of Y_k. We rewrite (13) as

$$V_k = z - \frac{X_{k,k}}{\sqrt{n}} - \frac{1}{\sqrt{n(n-1)}}\mathbf{u}_k^*\left(z_n I - \frac{Y_k}{\sqrt{n-1}}\right)^{-1}\mathbf{u}_k, \qquad \text{where } z_n := \frac{\sqrt{n}}{\sqrt{n-1}}z. \tag{14}$$

Hence,

$$\begin{aligned}
\mathbf{E}\left[\left|G_n(z) - \frac{1}{\mathbf{E}[V_1]}\right|^2\right] &= \mathbf{E}\left[\left|\frac{1}{n}\sum_{k=1}^n\left(\frac{1}{V_k} - \frac{1}{\mathbf{E}[V_k]}\right)\right|^2\right] \\
&\leq \mathbf{E}\left[\frac{1}{n}\sum_{k=1}^n\left|\frac{1}{V_k} - \frac{1}{\mathbf{E}[V_k]}\right|^2\right] \qquad \text{(by Cauchy-Schwarz)} \\
&\leq \frac{1}{(\Im z)^4}\mathbf{E}[|V_1 - \mathbf{E}[V_1]|^2] \qquad \text{by Lemma 38.}
\end{aligned}$$

[7] For a complex-valued random variable Z, by $\mathrm{Var}(Z)$ we mean $\mathbf{E}[|Z - \mathbf{E}[Z]|^2] = \mathbf{E}[|Z|^2] - |\mathbf{E}[Z]|^2$. This is consistent with the usual definition if Z is real-valued.

For a complex-valued random variable with finite second moment, $\mathbf{E}[|Z-c|^2]$ is minimized uniquely at $c = \mathbf{E}[Z]$. In particular we also have $|\mathbf{E}[Z]|^2 \leq \mathbf{E}[|Z|^2]$. Therefore, the above inequality implies the following two inequalities.

$$\mathrm{Var}(G_n(z)) \leq \frac{1}{(\Im z)^4}\mathbf{E}[|V_1 - \mathbf{E}[V_1]|^2] \tag{15}$$

$$\left|\overline{G}_n(z) - \frac{1}{\mathbf{E}[V_1]}\right|^2 \leq \frac{1}{(\Im z)^4}\mathbf{E}[|V_1 - \mathbf{E}[V_1]|^2]. \tag{16}$$

The next step is to compute $\mathbf{E}[V_1]$ and obtain a bound for $\mathrm{Var}(V_1)$. Firstly,

$$\begin{aligned}
\mathbf{E}[V_1] &= z - 0 - \frac{1}{\sqrt{n(n-1)}}\mathbf{E}\left[\mathbf{E}\left[\mathbf{u}_k^*\left(z_n I - \frac{Y_1}{\sqrt{n-1}}\right)^{-1}\mathbf{u}_k \,\middle|\, Y_1\right]\right] \\
&= z - \frac{\sqrt{n-1}}{\sqrt{n}}\frac{1}{n-1}\mathbf{E}\left[\mathrm{tr}\left(z_n I - \frac{Y_1}{\sqrt{n-1}}\right)^{-1}\right] \\
&= z - \sqrt{\frac{n-1}{n}}\,\overline{G}_{n-1}(z_n). \qquad (17)
\end{aligned}$$

Now, to estimate $\mathrm{Var}(V_1)$, recall that $X_{1,1}$, $\mathbf{u}_1$ and Y_1 are all independent. Write $A = (zI - \frac{Y_1}{\sqrt{n}})^{-1}$ and $B = (z_n I - \frac{Y_1}{\sqrt{n-1}})^{-1}$ and observe that if θ_j are eigenvalues of Y_1 then the eigenvalues of A and B are $(z-\theta_j/\sqrt{n})^{-1}$ and $(z-\theta_j/\sqrt{n-1})^{-1}$ both of which are bounded in absolute value by $(\Im z)^{-1}$.

Write $\mathrm{Var}(V_1)$ as $\mathbf{E}\left[\mathrm{Var}(V_1 \,|\, Y_1)\right] + \mathrm{Var}\left(\mathbf{E}[V_1 \,|\, Y_1]\right)$. We evaluate the two individually as follows. Using the expression (13) and part (b) of Lemma 38 for $\mathrm{Var}(V_1 \,|\, Y_1)$ we get

$$\mathbf{E}\left[\mathrm{Var}(V_1 \,|\, Y_1)\right] = \mathbf{E}\left[n^{-1} + n^{-2}(2+m_4)\mathrm{tr}(A^*A)\right] \leq n^{-1} + m_4 n^{-1}(\Im z)^{-2}.$$

Using the expression (14) we get $\mathbf{E}[V_1 \,|\, Y_1] = z - \sqrt{\frac{n-1}{n}}G_{n-1}(z_n)$ and hence

$$\mathrm{Var}\left(\mathbf{E}[V_1 \,|\, Y_1]\right) = \frac{n-1}{n}\mathrm{Var}\left(G_{n-1}(z_n)\right) \leq \mathrm{Var}\left(G_{n-1}(z_n)\right).$$

Add this to the inequality for $\mathbf{E}\left[\mathrm{Var}(V_1 \,|\, Y_1)\right]$ gives a bound for $\mathrm{Var}(V_1)$ which when inserted into (15) gives

$$\mathrm{Var}(G_n(z)) \leq \frac{1}{(\Im z)^4}\left(\frac{1}{n} + \frac{m_4}{n(\Im z)^2} + \mathrm{Var}\left(G_{n-1}(z_n)\right)\right).$$

Let $V_n := \sup\{\mathrm{Var}(G_n(z)) \,:\, \Im z \geq 2\}$. Observe that $V_n \leq 2^{-2}$ as $|G_n(z)| \leq (\Im z)^{-1}$, in particular V_n is finite. Since $\Im z_n > \Im z$, we arrive at the recursive inequality

$$V_n \leq \frac{1}{2^4 n} + \frac{m_4}{2^6 n} + \frac{1}{2^4}V_{n-1} \leq \frac{A}{n} + \frac{1}{2}V_{n-1}$$

where $A = 2^{-2} + 2^{-6}m_4$. We increased the first term from 2^4 to 2^{-2} so that $V_1 \le A$ also. Iterating this inequality gives

$$\begin{aligned} V_n &\le \frac{A}{n} + \frac{A}{2(n-1)} + \frac{A}{2^2(n-2)} + \ldots + \frac{A}{2^{n-2}2} + \frac{A}{2^{n-1}} \\ &\le \frac{A}{n/2} \sum_{k=0}^{n/2-1} \frac{1}{2^k} + \frac{A}{2^{n/2}} \frac{n}{2} \\ &\le \frac{5A}{n} \qquad (\text{for } n \ge 10). \end{aligned}$$

Insert this into (15) and (16) and use (17) to get

$$\sup_{\Im z \ge 2} \mathrm{Var}(G_n(z)) \le \frac{1}{n} \tag{18}$$

$$\sup_{\Im z \ge 2} \left| \overline{G}_n(z) - \frac{1}{z - \sqrt{(n-1)/n}\, \overline{G}_{n-1}(z_n)} \right|^2 \le \frac{1}{n}. \tag{19}$$

Convergence of $\overline{L}_n$ to semicircle: $\overline{L}_n$ is a sequence of probability measure with Stieltjes transforms $\overline{G}_n$. Let μ be any subsequential limit of $\overline{L}_n$, a priori allowed to be a defective measure. By (19) G_μ must satisfy $G_\mu(z)(z - G_\mu(z)) = 1$ for all z with $\Im z \ge 2$ (why? Justification is needed to claim that $\overline{G}_{n-1}(z_n) \to G_\mu(z)$, but one can argue this by using equicontinuity of $\overline{G}_n$ as in the next paragraph). Thus, $G_\mu(z) = (z \pm \sqrt{z^2-4})/2$. Since G_μ must be analytic in z and $G_\mu(z) \sim \mu(\mathbb{R})z^{-1}$ as $z \to \infty$, the branch of square root is easily fixed. We get

$$G_\mu(z) = \frac{z + \sqrt{z^2-4}}{2}, \qquad \text{for } \Im z \ge 2$$

where the square root is the branch $\sqrt{re^{i\theta}} = \sqrt{r}e^{i\theta/2}$ with $\theta \in (-\pi, \pi)$. By exercise 36 this is precisely the Stieltjes transform of the semicircle distribution on $[-2,2]$. Thus all subsequential limits of $\overline{L}_n$ are the same and we conclude that $\overline{L}_n \to \mu_{s.c}$.

Convergence of L_n to semicircle: Without loss of generality, let X_n be defined on the same probability space for all n.[8] If $\sum 1/n_k < \infty$, then by (18) it follows that for fixed z with $\Im z \ge 2$ we have $G_{n_k}(z) - \overline{G}_{n_k}(z) \overset{a.s.}{\to} 0$. Take intersection over a countable dense subset S of z and invoke the convergence of $\overline{G}_n$ to conclude that $G_{n_k}(z) \to G_{s.c}(z)$ for all $z \in S$, almost surely. For a Stieltjes transform G, we have the inequality $|G'(z)| \le (\Im z)^{-2}$, from which we see that G_n are equicontinuous on $\{\Im z \ge 2\}$. Therefore we get $G_{n_k}(z) \to G(z)$ for all z with $\Im z \ge 2$, almost surely. Hence $L_{n_k} \overset{a.s.}{\to} \mu_{s.c}$.

Now, given any subsequence $\{n_k\}$, choose a further subsequence $\{n_{k_\ell}\}$ such that $\sum 1/n_{k_\ell} < \infty$. Then $L_{n_{k_\ell}} \overset{a.s.}{\to} \mu_{s.c}$. Thus every subsequence has an almost sure convergent sub-sub-sequence. Therefore $L_n \overset{P}{\to} \mu$. ■

[8] The strategy used here is as follows. To show that real-valued random variables Y_n converge in probability to zero, we may first of all construct random variables Z_n on the same probability space so that $Z_n \overset{d}{=} Y_n$ and then show that Z_n converge in probability to zero. And for the latter, it suffices to show that any subsequence has a further subsequence that converges *almost surely* to zero.

Remark 40. If we had used Lemma 37 with $p = 4$ instead of $p = 2$ (which would force the assumption that $X_{i,j}$ have finite eighth moment), then we could get n^{-2} as a bound for $\mathbf{E}\left[\left|G_n(z) - \frac{1}{\mathbf{E}[V_1]}\right|^2\right]$. Therefore we would get almost sure convergence.

In fact, one can conclude almost sure convergence assuming only finite second moment! This requires us to use $p = 1$, but then we are faced with estimating $\mathbf{E}[|V_1 - \mathbf{E}[V_1]|]$ which is more complicated than estimating the variance. Lastly, Stieltjes' transform methods are very powerful, and can be used to prove rates of convergence in Wigner's semicircle law.

Exercise 41. Prove Theorem 24 by Stieltjes transform methods. Mainly, work out the heuristic steps in the proof and arrive at an equation for the Stieltjes transform of the limiting measure and show that the equation is satisfied uniquely by the Stieltjes transform of the Marcenko-Pastur law. The full details will involve similar technicalities and may be omitted.

11. Chatterjee's invariance principle

We have seen in a first course in probability that the sum of n i.i.d random variables with zero mean, scaled by $\sqrt{n}$, converges to Gaussian law provided the random variables have finite variance (these distributions are therefore said to be in the normal domain of attraction). And now we have Wigner's semicircle law which states that the spectrum of $X_n/\sqrt{n}$ converges to the semicircle law whenever X_n is a Wigner matrix with entries having finite variance. On the one hand this does not sound surprising. However, it is entirely unclear why finite variance condition which worked for sums of random variables should also be the right condition for eigenvalues of Wigner matrices. Chatterjee's invariance explains this phenomenon of invariance in much greater generality.[9] In this section we state the invariance principle in general and apply it to random matrices in the next section.

Theorem 42. *Let X_k, $k \leq n$ be independent real valued random variables with zero mean, unit variance and finite third moments satisfying $\max_{k\leq n} \mathbf{E}|X_k|^3 \leq \gamma$ for some $\gamma < \infty$. Let Y_k, $k \leq n$ be i.i.d $N(0,1)$ variables. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a C^3 function and define $U = g(f(X_1, \ldots, X_n))$ and $V = g(f(Y_1, \ldots, Y_n))$. Then for any $g \in C_b^3(\mathbb{R})$, we have*

$$\left|\mathbf{E}[g(U)] - \mathbf{E}[g(V)]\right| \leq \frac{n}{3}\gamma C(g)\lambda_3(f).$$

where $C(g) = \max_{i\leq 3} \|g^{(i)}\|_\infty$ and $\lambda_3(f) = \max\{|\partial_k^r f(x)|^{3/r} : k \leq n,\ 1 \leq r \leq 3,\ x \in \mathbb{R}^n\}$.

Let us parse the statement. If for every $g \in C_b^3(\mathbb{R})$ we had $\mathbf{E}[g(U)] = \mathbf{E}[g(V)]$, then U must have the same distribution as V, since C_b^3 is a measure-determining class (for example, the characterisitic functions would have to coincide). Hence if $\mathbf{E}[g(U)]$ and $\mathbf{E}[g(V)]$ are close for very g, then U and V must be close in distribution.

[9] Sourav Chatterjee **?** resurrected and developed an old method of Lindeberg to prove this general invariance principle. At once elementary and powerful, it has found many applications and has drastically simplified proofs in many cases.

The theorem asserts gives a bound for $\mathbf{E}[g(U)] - \mathbf{E}[g(V)]$ in terms of $C(g)$ and $\lambda_3(f)$. For fixed g, the constant $C(g)$ may be ignored. What does $\lambda_3(f)$ signify? It is a bound on iterated partial derivatives of f. If f is a function that does not depend too much on any one of the variables, then $\lambda_3(f)$ is small, while if it does depend too much on one or more of the variables, $\lambda_3(f)$ is large, giving a weak bound. The following example illustrates this well.

Example 43. As a quick application, consider $f(x_1,\ldots,x_n) = \frac{1}{\sqrt{n}}\sum_{k=1}^n x_k$. Then, $\lambda_3(f) = n^{-3/2}$. Hence for any $g \in C_b^3(\mathbb{R})$, Theorem 42 gives

$$\left|\, \mathbf{E}\left[g\left(\frac{X_1+\ldots+X_n}{\sqrt{n}}\right)\right] - \mathbf{E}\left[g\left(\frac{Y_1+\ldots+Y_n}{\sqrt{n}}\right)\right] \,\right| \leq \frac{\gamma C(g)}{3}\frac{1}{\sqrt{n}}.$$

But $(Y_1+\ldots+Y_n)/\sqrt{n} \stackrel{d}{=} Y_1$. Thus letting $S_n = X_1+\ldots+X_n$, we get for any $g \in C_b^3$ that

$$|\mathbf{E}[g(S_n/\sqrt{n})] - \mathbf{E}[g(Y_1)]| \leq \frac{\gamma C(g)}{3}\frac{1}{\sqrt{n}} \to 0$$

as $n \to \infty$. If we knew this for every $g \in C_b(\mathbb{R})$ that would be the definition of $\frac{S_n}{\sqrt{n}} \stackrel{d}{\to} Y_1$. It is an easy exercise to check that knowing this only for C_b^3 functions also suffices (why?). Thus we have proved CLT under third moment assumptions without recourse to characteristic functions! This was Lindeberg's proof of the CLT. Note that X_k are assumed to be independent, not necessarily identical in distribution.

In contrast, consider the function $f(x_1,\ldots,x_n) = c_1x_1+\ldots+c_nx_n$ where $\sum c_k^2 = 1$ so that $f(Y_1,\ldots,Y_n) \stackrel{d}{=} Y_1$ still. However $\lambda_3(f) = \max_k |c_k|^3$. Thus, the bound is $n\max_k |c_k|^3$ which may not go to zero as $n \to \infty$, for example if $c_k = 2^{-k/2}$. In this case, X_1 has too much influence on the outcome.

PROOF OF THEOREM 42. . Define the vectors

$$\begin{aligned} W_k &= (X_1,\ldots,X_{k-1},Y_k,\ldots,Y_n), \\ W_k^0 &= (X_1,\ldots,X_{k-1},0,Y_{k+1},\ldots,Y_n). \end{aligned}$$

Then, writing $h = g\circ f$,

$$\begin{aligned} U - V &= \sum_{k=0}^n h(W_{k+1}) - h(W_k) \\ &= \sum_{k=0}^n h(W_{k+1}) - h(W_k^0) \;-\; \sum_{k=0}^n h(W_k) - h(W_k^0). \end{aligned}$$

By Taylor expansion we write the k^{th} summands as

$$\begin{aligned} h(W_{k+1}) - h(W_k^0) &= \partial_k h(W_k^0)X_k + \partial_k^2 h(W_k^0)\frac{X_k^2}{2} + \partial_k^3 h(W_k^*)\frac{X_k^3}{6} \\ h(W_k) - h(W_k^0) &= \partial_k h(W_k^0)Y_k + \partial_k^2 h(W_k^0)\frac{Y_k^2}{2} + \partial_k^3 h(W_k^\#)\frac{Y_k^3}{6} \end{aligned}$$

where $W_k^* \in [0,X_k]$ and $W_k^\# \in [0,Y_k]$. Observe that X_k and Y_k are independent of W_k^0. Take expectations in the above equations and subtract the second from the hirst. As the first two moments of X_k

match with those of Y_k, the first two terms cancel and we get

$$\mathbf{E}\left[h(W_{k+1})\right] - \mathbf{E}\left[h(W_k)\right] = \frac{1}{6}\mathbf{E}\left[X_k^3 \partial_k^3 h(W_k^*)\right] + \frac{1}{6}\mathbf{E}\left[Y_k^3 \partial_k^3 h(W_k^{\#})\right].$$

As $\partial_k^3 h(x) = g'''(f(x))(\partial_k f(x))^3 + 3g''(f(x))\partial_k^2 f(x)\partial_k f(x) + g'(f(x))\partial_k^3 f(x)$ we get the bound $|\partial_k^3 h(x)| \leq C(g)\lambda_3(f)$ where $\gamma, C(g), \lambda_3(f)$ are as defined in the statement of the theorem (if necessary increase γ so that $\mathbf{E}[|Y_1|^3] \leq \gamma$). Thus we get

$$|\mathbf{E}\left[h(W_{k+1})\right] - \mathbf{E}\left[h(W_k)\right]| \leq \frac{1}{3}\gamma C(g)\lambda_3(f).$$

Sum over k to get $\mathbf{E}[|U-V|] \leq \frac{n}{3}\gamma C(g)\lambda_3(f)$ as claimed.

■

For simplicity we assumed the existence of third moments in Theorem 42. The following exercises show how to eliminate this condition.

Exercise 44. Let X_k be independent random variables having mean zero and unit variance. Let Y_k be i.i.d $N(0,1)$. Let f and g be as in Theorem 42. Fix any constant $A > 0$. Then,

$$\begin{aligned}\left|\mathbf{E}\left[g(f(X))\right] - \mathbf{E}\left[g(f(Y))\right]\right| &\leq C(g)\lambda_3(f)\left(\sum_{k=1}^n \mathbf{E}[|X_k|^3 \mathbf{1}_{|X_k|\leq A}] + \sum_{k=1}^n \mathbf{E}[|Y_k|^3 \mathbf{1}_{|Y_k|\leq A}]\right) \\ &\quad + C'(g)\lambda_2(f)\left(\sum_{k=1}^n \mathbf{E}[|X_k|^2 \mathbf{1}_{|X_k|>A}] + \sum_{k=1}^n \mathbf{E}[|Y_k|^2 \mathbf{1}_{|Y_k|>A}]\right)\end{aligned}$$

where $C(g)$ and $\lambda_3(f)$ are as before and $C'(g) = \max_{i\leq 2} \|g^{(i)}\|_\infty$ and $\lambda_2(f) = \max\{|\partial_k^r f(x)|^{2/r} : k \leq n,\ 1 \leq r \leq 2,\ x \in \mathbb{R}^n\}$. **[Hint:** Follow the same steps as before, except that in addition to the third order Taylor expansion, use also the second order Taylor expansions

$$\begin{aligned}h(W_{k+1}) - h(W_k^0) &= \partial_k h(W_k^0)X_k + \partial_k^2 h(\tilde{W}_k^*)\frac{X_k^2}{2}, \\ h(W_k) - h(W_k^0) &= \partial_k h(W_k^0)Y_k + \partial_k^2 h(\tilde{W}_k^{\#})\frac{Y_k^2}{2}.\end{aligned}$$

Use these second order expansions on the event $|X_k| > A$ and $|Y_k| > A$, respectively, and the third order Taylor expansions when $|X_k| \leq A$ and $|Y_k| \leq A$. Add them up and follow the same steps as before with minor modifications.]

Exercise 45. As an application of Exercise 44, prove the Lindeberg-Feller central limit theorem for triangular arrays. In particular, we do not need identical distribution and we do not need moments higher than 2 (the Lindeberg condition replaces it).

12. Wigner's semicircle law using invariance principle

For a vector $\mathbf{t} \in \mathbb{R}^{\mathbf{n(n+1)/2}}$ (indexed by (i,j), $1 \leq i \leq j \leq n$), let $M(\mathbf{t})$ be the real symmetric matrix with entries $M_{i,j} = M_{j,i} = t_{i,j}/\sqrt{n}$. Fix $z \in \mathbb{H}$ and define $f : \mathbb{R}^{n(n+1)/2} \to \mathbb{C}$ by $f(\mathbf{t}) = \mathbf{n^{-1}}\mathrm{tr}(\mathbf{zI} - \mathbf{M(t)})^{-\mathbf{1}}$.

Let $\mathbf{X} = (X_{i,j})_{i\leq j}$ where $X_{i,j}$ are i.i.d real-valued with mean zero and variance one and let $\mathbf{Y} = (Y_{i,j})_{i\leq j}$ where $Y_{i,j}$ are i.i.d N(0,1). Then $M(\mathbf{X})$ and $M(\mathbf{Y})$ are scaled Wigner matrices.

For simplicity we assume that $\gamma_3 := \mathbf{E}[|X_{i,j}|^3] < \infty$. Let $g \in C_b^{(3)}(\mathbb{R})$. We apply Theorem 42 to this situation. A small issue is that f is complex valued, but one can apply the theorem separately to $\Re f$ and $\Im f$ and put them together. We ignore this issue and just apply the bound in Theorem 42 directly to f and leave the rest as exercise. The only thing needed is to compute $\lambda_3(f)$. Let $H_{i,j}$ denote the Hermitian matrix which has 1 in the (i,j) and $(j,,i)$ slots, and zeros elsewhere. As $f(\mathbf{t}) = \mathbf{n^{-1}}\mathrm{tr}(\mathbf{zI} - \mathbf{M(t)})^{-\mathbf{1}}$ and $\partial_{i,j}M(\mathbf{t}) = \mathbf{n^{-1/2}H_{i,j}}$, we get

$$\begin{aligned}
\partial_{i,j}f(\mathbf{t}) &= n^{-1}\mathrm{tr}\left\{(zI - M(\mathbf{t}))^{-\mathbf{2}}\partial_{\mathbf{i,j}}\mathbf{M(t)}\right\} = n^{-3/2}\mathrm{tr}\left\{(zI - M(\mathbf{t}))^{-\mathbf{2}}\mathbf{H_{i,j}}\right\},\\
\partial^2_{i,j}f(\mathbf{t}) &= n^{-2}\mathrm{tr}\left\{(zI - M(\mathbf{t}))^{-\mathbf{3}}\partial_{\mathbf{i,j}}\mathbf{M(t)H_{i,j}}\right\} = n^{-5/2}\mathrm{tr}\left\{(zI - M(\mathbf{t}))^{-\mathbf{3}}\mathbf{H^2_{i,j}}\right\},\\
\partial^3_{i,j}f(\mathbf{t}) &= n^{-7/2}\mathrm{tr}\left\{(zI - M(\mathbf{t}))^{-\mathbf{4}}\mathbf{H^3_{i,j}}\right\}.
\end{aligned}$$

If M is any square matrix, let $M^*M = \sum \theta_k \mathbf{v}_k\mathbf{v}_k^*$ be the spectral decomposition of M^*M. For any matrix A, we then have

$$|\mathrm{tr}(MA)|^2 \le \mathrm{tr}(A^*M^*MA) = \sum_k \theta_k \|A^*\mathbf{v}_k\|^2 \le \theta_{max}\sum_k \|A^*\mathbf{v}_k\|^2 = \theta_{\max}\mathrm{tr}(A^*A).$$

We apply this with $(zI - M(\mathbf{t}))^{-\mathbf{p-1}}$ in place of M and $H_{i,j}^p$ in place of A (for $p = 1,2,3$). Then $\theta_{\max} \le (\Im z)^{-2p-2}$ and $\mathrm{tr}(H_{i,j}^{2p})$ is bounded by a constant for $p \le 3$. Hence we get

$$|\partial_{i,j}f(\mathbf{t})| \le \mathbf{Cn^{-3/2}}(\Im \mathbf{z})^{-\mathbf{2}}, \qquad |\partial^{\mathbf{2}}_{\mathbf{i,j}}\mathbf{f(t)}| \le \mathbf{Cn^{-5/2}}(\Im \mathbf{z})^{-\mathbf{4}}, \qquad |\partial^{\mathbf{3}}_{\mathbf{i,j}}\mathbf{f(t)}| \le \mathbf{Cn^{-7/2}}(\Im \mathbf{z})^{-\mathbf{4}}.$$

Thus, $\lambda_3(f) \le Cn^{-7/2}$. Theorem 42 implies

$$\left|\mathbf{E}\left[g(G_{\mathbf{X}}(z))\right] - \mathbf{E}\left[g(G_{\mathbf{Y}}(z))\right]\right| \le C\gamma_3 n^{-3/2}.$$

Thus, if $G_{\mathbf{Y}}(z) \xrightarrow{P} G_{s.c}(z)$, then so does $G_{\mathbf{X}}(z)$.

13. Summary

We have essentially seen one theorem in random matrix theory, the Wigner's semicircle law (and Marchenko-Pastur law, if you solved the exercises). This is the first nontrivial result in the subject, but another reason for spending time on this is to introduce the more widely applicable techniques of moment method, stieltjes' transforms and invariance principle. Put together, we have proved Theorem 15 under the assumption that $X_n = (X_{i,j}^{(n)})_{i,j\le n}$ has independent (for $i \le j$) with finite first and second moment and satisfying the Pastur condition

$$\frac{1}{n^2}\sum_{i,j=1}^{n} \mathbf{E}\left[|X_{,j}|^2\mathbf{1}_{|X_{i,j}|\ge \delta\sqrt{n}}\right] \to 0 \quad \text{for all } \delta > 0.$$

A similar result holds for Wishart matrices. Let us remark on some other models of random matrices.

(1) Jacobi random matrices: Let $X_{m\times n}$ and $Y_{m\times n}$ be independent matrices with i.i.d entries. Then $A = (XX^* + YY^*)^{-1}XX^*$ is a random Hermitian matrix. Let $m \le n$ and $m/n \to c$ (positive and finite) as before. *Without any scaling*, the ESD L_{A_n} converges to a Beta distribution

with parameters depending on c. This can be handled by the methods given earlier. Together with the Wigner and Wishart matrices, this is one of the classical models of random Hermitian matrices. In all these cases, the Gaussian versions have particularly nice properties, for example, the exact eigenvalue density may be found.

(2) Random Toeplitz matrices: A matrix $T_{n\times n}$ is said to be *Toeplitz* if $T_{i,j}$ depends only on $j-i$. If we pick i.i.d real valued random variables $X_0,\ldots,X_{n-1}$ and define an $n\times n$ matrix T with $T_{i,j}=X_{|j-i|}$, then we get a random, real symmetric Toeplitz matrix. If X_k have mean zero and finite variance, the ESD of $T/\sqrt{n}$ converges to a probability measure on the line. This has been shown by the method of moments and in some sense the moments of the limit distribution are understood. However, it is not known at present whether the limit distribution has a density!

(3) Other models of random matrices with various structures like Toeplitz are being studied. For example, *Hankel matrices* have $H_{i,j}$ depending on $i+j$ only. Real symmetric Hankel matrices with $H_{i,j}=X_{|i+j|}$, where X_k are i.i.d with zero mean and finite variance have been looked at. $H/sqrtn$ has a limit distribution which is again not fully understood.

(4) Consider a Wigner matrix, but drop the finite variance condition. For example, $X_{i,j}$, for $i\leq j$ could be i.i.d Cauchy random variables. What is the right scaling, and what is the limit distribution? Again, some results have been found in very recent time (the scaling is by $n^{-1/\alpha}$ if the entries fall in the Stable(α) domain of attraction, and then the ESD converge to a (nonrandom) probability measure. Various properties of the limit measure are yet to be understood.

CHAPTER 3

GOE and GUE

We quickly recall that a GUE matrix can be defined in the following three equivalent ways. We leave it to the reader to make the three analogous statements for GOE.

In the previous chapters, GOE and GUE matrices appeared merely as special cases of Wigner matrices for which computations were easier. However they have a great many neat properties not shared by other Wigner matrices. The main fact is that the exact density of eigenvalues of GOE and GUE can be found explicitly! And even more surprisingly, these exact densities have a nice structure that make them amenable to computations. Many results that are true for general Wigner matrices are much harder to prove in general but fairly easy for these two cases. Crucial to the "integrability" properties of GOE and GUE are their invariance under orthogonal and unitary conjugations respectively.

Exercise 46. (a) Let X and Y be $n \times n$ GUE and GOE matrices respectively. Then, for any fixed $U \in \mathcal{U}(n)$ and $P \in O(n)$, we have $U^*XU \stackrel{d}{=} X$ and $P^tYP \stackrel{d}{=} Y$.

(b) If X is a random matrix such that $X_{i,j}$, $i \leq j$ are independent real valued entries and suppose that $PXP^t \stackrel{d}{=} X$ for all $P \in O(n)$, then show that X has the same distribution as $c\tilde{X}$ where c is a constant and $\tilde{X}$ is a GOE matrix. The analogous statement for unitary invariance is also true.

Remark 47. This is analogous to the following well known fact. Let X be a random vector in $\mathbb{R}^n$. Then the following are equivalent.

(1) $X \sim N_n(0, \sigma^2 I)$ for some σ^2.
(2) X_k are independent and $PX \stackrel{d}{=} X$ for any $P \in O(n)$.

To see that the second implies the first, take for P an orthogonal matrix whose first column is $(1/\sqrt{2}, 1/\sqrt{2}, 0, \ldots, 0)$ to get $X_1 \stackrel{d}{=} (X_1 + X_2)/\sqrt{2}$. Further, X_1, X_2 are i.i.d - independence is given, and choosing P to be a permutation matrix we get identical distributions. It is well known that the only solutions to this distributional equation are the $N(0, \sigma^2)$ distributions. If not convinced, use characteristic functions or otherwise show this fact.

What is the use of unitary or orthogonal invariance? Write the spectral decomposition of a GUE matrix $X = VDV^*$. For any fixed $U \in \mathcal{U}(n)$, then $UXU^* = (UV)D(UV)^*$. By the unitary invariance, we see that VDV^* has the same distribution as $(UV)D(UV)^*$. This suggests that V

and D are independent. The only hitch in this reasoning is that the spectral decomposition is not exactly unique, but it can be taken care of[1]

1. Tridiagonalization

Let A be an $n \times n$ GOE. Write it as

$$A = \begin{bmatrix} a & \mathbf{u}^t \\ \mathbf{u} & B \end{bmatrix}$$

so that $a \sim N(0,2)$, $\mathbf{u} \sim N_{n-1}(0, I)$, $B \sim \text{GOE}_{n-1}$, and all three are independent. Condition on $\mathbf{u}$. Then pick any orthogonal matrix $P \in O(n-1)$ such that $P\mathbf{u} = \|u\|\mathbf{e}_1$. To be specific, we can take the transformation defined by

$$P\mathbf{v} = \mathbf{v} - 2\frac{\langle \mathbf{v}, \mathbf{w} \rangle}{\langle \mathbf{w}, \mathbf{w} \rangle}\mathbf{w}, \quad \text{with } \mathbf{w} = \mathbf{u} - \mathbf{e}_1.$$

For any $\mathbf{w} \neq 0$, the transformation defined on the left is the reflection across the hyperplane perpendicular to $\mathbf{w}$. These are also referred to as *Householder reflections*. Check that P is indeed unitary and that $P\mathbf{u} = \mathbf{e}_1$.

Since P depends on $\mathbf{u}$ and B is independent of U, the orthogonal invariance of GOE shows that $A_1 := P^t B P \stackrel{d}{=} B$, that is A_1 is a GOE matrix. Also A_1 is independent of $\mathbf{u}$ and a. Thus,

$$C := \begin{bmatrix} 1 & \mathbf{0}^t \\ \mathbf{0} & P^t \end{bmatrix} \begin{bmatrix} a & \mathbf{u}^t \\ \mathbf{u} & B \end{bmatrix} \begin{bmatrix} 1 & \mathbf{0}^t \\ \mathbf{0} & P \end{bmatrix} = \begin{bmatrix} a & r_1 \mathbf{e}_1^t \\ r_1 \mathbf{e}_1 & A_1 \end{bmatrix}$$

where $A_1 \sim GOE_{n-1}$, $a \sim N(0,1)$ and $r_1 = \|u\|$ are all independent. Since C is an orthogonal conjugation of A, the eigenvalues of A and C are exactly the same. Observe that $C_{j,1} = C_{1,j} = 0$ for $2 \leq j \leq n$. Note that $r_1^2 = \|u\|^2$ has χ^2_{n-1} distribution.

Now A_1 is a GOE matrix of one less order. We can play the same game with A_1 and get a matrix D which is conjugate to A_1 but has $D_{1,j} = D_{j,1} = 0$ for $2 \leq j \leq n-1$. Combining with the previous one, we get

$$C_2 := \begin{bmatrix} a & r_1 & \mathbf{0}^t \\ r_1 & a' & r_2 \mathbf{e}_1^t \\ \mathbf{0} & r_2 \mathbf{e}_1 & D \end{bmatrix}$$

with the following properties. C_2 is conjugate to A and hence has the same eigenvalues. $D \sim \text{GOE}_{n-2}$, $a, a' \sim N(0,2)$, $r_1^2 \sim \chi^2_{n-1}$, $r_2^2 \sim \chi^2_{n-2}$, and all these are independent.

[1]The eigenspace for a given eigenvalue is well-defined. This is the source of non-uniqueness. The set S of Hermitian matrices having distinct eigenvalues is a dense open set in the space of all Hermitian matrices. Therefore, almost surely, a GUE matrix has no eigenvalues of multiplicity more than one (explain why). However, even when the eigenspace is one dimensional, we can multiply the eigenvector by $e^{i\theta}$ for some $\theta \in \mathbb{R}$ and that leads to non-uniqueness. To fix this, let $\mathcal{D}(n)$ be the group of $n \times n$ diagonal unitary matrices and consider the quotient space $Q = \mathcal{U}(n)/\mathcal{D}(n)$ consisting of right cosets. Then, the mapping $X \to ([V], D)$ is one to one and onto on S. Now observe that for any U, $([UV], D) \stackrel{d}{=} ([V], D)$ and hence $[V]$ and D are independent.

It is clear that this procedure can be continued and we end up with a tridiagonal matrix that is orthogonally conjugate to A and such that

$$T_n = \begin{bmatrix} a_1 & b_1 & 0 & 0 & \dots & 0 \\ b_1 & a_2 & b_2 & 0 & \dots & 0 \\ 0 & b_2 & a_3 & b_3 & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & b_{n-2} & a_{n-1} & b_{n-1} \\ 0 & \dots & 0 & 0 & b_{n-1} & a_n \end{bmatrix} \tag{20}$$

where $a_k \sim N(0,2)$, $b_k^2 \sim \chi^2_{n-k}$, and all these are independent.

Exercise 48. If A is an $n \times n$ GUE matrix, show that A is conjugate to a tridiagonal matrix T as in (20) where a_k, b_k are all independent, $a_k \sim N(0,1)$ and $b_k^2 \sim \text{Gamma}(n-k,1)$.

Recall that χ^2_p is the same as $\text{Gamma}(\frac{p}{2}, \frac{1}{2})$ or equivalently, the distribution of $2Y$ where $Y \sim \text{Gamma}(\frac{p}{2})$. Thus, we arrive at the following theorem[2].

Theorem 49. *Let T be a tridiagonal matrix as in* (20) *where a_k, b_k are all independent.*

1. *If $a_k \sim N(0,2)$ and $b_k^2 \sim \chi^2_{n-k}$, then the vector of eigenvalues of T has the same distribution as the vector of eigenvalues of a GOE_n matrix.*
2. *If $a_k \sim N(0,2)$ and $b_k^2 \sim \chi^2_{2(n-k)'}$ then the vector of eigenvalues of T has the same distribution as the eigenvalues of a GUE_n matrix scaled by a factor of $\sqrt{2}$.*

2. Tridiagonal matrices and probability measures on the line

Our objective is to find eigenvalue density for certain random matrices, and hence we must find $n-1$ auxiliary parameters in addition to the n eigenvalues (since there are $2n-1$ parameters in the tridiagonal matrix) to carry out the Jacobian computation. The short answer is that if UDU^* is the spectral decomposition of the tridiagonal matrix, then $p_j = |U_{1,j}|^2$, $1 \le j \le n-1$ are the right parameters to choose. However, there are many conceptual reasons behind this choice and we shall spend the rest of this section on these concepts.

Fix $n \ge 1$ and write $T = T(a,b)$ for the real symmetric $n \times n$ tridiagonal matrix with diagonal entries $T_{k,k} = a_k$ for $1 \le k \le n$ and $T_{k,k+1} = T_{k+1,k} = b_k$ for $1 \le k \le n-1$.

Let $\mathcal{T}_n$ be the space of all $n \times n$ real symmetric tridiagonal matrices and let $\mathcal{T}_n^0$ be those $T(a,b)$ in $\mathcal{T}_n$ with n distinct eigenvalues. Let $\mathcal{P}_n$ be the space of all probability measures on $\mathbb{R}$ whose support consists of at most n distinct points and let $\mathcal{P}_n^0$ be those elements of $\mathcal{P}_n$ whose support has exactly n distinct points.

Tridiagonal matrix to probability measure: Recall that the *spectral measure* of a Hermitian operator T at a vector $\mathbf{v}$ is the unique measure ν on $\mathbb{R}$ such that $\langle T^p \mathbf{v}, \mathbf{v} \rangle = \int x^p \nu(dx)$ for all $p \ge 0$.

[2]The idea of tridiagonalizing the GOE and GUE matrices was originally due to Hale Trotter **?**. Part of his original motivation was to give a simple proof of the semicircle law for GOE and GUE matrices.

For example, if T is a real symmetric matrix, write its spectral decomposition as $T = \sum_{k=1}^{n} \lambda_k \mathbf{u}_k \mathbf{u}_k^*$. Then $\{\mathbf{u}_k\}$ is an ONB of $\mathbb{R}^n$ and λ_k are real. In this case, the spectral decomposition of T at any $\mathbf{v} \in \mathbb{R}^n$ is just $\nu = \sum_{k=1}^{n} |\langle \mathbf{v}, \mathbf{u}_k \rangle|^2 \delta_{\lambda_k}$. Thus $\nu \in \mathcal{P}_n$ (observe that the support may have less than n points as eigenvalues may coincide). In particular, the spectral measure of T at $\mathbf{e}_1$ is $\sum p_j \delta_{\lambda_j}$ where $p_j = |U_{1,j}|^2$ (here $U_{1,j}$ is the first co-ordinate of $\mathbf{u}_j$).

Given a real symmetric tridiagonal matrix T, let ν_T be the spectral measure of T at the standard unit vector $\mathbf{e}_1$. This gives a mapping from $\mathcal{T}_n$ into $\mathcal{P}_n$ which maps T_n^0 into P_n^0.

Probability measure to Tridiagonal matrix: Now suppose a measure $\mu \in \mathcal{P}_n^0$ is given. Write $\mu = p_1 \delta_{\lambda_1} + \ldots + p_n \delta_{\lambda_n}$ where λ_j are distinct real numbers and $p_j > 0$. Its moments are given by $\alpha_k = \sum p_j \lambda_j^k$. Let $h_k(x) = x^k$, so that $\{h_0, h_1, \ldots, h_{n-1}\}$ is a basis for $L^2(\mu)$ (how do you express h_n as a linear combination of $h_0, \ldots, h_{n-1}$?).

Apply Gram-Schmidt to the sequence $h_0, h_1, \ldots$ by setting $\phi_0 = \psi_0 = h_0$, and for $k \geq 1$ inductively by

$$\psi_k = h_k - \sum_{j=0}^{k-1} \langle h_k, \phi_j \rangle \phi_j, \qquad \phi_k = \frac{\psi_k}{\|\psi_k\|_{L^2(\mu)}}.$$

This process is stopped when $\|\psi_k\| = 0$. Here are some elementary observations.

(a) Since $\{h_0, \ldots, h_{n-1}\}$ is a linear basis for $L^2(\mu)$, it follows that $\{\phi_0, \ldots, \phi_{n-1}\}$ are well-defined and form an ONB for $L^2(\mu)$.

(b) For $0 \leq k \leq n-1$, ϕ_k is a polynomial of degree k and is orthogonal to all polynomials of degree less than k.

(c) As h_n is a linear combination of $h_0, \ldots, h_{n-1}$ (in $L^2(\mu)$), we see that ψ_n is well-defined but $\|\psi_n\| = 0$ and hence ϕ_n is not defined. Note that $\|\psi_n\| = 0$ means that $\psi_n(\lambda_k) = 0$ for all $k \leq n$, not that ψ_n is the zero polynomial. In fact, ψ_n is monic, has degree n and vanishes at λ_k, $k \leq n$, which implies that $\psi_n(\lambda) = \prod_{j=1}^{n} (\lambda - \lambda_j)$.

Fix $0 \leq k \leq n-1$ and expand $x\phi_k(x)$ as

$$x\phi_k(x) \overset{L^2(\mu)}{=} \sum_{j=0}^{n} c_{k,j} \phi_j(x), \qquad c_{k,j} = \int x\phi_k(x)\phi_j(x) d\mu(x).$$

Now, $x\phi_j(x)$ has degree less than k if $j < k$ and $x\phi_k(x)$ has degree less than j if $k < j$. Hence, $c_{k,j} = 0$ if $j \leq k-2$ or if $j \geq k+2$. Further, $c_{k,k+1} = c_{k+1,k}$ as both are equal to $\int x\phi_k(x)\phi_{k+1}(x) d\mu(x)$. Thus, we get the *three term recurrences*

$$\begin{aligned} x\phi_k(x) &\overset{L^2(\mu)}{=} b_{k-1}\phi_{k-1}(x) + a_k\phi_k(x) + b_k\phi_{k+1}(x), \qquad 0 \leq k \leq n \\ \text{where} \qquad & a_k = \int x\phi_k(x)^2 d\mu(x),\ b_k = \int x\phi_k(x)\phi_{k+1}(x) d\mu(x). \end{aligned} \tag{21}$$

We adopt the convention that ϕ_{-1}, ϕ_n, b_{-1} and b_{n-1} are all zero, so that these recurrences also hold for $k = 0$ and $k = n$. Since ϕ_k all have positive leading co-efficients, it is not hard to see that b_k is nonnegative.

From $\mu \in \mathcal{P}_n^0$ we have thus constructed a tridiagonal matrix $T_\mu := T(a,b) \in \mathcal{T}_n$ (caution: here we have indexed a_k, b_k starting from $k = 0$). If $\mu \in \mathcal{P}_m^0$ for some $m < n$, the T_μ constructed as before will

have size $m \times m$. Extend this by padding $n-m$ columns and rows of zeros to get a real symmetric tridiagonal matrix (we abuse notation and denote it as T_μ again) in $\mathcal{T}_n$. Thus we get a mapping $\mu \to T_\mu$ from $\mathcal{P}_n$ into $\mathcal{T}_n$.

The following lemma shows that $T \to \nu_T$ is a bijection, and relates objects defined on one side (matrix entries, characteristic polynomials, eigenvalues) to objects defined on the other side (the support $\{\lambda_j\}$, the weights p_j, associated orthogonal polnomials).

Lemma 50. *Fix $n \geq 1$.*

(a) The mapping $T \to \nu_T$ is a bijection from $\mathcal{T}_n^0$ into $\mathcal{P}_n^0$ whose inverse is $\mu \to T_\mu$.

(b) Let $T = T(a,b)$ and let $\mu = \nu_T$. For $0 \leq k \leq n-1$ P_k be the characteristic polynomial of the top $k \times k$ principal submatrix of T and let ψ_k, ϕ_k be as constructed earlier. Then $\psi_k = P_k$ for $k \leq n$ and hence there exist constants d_k such that $\phi_k = d_k P_k$ (for $k \leq n-1$).

(c) The zeros of ϕ_n are precisely the eigenvalues of T.

(d) If $T = T(a,b)$ and $\nu_T = \sum_{k=1}^n p_j \delta_{\lambda_j}$, then

$$\prod_{k=1}^n b_k^{2(n-k+1)} = \prod_{k=1}^n p_k \prod_{i<j} |\lambda_i - \lambda_j|^2. \tag{22}$$

In particular, $\mathcal{T}_n^0$ gets mapped into $\mathcal{P}_n^0$ (but not onto).

PROOF. (a) Let $\mu = \sum_{j=1}^n p_j \delta_{\lambda_j} \in \mathcal{P}_n$ and let $T = T_\mu$. For $0 \leq k \leq n-1$, let

$$\mathbf{u}_k = (\sqrt{p_1}\phi_0(\lambda_k), \ldots, \sqrt{p_n}\phi_{n-1}(\lambda_k))^t.$$

The three-term recurrences can be written in terms of T as $T\mathbf{u}_k = \lambda_k \mathbf{u}_k$. Thus, $\mathbf{u}_k$ is an eigenvector of T with eigenvalue λ_k. If U is the matrix with columns $\mathbf{u}_k$, then the rows of U are orthonormal because ϕ_k are orthogonal polynomials of μ. Thus $UU^* = I$ and hence also $U^*U = I$, that is $\{\mathbf{u}_k\}$ is an ONB of $\mathbb{R}^n$.

Consequently, $T = \sum_{k=1}^n \lambda_k \mathbf{u}_k \mathbf{u}_k^*$ is the spectral decomposition of T. In particular,

$$T^p \mathbf{e}_1 = \sum_{k=1}^n |u_{k,1}|^2 \lambda_k^p = \sum_{k=1}^n p_k \lambda_k^p$$

because $u_{k,1} = \sqrt{p_k}\phi_0(\lambda_k) = \sqrt{p_k}$ (as $h_0 = 1$ is already of unit norm in $L^2(\mu)$ and hence after Gram-Schmidt $\phi_0 = h_0$). Thus, $\langle T^p \mathbf{e}_1, \mathbf{e}_1 \rangle = \int x^p \mu(dx)$ which shows that $\nu_T = \mu$. This proves the first part of the lemma.

(b) We saw earlier that ϕ_n is zero in $L^2(\mu)$. Hence $\phi_n(\lambda_j) = 0$ for $1 \leq j \leq n$. Thus, ϕ_n and P_n are non-zero polynomials of degree n both of which vanish at the same n points. Hence, $\phi_n = d_n P_n$ for some constant P_n.

If S is the top $k \times k$ principal submatrix of T, then it is easy to see that the first k orthogonal polynomials of ν_S are the same as $\phi_0, \ldots, \phi_k$ (which were obtained as orthogonal polynomials of ν_T). This is easy to see from the three-term recurrences. Thus the above fact shows $\phi_k = d_k P_k$ for some constant d_k.

(c) By the first part, $\nu_T = \sum p_j \delta_{\lambda_j}$ where λ_j are the eigenvalues of T and $p_j > 0$. The footnote on the previous page also shows that ϕ_n vanishes at λ_j, $j \le n$. Since it has degree n, ϕ_n has only these zeros.

(d) This proof is taken from Forrester's book. Let T_k denote the bottom $(n-k)\times(n-k)$ principal submatrix of T. Let Q_k be its characteristic polynomial and let $\lambda_j^{(k)}$, $1 \le j \le n-k$ be its eigenvalues. In particular, $T_0 = T$.

If $T = \sum \lambda_k \mathbf{u}_k \mathbf{u}_k^*$ is the spectral decomposition of T and λ is not an eigenvalue of T, then $(\lambda I - T)^{-1} = \sum (\lambda - \lambda_k)^{-1} \mathbf{u}_k \mathbf{u}_k^*$. Hence, $(\lambda I - T)^{1,1} = \langle (\lambda I - T)^{-1} \mathbf{e}_1, \mathbf{e}_1 \rangle = \sum_j p_j (\lambda - \lambda_j)^{-1}$ for $\lambda \notin \{\lambda_j\}$. But we also know that $(\lambda I - T)^{1,1}$ is equal to $\det(\lambda I - T_1)/\det(\lambda I - T) = Q_1(\lambda)/Q_0(\lambda)$. Let λ approach λ_k to see that

$$p_k = \lim_{\lambda \to \lambda_k} (\lambda - \lambda_k)(\lambda I - T)^{1,1} = \lim_{\lambda \to \lambda_k} (\lambda - \lambda_k) \frac{Q_1(\lambda)}{Q_0(\lambda)} = \frac{Q_1(\lambda_k)}{\prod_{j=1 j \ne k}^{n} (\lambda_k - \lambda_j)}.$$

Take product over k to get (the left side is positive, hence absolute values on the right)

$$\prod_{k=1}^{n} p_k \prod_{i<j} (\lambda_i - \lambda_j)^2 = \prod_{k=1}^{n} |Q_1(\lambda_k^{(0)})|. \tag{23}$$

Let A be any $n \times n$ matrix with characteristic polynomial χ_A and eigenvalues λ_i. Let B be an $m \times m$ matrix with characteristic polynomial χ_B and eigenvalues μ_j. Then we have the obvious identity

$$\prod_{i=1}^{n} |\chi_B(\lambda_i)| = \prod_{i=1}^{n} \prod_{j=1}^{m} |\mu_j - \lambda_i| = \prod_{j=1}^{m} |\chi_A(\mu_j)|.$$

Apply to T_0 and T_1 to get $\prod_{k=1}^{n} |Q_1(\lambda_k^{(0)})| = \prod_{k=1}^{n-1} |Q_0(\lambda_k^{(1)})|$. But by expanding $\det(\lambda I - T)$ by the first row, we also have the identity

$$Q_0(\lambda) = (\lambda - a_1) Q_1(\lambda) - b_1^2 Q_2(\lambda).$$

Therefore $Q_0(\lambda_k^{(1)}) = b_1^2 Q_2(\lambda_k^{(1)})$ for $k \le n-1$. Thus $\prod_{k=1}^{n} |Q_1(\lambda_k^{(0)})| = b_1^{2n-2} \prod_{k=1}^{n-1} Q_2(\lambda_k^{(1)})$. The right side is of a similar form to the left side, with matrix size reduced by one. Thus, inductively we get $\prod_{k=1}^{n} |Q_1(\lambda_k^{(0)})| = \prod_{k=1}^{n-1} b_k^{2n-2k}$. Plugging into (23) we get the statement of the lemma. ■

3. Tridiagonal matrix generalities

Fix $n \ge 1$ and write $T = T(a,b)$ for the real symmetric $n \times n$ tridiagonal matrix with diagonal entries $T_{k,k} = a_k$ for $1 \le k \le n$ and $T_{k,k+1} = T_{k+1,k} = b_k$ for $1 \le k \le n-1$. Let $\mathcal{T}_n$ be the space of all $n \times n$ real symmetric tridiagonal matrices and let $\mathcal{T}_n^0$ be those $T(a,b)$ in $\mathcal{T}_n$ with b_k strictly positive. Let $\mathcal{P}_n$ be the space of all probability measures on $\mathbb{R}$ whose support consists of at most n distinct points and let $\mathcal{P}_n^0$ be those elements of $\mathcal{P}_n$ whose support has exactly n distinct points.

Given a real symmetric tridiagonal matrix T, let ν_T be the spectral measure of T at the standard unit vector $\mathbf{e}_1$[3]. This gives a mapping from $\mathcal{T}_n$ into $\mathcal{P}_n$. For future purpose, we also give the following idea to find eigenvalues of T.

Fix some $\lambda \in \mathbb{R}$ and suppose we want to find a vector $\mathbf{v}$ such that $T\mathbf{v} = \lambda\mathbf{v}$. This means

$$b_{k-1}v_{k-1} + a_k v_k + b_k v_{k+1} = \lambda v_k \implies v_{k+1} = \frac{\lambda v_k - b_{k-1}v_{k-1} - a_k v_k}{b_k}.$$

where we adopt the convention that $b_0 = 0$. We have also assumed that $b_k \neq 0$ for all k (if $b_k = 0$, the matrix splits into a direct sum of two matrices). Thus, we set $\mathbf{v}_1 = x$ to be arbitrary (non-zero)and solve for $v_1, v_2, \ldots$ successively. Denote these as $v_1(x), v_2(x), \ldots$. Therefore,

Now suppose a measure $\mu \in \mathcal{P}_n^0$ is given. We can construct a tridiagonal matrix T as follows. Write $\mu = p_1\delta_{\lambda_1} + \ldots + p_n\delta_{\lambda_n}$ where λ_j are distinct real numbers and $p_j > 0$. The moments are given by $\alpha_k = \sum p_j\lambda_j^k$. Let $h_k(x) = x^k$, so that $\{h_0, h_1, \ldots, h_{n-1}\}$ is a basis for $L^2(\mu)$. [Q: How do you express h_n as a linear combination of $h_0, \ldots, h_{n-1}$?].

Apply Gram-Schmidt to the sequence $h_0, h_1, \ldots$ to get an orthonormal basis $\{\phi_k : 0 \leq k \leq n-1\}$ of $L^2(\mu)$. It is easy to see that ϕ_k is a polynomial of degree exactly k, and is orthogonal to all polynomials of degree less than k. Fix any k and write

$$x\phi_k(x) \overset{L^2(\mu)}{=} \sum_{j=0}^{n} c_{k,j}\phi_j(x), \qquad c_{k,j} = \int x\phi_k(x)\phi_j(x)d\mu(x).$$

Now, $x\phi_j(x)$ has degree less than k if $j < k$ and $x\phi_k(x)$ has degree less than j if $k < j$. Hence, $c_{k,j} = 0$ if $j \leq k-2$ or if $j \geq k+2$. Further, $c_{k,k+1} = c_{k+1,k}$ as both are equal to $\int x\phi_k(x)\phi_{k+1}(x)d\mu(x)$. Thus, we get the *three term recurrences*

$$\begin{aligned} x\phi_k(x) &\overset{L^2(\mu)}{=} b_{k-1}\phi_{k-1}(x) + a_k\phi_k(x) + b_k\phi_{k+1}(x), \qquad 0 \leq k \leq n \\ \text{where} \qquad & a_k = \int x\phi_k(x)^2 d\mu(x),\ b_k = \int x\phi_k(x)\phi_{k+1}(x)d\mu(x). \end{aligned} \tag{24}$$

We adopt the convention that ϕ_{-1}, ϕ_n, b_{-1} and b_{n-1} are all zero, so that these recurrences also hold for $k = 0$ and $k = n$.

From $\mu \in \mathcal{P}_n^0$ we have thus constructed a tridiagonal matrix $T_\mu := T(a,b) \in \mathcal{T}_n$ (caution: here we have indexed a_k, b_k starting from $k = 0$). If $\mu \in \mathcal{P}_m^0$ for some $m < n$, the T_μ constructed as before will have size $m \times m$. Extend this by padding $n-m$ columns and rows of zeros to get a real symmetric tridiagonal matrix (we abuse notation and denote it as T_μ again) in $\mathcal{T}_n$. Thus we get a mapping $\mu \to T_\mu$ from $\mathcal{P}_n$ into $\mathcal{T}_n$.

Lemma 51. *Fix $n \geq 1$.*

(a) The mapping $T \to \nu_T$ is a bijection from $\mathcal{T}_n$ into $\mathcal{P}_n$ whose inverse is $\mu \to T_\mu$.

[3]The *spectral measure* of a Hermitian operator T at a vector $\mathbf{v}$ is the unique measure ν on $\mathbb{R}$ such that $\langle T^p\mathbf{v}, \mathbf{v}\rangle = \int x^p \nu(dx)$ for all $p \geq 0$. For example, if T is a real symmetric matrix, write its spectral decomposition as $T = \sum_{k=1}^n \lambda_k \mathbf{u}_k\mathbf{u}_k^*$. Then $\{\mathbf{u}_k\}$ is an ONB of $\mathbb{R}^n$ and λ_k are real. In this case, the spectral decomposition of T at any $\mathbf{v} \in \mathbb{R}^n$ is just $\nu = \sum_{k=1}^n |\langle \mathbf{v}, \mathbf{u}_k\rangle|^2 \delta_{\lambda_k}$. Thus $\nu \in \mathcal{P}_n$ (observe that the support may have less than n points as eigenvalues may coincide). In particular, if $T = UDU^*$ the spectral measure of T at $\mathbf{e}_1$ is $\nu_T = \sum p_i\delta_{\lambda_i}$, where $p_i = |U_{1,i}|^2$.

(b) *Let $\mu = \nu_T$. Write P_k for the characteristic polynomial of the top $k \times k$ submatrix of T for $k \leq n$ and let ϕ_k be the orthogonal polynomials for μ as defined earlier. Then $\phi_k = d_k P_k$ for for constants d_k. In particular, zeros of ϕ_n are precisely the eigenvalues of T.*

(c) *If $T = T(a,b)$ and $\mu = \sum_{k=1}^n p_j \delta_{\lambda_j}$ correspond to each other in this bijection, then*

$$\prod_{k=1}^n b_k^{2(n-k+1)} = \prod_{k=1}^n p_k \prod_{i<j} |\lambda_i - \lambda_j|^2. \tag{25}$$

In particular, $\mathcal{T}_n^0$ gets mapped into $\mathcal{P}_n^0$ (but not onto).

PROOF. (a) Let $\mu = \sum_{j=1}^n p_j \delta_{\lambda_j} \in \mathcal{P}_n$ and let $T = T_\mu$. For $0 \leq k \leq n-1$, let

$$\mathbf{u}_k = (\sqrt{p_1}\phi_0(\lambda_k), \ldots, \sqrt{p_n}\phi_{n-1}(\lambda_k))^t.$$

The three-term recurrences can be written in terms of T as $T\mathbf{u}_k = \lambda_k \mathbf{u}_k$. Thus, $\mathbf{u}_k$ is an eigenvector of T with eigenvalue λ_k. If U is the matrix with columns $\mathbf{u}_k$, then the rows of U are orthonormal because ϕ_k are orthogonal polynomials of μ. Thus $UU^* = I$ and hence also $U^*U = I$, that is $\{\mathbf{u}_k\}$ is an ONB of $\mathbb{R}^n$.

Consequently, $T = \sum_{k=1}^n \lambda_k \mathbf{u}_k \mathbf{u}_k^*$ is the spectral decomposition of T. In particular,

$$T^p \mathbf{e}_1 = \sum_{k=1}^n |u_{k,1}|^2 \lambda_k^p = \sum_{k=1}^n p_k \lambda_k^p$$

because $u_{k,1} = \sqrt{p_k}\phi_0(\lambda_k) = \sqrt{p_k}$ ($h_0 = 1$ is already of unit norm in $L^2(\mu)$ and hence after Gram-Schmidt $\phi_0 = h_0$). Thus, $\langle T^p \mathbf{e}_1, \mathbf{e}_1 \rangle = \int x^p \mu(dx)$ which shows that $\nu_T = \mu$. This proves the first part of the lemma.

(b) By part (a), the coefficients in the three term recurrence (24) are precisely the entries of T. Note that the equality in (24) is in $L^2(\mu)$, which means the same as saying that equality holds for $x = \lambda_k$, $1 \leq k \leq n$.

Here is a way to find

$$T$$

(c) Let A be any $n \times n$ matrix with characteristic polynomial χ_A and eigenvalues λ_i. Let B be an $m \times m$ matrix with characteristic polynomial χ_B and eigenvalues μ_j. Then we have the obvious identity

$$\prod_{i=1}^n \chi_B(\lambda_i) = \prod_{i=1}^n \prod_{j=1}^m (\mu_j - \lambda_i) = (-1)^{mn} \prod_{j=1}^m \chi_A(\mu_j)$$

If b_k are all positive, the right hand side of (**??**) is non-zero and hence λ_k must be distinct. This shows that $\mathcal{T}_n^0$ gets mapped into $\mathcal{P}_n^0$. It is obviously not onto (why?). ■

Lemma 52. *For $T = T(a,b)$ having the spectral measure $\sum_{j=1}^{n+1} p_i \delta_{\lambda_j}$ at $\mathbf{e}_0$, we have the identity*

$$\prod_{k=0}^{n-1} b_k^{2(n+1-k)} = \prod_{i=1}^{n+1} p_i \prod_{i<j\leq n+1} (\lambda_i - \lambda_j)^2.$$

4. More on tridiagonal operators*

This section may be omitted as we shall not use the contents in this course. However, as we are this close to a very rich part of classical analysis, we state a few interesting facts. The following four objects are shown to be intimately connected.

(1) Positive measures $\mu \in \mathcal{P}(\mathbb{R})$ having all moments.
(2) Positive definite sequences $\alpha = (\alpha_k)_{k\geq 0}$ such that $(\alpha_{i+j})_{i,j\geq 0}$ is non-negative definite (that is every principal submatrix has non-negative determinant).
(3) Orthogonal polynomials. Given an inner product on the vector space of all polynomials, one can obtain an orthonormal basis $\{\phi_k\}$ by applying Gram-Schmidt process to the basis $\{h_k\}_{k\geq 0}$ where $h_k(x) = x^k$. The sequence $\{\phi_k\}$ (which may be finite) is called an orthogonal polynomial sequence.
(4) Real symmetric tridiagonal matrices. We now consider semi-infinite matrices, that is $T_{k,k} = a_k$, $T_{k,k+1} = T_{k+1,k} = b_k$, for $k \geq 0$. Finite matrices are a subset of these, by padding them with zeros at the end.

Measure to Positive definite sequences: If μ is a measure that has all moments, define the moment sequence $\alpha_k = \int x^k d\mu(x)$. Then for any for any $m \geq 1$ and any $\mathbf{u} \in \mathbb{R}^{m+1}$, we have

$$\mathbf{u}^t \left(\alpha_{i+j}\right)_{0\leq i,j\leq m} \mathbf{u} = \sum_{i,j\leq m} \alpha_{i+j} u_i u_j = \int \left| \sum_{i=0}^{n} u_i x^i \right|^2 \mu(dx) \geq 0.$$

Hence α is a positive definite sequence. It is easy to see that μ is finitely supported if and only $L^2(\mu)$ is finite dimnesional if and only if $(\alpha_{i+j})_{i,j\geq 0}$ has finite rank.

Positive definite sequence to orthogonal polynomials: Let α be a positive definite sequence For simplicity we assume that $(\alpha_{i+j})_{i,j\geq 0}$ is strictly positive definite. Then the formulas $\langle h_i, h_j\rangle = \alpha_{i+j}$ define a valid inner product on the vector space $\mathcal{P}$ of all polynomials. Complete $\mathcal{P}$ under this inner product to get a Hilbert space H.

In H, h_k are linearly independent and their span (which is $\mathcal{P}$) is dense. Hence, applying Gram-Schmidt procedure to the sequence $h_0, h_1, \ldots$ give a sequence of polynomials $\phi_0, \phi_2, \ldots$ which form an orthonormal basis for H. Clearly ϕ_k has degree k.

Orthogonal polynomials to tridiagonal matrices: Let ϕ_k be an infinite sequence of polynomials such that ϕ_k has degree exactly k. Then it is clear that ϕ_k are linearly independent, that h_k is a linear combination of $\phi_0, \ldots, \phi_k$.

Consider an inner product on $\mathcal{P}$ such that $\langle \phi_k, \phi_\ell\rangle = \delta_{k,\ell}$. The same reasoning as before gives the three term recurrences (24) for ϕ_k. Thus we get $a_k \in \mathbb{R}$ and $b_k > 0$, $k \geq 0$. Form the infinite real symmetric tridiagonal matrix $T = T(a,b)$.

Symmetric tridiagonal operators to measures: Let T be a semi-infinite real symmetric tridiagonal matrix. Let $\mathbf{e}_k$ be the co-ordinate vectors in $\ell^2(\mathbb{N})$. Let $D = \{\sum x_k \mathbf{e}_k \ : \ x_k \neq 0 \text{ finitely often}\}$. This is a dense subspace of $\ell^2(\mathbb{N})$. T is clearly well-defined and linear on D. It is symmetric in the sense that $\langle Tu, v\rangle = \langle u, Tv\rangle$ for all $u, v \in D$ and the inner product is in $\ell^2(\mathbb{N})$.

Suppose T' is a self-adjoing extension of T. That is, there is a subspace D' containing D and a linear operator $T' : D \to \mathbb{R}$ such that $T'|_D = T$ and such that T' is self-adjoint (we are talking about unbounded operators on Hilbert spaces, hence self-adjointness and symmetry are two distinct things, and this is not the place to go into the definitions. Consult for example, chapter 13 of Rudin's *Functional Analysis*). Then it is a fact that T' has a spectral decomposition. The spectral measure of T' at $\mathbf{e}_0$ is a measure. In general there can be more than one extension. If the extension is unique, then μ is uniquely defined.

This cycle of connections is quite deep. For example, if we start with any positive definite sequence α_k and go through this cycle, we get an OP sequence and a tridiagonal symmetric operator. The spectral measure of any self-adjoint extension of this operator has the moment sequence α_k. Further, there is a unique measure with moments α_k if and only if T has a unique self-adjoint extension!

Remark 53. In the above discussion we assumed that α is a strictly positive definite sequence, which is the same as saying that the measure does not have finite support or that the orthogonal polynomial sequence is finite or that the tridigonal matrix is essentially finite. If we start with a finitely supported measure, we can still go through this cycle, except that the Gram-Schmidt process stops at some finite n etc.

5. Exact distribution of eigenvalues of the tridiagonal matrix

We wish to find the joint density of eigenavalues of certain random tridiagonal matrices. For this, we have to arrange the eigenvalues as a vector in $\mathbb{R}^n$, and write the density with respect to Lebesgue measure on $\mathbb{R}^n$. There are two common ways to arrange eigenvalues as a vector. Firstly, in descending order to get a vector $\lambda^{\downarrow} = (\lambda_1, \ldots, \lambda_n)$ with $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$. Secondly, we can place them in exchangeable random order. This means that we pick a permutation $\pi \in \mathcal{S}_n$ uniformly at random (and independently of the our random matrix), and set $\lambda_{\text{ex}} = (\lambda_{\pi(1)}, \ldots, \lambda_{\pi(n)})$. Of course, if f is the density of $\lambda^{\downarrow}$ and g is the density of λ_{ex}, we can recover one from the other by the relationship

$$f(u_1, \ldots, u_n) = n! g(u_1, \ldots, u_n) \mathbf{1}_{u_1 < \ldots < u_n}$$

and the fact that g is symmetric in its arguments. We shall usually express the eigenvalues in exchangeable random order without explicitly saying so, but this is just a convention.

Theorem 54. *Let $T = T(a,b)$ be the $n \times n$ random, real symmetric matrix, with $a_k \sim N(0,1)$, $b_k^2 \sim \chi^2_{\beta(n-k)}$ and all these are independent. Then, the eigenvalues of T have joint density*[4]

$$\frac{1}{\hat{Z}_{\beta,n}} \exp\left\{-\frac{1}{4}\sum_{k=1}^{n}\lambda_k^2\right\} \prod_{i<j}|\lambda_i - \lambda_j|^\beta$$

where the normalization constant may be explicitly found as

$$\hat{Z}_{\beta,n} = \pi^{\frac{n}{2}} 2^{n+\frac{\beta}{4}n(n-1)} \frac{\Gamma(\frac{n\beta}{2}) \prod_{j=1}^{n-1} \Gamma(\frac{j\beta}{2})}{\Gamma\left(\frac{\beta}{2}\right)^n}.$$

Corollary 55. *The joint density of eigenvalues of the GOE matrix is*

$$\frac{1}{\hat{Z}_{n,1}} \exp\left\{-\frac{1}{4}\sum_{k=1}^{n}\lambda_k^2\right\} \prod_{i<j}|\lambda_i - \lambda_j|.$$

The joint density of eigenvalues of the GOE matrix is

$$\frac{1}{\tilde{Z}_{n,2}} \exp\left\{-\frac{1}{2}\sum_{k=1}^{n}\lambda_k^2\right\} \prod_{i<j}|\lambda_i - \lambda_j|^2$$

where $\tilde{Z}_{n,2} = \hat{Z}_{n,2} 2^{-n^2/2}$.

PROOF OF THE COROLLARY. By Theorem 49 it follows that the eigenvalues of a GOE matrix have the same distribution as the eigenvalues of the tridiagonal matrix in Theorem 54 with $\beta = 1$. This gives the first statement. The second is similar, except that there is a scaling by $\sqrt{2}$ involved in Theorem 49. ■

PROOF OF THEOREM 54. The joint density of (a,b) on $\mathbb{R}^n \times \mathbb{R}_+^{n-1}$ is

$$\begin{aligned} f(a,b) &= \prod_{k=1}^{n} \frac{e^{-\frac{1}{4}a_k^2}}{2\sqrt{\pi}} \prod_{k=1}^{n-1} \frac{e^{-\frac{1}{2}b_k^2} b_k^{\beta(n-k)-1}}{2^{\frac{\beta}{2}(n-k)-1}\Gamma(\beta(n-k)/2)} \\ &= \frac{1}{Z_{\beta,n}} \exp\left\{-\frac{1}{4}\mathrm{tr}\left(T^2\right)\right\} \prod_{k=1}^{n-1} b_k^{\beta(n-k)-1}. \end{aligned} \tag{26}$$

where the normalizing constant

$$Z_{\beta,n} = \pi^{\frac{n}{2}} 2^{1+\frac{\beta}{4}n(n-1)} \prod_{j=1}^{n-1} \Gamma(\beta j/2).$$

[4]The corollary here was proved by by Wigner (or Dyson? before 1960 anyway) and it was noticed that the density could be generalized for any $\beta > 0$. Whether the general β-density could be realized as that of eigenvalues of a random matrix was in the air. The idea that this could be done by considering these random tridiagonal matrix with independent entries, is due to Dumitriu and Edelman. This development has had far-reaching consequences in the study of random matrices. In short, the reason is that the β-density given here is complicated to analyze, although explicit, and the tridiagonal matrix itself can be used in the analysis, as it has independent entries.

Now, let ν be the spectral measure of T at the vector $\mathbf{e}_1$ (this corresponds to $\mathbf{e}_0$ of the previous section). Then $\nu = \sum_{j=1}^{n} p_j \delta_{\lambda_j}$. According to the previous section, λ_j are the eigenvalues of T while $p_j = |U_{1,j}|^2$ are elements of the first row of the eigenvector matrix.

Observe that almost surely none of the b_ks is zero, and hence by part (c) of Lemma 51, the eigenvalues of T are distinct. By part (a) of the same lemma, $\mathcal{T}_n^0$ is in bijection with $\mathcal{P}_n^0$ and hence we may parameterize the matrices by λ_k, $k \le n$ and p_k, $k \le n-1$. We shall also write p_n in many formulas, but it will always be understood to be $1 - p_1 - \ldots - p_{n-1}$. If we write $(a,b) = G(\lambda, p)$, then by the change of variable formula we get the density of (λ, p) to be

$$\begin{aligned} g(\lambda,p) &= f(G(\lambda,p))\,|\det(J_G(\lambda,p))| \qquad (J_G \text{ is the Jacobian of } G) \\ &= \frac{1}{Z_{\beta,n}} \exp\left\{-\frac{1}{4}\sum_{k=1}^{n}\lambda_k^2\right\} \prod_{k=1}^{n-1} b_k^{\beta(n-k)-1}\,|\det(J_G(\lambda,p))| \end{aligned} \tag{27}$$

It remains to find the Jacobian determinant of G and express the product term in terms of λ_k and p_k. For this we use the definition of spectral measure $\langle T^k \mathbf{e}_1, \mathbf{e}_1\rangle = \sum \lambda_j^k p_j$ for $k = 1, \ldots, 2n-1$. We get

$$\begin{aligned} \sum p_j\lambda_j = T_{1,1} = a_1 \qquad & \sum p_j\lambda_j^2 = (T^2)_{1,1} = b_1^2 + [\ldots] \\ \sum p_j\lambda_j^3 = (T^3)_{1,1} = a_2 b_1^2 + [\ldots] \qquad & \sum p_j\lambda_j^4 = (T^4)_{1,1} = b_2^2 b_1^2 + [\ldots] \\ \sum p_j\lambda_j^5 = (T^5)_{1,1} = a_3 b_2^2 b_1^2 + [\ldots] \qquad & \sum p_j\lambda_j^6 = (T^6)_{1,1} = b_3^2 b_2^2 b_1^2 + [\ldots] \\ \ldots\ldots \qquad & \ldots\ldots \end{aligned}$$

Here the $[\ldots]$ include many terms, but all the a_k, b_k that appear there have appeared in previous equations. For example, $(T^2)_{1,1} = b_1^2 + a_1^2$ and as a_1 appeared in the first equation, we have brushed it under $[\ldots]$.

Let $U = (u_1, \ldots, u_{2n-1})$ where $u_j = (T^j)_{1,1}$. The right hand sides of the above equations express U as $F(a,b)$ while the left hand sides as $U = H(\lambda, p)$. We find the Jacobian determinants of F and H as follows.

Jacobian of F: Note that u_{2k} is a function of $a_i, i \le k$ and $b_j, j \le k$ while u_{2k-1} is a function of $a_i, i \le k$ and $b_j, j \le k-1$. Thus, $J_F(a,b)$ is an upper triangular matrix and we see that

$$\det(J_F(a,b)) = 2^{n-1} \prod_{k=1}^{n-1} b_k^{4(n-k)-1}. \tag{28}$$

Jacobian of H: The equations above give the Jacobian of H (recall that $p_n = 1 - \sum_{j=1}^{n-1} p_j$)

$$J_H(\lambda,p) = \begin{bmatrix} p_1 & \ldots & p_n & \lambda_1 - \lambda_n & \ldots & \lambda_{n-1} - \lambda_n \\ 2p_1\lambda_1 & \ldots & 2p_n\lambda_n & \lambda_1^2 - \lambda_n^2 & \ldots & \lambda_{n-1}^2 - \lambda_n^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \\ (2n-1)p_1\lambda_1^{2n-2} & \ldots & (2n-1)p_n\lambda_n^{2n-2} & \lambda_1^{2n-1} - \lambda_n^{2n-1} & \ldots & \lambda_{n-1}^{2n-1} - \lambda_n^{2n-1} \end{bmatrix}.$$

To find its determinant, first factor out p_i from the i^{th} column, for $i \le n-1$. The resulting matrix is of the same form (as if $p_i = 1$ for all i) and its determinant is clearly a polynomial in $\lambda_1, \ldots, \lambda_n$.

It must also symmetric in λ_ks, because the original problem we started with was symmetric in λ_ks (can you infer symmetry directly from the above matrix?).

If $h := \lambda_1 - \lambda_n \to 0$, then $C_{n+1} = O(h)$, $C_1 - C_n = O(h)$. Further, it is easy to check that $C_{n+1} - h(C_1 + C_2)/2 = O(h^2)$. Thus for fixed λ_k, $k \geq 2$, the polynomial in λ_1 has (at least) a four fold zero at λ_n. By symmetry, the determinant has a factor $\Delta(\lambda)^4$. However, the determinant above and $\Delta(\lambda)^4 = \prod_{i<j}(\lambda_i - \lambda_j)^4$ are both polynomials of degree $4(n-1)$. Further, the coefficient of λ_1^{4n-4} in both is the same. Therefore we get

$$\det(J_H(a,b)) = \pm|\Delta(\lambda)|^4 \prod_{i=1}^{n} p_i. \tag{29}$$

From (28) and (29) we deduce that

$$|\det(J_G(\lambda,p))| = \pm \frac{\prod_{i=1}^{n} p_i \prod_{i<j} |\lambda_i - \lambda_j|^4}{2^{n-1} \prod_{k=1}^{n-1} b_k^{4(n-k)-1}}.$$

Substitute this in (27) to get

$$\begin{aligned} g(\lambda,q) &= \frac{1}{2^{n-1} Z_{\beta,n}} \exp\left\{-\frac{1}{4}\sum_{k=1}^{n}\lambda_k^2\right\} \prod_{i=1}^{n} p_i \prod_{i<j}|\lambda_i - \lambda_j|^4 \left(\prod_{k=1}^{n-1} b_k^{(n-k)}\right)^4 \\ &= \frac{1}{2^{n-1} Z_{\beta,n}} \exp\left\{-\frac{1}{4}\sum_{k=1}^{n}\lambda_k^2\right\} \left(\prod_{i=1}^{n} p_i\right)^{\frac{\beta}{2}-1} \prod_{i<j}|\lambda_i - \lambda_j|^\beta \end{aligned}$$

by part (c) of Lemma 51.

This gives the joint density of λ and p and we see that the two are independent. It remains to integrate out the p variables. But that is just a Dirichlet integral

$$\int_0^1 \int_0^{1-p_1} \cdots \int_0^{1-\sum_{i=1}^{n-2} p_i} \left(\prod_{i=1}^{n} p_i\right)^{\frac{\beta}{2}-1} dp_{n-1} \ldots dp_1 = \text{Dirichlet}(\beta/2, \ldots, \beta/2) = \frac{\Gamma(\beta/2)^n}{\Gamma(\beta n/2)}.$$

This completes the proof of the theorem. ■

6. Beta ensembles*

Consider n particles $(\lambda_1, \ldots, \lambda_n)$ with density

$$g(\lambda) = \frac{1}{\hat{Z}_{\beta,n}} \exp\left\{-\frac{1}{4}\sum_{k=1}^{n}\lambda_k^2\right\} \prod_{i<j}|\lambda_i - \lambda_j|^\beta$$

for $\beta > 0$. As we saw, this is the density of eigenvalues of random tridiagonal matrix T_β. What can we do with this density? Here are some features.

(1) *Repulsion of eigenvalues:* The density is $g(\lambda) = \exp\{-\sum V(\lambda_k)\}|\Delta(\lambda)|^\beta$ with $V(x) = x^2/4$ and where $\Delta(\lambda)$ is the Vandermonde factor. Without the Vandermonde factor (i.e., $\beta = 0$), this is the density of n i.i.d variables with density $\exp\{-V(x)\}$. But $\Delta(\lambda)$ vanishes whenever $\lambda_i - \lambda_j = 0$ for some $i \neq j$. This means that the eigenvalues tend to keep away from each

other. Further, the vanishing of $|\lambda_i - \lambda_j|^\beta$ increases with β which means that the repulsion increases with β. As $\beta \to \infty$, the density concentrates at a particular configuration, or "the particles freeze at the lowest energy configuration".

(2) *Gibbs interpretation of the density:* For convenience, scale the eigenvalues down by $\sqrt{\beta}$. Continue to denote the variables by λ_k. The resulting density is $f_\beta(\lambda) = g_\beta(\lambda\sqrt{\beta}) = \exp\{-\beta H_{n,\beta}(\lambda)\}$ where

$$H_n(\lambda) = \frac{1}{4}\sum_{k=1}^{n}\lambda_k^2 - \frac{1}{2}\sum_{i\neq j}\log|\lambda_i - \lambda_j|.$$

$H_n(\lambda)$ is called the energy of the configuration λ. According to Boltzmann, all systems in Statistical mechanics have this structure - the density is $\exp\{-\text{energy}\}$ where the energy (or *Hamiltonian*) varies from system to system and in fact characterizes the system.

In the case at hand, the energy has two terms. The function V is interpreted as a potential, a particle sitting at a location x will have potential energy $V(x)$. Further, there is pairwise interaction - if a particle is at location x and another at y, then they have an interaction potential of $-\log|x-y|$. This just means that they repel each other with force (which is the gradient of the interaction energy) $1/|x-y|$ (repulsion rather than attraction, because of the negative sign on $\log|x-y|$). This is precisely Coulomb's law, suitably modified because we are not in three dimensions. More physically, if one imagines infinite sheets of uniformly charged plates placed perpendicular to the x-axis, and a potential $V(x)$ is applied, then they repel each other by a force that is inverse of the distance.

Thus, they prefer to locate themselves at points $x_1, \ldots, x_n$ that minimizes the energy $H_n(x)$. However, if there is a positive temperature $1/\beta$, then they don't quite stabilize at the minimum, but have a probability to be at other locations, but with the density that decreases exponential with the energy. Thus the density is given exactly by the density $g_\beta(\lambda)$! This is called a *one-component plasma* on the line.

(3) Note that we ignored the normalization constants in the previous discussion. Many probability distributions that arise in probability are described by giving their density as $Z_\beta^{-1}\exp\{-\beta H(x)\}$ where $H(\cdot)$ is specified. The trouble is analyzing the system to make useful statements about a typical configuration sampled from this measure. As $Z_\beta = \int \exp\{-\beta H(x)\}dx$, we see that Z_β is like a Laplace trnaform of the function $H(x)$. Thus, if we can compute Z_β (for all β), one can deduce many things about the distribution. For example, the expected energy of a random sample from the given density is

$$\frac{1}{Z_\beta}\int H(x)\exp\{-\beta H(x)\}dx = \frac{1}{Z_\beta}\frac{\partial Z_\beta}{\partial \beta} = \frac{\partial}{\partial\beta}\log Z_\beta.$$

This is the reason why physicists lay great stress on finding the normalization constant Z_β, which they term the *partition function*. Generally speaking, computing Z_β is fairly impossible. The system that we have, the one with energy function H_n, is exceptional in that the partition function can be found explicitly, as we did in the previous section!

(4) The computation of the normalization constant from the previous section proves the following highly non-trivial integration formula (try proving it!)

$$\int_{\mathbb{R}^n} \exp\left\{-\frac{1}{4}\sum_{k=1}^{n}\lambda_k^2\right\}\prod_{i<j}|\lambda_i-\lambda_j|^{\beta}d\lambda_1\ldots d\lambda_n = \pi^{\frac{n}{2}}2^{n+\frac{\beta}{4}n(n-1)}\frac{\Gamma(\frac{n\beta}{2})\prod_{j=1}^{n-1}\Gamma(\frac{j\beta}{2})}{\Gamma\left(\frac{\beta}{2}\right)^n}. \tag{30}$$

This can be derived from a similar but more general integral of Selberg, who computed

$$S(\alpha,\beta,\gamma) = \int_{[0,1]^n}|\Delta(x)|^{2\gamma}\prod_{i=1}^{n}x_i^{\alpha-1}(1-x_i)^{\beta-1}dx$$

where $\Delta(x)=\prod_{i<j}|x_i-x_j|$ and α,β,γ are complex parameters satisfying some inequalities so that the integral converges[5].

But this does not cover the main questions one would like to answer when an explicit density $g(\lambda_1,\ldots,\lambda_n)$ is at hand. Observe that the labeling here is introduced for convenience, and what we care about is the empirical measure $L_n = n^{-1}\sum_{k=1}^{n}\delta_{\lambda_k}$. If λ has density $g_\beta(\lambda)$, what is $\mathbf{E}[L_n[a,b]]$ for any $a<b$? What about the variance $\mathrm{Var}(L_n[a,b])$? What is the typical spacing between one eigenvalue and the next? What is the chance that there in no eigenvalue in a given interval? Does L_n (perhaps after rescaling λ_k) converge to a fixed measure (perhaps the semicircle law) as $n\to\infty$?

The last question can actually be answered from the joint density, but the other questions are more "local". For example, if $I=[a,b]$, then by the exchangeability of λ_ks

$$\mathbf{E}[L_{n,\beta}(I)] = n\mathbf{P}(\lambda_1\in I) = n\int_I\left(\int_{\mathbb{R}^{n-1}} g_\beta(\lambda_1,\ldots,\lambda_n)d\lambda_n\ldots d\lambda_2\right)d\lambda_1$$

which involves integrating out some of the variables. Can we do this explicitly? It is not clear at all from the density g_β. In fact, there is no known method to do this, except for special values of β, especially $\beta=1,2,4$. Of these $\beta=2$ is particularly nice, and we shall concentrate on this case in the next few sections.

7. The special case of $\beta=2$

Consider the GUE ensemble density

$$g(\lambda) = \exp\left\{-\frac{1}{4}\sum_{k=1}^{n}\lambda_k^2\right\}\prod_{i<j}|\lambda_i-\lambda_j|^2$$

where Z_n is the normalizing constant. More generally, let μ be a Borel probability measure on $\mathbb{R}$ and consider the density f on $\mathbb{R}^n$ proportional to $|\Delta(x)|^2$ with respect to the measure $\mu^{\otimes n}$. All of what we say in this section will apply to this more general density[6]. This is symmetric in $\lambda_1,\ldots,\lambda_n$. The following lemma shows that it is possible to explicitly integrate out a subset of variables and

[5]More on the Selberg integral, its proofs and its consequences may be found in the book of Mehta or of Andrews, Askey and Roy.

[6]These are special cases of what are known as *determinantal point processes*.

get marginal densities of any subset of the co-ordinates. As we discussed earlier, this is crucial to computing local properties of the system of particles defined by the density f.

Observation: Let $p_0, p_1, \ldots, p_{n-1}$ be monic polynomials such that p_k has degree k. Then,

$$\Delta(x) = \det \begin{bmatrix} 1 & x_1 & x_1^2 & \ldots & x_1^{n-1} \\ 1 & x_2 & x_2^2 & \ldots & x_2^{n-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 & \ldots & x_n^{n-1} \end{bmatrix} = \det \begin{bmatrix} p_0(x_1) & p_1(x_1) & p_2(x_1)\ldots & p_{n-1}(x_1) \\ p_0(x_2) & p_1(x_2) & p_2(x_2)\ldots & p_{n-1}(x_2) \\ \vdots & \vdots & \vdots & \vdots \\ p_0(x_n) & p_1(x_n) & p_2(x_n)\ldots & p_{n-1}(x_n) \end{bmatrix}$$

as can be seen by a sequence of column operations. If ϕ_k is any polynomial with degree k and having leading coefficient c_k, then we get $\Delta(x) = C_n \det(A)$ where $a_{i,j} = \phi_j(x_i)$ with the index i running from 0 to $n-1$ and the index j from 1 to n. The constant $C_n = (c_0 c_1 \ldots c_{n-1})^{-1}$. Thus,

$$|\Delta(x)|^2 = C_n^2 \det\left(A\, A^t\right) = C_n^2 \det\left(K_n(x_i,x_j)\right)_{i,j\le n}$$

where $K_n(x,y) = \sum_{j=0}^{n-1} \phi_j(x)\phi_j(y)$. It turns out that choosing ϕ_j to be the orthogonal polynomials with respect to μ enables us to integrate out any subset of variables explicitly!

Lemma 56. *Let $(A, \mathcal{A}, \mu)$ be a measure space. Let ϕ_k, $1 \le k \le n$, be an orthonormal set in $L^2(\mu)$ and define $K(x,y) = \sum_{k=1}^{n} \phi_k(x)\overline{\phi}_k(y)$. Define $f : A^n \to \mathbb{R}$ by*

$$f(x) = (n!)^{-1} \det\left(K(x_i,x_j)\right)_{i,j\le n}.$$

(1) For any $m \le n$ and any λ_k, $k \le m-1$, we have

$$\int_{\mathbb{R}} \det\left(K(\lambda_i,\lambda_j)\right)_{i,j\le m} \mu(d\lambda_m) = (n-m+1)\det\left(K(\lambda_i,\lambda_j)\right)_{i,j\le m-1}.$$

(2) f is a probability density on A^n with respect to $\mu^{\otimes n}$. Further, if $(\lambda_1,\ldots,\lambda_n)$ is a random vector in $\mathbb{R}^n$ with density f, then λ_i are exchangeable, and for any $m \le n$, the density of $(\lambda_1,\ldots,\lambda_m)$ with respect to $\mu^{\otimes m}$ is given by

$$f_k(\lambda_1,\ldots,\lambda_m) = \frac{(n-k)!}{n!} \det\left(K(\lambda_i,\lambda_j)\right)_{i,j\le m}.$$

Corollary 57. *Let $\mu \in \mathcal{P}(\mathbb{R})$ have finite moments up to order $2n-2$ and let $\phi_0,\ldots,\phi_{n-1}$ be the first n orthogonal polynomials normalized so that $\int \phi_k\phi_\ell d\mu = \delta_{k,\ell}$. Then, the density $f(x) = Z_n^{-1}|\Delta(x)|^2$ on $\mathbb{R}^n$ with respect to a measure $\mu^{\otimes n}$ can be rewritten as $f(x) = (n!)^{-1}\det\left(K_n(\lambda_i,\lambda_j)\right)_{i,j\le n}$ where $K_n(x,y) = \sum_{j=0}^{n-1}\phi_j(x)\phi_j(y)$. Further, the marginal density of any k co-ordinates is given by $\frac{(n-k)!}{n!}\det\left(K(\lambda_i,\lambda_j)\right)_{i,j\le k}$.*

The corollary trivially follows from the lemma and the observations made before the theorem. We now prove the lemma.

PROOF. (1) We need two properties of the kernel K. Both follow from orthonormality of ϕ_ks.

(a) The *reproducing kernel property:* $\int K(x,y)K(y,z)\mu(dy) = K(x,z)$.

(b) $\int K(x,x)\mu(dx) = n$.

By expanding the determinant

$$\int_{\mathbb{R}} \det\left(K(\lambda_i,\lambda_j)_{i,j\le m}\right) d\lambda_m = \sum_{\pi\in\mathcal{S}_m} \operatorname{sgn}(\pi) \int_{\mathbb{R}} \prod_{i=1}^{m} K(\lambda_i,\lambda_{\pi(i)}) d\lambda_m.$$

Fix π. There are two cases.

Case 1: $\pi(m) = m$. then by property (b), the term becomes

$$\prod_{i=1}^{m-1} K(\lambda_i,\lambda_{\pi(i)}) \int K(\lambda_m,\lambda_m) d\lambda_m = n \prod_{i=1}^{m-1} K(\lambda_i,\lambda_{\sigma(i)}).$$

where $\sigma \in \mathcal{S}_{m-1}$ is defined by $\sigma(i) = \pi(i)$. Observe that $\operatorname{sgn}(\sigma) = \operatorname{sgn}(\pi)$.

Case 2: Fix $\pi(m) \neq m$. Let $p = \pi^{-1}(m)$ and $q = \pi(m)$ (thus $p, q < m$). By property (a) above,

$$\begin{aligned}\int_{\mathbb{R}} \prod_{i=1}^{m} K(\lambda_i,\lambda_{\pi(i)}) d\lambda_m &= \prod_{i\neq p,m} K(\lambda_i,\lambda_{\pi(i)}) \int_{\mathbb{R}} K(\lambda_p,\lambda_m) K(\lambda_m,\lambda_q) d\lambda_m \\ &= \prod_{i\neq m} K(\lambda_i,\lambda_{\sigma(i)})\end{aligned}$$

where $\sigma(i) = \pi(i)$ for $i \neq p$ and $\sigma(p) = q$. Then $\sigma \in \mathcal{S}_{m-1}$ and $\operatorname{sgn}(\sigma) = -\operatorname{sgn}(\pi)$.

Now consider any $\sigma \in \mathcal{S}_{m-1}$. It arises from one permutation π in Case 1, and from $m-1$ distinct π in Case 2. As the $\operatorname{sgn}(\sigma)$ has opposing signs in the two cases, putting them together, we see that $\int_{\mathbb{R}} \det\left(K(\lambda_i,\lambda_j)_{i,j\le n}\right) d\lambda_m$ is equal to

$$(n-(m-1)) \sum_{\sigma\in\mathcal{S}_{n-1}} \prod_{i=1}^{n-1} K(\lambda_i,\lambda_{\sigma(i)}) = (n-m+1) \det\left(K(\lambda_i,\lambda_j)_{i,j\le n-1}\right).$$

(2) Let $m < n$ and let $f_m(x_1,\ldots,x_m) = \int_{R^{n-m}} f(x_1,\ldots,x_n) d\mu(x_{m+1}) \ldots d\mu(x_n)$. Inductively applying the integration formula in part (2), we get

$$f_m(\lambda_1,\ldots,\lambda_m) = C_n^{-1}(n-m)! \det\left(K(\lambda_i,\lambda_j)_{i,j\le m}\right).$$

In particular, if we integrate out all variables, we get $C_n^{-1} n!$. Thus, we must have $C_n = n!$ for f to be a probability density (the positivity of f is clear because $(K(x_i,x_j))_{i,j\le n}$ is n.n.d, being of the form AA^t).

Plugging the value of C_n back into the expression for f_m shows that

$$f_m(\lambda_1,\ldots,\lambda_m) = \frac{(n-m)!}{n!} \det\left(K(\lambda_i,\lambda_j)_{i,j\le k}\right). \quad \blacksquare$$

These integration formulas are what make $\beta = 2$ special. None of this would work if we considered density proportional to $|\Delta(x)|^\beta$ with respect to $\mu^{\otimes n}$. As a corollary of these integration formulas, we can calculate the mean and variance of the number of points that fall in a given subset.

Proposition 58. *In the setting of Lemma 56, let $N(\cdot) = \sum_{k=1}^{n} \delta_{\lambda_k}$ be the unnormalized empirical measure. Let $I \subseteq A$ be a measurable subset. Then,*

(i) $\mathbf{E}\left[(N(I))_{m\downarrow}\right] = \int_{I^m} \det\left(K(x_i,x_j)_{i,j\le k}\right) d\mu(x)$ *where* $(k)_{m\downarrow} = k(k-1)\ldots(k-m+1)$.

(ii) $\mathbf{E}[N(I)] = \int_I K(x,x)d\mu(x)$ *and* $Var(N(I)) = \iint_{I^c} |K(x,y)|^{2d\mu(y)d\mu(x)}$.

(iii) Let T_I *be the integral operator on* $L^2(I,\mu)$ *with the kernel* K. *That is* $T_I f(x) = \int_I K(x,y)f(y)d\mu(y)$ *for* $x \in I$. *Let* $\theta_1, \theta_2, \ldots$ *be the non-zero eigenvalues of* T. *Then* $\theta_i \in (0,1]$ *and if* $\xi_i \sim Ber(\theta_i)$ *are independent, then* $N(I) \stackrel{d}{=} \xi_1 + \xi_2 + \ldots$.

PROOF. (i) Write $N(I) = \sum_{k=1}^n \mathbf{1}_{\lambda_k \in I}$. Use the exchangeability of λ_k to write

$$\mathbf{E}\left[(N(I))_{m\downarrow}\right] = \mathbf{E}\left[\sum_{\substack{i_1,\ldots,i_m \le n \\ \text{distinct}}} \mathbf{1}_{\lambda_{i_1}\in I}\mathbf{1}_{\lambda_{i_2}\in I}\ldots\mathbf{1}_{\lambda_{i_m}\in I}\right] = (n)_{\downarrow m}\mathbf{P}\left[\lambda_i \in I,\ 1 \le i \le m\right].$$

Using the density of $(\lambda_1, \ldots, \lambda_m)$ given in Lemma 56 we get

$$\mathbf{E}\left[(N(I))_{m\downarrow}\right] = \int_{I^m} \det\left(K(x_i,x_j)\right)_{i,j\le m} d\mu(x).$$

(ii) Apply the formula in part (i) with $m = 1$ to get $\mathbf{E}[N(I)] = \int_I K(x,x)d\mu(x)$. Expressing the variance of $N(I)$ in terms of $\mathbf{E}[N(I)]$ and $\mathbf{E}[N(I)(N(I)-1)]$ one arrives at

$$\mathrm{Var}(N(I)) = \int_I K(x,x)d\mu(x) - \int_I\int_I |K(x,y)|^2 d\mu(x)d\mu(y).$$

Write the first integral as $\int_I\int_A |K(x,y)|^2 d\mu(y)$ by the reproducing property of K. Subtracting the second term give $\int_I\int_{I^c} |K(x,y)|^2 d\mu(x)d\mu(y)$.

(iii) With $I = A$, we have $T_A f = \sum_{k=1}^n \langle f, \phi_k\rangle \phi_k$. Thus, T is a projection operator with rank n. Clearly, $0 \le T_I \le T_A$ from which it follows that $\theta_i \in [0,1]$ and at most n of them are nonzero. If ψ_i are the corresponding eigenfunctions, then it is easy to see that $K(x,y) = \sum_i \theta_i \psi_i(x)\overline{\psi}_i(y)$.

■

Remark 59. In random matrix theory, one often encounters the following situation. Let $\mu \in \mathcal{P}(\mathbb{C})$ such that $\int |z|^{2n-2}\mu(dz) < \infty$. On $\mathbb{C}^n$ define the density $f(x) \propto |\Delta(x)|^2$ with respect $\mu^{\otimes n}$. Then we can again orthogonalize $1, z, \ldots, z^{n-1}$ with respect to μ to get ϕ_k, $0 \le k \le n-1$ and the kernel $K(z,w) = \sum_{j=0}^{n-1} \phi_j(z)\overline{\phi}_j(w)$. The density can be rewritten as $f(x) = (n!)^{-1}\det\left(K(x_i,x_j)\right)_{i,j\le n}$. This is of course a special case of the more general situation outlined in Lemma 56, except that one needs to keep track of conjugates everywhere when taking inner products.

8. Determinantal point processes

Consider the density f_m of $(\lambda_1, \ldots, \lambda_m)$ as described in Lemma 56. Let us informally refer to it as the chance that λ_i falls at location x_i for $1 \le i \le m$. Then the chance that $L_n := \sum_{k=1}^n \delta_{\lambda_k}$ puts a point at each x_i, $i \le m$, is precisely $(n)_{m\downarrow} f_m(x_1, \ldots, x_m) = \det\left(K(x_i,x_j)\right)$.

For any random variable L taking values in the space of locally finite counting measures (eg., L_n), one can consider this chance (informally speaking), called the m^{th} joint intensity of L. If for every m, the joint intensities are given by $\det\left(K(x_i,x_j)\right)$ for some $K(x,y)$, then we say that L is a *determinantal point process*. A determinantal point process may have infinitely many points.

If a point process which has a fixed finite total number of points, then we can randomly arrange it as a vector and talk in terms of densities. But when we have infinitely many points, we cannot do this and instead talk in terms of joint intensities. Like densities, joint intensities may or may not exist. But if they do exist, they are very convenient to work with. In random matrix theory we usually get finite determinantal processes, but in the limit we often end up with infinite ones. Therefore, we shall now give precise definitions of point processes, joint intensities and determinantal processes[7]

Definition 60. Let A be a locally compact Polish space (i.e., a complete separable metric space) and let μ be a Radon measure on A. A *point process* L on A is a random integer-valued positive Radon measure on A. If L almost surely assigns at most measure 1 to singletons, we call it a *simple* point process;

Definition 61. If L is a simple point process, its *joint intensities* w.r.t. μ are functions (if any exist) $p_k : A^k \to [0,\infty)$ for $k \geq 1$, such that for any family of mutually disjoint subsets $I_1, \ldots, I_k$ of A,

$$\mathbf{E}\left[\prod_{j=1}^{k} L(I_j)\right] = \int_{I_1 \times \ldots \times I_k} p_k(x_1, \ldots, x_k) d\mu(x_1) \ldots d\mu(x_k). \tag{31}$$

In addition, we shall require that $p_k(x_1, \ldots, x_k)$ vanish if $x_i = x_j$ for some $i \neq j$.

Definition 62. A point process L on A is said to be a *determinantal process* with kernel K if it is simple and its joint intensities with respect to the measure μ satisfy

$$p_k(x_1, \ldots, x_k) = \det\left(K(x_i, x_j)\right)_{1 \leq i,j \leq k}, \tag{32}$$

for every $k \geq 1$ and $x_1, \ldots, x_k \in A$.

Exercise 63. When λ has density as in Lemma 56, check that the point process $L = \sum_{k=1}^{n} \delta_{\lambda_k}$ is a determinantal point process with kernel K as per the above definition.

9. One dimensional ensembles

Let $V : \mathbb{R} \to \mathbb{R}$ be a function that increases fast enough at infinity so that $\int e^{-\beta V(x)} dx < \infty$ for all $\beta > 0$. Then, define the probability measure $\mu_n(dx) = e^{-nV(x)}/Z_n$ and the let λ be distributed according to the measure $Z_{n,\beta}^{-1} |\Delta(x)|^{\beta} e^{-n\sum_{k=1}^{n} V(x_k)}$. Under some conditions on V, the empirical measure of λ converges to a fixed measure $\mu_{V,\beta}$. Then one asks about

We will now concentrate on two particular examples of $\beta = 2$ ensembles.

(1) The GUE (scaled by $\sqrt{2}$). The density is $Z_n^{-1} |\Delta(\lambda)|^2 \exp\{-\sum_{k=1}^{n} \lambda_k^2/4\}$. To write it in determinant form, we define μ as the $N(0,2)$ distribution, that is $\mu(dx) = (2\sqrt{\pi})^{-1} e^{-x^2/4} dx$. Let H_k, $k \geq 0$, be the orthogonal polynomials with respect to μ obtained by applying Gram-Schmidt to the monomials $1, x, x^2, \ldots$. H_k are called *Hermite polynomials*. The kernel is

[7] For more detailed discussion on joint intensities, consult chapter 1 of the book **?** available at http://math.iisc.ernet.in/ manju/GAF_book.pdf. Chapter 4 of the same book has discussion and examples of determinantal point processes.

$K_n(x,y) = \sum_{k=0}^{n-1} H_k(x)H_k(y)$. We have chosen them to be orthonormal, $\int H_k(x)H_\ell(x)d\mu(x) = \delta_{k,\ell}$. Hermite polynomials are among the most important special functions in mathematics.

(2) The CUE (*circular unitary ensemble*). Let μ be the uniform measure on S^1 and on $(S^1)^n$ define the density $f(x) = |\Delta(x)|^2$ with respect to $\mu^{\otimes n}$. In this case $z^k = e^{ikt}$ are themselves orthonormal, but it is a bit more convenient to take $\phi_k(t) = e^{-i(n-1)t/2}e^{ikt}$. Then, the kernel is

$$e^{-i(n-1)s/2}e^{i(n-1)t/2}\frac{1-e^{in(s-t)}}{1-e^{i(s-t)}} = D_n(s-t), \qquad D_n(u) := \frac{\sin(nu/2)}{\sin(u/2)}.$$

D_n is the well-known Dirichlet kernel (caution: what is usually called D_n is our D_{2n+1}). We shall later see that the eigenvalues of a random unitary matrix sampled from the Haar measure are distributed as CUE.

The GUE and CUE are similar in the local structure of eigenvalues. However, there are edge phenomena in GUE but none in CUE. However, all calculations are simpler for the CUE as the kernel is even simpler than the GUE kernel. The difficulty is just that we are less familiar with Hermite polynomials than with monomials. Once we collect the facts about Hermite functions the difficulties mostly disappear. The study of the edge is rather difficult, nevertheless.

10. Mean and Variance of linear statistics in CUE

Let λ be distributed according to CUE_n. Let $h : S^1 \to \mathbb{R}$ be a bounded measurable function. Let $N_n(h)$ be the linear statistic $\sum_{k=1}^n h(\lambda_k)$. Then,

$$\mathbf{E}[N_h] = \int h(x)K(x,x)d\mu(x) = n\int_0^{2\pi} h(t)\frac{dt}{2\pi}.$$

Actually this holds for any rotation invariant set of points on the circle. In particular, $\mathbf{E}[N_n(I)] = |I|/2\pi$.

The variance is considerably more interesting. Write the Fourier series of $h(t) = \sum_{k\in\mathbb{Z}} a_k e^{ikt}$ where $a_k = \int_0^{2\pi} h(t)e^{-ikt}\frac{dt}{2\pi}$. This equality is in L^2. Then,

$$\mathrm{Var}(N_n(h)) = \frac{1}{2}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi}(h(t)-h(s))^2 K_n(s,t)K_n(t,s)\frac{dt\,ds}{4\pi^2}.$$

We write

$$(h(t)-h(s))^2 = \sum_{k,\ell\in\mathbb{Z}} a_k\bar{a}_\ell(e^{ikt}-e^{iks})(e^{-i\ell t}-e^{-i\ell s}), \qquad K_n(t,s)K_n(s,t) = \sum_{p,q=0}^{n-1} e^{i(p-q)t}e^{i(q-p)s}.$$

Hence,

$$\begin{aligned}
\mathrm{Var}(N_n(h)) &= \frac{1}{2}\sum_{k,\ell\in\mathbb{Z}}\sum_{p,q=0}^{n-1} a_k\overline{a}_\ell \int_{-\pi}^{\pi}\int_{-\pi}^{\pi}(e^{ikt-i\ell t}+e^{iks-i\ell s}-e^{iks-i\ell t}-e^{ikt-i\ell s})e^{i(p-q)t}e^{i(q-p)s}\frac{dt\ ds}{4\pi^2}\\
&= \frac{1}{2}\sum_{k,\ell\in\mathbb{Z}}\sum_{p,q=0}^{n-1} a_k\overline{a}_\ell\left\{\delta_{k-\ell+p-q}\delta_{q-p}+\delta_{p-q}\delta_{k-\ell+q-p}-\delta_{k+p-q}\delta_{-\ell+q-p}-\delta_{-\ell+p-q}\delta_{k+q-p}\right\}\\
&= \frac{1}{2}\sum_{k,\ell\in\mathbb{Z}}\sum_{p,q=0}^{n-1} a_k\overline{a}_\ell\delta_{k-\ell}\left\{\delta_{p-q}+\delta_{p-q}-\delta_{k+p-q}-\delta_{k+q-p}\right\}\\
&= \frac{1}{2}\sum_{k\in\mathbb{Z}}|a_k|^2\sum_{p,q=0}^{n-1}\left\{2\delta_{p-q}-\delta_{k+p-q}-\delta_{k+q-p}\right\}\\
&= \sum_{k\in\mathbb{Z}}|a_k|^2\left(n-(n-|k|)_+\right).
\end{aligned}$$

Remark 64. The variance can be written as $\sum_{|k|\le n}|k||\hat{h}(k)|^2+n\sum_{|k|>n}|a_k|^2$ where $\hat{h}(k)=a_k$. The first sum is the contribution of low frequencies in h while the second gives the contribution of the high frequencies. For smooth functions, the high frequency Fourier co-efficients will be small and the first term dominates. For more wild functions, the second sum becomes significant. We shall consider two cases next.

Case 1: $h\in H^{1/2}$ which by definition means that $\|h\|_{H^{1/2}}^2:=\sum_{k\in\mathbb{Z}}|k||a_k|^2<\infty$. Observe that if $h\in C^r$, then $h^{(r)}$ has the Fourier series $\sum_{k\in\mathbb{Z}}(-ik)^2a_ke^{-ikt}$ and hence $\sum|k|^{2r}|a_k|^2=\|h^{(r)}\|_{L^2}^2$. Thus $H^{1/2}$ can be roughly called those functions hat have half a derivative. Indeed, one can also write the norm in a different way as

Exercise 65. Show that $\|h\|_{H^{1/2}}^2=\int_{-\pi}^{\pi}\int_{-\pi}^{\pi}\frac{(h(t)-h(s))^2}{(t-s)^2}\frac{dt\ ds}{4\pi^2}$.

If $h\in H^{1/2}$, we use the inequality $n-(n-k)_+\le|k|$ to see that

$$\mathrm{Var}(N_n(h))\le\frac{1}{2}\sum_{k\in\mathbb{Z}}|a_k|^2|k|=\|h\|_{H^{1/2}}^2.$$

This means that even as the expectation grows linearly in n, the variance stays bounded! Further, for each k fixed, $n-(n-k)_+\to|k|$ as $n\to\infty$, and hence by DCT $\mathrm{Var}(N_n(h))\to\|h\|_{H^{1/2}}^2$ as $n\to\infty$.

Case 2: h is the indicator of an arc $I=[a,b]$. Then $N_n(h)=N_n(I)$. We assume that the arc is proper (neither I or I^c is either empty or a singleton). The Fourier coefficients are given by $a_k=\int_a^b e^{-ikt}\frac{dt}{2\pi}=\frac{i(e^{-ikb}-e^{-ika})}{2\pi k}$. Evidently h is not in $H^{1/2}$.

We work out the special case when $I = [-\pi/2, \pi/2]$. Then $a_k = \frac{\sin(\pi k/2)}{\pi k}$ which is zero if k is even and equal to $\frac{(-1)^{j-1}}{\pi(2j+1)}$ if $k = 2j+1$. Thus,

$$\begin{aligned} \mathrm{Var}(N_n) &= 2\sum_{j=0}^{\infty} \frac{n-(n-2j-1)_+}{\pi^2(2j+1)^2} \\ &= 2\sum_{j=0}^{\lfloor \frac{n-1}{2} \rfloor} \frac{2j+1}{\pi^2(2j+1)^2} + 2\sum_{j>\lfloor \frac{n-1}{2} \rfloor} \frac{n}{\pi^2(2j+1)^2}. \end{aligned}$$

As $\sum_{j>m} j^{-2} = O(m^{-1})$, the second term is $O(1)$. The first term is easily seen to be $\frac{2}{\pi^2}\frac{1}{2}\log n + O(1)$. Thus, $\mathrm{Var}(N_n) = \frac{1}{\pi^2}\log n + O(1)$. Although it increases to infinity with n, the variance grows remarkably slower than the mean. Compare with sums of independent random variables where the mean and variance are both of order n. In the next exercise, take $I = [-\alpha, \alpha]$ without losing generality and show that the variance is asymptotically the same.

Exercise 66. Let f be a 2π-periodic function on $\mathbb{R}$ such that $\|f\|^2 := (2\pi)^{-1}\int_{-\pi}^{\pi}|f|^2$ is finite. Let $\hat{f}(k) := \int_{-\pi}^{\pi} f(t)e^{-ikt}\frac{dt}{2\pi}$ denote its Fourier coefficients. Let $f_\tau(t) = f(t-\tau)$ be the translates of f for any $\tau \in \mathbb{R}$.

(1) Use the Plancherel theorem to show that $4\sum_{k\in\mathbb{Z}} |\hat{f}(k)|^2 \sin^2(k\tau) = \|f_\tau - f_{-\tau}\|^2$. **[Hint:** $\hat{f}_\tau(k) = e^{-ik\tau}\hat{f}(k)$**]**

(2) Let $f(t) = t$ on $[-\pi, \pi]$ and extended periodically. Show that $\hat{f}(k) = \frac{(-1)^k}{k}$ and hence conclude that for $\tau \in [0, \pi]$

$$\sum_{k=1}^{\infty} \frac{\sin^2(k\tau)}{k^2} = \tau(\pi - \tau).$$

(3) Fix $\tau \in [0, \pi]$ and let $A_n = \sum_{k=1}^{n} \frac{\sin^2(k\tau)}{k}$ and $B_n = \sum_{k=1}^{n} \frac{\cos^2(k\tau)}{k}$. Show that $A_n + B_n = \log n + O(1)$ and $B_n - A_n = O(1)$ as $n \to \infty$. Conclude that both A_n and B_n are equal to $\frac{1}{2}\log n + O(1)$.

(4) Deduce that $\mathrm{Var}(N_n(I)) = \frac{1}{\pi^2}\log n + O(1)$ as $n \to \infty$ for any proper arc I (proper means $0 < |I| < 2\pi$).

Observe that the constant in front of $\log n$ does not depend on the length of the interval. Essentially the entire contribution to the variance comes from a few points falling inside or outside the interval at the two endpoints. Points which "were supposed to fall" deep in the interior of I (or deep in I^c) have almost no chance of falling outside of I (outside of I^c, respectively) and do not contribute to the variance. This shows the remarkable rigidity of the CUE.

Proposition 67. *In the setting of the previous discussion, for any proper arc I, as $n \to \infty$,*

$$\frac{N_n(I) - \frac{|I|}{2\pi}}{\pi^{-1}\sqrt{\log n}} \xrightarrow{d} N(0,1).$$

PROOF. Fix an arc I. By part (c) of Lema 56, $N_n(I)$ is a sum of independent Bernoulli random variables. By the Lindeberg Feller CLT for triangular arrays, any sum of independent Bernoullis converges to N(0,1) after subtracting the mean and dividing by the standard deviation, provided

the variance of the random variable goes to infinity. As $\mathrm{Var}(N_n(I)) \sim c\log n$, this applies to our case. ■

Next we compute the covariance between $N(I)$ and $N(J)$. We take I and J to be disjoint. Then, $\mathrm{Cov}(N(I),N(J)) = -\int_I \int_J |K(x,y)|^2 d\mu(x) d\mu(y)$.

11. Fredholm determinants and hole probabilities

Let $(A,\mathcal{A},\mu)$ be a probability space. Let $K : A^2 \to \mathbb{R}$ or $\mathbb{C}$ be a kernel such that $\|K\| := \sup_{x,y} |K(x,y)| < \infty$. Let T be the integral operator with kernel K.

Definition 68. The *Fredholm determinant* of the operator $I-T$ which we shall also call the Fredholm determinant associated to the kernel K is defined as

$$\Delta(K) := \sum_{m=0}^{\infty} \frac{(-1)^m}{m!} \int_{A^m} \det\left(K(x_i,x_j)\right)_{i,j\le m} d\mu(x_1)\ldots d\mu(x_m).$$

Recall the *Hadamard inequality* for matrices which says that if M is a square matrix with columns u_k, $k \le n$, then $|\det(M)| \le \prod_{j=1}^n \|u_j\|$. Therefore, $|\det\left(K(x_i,x_j)\right)_{i,j\le m}| \le (\|K\|\sqrt{m})^m$ for any m and any $x_1,\ldots,x_m$. This shows that $\Delta(K)$ is well-defined for any K with $\|K\| < \infty$.

Remark 69. Let M be an $n\times n$ matrix with eigenvalues θ_j, $j \le n$. Then, we leave it as an exercise to show the identity

$$\sum_{1\le i_1<i_2<\ldots<i\le m} \theta_{i_1}\theta_{i_2}\ldots\theta_{i_m} = \sum_{1\le i_1<i_2<\ldots<i\le m} \det\left(M_{i_p,i_q}\right)_{p,q\le m}$$

for any $m \ge 1$. For $m=1$ this is just the identity $\sum\theta_i = \sum_i M_{i,i}$. For any $m \ge 1$, one can think of the identity as being exactly the same identity, applied to a different matrix. If M acts on a vector space V, then one can define the operator $M^{\wedge k}$ on the alternating tensor power $V^{\wedge k}$ as $\langle M(e_{j_1}\wedge \ldots\wedge e_{j_k}), e_{i_1}\wedge\ldots\wedge e_{i_k}\rangle = \det\left(M_{i_p,j_q}\right)_{p,q\le k}$. This has eigenvalues $\theta_{i_1}\theta_{i_2}\ldots\theta_{i_k}$ where $i_1<i_2<\ldots<i_k$. Expressing $\mathrm{tr}(M^{\wedge k})$ in two ways gives the above identity.

Anyhow, from this identity, we get the following expression for $\det(I-M) = \prod_{j=1}^n(1-\theta_j)$.

$$\begin{aligned}
\det(I-M) &= \prod_{j=1}^n (1-\theta_j) \\
&= 1 - \sum_i \theta_i + \sum_{i<j}\theta_i\theta_j - \sum_{i<j<k}\theta_i\theta_j\theta_k + \ldots \\
&= 1 - \sum_i M_{i,i} + \frac{1}{2}\sum_{i,j}\det\begin{bmatrix} M_{i,i} & M_{i,j} \\ M_{j,i} & M_{j,j}\end{bmatrix} - \frac{1}{6}\sum_{i,j,k}\det\begin{bmatrix} M_{i,i} & M_{i,j} & M_{i,k} \\ M_{j,i} & M_{j,j} & M_{j,k} \\ M_{k,i} & M_{k,j} & M_{k,k}\end{bmatrix} + \ldots
\end{aligned}$$

Thus, $\det(I-M)$ is exactly what we defined as $\Delta(K)$, provided we take $A=[n]$ and $K(i,j) = M_{i,j}$. With the usual philosophy of regarding an integral kernel as a matrix $(K(x,y))_{x,y}$, we arrive at the definition of the Fredholm determinant. The following exercise is instructive in this respect.

Exercise 70. Let T be the integral operator with a Hermitian kernel K with $\|K\| < \infty$. Let θ_j be the eigenvalues of T. Then, for any $m \geq 1$, we have

$$\sum_{i_1<i_2<\ldots<i_m} \theta_{i_1}\theta_{i_2}\ldots\theta_{i_m} = \frac{1}{m!}\int_{A^m} \det\left(K(x_i,x_j)\right)_{i,j\leq m} d\mu(x_1)\ldots d\mu(x_m).$$

We shall need the following simple lemma later[8].

Lemma 71. *Let K and L be two kernels on $L^2(A,\mathcal{A},\mu)$ such that $C = \max\{\|K\|,\|L\|\} < \infty$. Then,*

$$|\Delta(K) - \Delta(L)| \leq \|K-L\| \left(\sum_{m=0}^{\infty} \frac{m(C\sqrt{m})^{m-1}}{m!}\right).$$

PROOF. Fix $m \geq 1$ and $x_1,\ldots,x_m \in A$. Let $X_0 = (K(x_i,x_j))_{i,j\leq m}$ and $X_m = (L(x_i,x_j))_{i,j\leq m}$. For $1 \leq k \leq m$, let X_k be the matrix whose first k rows are those of X_0 and the rest are those of X_m. Then, $\det(X_0) - \det(X_m) = \sum_{k=1}^{m-1} \det(X_{k-1}) - \det(X_k)$. Using Hadamard's inequality we see that $|\det(X_{k-1}) - \det(X_k)|$ is bounded by $(C\sqrt{m})^{m-1}\|K-L\|$. Thus

$$|\det\left(K(x_i,x_j)\right)_{i,j\leq m} - \det\left(L(x_i,x_j)\right)_{i,j\leq m}| \leq m(C\sqrt{m})^{m-1}\|K-L\|.$$

Integrate over x_js and then sum over m (after multiplying by $(-1)^{m-1}/m!$ to get the claimed result. ■

The importance of Fredholm determinants for us comes from the following expression for "hole probabilities" or "gap probabilities" in determinantal processes.

Proposition 72. *Let $(A,\mathcal{A},\mu)$ be a probability space and let K be a finite rank projection kernel (that is $K(x,y) = \sum_{j=1}^{n} \phi_j(x)\overline{\phi}_j(y)$ for some orthonormal set $\{\phi_j\}$). Let λ have density $(n!)^{-1}\det\left(K(x_i,x_j)\right)_{i,j\leq n}$. Let $I \subseteq A$ be a measurable subset of A. Then $\mathbf{P}(N(I)=0) = \Delta(K_I)$, where K_I is the kernel K restricted to $I\times I$.*

PROOF. From part (c) of Lemma **??**, we know that $\mathbf{P}(N(I)=0) = \prod_j(1-\theta_j)$ where θ_j are the eigenvalues of the integral operator T_I with kernel K_I. Hence,

$$\begin{aligned}\mathbf{P}(N(I)=0) &= 1 - \sum_i \theta_i + \sum_{i<j}\theta_i\theta_j - \sum_{i<j<k}\theta_i\theta_j\theta_k + \ldots \\ &= \sum_{m=0}^{\infty} \frac{(-1)^m}{m!}\int_{I^m} \det\left(K(x_i,x_j)\right)_{i,j\leq m} d\mu(x_1)\ldots d\mu(x_m)\end{aligned}$$

by Exercise 70. The last expression is $\Delta(K_I)$ by definition. ■

[8]We have borrowed much of this section from the book of Anderson, Guionnet and Zeitouni **?** where the reader may find more about these objects. F.Riesz and Sz. Nagy's great book on Functional analysis is another good reference for Fredholm's work in functional analysis.

12. Gap probability for CUE

Let λ be distributed as CUE_n ensemble and unwrap the circle onto the interval $[-\pi,\pi]$. Thus λ follows the measure on $[-\pi,\pi]^n$ given by

$$\frac{1}{n!}\det\left(K_n(t_i,t_j)\right)_{i,j\le n}\frac{dt_1\dots dt_n}{(2\pi)^n}, \qquad \text{where } K_n(t,s)=\frac{\sin\left(\frac{n}{2}(s-t)\right)}{\sin\left(\frac{s-t}{2}\right)}.$$

Scale up by a factor of $n/2$ to get $\tilde{\lambda}=n\lambda/2$ which follows the measure (on $[-n\pi/2,n\pi/2]^n$)

$$\frac{1}{n!}\det\left(\tilde{K}_n(t_i,t_j)\right)_{i,j\le n}\frac{dt_1\dots dt_n}{(2\pi)^n}, \qquad \text{where } \tilde{K}_n(t,s)=\frac{2}{n}K_n\left(\frac{2t}{n},\frac{2s}{n}\right).$$

Then,

$$\tilde{K}_n(t,s)=\frac{2\sin(s-t)}{n\sin\left(\frac{s-t}{n}\right)}\to K(t,s):=\frac{2\sin(s-t)}{s-t}.$$

It is also easy to see that the convergence is uniform over (t,s) in any compact subset of $\mathbb{R}^2$. Further, $\|\tilde{K}_n\|\le 2$ and $\|K\|\le 2$. Thus, by Lemma 71, we see that $\Delta(\tilde{K}_{n,I})\to\Delta(K_I)$ for any compact interval I. By Proposition prop:holefordeterminantal this shows that for any $a<0<b$,

$$\mathbf{P}\left(\lambda_i\notin\left[\frac{2a}{n},\frac{2b}{n}\right]\ \forall i\le n\right)=\mathbf{P}\left(\tilde{\lambda}_i\notin[-a,b]\ \forall i\le n\right)\to\Delta(K_{[a,b]})$$

as $n\to\infty$. This gives the asymptotics of gap probabilities in CUE. Some remarks are due.

Of course, it is incorrect to say that we have calculated the gap probability unless we can produce a number or decent bounds for this probability. For example, we could define $F(t):=\Delta(K_{[-t,t]})$ which is the asymptotic probability that the nearest eigenvalue to 0 in CUE_n is at least $2t/n$ away. Can we find $F(t)$? All we need to so is study the kernel K (called the *sine kernel*) and deduce $F(t)$ from it. This is not trivial, but has been done by ???? They show that $F(t)$ can be characterized in terms of the solution to a certain second order ODE, called the ?????? We do not prove this result in this course.

Secondly, we considered only the gap probability, but we could also consider the distributional limit of the whole point process $\tilde{L}_n:=\sum_k\delta_{\tilde{\lambda}_k}$. But then we must employ the language of Section **??**. In that language, it is not difficult to show that the convergence of $\tilde{K}_n$ to K implies that $\tilde{L}_n$ converges in distribution to L, the determinantal point process with kernel K. The latter is a stationary point process on the line (and hence has infinitely many points, almost surely). Basically this distributional convergence is the statement that all the joint intensities $\det\left(K_n(x_i,x_j)\right)_{i,j\le m}$ converge to the corresponding quantities $\det\left(K(x_i,x_j)\right)_{i,j\le m}$. However, note that the distributional convergence does not automatically imply convergence of the gap probability, because the latter is expressed as a series involving joint intensities of all orders. That is why we had to establish Lemma 71 first.

13. Hermite polynomials

Our next goal is to prove results for GUE analogous to those that we found for CUE. Additionally, we would also like to study the edge behaviour in GUE, for which there is no analogue in

CUE. In this section we shall establish various results on Hermite polynomials that will be needed in carrying out this programme.

For $n \geq 0$, define $\tilde{H}_n(x) := (-1)^n e^{-x^2/2} \frac{\partial^n}{\partial x^n} e^{-x^2/2}$. It is easily seen that $\tilde{H}_n$ is a monic polynomial of degree n. It is also easy to see that the coefficients of x^{n-1}, x^{n-3} etc. are zero. Consider

$$\begin{aligned}
\int H_n(x) H_m(x) e^{-x^2/2} \frac{dx}{\sqrt{2\pi}} &= (-1)^n \int H_n(x) \frac{\partial^n}{\partial x^n} e^{-x^2/2} \frac{dx}{\sqrt{2\pi}} \\
&= \int e^{-x^2/2} \frac{\partial^n}{\partial x^n} H_n(x) \frac{dx}{\sqrt{2\pi}} \\
&= \begin{cases} 0 & \text{if } n < m \text{ because } H_n \text{ has degree only } n. \\ n! & \text{if } m = n. \end{cases}
\end{aligned}$$

Thus $H_n(x) := \frac{1}{\sqrt{n!}} \tilde{H}_n(x)$ define an orthonormal sequence of polynomials with respect to N(0,1) measure called *Hermite polynomials*. Let $\psi_n(x) = (2\pi)^{-1/4} e^{-x^2/4} H_n(x)$ be the *Hermite functions*. Then $\{\psi_n : n \geq 0\}$ for an ONB for $L^2(\mathbb{R}, \text{Lebesgue})$. The following properties may be derived easily (or look up any book on special functions, for example, Andrews, Askey and Roy **?**).

Exercise 73. (1) $\left(-\frac{\partial}{\partial x} + x\right) \tilde{H}_n(x) = \tilde{H}_{n+1}(x)$ and hence also $\left(-\frac{\partial}{\partial x} + x\right) H_n(x) = \sqrt{n+1} H_{n+1}(x)$.

(2) *Hermite functions are eigenfunctions of the Hermite operator:* $\left(-\frac{\partial}{\partial x} + \frac{x}{2}\right) \psi_n(x) = \sqrt{n+1} \psi_{n+1}(x)$ and $\left(\frac{\partial}{\partial x} + \frac{x}{2}\right) \psi_n(x) = \sqrt{n} \psi_{n-1}(x)$. Consequently,

$$\left(-\frac{\partial^2}{\partial x^2} + \frac{x^2}{4}\right) \psi_n(x) = \left(n + \frac{1}{2}\right) \psi_n(x). \tag{33}$$

(3) *Three term recurrence:* $x\tilde{H}_n(x) = n\tilde{H}_{n-1}(x) + \tilde{H}_{n+1}(x)$. Consequently, $xH_n(x) = \sqrt{n} H_{n-1}(x) + \sqrt{n+1} H_{n+1}(x)$.

We now derive two integral representations for Hermite polynomials. Observe that $\frac{\partial^n}{\partial x^n} e^{-(x-w)^2/2} \Big|_{w=0} = (-1)^n \frac{\partial^n}{\partial x^n} e^{-x^2/2}$. Therefore, fixing x, we get the power series expansion $e^{-(x-w)^2} = \sum_{n=0}^{\infty} H_n(x) w^n / n!$ which simplifies to $e^{xw - \frac{w^2}{2}} = \sum_{n=0}^{\infty} H_n(x) w^n / n!$. Thus,

$$H_n(x) = \frac{1}{2\pi i} \int_\gamma \frac{e^{xw - \frac{w^2}{2}}}{w^{n+1}} dw, \qquad \text{for any closed curve } \gamma \text{ with } \text{Ind}_\gamma(0) = 1. \tag{34}$$

A second integral representation will be obtained from the well-known identity

$$e^{-x^2/2} = \int_{\mathbb{R}} e^{-itx} e^{-t^2/2} \frac{dt}{\sqrt{2\pi}} = \int_{\mathbb{R}} \cos(tx) e^{-t^2/2} \frac{dt}{\sqrt{2\pi}}.$$

Differentiate n times with respect to x to get

$$\tilde{H}_n(x) = \begin{cases} (-1)^m \int_{\mathbb{R}} \cos(tx) x^n e^{-t^2/2} \frac{dt}{\sqrt{2\pi}} & \text{if } n = 2m. \\ (-1)^{m-1} \int_{\mathbb{R}} \sin(tx) x^n e^{-t^2/2} \frac{dt}{\sqrt{2\pi}} & \text{if } n = 2m-1. \end{cases} \tag{35}$$

We end the section with the Christoffel-Darboux formula.

Lemma 74. *Let μ be a probability measure on $\mathbb{R}$ with infinite support. Let p_k be the orthogonal polynomials with respect to μ normalized so that $p_n(x) = \kappa_n x^n + \ldots$ with $\kappa_n > 0$. Then,*

$$\sum_{k=0}^{n-1} p_k(x)p_k(y) = \frac{\kappa_{n-1}}{\kappa_n} \frac{p_n(x)p_{n-1}(y) - p_{n-1}(x)p_n(y)}{x-y}.$$

For $x = y$ the right hand should be interpreted as $\frac{\kappa_{n-1}}{\kappa_n}(p_n(x)p'_{n-1}(x) - p'_n(x)p_{n-1}(x))$.

PROOF. Write the three term recurrence

$$xp_n(x) = b_{n-1}p_{n-1}(x) + a_n p_n(x) + b_n p_{n+1}(x).$$

Multiply by $p_n(y)$ to get the equation

$$xp_n(x)p_n(y) = b_{n-1}p_{n-1}(x)p_n(y) + a_n p_n(x)p_n(y) + b_n p_{n+1}(x)p_n(y).$$

Write this same equation with x and y reversed and subtract from the above equation to get

$$(x-y)p_n(x)p_n(y) = -b_{n-1}(p_{n-1}(y)p_n(x) - p_{n-1}(x)p_n(y)) + b_n(p_n(y)p_{n+1}(x) - p_n(x)p_{n+1}(y)).$$

Put k in place of n and sum over $0 \leq k \leq n-1$ to get the identity

$$\sum_{k=0}^{n-1} p_k(x)p_k(y) = b_{n-1} \frac{p_n(x)p_{n-1}(y) - p_{n-1}(x)p_n(y)}{x-y}.$$

In the original three term recurrence equate the coefficients of x^{n+1} to see that $b_n \kappa_{n+1} = \kappa_n$. This completes the proof. ■

Corollary 75. *For any $n \geq 1$, we have*

$$\sum_{k=0}^{n-1} \psi_k(x)\psi_k(y) = \sqrt{n} \frac{\psi_n(x)\psi_{n-1}(y) - \psi_{n-1}(x)\psi_n(y)}{x-y}.$$

The corollary follows immediately from the lemma. The importance for us is that it makes it very clear that analysis of the GUE for large n depends on understanding ψ_n (or equivalently, understanding H_n) for large n.

Remark 76. Below are supposed to be three sections on

(1) Deriving the semi-circle law from the exact GUE density using properties of Hermite polynomials, Hermite polynomials,
(2) Getting the bulk scaling limit using Laplace's method and
(3) Getting the edge-scaling limit using the saddle-point method.

But never got to cover all of it in class or to write notes for them.

CHAPTER 4

Elements of free probability theory

1. Cumulants and moments in classical probability

Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space. For random variables X_i on this probability space, define $m_n[X_1, \ldots, X_n] = \mathbf{E}[\prod_{j=1}^n X_j]$ whenever the expectation exists. We will also write $m_0 = 1$. The function $m_\cdot[\cdot]$ is called the *moment* function.

Let $\mathcal{P}_n$ denote the set of all set-partitions of $[n]$. For example, $\mathcal{P}_3$ consists of the five partitions $\{\{1,2,3\}\}$, $\{\{1,2\},\{3\}\}$, $\{\{1,3\},\{2\}\}$, $\{\{2,3\},\{1\}\}$ and $\{\{1\},\{2\},\{3\}\}$. The sets that make up a partition are referred to as *blocks*. Note that the order of the blocks, or of the elements in individual blocks are irrelevant (in other words, the partition $\{\{3\},\{2,1\}\}$ is the same as $\{\{1,2\},\{3\}\}$). For a partition Π we denote the number of blocks by ℓ_Π and the individual blocks by Π_j, $1 \leq j \leq \ell_\Pi$. If we ever need to be more definite, we shall let Π_1 be the block containing 1, Π_2 to be the block containing the least element not in Π_1 etc.

Definition 77. Define the *cumulant function* $\kappa_n[X_1, \ldots, X_n]$ by the equations

$$m_n[X_1, \ldots, X_n] = \sum_{\Pi \in \mathcal{P}_n} \prod_{j=1}^{\ell_\Pi} \kappa_{|\Pi_j|}[X[\Pi_j]]. \tag{36}$$

Here if $\Pi_j = \{k_1, \ldots, k_r\}$ with $k_1 < k_2 < \ldots < k_r$, then $|\Pi_j| := r$ and $[X[\Pi_j]]$ is the short form for $[X_{k_1}, \ldots, X_{k_r}]$.

Rewrite the first three equations as

$$\kappa_1[X] = m_1[X], \qquad \kappa_2[X,Y] = m_2[X,Y] - \kappa_1[X]\kappa_1[Y]$$
$$\kappa_3[X,Y,Z] = m_3[X,Y,Z] - \kappa_2[X,Y]\kappa_1[Z] - \kappa_2[X,Z]\kappa_1[Y] - \kappa_2[Y,Z]\kappa_1[X] + \kappa_1[X]\kappa_1[Y]\kappa_1[Z]$$

It is clear that we can define κ_1 from the first equation, κ_2 from the second and so on, inductively.

For any $\Pi \in \mathcal{P}_n$, introduce the notation

$$m_\Pi[X_1, \ldots, X_n] = \prod_{j=1}^{\ell_\Pi} m_{|\Pi_j|}[X[\Pi_j]], \qquad \kappa_\Pi[X_1, \ldots, X_n] = \prod_{j=1}^{\ell_\Pi} \kappa_{|\Pi_j|}[X[\Pi_j]].$$

In this notation, the equations defining cumulants may be written as $m_n[X] = \sum_{\Pi \in \mathcal{P}_n} \kappa_\Pi[X]$ where $X = (X_1, \ldots, X_n)$.

Exercise 78. Show that $\kappa_n[X] = \sum_{\Pi \in \mathcal{P}_n} (-1)^{\ell_\Pi - 1} (\ell_\Pi - 1)! m_\Pi[X]$.

The following lemma collects some basic properties of cumulants.

Lemma 79. *(1) Cumulant function is multilinear:* $\kappa_n[cX_1 + dX_1', X_2, \ldots, X_n] = c\kappa_n[X_1, X_2, \ldots, X_n] + d\kappa_n[X_1', X_2, \ldots, X_n]$ *and similarly in each of the other co-ordinates. Further,* κ_n *is symmetric in its arguments. For* $\Pi \in \mathcal{P}_n$, κ_Π *is multilinear but not necessarily symmetric.*

(2) Assume that $X = (X_1, \ldots, X_d)$ *is such that* $\mathbf{E}[e^{\langle \mathbf{t}, \mathbf{X}\rangle}] < \infty$ *for* $\mathbf{t}$ *in a neighbourhood of* 0 *in* $\mathbb{R}^d$. *Let* $\phi_X(\mathbf{t}) = \mathbf{E}\left[\mathbf{e}^{\langle \mathbf{t}, \mathbf{X}\rangle}\right]$ *and* $\psi_X(\mathbf{t}) = \log \mathbf{E}\left[\mathbf{e}^{\langle \mathbf{t}, \mathbf{X}\rangle}\right]$. *Then,*

$$\phi_X(\mathbf{t}) = \sum_{n=0}^{\infty} \sum_{i_1, \ldots, i_n = 1}^{d} \frac{t_{i_1} \ldots t_{i_n}}{n!} m_n[X_{i_1}, \ldots, X_{i_n}],$$

$$\psi_X(\mathbf{t}) = \sum_{n=1}^{\infty} \sum_{i_1, \ldots, i_n = 1}^{d} \frac{t_{i_1} \ldots t_{i_n}}{n!} \kappa_n[X_{i_1}, \ldots, X_{i_n}].$$

(3) Let $U = (X_1, \ldots, X_k)$ *and* $V = (X_{k+1}, \ldots, X_d)$. *Then, the following are equivalent.*

(i) U *and* V *are independent.*

(ii) $\kappa_n[X_{i_1}, \ldots, X_{i_n}] = 0$ *for any* $n \geq 1$ *and any* $i_1, \ldots, i_n \in [d]$ *whenever there is least one* p *such that* $i_p \leq k$ *and at least one* q *such that* $i_q > k$.

PROOF. (1) Obvious.

(2) Expand $e^{\langle \mathbf{t}, \mathbf{X}\rangle} = \sum_n \langle \mathbf{t}, \mathbf{X}\rangle^{\mathbf{n}}/\mathbf{n}!$ and $\langle \mathbf{t}, \mathbf{X}\rangle^{\mathbf{n}} = \sum_{\mathbf{i_1}, \ldots, \mathbf{i_n} = \mathbf{1}}^{\mathbf{d}} \mathbf{t_{i_1}} \ldots \mathbf{t_{i_n}} \mathbf{X_{i_1}} \ldots \mathbf{X_{i_n}}$. Taking expectations gives the expansion for $\phi_X(t)$. To get the expansion for $\Psi_X(\mathbf{t})$, let $\psi(\mathbf{t}) = \sum_{\mathbf{n}=\mathbf{1}}^{\infty} \sum_{\mathbf{i_1}, \ldots, \mathbf{i_n} = \mathbf{1}}^{\mathbf{d}} \frac{\mathbf{t_{i_1}} \ldots \mathbf{t_{i_n}}}{\mathbf{n}!} \kappa_{\mathbf{n}}[\mathbf{X_{i_1}}, \ldots, \mathbf{X_{i_n}}]$ and consider

$$e^{\psi(\mathbf{t})} = \sum_{n=1}^{\infty} \frac{1}{n!} \sum_{k_1, \ldots, k_n = 1}^{d} \kappa_{k_1}$$

(3) $U = (X_1, \ldots, X_m)$ is independent of $V = (X_{m+1}, \ldots, X_n)$ if and only if $\psi_{(U,V)}(t, s) = \psi_U(t) + \psi_V(s)$ for all $t \in \mathbb{R}^m$, $s \in \mathbb{R}^{n-m}$. By part (b), ψ_U (respectively, ψ_V) has an expansion involving $\kappa_k[X_{i_1}, \ldots, X_{i_k}]$ where $i_1, \ldots, i_k \leq m$ (respectively, $i_1, \ldots, i_k > m$). However, $\psi_{(U,V)}$ has coefficients $\kappa_k[X_{i_1}, \ldots, X_{i_k}]$ where i_r range over all of $[n]$. Thus, U and V are independent if and only if $\kappa_k[X_{i_1}, \ldots, X_{i_k}] = 0$ whenever there are p, q such that $i_p \leq m$ and $i_q > m$. This proves the equivalence of the two statements. ■

Part (c) of the lemma is the reason why cumulants are useful in studying independent random variables. We shall illustrate this by a quick proof of the central limit theorem (for a restricted class of random variables). However, first we make a few remarks on cumulants of one random variable which the reader may be familiar with.

Let X be a real-valued random variable such that $\mathbf{E}[e^{tX}] < \infty$ for t in a neighbourhood of zero. Then $\phi_X(t) = \sum_{n=0}^{\infty} m_n(X) t^n/n!$ and $\psi_X(t) = \sum_{n=1}^{\infty} \kappa_n(X) t^n/n!$ where $m_n = m_n[X, \ldots, X]$ and $\kappa_n[X, \ldots, X]$. The relationship between moments and cumulants becomes

$$m_n(X) = \sum_{\Pi \in \mathcal{P}_n} \prod_{j=1}^{\ell_\Pi} \kappa_{|\Pi_j|}(X), \qquad \kappa_n(X) = \sum_{\Pi \in \mathcal{P}_n} (-1)^{\ell_\Pi - 1} (\ell_\Pi - 1)! \prod_{j=1}^{\ell_\Pi} \kappa_{|\Pi_j|}(X).$$

The cumulant sequence (or the moment sequence) determines the moment generating function and hence the distribution of X. Thus knowing the cumulant sequence is sufficient to answer

every question about X (in principle). Of course, a quantity like $\mathbf{P}(1 < X < 2)$ is not easy to express in terms of cumulants, so the "in principle" phrase must be taken seriously. There is an additional issue of loss of generality in considering only random variables with moments. For these reasons usually one does not base probability theory on moments or cumulants exclusively. However, there are features that can be captured well in terms of cumulants. Independence is one of them, as part (c) of Lemma 79 shows.

Summing of independent random variables is also captured nicely in terms of cumulants. Indded, if X and Y are independent random variables, by part (a) of Lemma 79 we can write $\kappa_n(X+Y) = \kappa_n[X+Y,\ldots,X+Y]$ as a sum of 2^n terms. By part (c) of the same lemma, using independence, all but two of these vanish and we get $\kappa_n(X+Y) = \kappa_n(X) + \kappa_n(Y)$. A particular case is when $Y = c$, a constant, in which case $\kappa_n(X+c) = \kappa_n(X) + c\delta_{n,1}$. Observe that in contrast, $m_n(X+c)$ has a relatively more complicated expression in terms of moments of X.

Exercise 80. (1) If $X \sim N(\mu, \sigma^2)$, then $\kappa_1[X] = \mu$, $\kappa_2[X] = \sigma^2$ and $\kappa_n[X] = 0$ for $n \geq 3$.

(2) Conversely, if $\kappa_n[X] = 0$ for all $n \geq 3$, then $X \sim N(\kappa_1, \kappa_2)$.

(3) If X, Y are i.i.d random variables and $X + Y \stackrel{d}{=} \sqrt{2}X$, show that $X \sim N(0, \sigma^2)$ for some σ^2.

Example 81. Let $X \sim \exp(1)$. Then $\phi_X(t) = (1-t)^{-1} = \sum_{n\geq 0} t^n$ for $t < 1$. Hence $m_n = n!$. $\psi_X(t) = -\log(1-t) = \sum_{n\geq 1} n^{-1}t^n$ which shows that $\kappa_n = (n-1)!$. If $Y \sim \text{Gamma}(\nu, 1)$ then for integer values of ν it is a sum of i.i.d exponentials, hence $\kappa_n(Y) = \nu(n-1)!$. It may be verified directly that this is also true for any $\nu > 0$.

Example 82. Let $X \sim \text{Pois}(1)$. Then $\mathbf{E}[e^{tX}] = e^{-1+e^t}$. Expanding this, one can check that $m_n = e^{-1}\sum_{k=0}^{\infty} \frac{k^n}{k!}$. It is even easier to see that $\psi_X(t) = -1 + e^t$ and hence $\kappa_n = 1$ for all $n \geq 1$ and hence also $\kappa_\Pi = 1$. But then, the defining equation for cumulants in terms of moments shows that $m_n = \sum_{\Pi \in \mathcal{P}_n} \kappa_\Pi = |\mathcal{P}_n|$. Thus as a corollary, we have the non-trivial relation $|\mathcal{P}_n| = e^{-1}\sum_{k=0}^{\infty} \frac{k^n}{k!}$, known as Dobinsky's formula.

Remark 83. The relationship between m_n and κ_n just comes from the connection that $\log\phi = \psi$ where $m_n/n!$ are the coefficient of ϕ and $\kappa_n/n!$ are coefficients of ψ. The same is true for coefficients of any two power series related this way. A closer look at the expressions for m_n in terms of κ_n or the reverse one shows that if m_n counts some combinatorial objects, then κ_n counts the connected pieces of the same combinatorial object.

For example, in Example 81, $m_n = n!$ counts the number of permutations on n letters while $\kappa_n = (n-1)!$ counts the number of cyclic permutations. As any permutation may be written as a product of disjoint cycles, it makes sense to say that cycles are the only connected permutations.

In Example 82, $m_n = |\mathcal{P}_n|$ while $\kappa_n = 1$. Indeed, the only "connected partition" is the one having only one block $\{1, 2, \ldots, n\}$.

In case of $N(0,1)$, we know that m_n counts the number of matching of $[n]$. What are connected matchings? If $n > 2$, there are no connected matchings! Hence, $\kappa_n = 0$ for $n \geq 3$.

Now we turn to the promised proof of CLT. By part (c) of Exercise 80, if $S_n/\sqrt{n}$ were to converge to a limit, then it is easy to see that the limit random would have to satisfy the recursive distributional equation $U+V \stackrel{d}{=} \sqrt{2}U$ where U,V are i.i.d copies of the limit variable and hence $U \sim N(0,\sigma^2)$. Using cumulants we can actually show that this is the case.

PROOF OF CENTRAL LIMIT THEOREM ASSUMING MGF EXISTS. Suppose $X_1, X_2, \ldots$ are i.i.d with zero mean and unit variance and such that the mgf of X_1 exists in a neighbourhood of zero, then for any fixed $p \geq 1$,

$$\kappa_p[S_n/\sqrt{n}] \;=\; n^{-\frac{p}{2}}\kappa[S_n,\ldots,S_n] \;=\; n^{-\frac{p}{2}} \sum_{\mathbf{i}\in[n]^p} \kappa[X_{i_1},\ldots,X_{i_p}]$$

by multilinearity of cumulants. If $X_{i_r} \neq X_{i_s}$, the corresponding summand will vanish by the independence of X_js. Therefore,

$$\kappa_p[S_n/\sqrt{n}] = n^{-\frac{p}{2}} \sum_{j=1}^{n} \kappa[X_j, X_j, \ldots, X_j] = n^{-\frac{p}{2}+1}\kappa_p[X_1]$$

which goes to zero for $p \geq 3$. As the first two cumulants are 0 and 1 respectively, we see that the cumulants of $S_n/\sqrt{n}$ converge to cumulants of $N(0,1)$ and hence the moments converge also. Thus, $S_n/\sqrt{n}$ converges in distribution to $N(0,1)$. ■

2. Non-commutative probability spaces

We define[1] three notions of non-commutative probability space, of which the first one is sufficient for our purposes. In the next section we shall introduce the notion of independence in such spaces.

Definition 84. A *non-commutative probability space* is a pair $(\mathcal{A},\phi)$ where $\mathcal{A}$ is a *unital algebra* over complex numbers and ϕ is a linear functional on $\mathcal{A}$ such that $\phi(\mathbf{1}) = \mathbf{1}$.

A unital algebra $\mathcal{A}$ is a vector space over $\mathbb{C}$ endowed with a multiplication operation $(a,b) \to ab$ which is assumed to be associative and also distributive over addition and scalar multiplication. In addition we assume that there is a *unit*, denoted 1, such that $a\mathbf{1} = a = \mathbf{1}a$ for all $a \in \mathcal{A}$.

Example 85. Let $\mathcal{A}$ be the space of all polynomials in one variable with complex coefficients. This is a unital algebra with the obvious operations. Fix a complex Borel measure μ on $\mathbb{R}$ such that $\mu(\mathbb{R}) = 1$. Define $\phi(P) = \int P(x)\mu(dx)$ for any $P \in \mathcal{A}$. Then, $(\mathcal{A},\phi)$ is a (commutative!) ncps. This leads us to a smaller class of ncps. If we considered polynomials in three variables and μ a measure on $\mathbb{R}^3$, we would again get a ncps. The difference is that in one dimension, at least if μ is compactly supported, then $(\mathcal{A},\phi)$ has all the information in the classical measure space $(\mathbb{R},\mathcal{B}_{\mathbb{R}},\mu)$.

[1]Much of our presentation of free probability is taken from three sources. The St. Flour lecture notes of Voiculescu **?**, various lecture notes of Roland Speicher available on his home page, and the book of Anderson Guionnet and Zeitouni.

In the example above, of particular interest are probability measures. We have assumed that $\mu(\mathbb{R}) = 1$, but positivity is an extra condition which can be framed by saying that $\phi(P) \geq 0$ if $P(x) \geq 0$ for all $x \in \mathbb{R}$. Observe that there is no clear way to introduce the notion of positivity in a general unital algebra. This leads us to a smaller sub-class of ncps.

Definition 86. Let $\mathcal{A}$ be a C^*-algebra[2] with a unit Let $\phi : \mathcal{A} \to \mathbb{C}$ be a linear functional such that $\phi(aa^*) \geq 0$ for all $a \in \mathcal{A}$ (we say that ϕ is a positive linear functional). Assume also that $\phi(\mathbf{1}) = 1$. Then, we say that ϕ is a state. $(\mathcal{A}, \phi)$ is called a C^*-probability space.

Observe that ϕ is necessarily bounded. In fact, for any self-adjoint a, $a - \|a\|1$ and $\|a\|1 - a$ are non-negative elements (can be written as b^*b for some b). Hence $|\phi(a)| \leq \|a\|$ as $\phi(1) = 1$. If a is any element of the algebra, it can be written in a unique way as $x + iy$ where x, y are self-adjoint and hence $|\phi(a)| \leq 2$.

Example 87. Let $\mathcal{A} := \mathcal{B}(H)$ be the algebra of bounded linear operators on a Hilbert space H. This is a C^*-algebra where the identity I is the unit and taking adjoints is the involution. Let $u \in H$ be a unit vector and define $\phi(T) = \langle Tu, u\rangle$. Then, ϕ is a linear functional and $\phi(I) = 1$. Further, $\phi(T^*T) = \|Tu\|^2 \geq 0$. Thus, $(\mathcal{A}, \phi)$ is a C^*-probability space. Here multiplication is truly non-commutative.

If $\psi(T) = \langle Tv, v\rangle$ for a different unit vector v, then for $0 < s < 1$, the pair $(\mathcal{A}, s\phi + (1-s)\psi)$ is also a C^*-probability space. ϕ is called a pure state while $s\phi + (1-s)\psi$ is called a mixed state. Any closed subalgebra of $\mathcal{B}(H)$ that is closed under adjoints is also a C^*-algebra. We only consider those that contain the unit element.

Example 88. Let K be a compact metric space and let $\mathcal{A} = C(K)$ (continuous complex-valued functions). The operations are obvious (involution means taking the conjugate of a function). Let μ be any Borel probability measure on K and define $\phi(f) = \int_K f d\mu$. Then $(\mathcal{A}, \phi)$ is a C^*-probability space.

Example 89. The same applies to $C_b(\mathbb{R})$ and $\phi(f) = \int f d\mu$ for some Borel probability measure μ. It is a commutative C^*-algebra. In fact this is not different from the previous example, as $C_b(\mathbb{R}) = C(K)$ where K is the Stone-Cech compactification of $\mathbb{R}$.

As these examples show, a C^*-probability space generalizes the idea of presenting a probability measure on $\mathbb{R}$ by giving the integrals of all bounded continuous functions which is more than giving the integral of polynomials only. However, for later purposes, it is useful to remark that C^*-probability space is like the algebra of *complex-valued* random variables, not real valued ones. A third level is to specify a probability measure μ by giving the integrals of bounded measurable functions.

[2]By definition, this means that $\mathcal{A}$ has three structures. (a) That of a complex Banach space, (b) that of an algebra and finally, (c) an *involution* $* : \mathcal{A} \to \mathcal{A}$. These operations respect each other as follows. The algebra operations are continuous and respect the norm in the sense that $\|ab\| \leq \|a\|\|b\|$. The involution is idempotent ($(a^*)^* = a$) and satisfies $(ab)^* = b^*a^*$. In addition it is norm-preserving, and conjugate linear (and hence also continuous). Lastly, we have the identity $\|aa^*\| = \|a\|^2$ for all $a \in \mathcal{A}$. We say that a is Hermitian if $a^* = a$ and that a is positive is $a = bb^*$ for some $b \in \mathcal{A}$.

Definition 90. Let H be a Hilbert space and let $\mathcal{A} \subseteq \mathcal{B}(H)$ be a W^*-algebra[3] We assume that it contains the identity. Let u b a unit vector in H and define $\phi(T) = \langle Tu, u\rangle$ for $T \in \mathcal{A}$ (a pure state). Then we say that $(\mathcal{A}, \phi)$ is a W^*-probability space.

Example 91. (1) Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space and let $\mathcal{A} = L^\infty(\mathbf{P})$. We can think of $\mathcal{A}$ as a subalgebra of $\mathcal{B}(L^2(\mu))$ by the map $M : \mathcal{A} \to \mathcal{B}(L^2(\mu))$ by $f \to M_f$ where $M_f(g) = f \cdot g$. Then we leave it as an exercise to check that $\mathcal{A}$ is closed under weak operator topology.

Let $\mathbf{1}$ be the constant random variable 1. Then $\mathcal{A}$ is a unital algebra. Let $\phi(X) := \mathbf{E}[X] = \langle M_X\mathbf{1}, \mathbf{1}\rangle$ for $X \in \mathcal{A}$. This satisfies the definition of a n.c.p.s. Of course $(\mathcal{A}, \phi)$ is commutative and not of the main interest to us here, but this example explains the phrase "probability space" in the n.c.p.s. In this case there is a notion of positivity, and $\phi(X) \geq 0$ for $X \geq 0$.

(2) The example 87 is a W^*-probability space too. Subalgebras of $\mathcal{B}(H)$ that are closed in weak operator topology are also W^*-probability spaces.

Example 92 (The prime example - 1)**.** Let $\mathcal{M}_n$ be the space of $n \times n$ complex matrices. This is a W^*-algebra (it is $\mathcal{B}(H)$ where $H = \mathbb{C}^n$). If $\mathbf{e}_k$ is the k^{th} standard co-ordinate vector, then $\phi_k(T) = \langle T\mathbf{e}_k, \mathbf{e}_k\rangle$ defines a pure state on $\mathcal{M}_n$. Average over k to get a new positive linear functional $\hat{\text{tr}}_n(T) := n^{-1}\text{tr}(T)$. In other words, $\hat{\text{tr}}$ is the mean of the ESD of T.

Example 93 (The prime example - 2)**.** Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space and let $\mathcal{A} = L^\infty(\mathbf{P}) \otimes \mathcal{M}_n$ be the space of all random matrices $X = (X_{i,j})_{i,j\leq n}$ where $X_{i,j}$ are bounded, complex-valued random variables on Ω. Then, define $\phi_n(X) = \mathbf{E}[\hat{\text{tr}}(X)]$, the mean of the expected ESD. Then $(\mathcal{A}, \phi)$ is a ncps, in fact a C^* probability space.

Boundedness of entries is too restrictive as it does not even allow GUE matrices. Instead, we may consider the space $\mathcal{A}$ of random matrices $X = (X_{i,j})_{i,j\leq n}$ where $X_{i,j} \in \bigcap_{p<\infty} L^p(\Omega, \mathcal{F}, \mathbf{P})$. Define $\phi_n(X) = \mathbf{E}[\hat{\text{tr}}(X)]$ as before. This is a non-commutative probability space, although not a C^* probability space.

3. Distribution of non-commutative random variables and Free independence

Let $(\mathcal{A}, \phi)$ be a ncps. Any element $a \in \mathcal{A}$ are referred to as a *non-commutative random variable* and $\phi(a)$ as its *non-commutative expectation*.

Define the non-commutative moment function as $m_n[a_1, \ldots, a_n] = \phi(a_1a_2 \ldots a_n)$. As in the classical case, $m_n[\cdot]$ is multilinear, but not symmetric because of non-commutativity. If $a_1, \ldots, a_k$ are ncrvs on the same ncps, then the collection of all moments $\{m_n[a_{i_1}, \ldots, a_{i_n}] \ : \ 1 \leq i_1, \ldots, i_n \leq k\}$ is called the *joint distribution* of $a_1, \ldots, a_n$. For one variable, this is just the collection of moments $\phi(a^n)$, $n \geq 1$.

In classical probability, the distribution of a bounded real-valued random variable X can be recovered from its moments $\mathbf{E}[X^n]$, $n \geq 1$. However, for a complex-valued random variable (even

[3]This means that $\mathcal{A}$ is a C^*-subalgebra of $\mathcal{B}(H)$ and in addition is closed under weak operator topology. That is, if $T \in \mathcal{B}(H)$ and T_α is a net in $\mathcal{A}$ such that $\langle T_\alpha u, v\rangle \to \langle Tu, v\rangle$ for all $u, v \in H$, then $T \in \mathcal{A}$.

if bounded), one needs joint moments of the real and imaginary parts of X, or equivalently, that of X and $\overline{X}$, to recover the distribution of X. This motivates the following definition.

In a C^* or W^* probability space, the joint distribution of a and a^* is called the $*$-distribution of a. Observe that this involves specifying $\phi(P(a,a^*))$ for any non-commutative polynomial P (with complex coefficients) in two variables. Similarly one defines the $*$-distribution for more than one variable. As we remarked earlier, an element of a C^*-probability space is analogous to a complex valued random variable. For a probability measure on the complex plane, the moments , $\{\int z^n \mu(dz) \,:\, n \geq 1\}$ does not determine the measure. For example, any radially symmetric μ has $\int z^n \mu(dz) = 0$ for $n \geq 1$. Instead, one should specify the joint moments of the real and imaginary parts, or equivalently, $\int z^m \bar{z}^n \mu(dz)$. Thus, the $*$-distribution is what corresponds to the distribution of a complex-valued random variable.

In the special, but important case when a is *Hermitian* (to be considered analogous to real-valued random variables), the the $*$-distribution is the same as the distribution of a. Further, the following fact is important.

Proposition 94. *If a is a self-adjoint element of a C^*-probability space, then there exists a unique Borel probability measure μ_a on $\mathbb{R}$ such that $m_n(a) = \int x^n \mu_a(dx)$.*

Assuming the fact, by abuse of terminology we may refer to μ_a as the distribution of a. Thus, for self-adjoint elements of a C^*-probability space, the distribution refers to a p=classical probability measure on $\mathbb{R}$. Observe that this does not hold for non self-adjoint elements, or for joint distribution of several ncrvs.

PROOF OF 94. $m_n[a] = \phi(a^n)$. Let $P(a) = \sum_{k=0}^n c_k a^k$. By the positivity of ϕ, we see that

$$0 \leq \phi\left(P(a)P(a)^*\right) = \sum_{k,\ell=0}^{n} c_k \bar{c}_\ell \phi(a^{k+\ell})$$

which means that the infinite matrix $\left(\phi(a^{i+j})\right)_{i,j\geq 0}$ is a positive definite matrix. Therefore, there exists at least one probability measure μ with moments $\phi(a^n)$. However, by the boundedness of ϕ (we showed earlier that $\|\phi\| \leq 2$) and the properties of norm in a C^*-algebra, we see that $\phi(a^n) \leq 2\|a^n\| \leq 2\|a\|^n$. Thus, the moments of μ satisfy $\int x^n \mu(dx) \leq 2\|a\|^n$. This implies that μ must be compactly supported in $[-\|a\|, \|a\|]$. Since the moments of a compactly supported measure determines the measure, we also see that μ is unique. ■

Remark 95. Alternately, restrict to the example of a C^*-probability space given in 87. Then a is a self-adjoint operator on H and by the spectral theorem, there is a spectral measure of a at the vector u satisfying $\int x^n \mu(dx) = \langle a^n u, u\rangle = \phi(a^n)$. This is the μ we require. Since we know that the spectral measure is supported on the spectrum, and the spectrum is contained in $B(0, \|a\|)$ and the spectrum of a self-adjoint element is real, it follows that μ is supported on $[-\|a\|, \|a\|]$.

We now illustrate with an example.

Example 96. Let $H = \ell^2(\mathbb{N})$ and $\mathbf{e}_0 := (1,0,0,\ldots)$. Let $\mathcal{A} = \mathcal{B}(H)$ and $\phi(T) = \langle T\mathbf{e}_0, \mathbf{e}_0\rangle$. Now let $L(x_0,x_1,\ldots) = (x_1,x_2,\ldots)$ define the left-shift operator. Its adjoint is the right shift operator $L^*(x_0,x_1,\ldots) = (0,x_0,x_1,x_2,\ldots)$. It is easy to see that $\phi(L^n) = \phi(L^{*n}) = 1$ for $n = 0$ and equal to 0 for $n \geq 1$. Let $S = L + L^*$, a self-adjoint variable. Then $\phi(S^n) = \langle (L+L^*)^n \mathbf{e}_0, \mathbf{e}_0\rangle$. It is easy to check that the latter is zero for n odd and is equal to the Catalan number $C_k = \frac{1}{k+1}\binom{2k}{k}$ for $n = 2k$. These are the (classical) moments of the semicircle law supported on $[-2,2]$. Hence the non-commutative distribution of S is $\mu_{s.c}$.

If we define $\psi(T) = \langle T\mathbf{e}_1, \mathbf{e}_1\rangle$ where $\mathbf{e}_1 = (0,1,0,\ldots)$, can you find the distribution of S in the new ncps $(\mathcal{A}, \psi)$?

Example 97. Let $H = \ell^2(\mathbb{Z})$ and let $\mathbf{e}_0$ be the vector $\mathbf{e}_0(k) = \delta_{k,0}$. Then define the left shift operator L and its adjoint L^* (the right shift operator) in the obvious way. Again, $m_n(L) = m_n(L^*) = \delta_{n,0}$. Let $S = L + L^*$. Now, it is easy to check that $m_n(S)$ is $\binom{2k}{k}$ if $n = 2k$ and equal to zero if n is odd. These are the moments of the *arcsine* distribution with density $\frac{1}{\pi\sqrt{4-x^2}}$ on $[-2,2]$. Hence S has arc-sine distribution on $[-2,2]$.

4. Free independence and free cumulants

Independence is a central concept in probability theory. What is the analogue in the non-commutative setting? There is more than one possible notion of independence in non-commutative probability spaces, but there is a particular one that relates to random matrix theory.

Definition 98. Let $(\mathcal{A}, phi)$ be a ncps and let $\mathcal{A}_i$ be a collection of unital subalgebras of $\mathcal{A}$. We say that $\mathcal{A}_i$ are *freely independent* if $\phi(a_1 a_2 \ldots a_n) = 0$ for any $n \geq 1$ and any $a_i \in \mathcal{A}_{k_i}$ where $k_1 \neq k_2 \neq k_3 \ldots \neq k_n$ (consecutive elements come from different subalgebras). Elements $b_1, b_2, \ldots$ are said to be freely independent if the unital subalgebras generated by b_1, by b_2 etc., are freely independent.

Example 99. So far, classical probability spaces were special cases of non-commutative probability spaces. However, classically independent random variables are almost never freely independent. For example, if X, Y are random variables on $(\Omega, \mathcal{F}, \mathbf{P})$, for them to be freely independent we must have $\mathbf{E}[XYXY] = 0$ by this happens if and only if at least one of X and Y is degenerate at zero.

Example 100. We construct two non-trivial variables that are freely independent. Let $H = \mathbb{C}^2$ with orthonormal basis $\mathbf{e}_1, \mathbf{e}_2$. Then for $n \geq 2$ we define $H^{\otimes n}$ as a 2^n-dimensional space whose basis elements we denote by $\mathbf{e}_{i_1} \otimes \mathbf{e}_{i_2} \otimes \ldots \otimes \mathbf{e}_{i_n}$ where $i_1, \ldots, i_n \in \{1,2\}$. Let $H^{\otimes 0} = \mathbb{C}$ with orthonormal basis $\mathbf{e}_0 = 1$ (thus $\mathbf{e}_0 = \pm 1$).Then set $\mathcal{H} := \bigoplus_{n\geq 0} H^{\otimes n}$. $\mathcal{H}$. This is called the *full Fock space* corresponding to H and clearly $\{\mathbf{e}_{i_1} \otimes \mathbf{e}_{i_2} \otimes \ldots \otimes \mathbf{e}_{i_n} \ : \ n \geq 1,\ i_k = 1,2\} \cup \{\mathbf{e}_0\}$. It is evident how to generalize this definition for any Hilbert space H, not just $\mathbb{C}^2$.

Define the state $\phi(T) = \langle T\mathbf{e}_0, \mathbf{e}_0\rangle$ for $T \in \mathcal{B}(\mathcal{H})$. This is a C^*-probability space.

We define $L_1, L_2 \in \mathcal{B}(\mathcal{H})$ as follows. Let $L_1(\mathbf{e}_{i_1} \otimes \mathbf{e}_{i_2} \otimes \ldots \otimes \mathbf{e}_{i_n}) = \mathbf{e}_1 \otimes \mathbf{e}_{i_1} \otimes \ldots \otimes \mathbf{e}_{i_n}$ and extend linearly to $\mathcal{H}$. Likewise define L_2 using $\mathbf{e}_2$. The adjoints are given by

$$L_1^*(\mathbf{e}_{i_1} \otimes \mathbf{e}_{i_2} \otimes \ldots \otimes \mathbf{e}_{i_n}) = \begin{cases} \mathbf{e}_{i_2} \otimes \ldots \otimes \mathbf{e}_{i_n} & \text{if } i_1 = 1. \\ 0 & \text{otherwise} \end{cases}$$

and likewise for L_2^*. By the same logic as in example 97 it is easy to see that the non-commutative distribution of $T := L_1 + L_1^*$ and $S := L_2 + L_2^*$ are both semicircle distribution on $[-2,2]$. We now claim that they are freely independent. In fact the algebras $\mathcal{A}_1 = \langle L_1, L_1^* \rangle$ and $\mathcal{A}_2 = \langle L_2, L_2^* \rangle$ are freely independent.

We shall only consider the simplest non-trivial example and leave the full proof to the reader. Since $\phi(T) = \phi(S) = 0$, we must show that $\phi(TSTS) = 0$. For this, consider $\langle (L_1 + L_1^*)(L_2 + L_2^*)(L_1 + L_1^*)(L_2 + L_2^*)\mathbf{e}_0, \mathbf{e}_0 \rangle$, expand the product and observe that each term vanishes.

I have not written the next few sections fully or properly. Please refer to the books of Anderson, Guionnet and Zeitouni or the various lecture notes of Roland Speicher available on his homepage. If I find time, I shall write this stuff and post it here. For now, just a summary of what we covered in class.

Topics covered next:

(i) Free cumulants defined through free moments by a similar formula to the classical case, but summing only over non-crossing partitions.

(ii) Free independence is equivalent to vanishing of mixed cumulants.

(iii) Free central limit theorem - once the previous section is in place, this follows by copying word by word the proof of classical CLT using cumulants.

(iv) Relationship to random matrix theory - Random matrices $X = (X_{i,j})_{i,j \le n}$ where $X_{i,j}$ are random variables on $(\Omega, \mathcal{F}, \mathcal{P})$ can be considered also as elements of the non-commutative probability space as described in Example 93.

(v) The crucial connecting fact is that in many cases, large random matrices that are independent in the classical sense, are asymptotically (as the matrix size grows) freely independent. In particular this holds for the following pairs of random matrices.

   (a) Let D be a real diagonal whose ESD converges to a compactly supported measure on $\mathbb{R}$. Let $X^{(i)}$ be (scaled by $1/\sqrt{n}$) independent Wigner matrices with entries that have all moments. Then $D, X^{(1)}, X^{(2)}, \ldots$ are freely independent.

   (b) Let A_n and B_n be fixed sequences of real diagonal matrices. Let U_n be a Haar-distribute unitary matrix. Then A_n and $U_n^* B_n U_n$ are freely independent.

In all these cases, a particular consequence is that the ESD of the sum converges to the free convolution of the two individual limits.

5. Free cumulants

6. Free central limit theorem

We have said before that the semicircle plays a role in free probability very analogous to the Gaussian in classical probability. Now we prove a free version of the central limit theorem. Suppose a_k are freely independent and identically distributed elements in an algebra $\mathcal{A}$. Does $(a_1+\ldots+a_n)/\sqrt{n}$ converge in distribution to some variable? Firstly note that $\kappa_2[a_1+\ldots+a_n] = n\kappa_2[a_1]$ and hence $\sqrt{n}$ is the right scaling factor. Secondly, if we assume that $(a_1+\ldots+a_n)/\sqrt{n}$ does converge in distribution to some variable a, then for two freely independent copies a,b of this variable $a+b$ must have the same distribution as $\sqrt{2}a$. Just as we saw earlier for classical random variables, this forces the free cumulants to satisfy the relationship $2^{\frac{p}{2}}\kappa_p[a] = 2\kappa_p[a]$ which implies $\kappa_p[a] = 0$ for $p \neq 2$ which implies that a is a semicircular variable. Now we actually prove that the convergence does happen.

Theorem 101. *Let a, a_k be freely independent, identically distributed self-adjoint variables in a non-commutative probability space $(\mathcal{A},\phi)$ with $\kappa_2[a] > 0$. Then,*

$$\frac{a_1+\ldots+a_n-n\kappa_1[a]}{\sqrt{n}\sqrt{\kappa_2[a]}} \xrightarrow{d} \mu_{s.c.},$$

the standard semicircle law supported on $[-2,2]$.

PROOF. Without loss of generality assume that $\kappa_1[a] = 0$ and $\kappa_2[a] = 1$. The proof is word for word the same as we gave for classical CLT using cumulants (wisely we did not even change the notation for cumulants!). We conclude that $\kappa_p[S_n/\sqrt{n}] \to \delta_{p,2}$. The only non-commutative variable whose free cumulants are $\delta_{p,2}$ is the standard semicircle law. Hence the conclusion. ■

7. Random matrices and freeness

We have now seen Voiculescu's world of free probability with objects and theorems analogous to those in classical probability theory (we saw only a tiny sample of this. There is a free version of nearly everything, free Poisson, free Brownian motion, free Lévy process, free entropy, ... even free graduate students).

Apart from analogy, there is connection between the classical and free worlds, and that is provided by random matrix theory. Indeed, one of our motivations for introducing free probability theory is to explain the occurrence of semicircle law and other limit laws in random matrices, from a more conceptual algebraic framework. The essential connection is in the following theorem (and other such statements asserting free independence of classically independent large random matrices).

Theorem 102. *Consider $M_n(\mathbb{C}) \otimes L^\infty(\mathbf{P})$, the algebra of $n \times n$ random complex matrices with the state $\phi(A) = n^{-1}\mathbf{E}[tr(A)]$. Let $X_n = (X_{i,j})$ and $Y_n = (Y_{i,j})_{i,j \le n}$ be random Hermitian matrices on a common probability space taking values in $M_n(\mathbb{C})$. We consider two scenarios.*

(1) X_n and Y_n are Wigner matrices with $X_{1,1}$ and $X_{1,2}$ having exponential tails.

(2) $X_n = A_n$ and $Y_n = U_n B_n U_n^*$ where A_n, B_n are real diagonal matrices and U_n is a Haar distributed unitary matrix. We assume that the ESD of A_n and B_n are tight????

In either of these two situations, X_n and Y_n are asymptotically freely independent.

Now suppose X_n and Y_n are independent copies of GOE matrix. By properties of normals, $X_n + Y_n$ has the same distribution as $\sqrt{2}X_n$.

8. Spectrum of the sum of two matrices and free convolution

Let a, b be two self-adjoint, freely independent variables in a non-commutative probability space $(\mathcal{A}, \phi)$. Then, $\kappa_n[a+b] = \kappa_n[a] + \kappa_n[b]$. Hence the distribution of a and b determine the distribution of $a+b$. The procedure to find the distribution of $a+b$ is as follows.

(1) Let μ and ν be the distributions of a and b respectively. This means $\phi(a^n) = \int x^n \mu(dx)$ and $\phi(b^n) = \int x^n \nu(dx)$ for all n.

(2) From the moments $m_n(a) := \phi(a^n)$ and $m_n(b) = \phi(b^n)$ find the free cumulants $\kappa_n[a]$ and $\kappa_n[b]$. This can be done using the relations (**??**).

(3) Find $\kappa_n := \kappa_n[a] + \kappa_n[b]$ and insert into formulas (**??**) to find m_n.

(4) Find the measure θ whose moments are m_n. Then θ is the distribution of $a+b$.

An analogous procedure can be described in classical probability, to find the sum of two independent random variables using their cumulants. But there are also other useful techniques for dealing with sums of random variables such as the characteristic function (which is multiplicative under independence) or the logarithm of the characteristic function (which is additive). There are also such analytic objects associated to non-commutative random variables, which we describe now.

Let μ be a compactly supported on $\mathbb{R}$ with Stieltjes' transform $G_\mu(z) = \int (z-x)^{-1}\mu(dx)$ for the Stieltjes' transform of μ. From properties of Stieltjes transforms, we know that knowing G_a in a neighbourhood of ∞ one can recover all the moments of μ and hence recover μ itself. Further, G_μ is one-one in a neighbourhood of ∞ and has an analytic inverse K_μ defined in a neighbourhood of 0. Since $G_\mu(z) = z^{-1} + m_1 z^{-2} + \dots$ (where m_k are the moments of μ) for z close to ∞, we see that $K_\mu(w) = w^{-1} + R_\mu(w)$ for some analytic function R (defined in a neighbourhood of 0). R_μ is called the *R-transform* of μ.

Lemma 103. *$R_\mu(w) = \sum_{n=1}^\infty \kappa_n^\mu w^{n-1}$, where κ_n^μ are the free cumulants of μ.*

PROOF. Let $S(w) = \sum_{n=1}^\infty \kappa_n^\mu w^{n-1}$. We show that $G(w^{-1} + S(w)) = w$ for w close to 0 and this clearly implies that $S = R_\mu$. ■