LEAST SINGULAR VALUE, CIRCULAR LAW, AND LINDEBERG EXCHANGE

TERENCE TAO

1. The least singular value

This section[1] of the lecture notes is concerned with the behaviour of the *least singular value* $\sigma_n(M)$ of an $n \times n$ matrix M (or, more generally, the least non-trivial singular value $\sigma_p(M)$ of a $n \times p$ matrix with $p \leqslant n$). This quantity controls the invertibility of M. Indeed, M is invertible precisely when $\sigma_n(M)$ is non-zero, and the operator norm $\|M^{-1}\|_{\mathrm{op}}$ of M^{-1} is given by $1/\sigma_n(M)$. This quantity is also related to the *condition number* $\sigma_1(M)/\sigma_n(M) = \|M\|_{\mathrm{op}}\|M^{-1}\|_{\mathrm{op}}$ of M, which is of importance in numerical linear algebra. As we shall see in Section 2, the least singular value of M (and more generally, of the shifts $\frac{1}{\sqrt{n}}M - zI$ for complex z) will be of importance in rigorously establishing the *circular law* for iid random matrices M.

The least singular value

$$\sigma_n(M) = \inf_{\|x\|=1} \|Mx\|,$$

which sits at the "hard edge" of the spectrum, bears a superficial similarity to the operator norm

$$\|M\|_{\mathrm{op}} = \sigma_1(M) = \sup_{\|x\|=1} \|Mx\|$$

at the "soft edge" of the spectrum. For strongly rectangular matrices, the techniques that are useful to control the latter can also control the former, but the situation becomes more delicate for square matrices. For instance, the "epsilon net" method that is so useful for understanding the operator norm can control some "low entropy" portions of the infimum that arise from "structured" or "compressible" choices of x, but are not able to control the "generic" or "incompressible" choices of x, for which new arguments will be needed. Similarly, the moment method, can give the coarse order of magnitude (for instance, for rectangular matrices with $p = yn$ for $0 < y < 1$, it gives an upper bound of $(1 - \sqrt{y} + o(1))n$ for the singular value with high probability, thanks to the *Marcenko-Pastur law*), but again this method begins to break down for square matrices, although one can make some partial headway by considering *negative* moments such as $\operatorname{tr} M^{-2}$, though these are more difficult to compute than positive moments $\operatorname{tr} M^k$.

So one needs to supplement these existing methods with additional tools. It turns out that the key issue is to understand the distance between one of the n rows $X_1, \ldots, X_n \in \mathbb{C}^n$ of the matrix M, and the hyperplane spanned by the other $n-1$ rows. The reason for this is as follows. First suppose that $\sigma_n(M) = 0$, so that M is non-invertible, and

[1]The material in this section and the next is based on that in [39].

there is a linear dependence between the rows $X_1, \ldots, X_n$. Thus, one of the X_i will lie in the hyperplane spanned by the other rows, and so one of the distances mentioned above will vanish; in fact, one expects many of the n distances to vanish. Conversely, whenever one of these distances vanishes, one has a linear dependence, and so $\sigma_n(M) = 0$.

More generally, if the least singular value $\sigma_n(M)$ is small, one generically expects many of these n distances to be small also, and conversely. Thus, control of the least singular value is morally equivalent to control of the distance between a row X_i and the hyperplane spanned by the other rows. This latter quantity is basically the dot product of X_i with a unit normal n_i of this hyperplane.

When working with random matrices with jointly independent coefficients, we have the crucial property that the unit normal n_i (which depends on all the rows other than X_i) is *independent* of X_i, so even after conditioning n_i to be fixed, the entries of X_i remain independent. As such, the dot product $X_i \cdot n_i$ is a familiar scalar random walk, and can be controlled by a number of tools, most notably Littlewood-Offord theorems and the Berry-Esséen central limit theorem. As it turns out, this type of control works well except in some rare cases in which the normal n_i is "compressible" or otherwise highly structured; but epsilon-net arguments can be used to dispose of these cases[2].

These methods rely quite strongly on the joint independence on all the entries; it remains a challenge to extend them to more general settings. Even for Wigner matrices, the methods run into difficulty because of the non-independence of some of the entries (although it turns out one can understand the least singular value in such cases by rather different methods).

To simplify the exposition, we shall focus primarily here on just one specific ensemble of random matrices, the *Bernoulli ensemble* $M = (\xi_{ij})_{1 \leqslant i \leqslant p; 1 \leqslant j \leqslant n}$ of random sign matrices, where $\xi_{ij} = \pm 1$ are independent Bernoulli signs. However, the results can extend to more general classes of random matrices, with the main requirement being that the coefficients are jointly independent.

Throughout these notes, we use $X \ll Y$, $Y \gg X$, or $X = O(Y)$ to denote a bound of the form $|X| \leqslant CY$ for some absolute constant C; we take n as an asymptotic parameter, and write $X = o(Y)$ to denote a bound of the form $|X| \leqslant c(n)Y$ for some quantity $c(n)$ that goes to zero as n goes to infinity (holding other parameters fixed). If C or $c(n)$ needs to depend on additional parameters, we will denote this by subscripts, e.g. $X \gg_\delta Y$ denotes a bound of the form $X \geqslant c_\delta |Y|$ for some $c_\delta > 0$.

1.1. **The epsilon-net argument.** We begin by using the epsilon net argument to upper bound the operator norm:

Theorem 1.1 (Upper bound for operator norm)**.** *Let $M = (\xi_{ij})_{1 \leqslant i,j \leqslant n; 1 \leqslant j \leqslant n}$ be an $n \times n$ Bernoulli matrix. Then with exponentially high probability (i.e. $1 - O(e^{-cn})$ for some*

[2]This general strategy was first developed for the technically simpler singularity problem in [28], and then extended to the least singular value problem in [35].

$c > 0$), one has

$$\|M\|_{\mathrm{op}} = \sigma_1(M) \leqslant C\sqrt{n} \tag{1.1}$$

for some absolute constant C.

Proof. We use the "epsilon net argument". We write

$$\|M\|_{\mathrm{op}} = \sup_{x \in \mathbb{R}^n : \|x\| = 1} \|Mx\|.$$

Let Σ be a maximal 1/2-net of the unit sphere in $\mathbb{R}^n$, that is to say a maximal 1/2-separated subset of the sphere. Then we have

$$\|M\|_{\mathrm{op}} \leqslant \sup_{x \in \Sigma} \|Mx\| + \frac{1}{2}\|M\|_{\mathrm{op}}$$

and hence

$$\|M\|_{\mathrm{op}} \leqslant 2 \sup_{x \in \Sigma} \|Mx\|,$$

and so it suffices to show that

$$\mathbb{P}(\sup_{x \in \Sigma} \|Mx\| \leqslant \frac{C}{2}\sqrt{n})$$

is exponentially small in n. From the union bound, we can upper bound this by

$$\sum_{x \in \Sigma} \mathbb{P}(\|Mx\| \leqslant \frac{C}{2}\sqrt{n}).$$

The balls of radius 1/4 around each point in Σ are disjoint, and lie in the 1/4-neighbourhood of the sphere. From volume considerations we conclude that

$$|\Sigma| \leqslant O(1)^n \tag{1.2}$$

We set aside this bound as an "entropy cost" to be paid later, and focus on upper bounding, for each $x \in \Sigma$, the probability

$$\mathbb{P}(\|Mx\| \leqslant \frac{C}{2}\sqrt{n}).$$

If we let $Y_1, \ldots, Y_n \in \mathbb{R}^n$ be the rows of M, we can write this as

$$\mathbb{P}\left(\sum_{j=1}^n |Y_j \cdot x|^2 \leqslant \frac{C^2}{4} n\right).$$

By Markov's inequality, the only way that this event can hold is if we have

$$|Y_j \cdot x|^2 \leqslant \frac{C^2}{8}$$

for at least $n/2$ values of j. We do not know in advance what the set of j is for which this event holds. But the number of possible values of such sets of j is at most 2^n. Applying the union bound (and paying the entropy cost of 2^n) and using symmetry, we may thus bound the above probability by[3]

$$\leqslant 2^n \mathbb{P}(|Y_j \cdot x|^2 \leqslant \frac{C^2}{8} \text{ for } 1 \leqslant j \leqslant n/2).$$

[3]We will take n to be even for sake of notation, although it makes little essential difference.

Now observe that the random variables $Y_j \cdot x$ are independent, and so we can bound this expression by

$$\leqslant 2^n \mathbb{P}(|Y \cdot x| \leqslant \frac{C}{\sqrt{8}})^{n/2}$$

where $Y = (\xi_1, \ldots, \xi_n)$ is a random vector of iid Bernoulli signs. By the Chernoff inequality, this probability is $O(\exp(-cC^2))$ for some absolute constant $c > 0$. Taking C large enough, we obtain the claim. □

Now we use the same method to establish a lower bound in the rectangular case, first established in [30]:

Theorem 1.2 (Lower bound)**.** *Let $M = (\xi_{ij})_{1 \leqslant i \leqslant p; 1 \leqslant j \leqslant n}$ be an $n \times p$ Bernoulli matrix, where $1 \leqslant p \leqslant (1-\delta)n$ for some $\delta > 0$ (independent of n). Then with exponentially high probability, one has $\sigma_p(M) \gg_\delta \sqrt{n}$.*

To prove this theorem, we again use the "epsilon net argument". We write

$$\sigma_p(M) = \inf_{x \in \mathbb{R}^p : \|x\| = 1} \|Mx\|.$$

Let $\varepsilon > 0$ be a parameter to be chosen later. Let Σ be a maximal ε-net of the unit sphere in $\mathbb{R}^p$. Then we have

$$\sigma_p(M) \geqslant \inf_{x \in \Sigma} \|Mx\| - \varepsilon \|M\|_{\mathrm{op}}$$

and thus by (1.1), we have with overwhelming probability that

$$\sigma_p(M) \geqslant \inf_{x \in \Sigma} \|Mx\| - C\varepsilon\sqrt{n},$$

and so it suffices to show that

$$\mathbb{P}(\inf_{x \in \Sigma} \|Mx\| \leqslant 2C\varepsilon\sqrt{n})$$

is exponentially small in n. From the union bound, we can upper bound this by

$$\sum_{x \in \Sigma} \mathbb{P}(\|Mx\| \leqslant 2C\varepsilon\sqrt{n}).$$

The balls of radius $\varepsilon/2$ around each point in Σ are disjoint, and lie in the $\varepsilon/2$-neighbourhood of the sphere. From volume considerations we conclude that

$$|\Sigma| \leqslant O(1/\varepsilon)^p \leqslant O(1/\varepsilon)^{(1-\delta)n}. \tag{1.3}$$

We again set aside this bound as an "entropy cost" to be paid later, and focus on upper bounding, for each $x \in \Sigma$, the probability

$$\mathbb{P}(\|Mx\| \leqslant 2C\varepsilon\sqrt{n}).$$

If we let $Y_1, \ldots, Y_n \in \mathbb{R}^p$ be the rows of M, we can write this as

$$\mathbb{P}\left(\sum_{j=1}^n |Y_j \cdot x|^2 \leqslant 4C^2\varepsilon^2 n\right).$$

By Markov's inequality, the only way that this event can hold is if we have

$$|Y_j \cdot x|^2 \leqslant 8C^2\varepsilon^2/\delta$$

for at least $(1-\delta/2)n$ values of j. Applying the union bound as before (and paying the entropy cost of 2^n) and using symmetry, we may thus bound the above probability by

$$\leqslant 2^n \mathbb{P}(|Y_j \cdot x|^2 \leqslant 8C^2\varepsilon^2/\delta \text{ for } 1 \leqslant j \leqslant n/2).$$

Now observe that the random variables $Y_j \cdot x$ are independent, and so we can bound this expression by

$$\leqslant 2^n \mathbb{P}(|Y \cdot x| \leqslant \sqrt{8}C\varepsilon/\delta^{1/2})^{n/2}$$

where $Y = (\xi_1, \ldots, \xi_n)$ is a random vector of iid Bernoulli signs.

We write $x = (x_1, \ldots, x_n)$, so that $Y \cdot x$ is a random walk

$$Y \cdot x = \xi_1 x_1 + \cdots + \xi_n x_n.$$

To understand this walk, we apply (a slight variant) of the Berry-Esséen theorem:

Exercise 1.3. Show[4] that

$$\sup_t \mathbb{P}(|Y \cdot x - t| \leqslant r) \ll \frac{r}{\|x\|} + \frac{1}{\|x\|^3} \sum_{j=1}^n |x_j|^3$$

for any $r > 0$ and any non-zero x.

Conclude in particular that if

$$\sum_{j:|x_j| \leqslant \varepsilon} |x_j|^2 \geqslant \eta$$

for some $\eta > 0$, then

$$\sup_t \mathbb{P}(|Y \cdot x - t| \leqslant \sqrt{8}C\varepsilon) \ll_{\eta,\delta} \varepsilon.$$

(*Hint:* condition out all the x_j with $|x_j| > \varepsilon$.)

Let us temporarily call x *incompressible* if

$$\sum_{j:|x_j| \leqslant \varepsilon} |x_j|^2 \geqslant \eta$$

and *compressible* otherwise, where $\eta > 0$ is a parameter to be chosen later. If we only look at the incompressible elements of Σ, we can now bound

$$\mathbb{P}(\|Mx\| \leqslant 2C\varepsilon\sqrt{n}) \ll O_{\varepsilon,\delta}(\varepsilon)^n,$$

and comparing this against the entropy cost (1.3) we obtain an acceptable contribution for ε small enough (here we are crucially using the rectangular condition $p \leqslant (1-\delta)n$).

It remains to deal with the compressible vectors. Observe that such vectors lie within η of a *sparse* unit vector which is only supported in at most ε^{-2} positions. The ε-entropy of these sparse vectors (i.e. the number of balls of radius ε needed to cover this space) can easily be computed to be of polynomial size $O(n^{O_{\varepsilon,\eta}(1)})$ in n. Meanwhile, we have the following crude bound:

[4]Actually, for the purposes of this section, it would suffice to establish a weaker form of the Berry-Esséen theorem with $\sum_{j=1}^n |x_j|^3/\|x\|^3$ replaced by $(\sum_{j=1}^3 |x_j|^3/\|x\|^3)^c$ for any fixed $c > 0$. This can for instance be done using the Lindeberg exchange method, discussed in Section 3.

Exercise 1.4. For any unit vector x, show that

$$\mathbb{P}(|Y \cdot x| \leqslant \kappa) \leqslant 1 - \kappa$$

for $\kappa > 0$ small enough. (*Hint:* Use the *Paley-Zygmund inequality* $\mathbb{P}(Z \geqslant \theta \mathbb{E}Z) \geqslant (1-\theta)^2 \frac{(\mathbb{E}Z)^2}{\mathbb{E}(Z^2)}$, valid for any non-negative random variable Z of finite non-zero variance, and any $0 \leqslant \theta \leqslant 1$. Bounds on higher moments on $|Y \cdot x|$ can be obtained for instance using Hoeffding's inequality, or by direct computation.) Use this to show that

$$\mathbb{P}(\|Mx\| \leqslant 2C\eta\sqrt{n}) \ll \exp(-cn)$$

for all such x and ε sufficiently small, with $c > 0$ independent of ε and n.

Thus the compressible vectors give a net contribution of $O(n^{O_{\varepsilon,\eta}(1)}) \times \exp(-cn)$, which is acceptable. This concludes the proof of Theorem 1.2.

1.2. **Singularity probability.** Now we turn to square iid matrices. Before we investigate the size of the least singular value of M, we first tackle the easier problem of bounding the *singularity probability*

$$\mathbb{P}(\sigma_n(M) = 0),$$

i.e. the probability that M is not invertible. The problem of computing this probability exactly is still not completely settled. Since M is singular whenever the first two rows (say) are identical, we obtain a lower bound

$$\mathbb{P}(\sigma_n(M) = 0) \geqslant \frac{1}{2^n},$$

and it is conjectured that this bound is essentially tight in the sense that

$$\mathbb{P}(\sigma_n(M) = 0) = \left(\frac{1}{2} + o(1)\right)^n,$$

but this remains open; the best bound currently is [9], and gives

$$\mathbb{P}(\sigma_n(M) = 0) \leqslant \left(\frac{1}{\sqrt{2}} + o(1)\right)^n.$$

We will not prove this bound here, but content ourselves with a weaker bound, essentially due to Komlós [28]:

Proposition 1.5. *We have* $\mathbb{P}(\sigma_n(M) = 0) \ll 1/n^{1/2}$.

To show this, we need the following combinatorial fact, due to Erdös [14]:

Proposition 1.6 (Erdös Littlewood-Offord theorem)**.** *Let* $x = (x_1, \ldots, x_n)$ *be a vector with at least* k *nonzero entries, and let* $Y = (\xi_1, \ldots, \xi_n)$ *be a random vector of iid Bernoulli signs. Then* $\mathbb{P}(Y \cdot x = 0) \ll k^{-1/2}$.

Proof. By taking real and imaginary parts we may assume that x is real. By eliminating zero coefficients of x we may assume that $k = n$; reflecting we may then assume that all the x_i are positive. Observe that the set of $Y = (\xi_1, \ldots, \xi_n) \in \{-1, 1\}^n$ with $Y \cdot x = 0$

forms an *antichain*[5]. The product partial ordering on $\{-1,1\}^n$ is defined by requiring $(x_1,\dots,x_n) \leqslant (y_1,\dots,y_n)$ iff $x_i \leqslant y_i$ for all i. On the other hand, *Sperner's theorem* asserts that all anti-chains in $\{-1,1\}^n$ have cardinality at most $\binom{n}{\lfloor n/2 \rfloor}$. in $\{-1,1\}^n$ with the product partial ordering. The claim now easily follows from this theorem and Stirling's formula. □

Note that we also have the obvious bound

$$\mathbb{P}(Y \cdot x = 0) \leqslant 1/2 \tag{1.4}$$

for any non-zero x.

Now we prove the proposition. In analogy with the arguments of Section 1.1, we write

$$\mathbb{P}(\sigma_n(M) = 0) = \mathbb{P}(Mx = 0 \text{ for some nonzero } x \in \mathbb{C}^n)$$

(actually we can take $x \in \mathbb{R}^n$ since M is real). We divide into compressible and incompressible vectors as before, but our definition of compressibility and incompressibility is slightly different now. Also, one has to do a certain amount of technical maneuvering in order to preserve the crucial independence between rows and columns.

Namely, we pick an $\varepsilon > 0$ and call x *compressible* if it is supported on at most εn coordinates, and *incompressible* otherwise.

Let us first consider the contribution of the event that $Mx = 0$ for some nonzero compressible x. Pick an x with this property which is as sparse as possible, say k sparse for some $1 \leqslant k < \varepsilon n$. Let us temporarily fix k. By paying an entropy cost of $\lfloor \varepsilon n \rfloor \binom{n}{k}$, we may assume that it is the first k entries that are non-zero for some $1 \leqslant k \leqslant \varepsilon n$. This implies that the first k columns $Y_1,\dots,Y_k$ of M have a linear dependence given by x; by minimality, $Y_1,\dots,Y_{k-1}$ are linearly independent. Thus, x is uniquely determined (up to scalar multiples) by $Y_1,\dots,Y_k$. Furthermore, as the $n \times k$ matrix formed by $Y_1,\dots,Y_k$ has rank $k-1$, there is some $k \times k$ minor which already determines x up to constants; by paying another entropy cost of $\binom{n}{k}$, we may assume that it is the top left minor which does this. In particular, we can now use the first k rows $X_1,\dots,X_k$ to determine x up to constants. But the remaining $n-k$ rows are independent of $X_1,\dots,X_k$ and still need to be orthogonal to x; by Proposition 1.6, this happens with probability at most $O(\sqrt{k})^{-(n-k)}$, giving a total cost of

$$\sum_{1 \leqslant k \leqslant \varepsilon n} \binom{n}{k}^2 O(\sqrt{k})^{-(n-k)},$$

which by Stirling's formula is acceptable (in fact this gives an exponentially small contribution).

The same argument gives that the event that $y^* M = 0$ for some nonzero compressible y also has exponentially small probability. The only remaining event to control is the

[5]An *antichain* in a partially ordered set X is a subset S of X such that no two elements in S are comparable in the order.

event that $Mx = 0$ for some incompressible x, but that $Mz \neq 0$ and $y^*M \neq 0$ for all nonzero compressible z, y. Call this event E.

Since $Mx = 0$ for some incompressible x, we see that for at least εn values of $k \in \{1, \ldots, n\}$, the row X_k lies in the vector space V_k spanned by the remaining $n-1$ rows of M. Let E_k denote the event that E holds, and that X_k lies in V_k; then we see from double counting that

$$\mathbb{P}(E) \leqslant \frac{1}{\varepsilon n} \sum_{k=1}^{n} \mathbb{P}(E_k).$$

By symmetry, we thus have

$$\mathbb{P}(E) \leqslant \frac{1}{\varepsilon} \mathbb{P}(E_n).$$

To compute $\mathbb{P}(E_n)$, we freeze $X_1, \ldots, X_{n-1}$ consider a normal vector x to V_{n-1}; note that we can select x depending only on $X_1, \ldots, X_{n-1}$. We may assume that an incompressible normal vector exists, since otherwise the event E_n would be empty. We make the crucial observation that X_n is still independent of x. By Proposition 1.6 and (1.4), we thus see that the conditional probability that $X_n \cdot x = 0$, for fixed $X_1, \ldots, X_{n-1}$, is $O_\varepsilon(n^{-1/2})$. We thus see that $\mathbb{P}(E) \ll_\varepsilon 1/n^{1/2}$, and the claim follows.

Remark 1.7. Further progress has been made on this problem by a finer analysis of the concentration probability $\mathbb{P}(Y \cdot x = 0)$, and in particular in classifying those x for which this concentration probability is large (this is known as the *inverse Littlewood-Offord problem*). Important breakthroughs in this direction were made by Halász [24] (introducing Fourier-analytic tools) and by Kahn, Komlós, and Szemerédi [25] (introducing an efficient "swapping" argument). In [40] tools from additive combinatorics (such as Freiman's theorem) were introduced to obtain further improvements, leading eventually to the results from [9] mentioned earlier.

1.3. **Lower bound for the least singular value.** Now we return to the least singular value $\sigma_n(M)$ of an iid Bernoulli matrix, and establish a lower bound. Given that there are n singular values between 0 and $\sigma_1(M)$, which is typically of size $O(\sqrt{n})$, one expects the least singular value to be of size about $1/\sqrt{n}$ on the average. Another argument supporting this heuristic scomes from the following identity:

Exercise 1.8 (Negative second moment identity)**.** Let M be an invertible $n \times n$ matrix, let $X_1, \ldots, X_n$ be the rows of M, and let $R_1, \ldots, R_n$ be the columns of M^{-1}. For each $1 \leqslant i \leqslant n$, let V_i be the hyperplane spanned by all the rows $X_1, \ldots, X_n$ other than X_i. Show that $\|R_i\| = \operatorname{dist}(X_i, V_i)^{-1}$ and $\sum_{i=1}^n \sigma_i(M)^{-2} = \sum_{i=1}^n \operatorname{dist}(X_i, V_i)^2$.

From concentration of measure results such as Talagrand's inequality [38], we expect each $\operatorname{dist}(X_i, V_i)$ to be of size $O(1)$ on the average, which suggests that $\sum_{i=1}^n \sigma_i(M)^{-2} = O(n)$; this is consistent with the heuristic that the eigenvalues $\sigma_i(M)$ should be roughly evenly spaced in the interval $[0, 2\sqrt{n}]$ (so that $\sigma_{n-i}(M)$ should be about $(i+1)/\sqrt{n}$).

Now we give a rigorous lower bound:

Theorem 1.9 (Lower tail estimate for the least singular value)**.** *For any $\lambda > 0$, one has*

$$\mathbb{P}(\sigma_n(M) \leqslant \lambda/\sqrt{n}) \ll o_{\lambda\to 0}(1) + o_{n\to\infty;\lambda}(1)$$

where $o_{\lambda\to 0}(1)$ goes to zero as $\lambda \to 0$ uniformly in n, and $o_{n\to\infty;\lambda}(1)$ goes to zero as $n \to \infty$ for each fixed λ.

This is a weaker form of a result of Rudelson and Vershynin [36] (which obtains a bound of the form $O(\lambda) + O(c^n)$ for some $c < 1$), which builds upon the earlier works [35], [42], which obtained variants of the above result.

The scale $1/\sqrt{n}$ that we are working at here is too fine to use epsilon net arguments (unless one has a *lot* of control on the entropy, which can be obtained in some cases thanks to powerful inverse Littlewood-Offord theorems, but is difficult to obtain in general.) We can prove this theorem along similar lines to the arguments in the previous section; we sketch the method as follows. We can take λ to be small. We write the probability to be estimated as

$$\mathbb{P}(\|Mx\| \leqslant \lambda/\sqrt{n} \text{ for some unit vector } x \in \mathbb{C}^n).$$

We can assume that $\|M\|_{\mathrm{op}} \leqslant C\sqrt{n}$ for some absolute constant C, as the event that this fails has exponentially small probability.

We pick an $\varepsilon > 0$ (not depending on λ) to be chosen later. We call a unit vector $x \in \mathbb{C}^n$ *compressible* if x lies within a distance ε of a εn-sparse vector. Let us first dispose of the case in which $\|Mx\| \leqslant \lambda\sqrt{n}$ for some compressible x. By paying an entropy cost of $\binom{n}{\lfloor \varepsilon n\rfloor}$, we may assume that x is within ε of a vector y supported in the first $\lfloor \varepsilon n\rfloor$ coordinates. Using the operator norm bound on M and the triangle inequality, we conclude that

$$\|My\| \leqslant (\lambda + C\varepsilon)\sqrt{n}.$$

Since y has norm comparable to 1, this implies that the least singular value of the first $\lfloor \varepsilon n\rfloor$ columns of M is $O((\lambda+\varepsilon)\sqrt{n})$. But by Theorem 1.2, this occurs with probability $O(\exp(-cn))$ (if λ, ε are small enough). So the total probability of the compressible event is at most $\binom{n}{\lfloor \varepsilon n\rfloor} O(\exp(-cn))$, which is acceptable if ε is small enough.

Thus we may assume now that $\|Mx\| > \lambda/\sqrt{n}$ for all compressible unit vectors x; we may similarly assume that $\|y^* M\| > \lambda/\sqrt{n}$ for all compressible unit vectors y. Indeed, we may also assume that $\|y^* M_i\| > \lambda/\sqrt{n}$ for every i, where M_i is M with the i^{th} column removed.

The remaining case is if $\|Mx\| \leqslant \lambda/\sqrt{n}$ for some incompressible x. Let us call this event E. Write $x = (x_1, \ldots, x_n)$, and let $Y_1, \ldots, Y_n$ be the column of M, thus

$$\|x_1 Y_1 + \cdots + x_n Y_n\| \leqslant \lambda/\sqrt{n}.$$

Letting W_i be the subspace spanned by all the $Y_1, \ldots, Y_n$ except for Y_i, we conclude upon projecting to the orthogonal complement of W_i that

$$|x_i| \operatorname{dist}(Y_i, W_i) \leqslant \lambda/\sqrt{n}$$

for all i (compare with Exercise 1.8). On the other hand, since x is incompressible, we see that $|x_i| \geqslant \varepsilon/\sqrt{n}$ for at least εn values of i, and thus

$$\operatorname{dist}(Y_i, W_i) \leqslant \lambda/\varepsilon. \tag{1.5}$$

for at least εn values of i. If we let E_i be the event that E and (1.5) both hold, we thus have from double-counting that

$$\mathbb{P}(E) \leqslant \frac{1}{\varepsilon n} \sum_{i=1}^{n} \mathbb{P}(E_i)$$

and thus by symmetry

$$\mathbb{P}(E) \leqslant \frac{1}{\varepsilon} \mathbb{P}(E_n)$$

(say). However, if E_n holds, then setting y to be a unit normal vector to W_i (which is necessarily incompressible, by the hypothesis on M_i), we have

$$|Y_i \cdot y| \leqslant \lambda/\varepsilon.$$

Again, the crucial point is that Y_i and y are independent. The incompressibility of y, combined with a Berry-Esséen type theorem, then gives

Exercise 1.10. Show that

$$\mathbb{P}(|Y_i \cdot y| \leqslant \lambda/\varepsilon) \ll \varepsilon^2$$

(say) if λ is sufficiently small depending on ε, and n is sufficiently large depending on ε.

This gives a bound of $O(\varepsilon)$ for $\mathbb{P}(E)$ if λ is small enough depending on ε, and n is large enough; this gives the claim.

Remark 1.11. A variant of these arguments, based on inverse Littlewood-Offord theorems rather than the Berry-Esséen theorem, gives the variant estimate

$$\sigma_n(\frac{1}{\sqrt{n}} M_n - zI) \geqslant n^{-A} \tag{1.6}$$

with high probability for some $A > 0$, and any z of polynomial size in n. There are several results of this type, with overlapping ranges of generality (and various values of A) [23, 34, 41], and the exponent A is known to degrade if one has too few moment assumptions on the underlying random matrix M. This type of result (with an unspecified A) is important for the circular law, discussed in the next section.

1.4. **Upper bound for the least singular value.** One can complement the lower tail estimate with an upper tail estimate:

Theorem 1.12 (Upper tail estimate for the least singular value)**.** *For any $\lambda > 0$, one has*

$$\mathbb{P}(\sigma_n(M) \geqslant \lambda/\sqrt{n}) \ll o_{\lambda\to\infty}(1) + o_{n\to\infty;\lambda}(1). \tag{1.7}$$

We prove this using an argument of Rudelson and Vershynin [37]. Suppose that $\sigma_n(M) > \lambda/\sqrt{n}$, then

$$\|y^* M^{-1}\| \leqslant \sqrt{n}\|y\|/\lambda \tag{1.8}$$

for all y.

Next, let $X_1, \ldots, X_n$ be the rows of M, and let $R_1, \ldots, R_n$ be the columns of M^{-1}, thus $R_1, \ldots, R_n$ is a *dual basis*for $X_1, \ldots, X_n$. From (1.8) we have

$$\sum_{i=1}^n |y \cdot R_i|^2 \leqslant n\|y\|^2/\lambda^2.$$

We apply this with y equal to $X_n - \pi_n(X_n)$, where π_n is the orthogonal projection to the space V_{n-1} spanned by $X_1, \ldots, X_{n-1}$. On the one hand, we have

$$\|y\|^2 = \mathrm{dist}(X_n, V_{n-1})^2$$

and on the other hand we have for any $1 \leqslant i < n$ that

$$y \cdot R_i = -\pi_n(X_n) \cdot R_i = -X_n \cdot \pi_n(R_i)$$

and so

$$\sum_{i=1}^{n-1} |X_n \cdot \pi_n(R_i)|^2 \leqslant n\,\mathrm{dist}(X_n, V_{n-1})^2/\lambda^2. \tag{1.9}$$

If (1.9) holds, then $|X_n \cdot \pi_n(R_i)|^2 = O(\mathrm{dist}(X_n, V_{n-1})^2/\lambda^2)$ for at least half of the i, so the probability in (1.7) can be bounded by

$$\ll \frac{1}{n} \sum_{i=1}^{n-1} \mathbb{P}(|X_n \cdot \pi_n(R_i)|^2 = O(\mathrm{dist}(X_n, V_{n-1})^2/\lambda^2))$$

which by symmetry can be bounded by

$$\ll \mathbb{P}(|X_n \cdot \pi_n(R_1)|^2 = O(\mathrm{dist}(X_n, V_{n-1})^2/\lambda^2)).$$

Let $\varepsilon > 0$ be a small quantity to be chosen later. From Talagrand's inequality [38] we know that $\mathrm{dist}(X_n, V_{n-1}) = O_\varepsilon(1)$ with probability $1 - O(\varepsilon)$, so we obtain a bound of

$$\ll \mathbb{P}(X_n \cdot \pi_n(R_1) = O_\varepsilon(1/\lambda)) + O(\varepsilon).$$

Now a key point is that the vectors $\pi_n(R_1), \ldots, \pi_n(R_{n-1})$ depend only on $X_1, \ldots, X_{n-1}$ and not on X_n; indeed, they are the dual basis for $X_1, \ldots, X_{n-1}$ in V_{n-1}. Thus, after conditioning $X_1, \ldots, X_{n-1}$ and thus $\pi_n(R_1)$ to be fixed, X_n is still a Bernoulli random vector. Applying a Berry-Esséen inequality, we obtain a bound of $O(\varepsilon)$ for the conditional probability that $X_n \cdot \pi_n(R_1) = O_\varepsilon(1/\lambda)$ for λ sufficiently small depending on ε, unless $\pi_n(R_1)$ is compressible (in the sense that, say, it is within ε of an εn-sparse vector). But this latter possibility can be controlled (with exponentially small probability) by the same type of arguments as before; we omit the details.

1.5. **Asymptotic for the least singular value.** The distribution of singular values of a Gaussian random matrix can be computed explicitly. In particular, if M is a real Gaussian matrix (with all entries iid with distribution $N(0,1)_{\mathbb{R}}$), it was shown in [12] that $\sqrt{n}\sigma_n(M)$ converges in distribution to the distribution $\mu_E := \frac{1+\sqrt{x}}{2\sqrt{x}} e^{-x/2-\sqrt{x}}\,dx$ as $n \to \infty$. It turns out that this result can be extended to other ensembles with the same mean and variance. In particular, we have the following result from [43]:

Theorem 1.13. *If M is an iid Bernoulli matrix, then $\sqrt{n}\sigma_n(M)$ also converges in distribution to μ_E as $n \to \infty$. (In fact there is a polynomial rate of convergence.)*

This should be compared with Theorems 1.9, 1.12, which show that $\sqrt{n}\sigma_n(M)$ have a tight sequence of distributions in $(0, +\infty)$. The arguments from [43] thus provide an alternate proof of these two theorems. The same result in fact holds for all iid ensembles obeying a finite moment condition.

The arguments used to prove Theorem 1.13 do not establish the limit μ_E directly, but instead use the result of [12] as a black box, focusing instead on establishing the *universality* of the limiting distribution of $\sqrt{n}\sigma_n(M)$, and in particular that this limiting distribution is the same whether one has a Bernoulli ensemble or a Gaussian ensemble.

The arguments are somewhat technical and we will not present them in full here, but instead give a sketch of the key ideas.

In previous sections we have already seen the close relationship between the least singular value $\sigma_n(M)$, and the distances $\mathrm{dist}(X_i, V_i)$ between a row X_i of M and the hyperplane V_i spanned by the other $n-1$ rows. It is not hard to use the above machinery to show that as $n \to \infty$, $\mathrm{dist}(X_i, V_i)$ converges in distribution to the absolute value $|N(0,1)_{\mathbb{R}}|$ of a Gaussian regardless of the underlying distribution of the coefficients of M (i.e. it is asymptotically universal). The basic point is that one can write $\mathrm{dist}(X_i, V_i)$ as $|X_i \cdot n_i|$ where n_i is a unit normal of V_i (we will assume here that M is non-singular, which by previous arguments is true asymptotically almost surely). The previous machinery lets us show that n_i is incompressible with high probability, and then claim then follows from the Berry-Esséen theorem.

Unfortunately, despite the presence of suggestive relationships such as Exercise 1.8, the asymptotic universality of the distances $\mathrm{dist}(X_i, V_i)$ does not directly imply asymptotic universality of the least singular value. However, it turns out that one can obtain a higher-dimensional version of the universality of the scalar quantities $\mathrm{dist}(X_i, V_i)$, as follows. For any small k (say, $1 \leqslant k \leqslant n^c$ for some small $c > 0$) and any distinct $i_1, \ldots, i_k \in \{1, \ldots, n\}$, a modification of the above argument shows that the covariance matrix

$$(\pi(X_{i_a}) \cdot \pi(X_{i_b}))_{1 \leqslant a,b \leqslant k} \tag{1.10}$$

of the orthogonal projections $\pi(X_{i_1}), \ldots, \pi(X_{i_k})$ of the k rows $X_{i_1}, \ldots, X_{i_k}$ to the complement $V_{i_1,\ldots,i_k}^{\perp}$ of the space $V_{i_1,\ldots,i_k}$ spanned by the other $n-k$ rows of M, is also universal, converging in distribution to the covariance[6] matrix $(G_a \cdot G_b)_{1 \leqslant a,b \leqslant k}$ of k iid Gaussians $G_a \equiv N(0,1)_{\mathbb{R}}$ (note that the convergence of $\mathrm{dist}(X_i, V_i)$ to $|N(0,1)_{\mathbb{R}}|$ is the $k = 1$ case of this claim). The key point is that one can show that the complement $V_{i_1,\ldots,i_k}^{\perp}$ is usually "incompressible" in a certain technical sense, which implies that the projections $\pi(X_{i_a})$ behave like iid Gaussians on that projection thanks to a multidimensional Berry-Esséen theorem.

On the other hand, the covariance matrix (1.10) is closely related to the inverse matrix M^{-1}:

Exercise 1.14. Show that (1.10) is also equal to A^*A, where A is the $n \times k$ matrix formed from the $i_1, \ldots, i_k$ columns of M^{-1}.

[6]These covariance matrix distributions are also known as *Wishart distributions*.

In particular, this shows that the singular values of k randomly selected columns of M^{-1} have a universal distribution.

Recall our goal is to show that $\sqrt{n}\sigma_n(M)$ has an asymptotically universal distribution, which is equivalent to asking that $\frac{1}{\sqrt{n}}\|M^{-1}\|_{\mathrm{op}}$ has an asymptotically universal distribution. The goal is then to extract the operator norm of M^{-1} from looking at a random $n \times k$ minor B of this matrix. This comes from the following application of the second moment method:

Exercise 1.15. Let A be an $n \times n$ matrix with columns $R_1, \ldots, R_n$, and let B be the $n \times k$ matrix formed by taking k of the columns $R_1, \ldots, R_n$ at random. Show that

$$\mathbb{E}\|A^*A - \frac{n}{k}B^*B\|_F^2 \leqslant \frac{n}{k}\sum_{k=1}^{n}\|R_k\|^4,$$

where $\|\|_F$ is the Frobenius norm $\|A\|_F := \mathrm{tr}(A^*A)^{1/2}$.

Recall from Exercise 1.8 that $\|R_k\| = 1/\operatorname{dist}(X_k, V_k)$, so we expect each $\|R_k\|$ to have magnitude about $O(1)$. As such, we expect $\sigma_1((M^{-1})^*(M^{-1})) = \sigma_n(M)^{-2}$ to differ by $O(n^2/k)$ from $\frac{n}{k}\sigma_1(B^*B) = \frac{n}{k}\sigma_1(B)^2$. In principle, this gives us asymptotic universality on $\sqrt{n}\sigma_n(M)$ from the already established universality of B.

There is one technical obstacle remaining, however: while we know that each $\operatorname{dist}(X_k, V_k)$ is distributed like a Gaussian, so that each individual R_k is going to be of size $O(1)$ with reasonably good probability, in order for the above exercise to be useful, one needs to bound *all* of the R_k *simultaneously* with high probability. A naive application of the union bound leads to terrible results here. Fortunately, there is a strong correlation between the R_k: they tend to be large together or small together, or equivalently that the distances $\operatorname{dist}(X_k, V_k)$ tend to be small together or large together. Here is one indication of this:

Lemma 1.16. *For any $1 \leqslant k < i \leqslant n$, one has*

$$\operatorname{dist}(X_i, V_i) \geqslant \frac{\|\pi_i(X_i)\|}{1 + \sum_{j=1}^{k}\frac{\|\pi_i(X_j)\|}{\|\pi_i(X_i)\|\operatorname{dist}(X_j, V_j)}},$$

where π_i is the orthogonal projection onto the space spanned by $X_1, \ldots, X_k, X_i$.

Proof. We may relabel so that $i = k + 1$; then projecting everything by π_i we may assume that $n = k + 1$. Our goal is now to show that

$$\operatorname{dist}(X_n, V_{n-1}) \geqslant \frac{\|X_n\|}{1 + \sum_{j=1}^{n-1}\frac{\|X_j\|}{\|X_n\|\operatorname{dist}(X_j, V_j)}}.$$

Recall that $R_1, \ldots, R_n$ is a dual basis to $X_1, \ldots, X_n$. This implies in particular that

$$x = \sum_{j=1}^{n}(x \cdot X_j)R_j$$

for any vector x; applying this to X_n we obtain

$$X_n = \|X_n\|^2 R_n + \sum_{j=1}^{n-1} (X_j \cdot X_n) R_j$$

and hence by the triangle inequality

$$\|X_n\|^2 \|R_n\| \leqslant \|X_n\| + \sum_{j=1}^{n-1} \|X_j\| \|X_n\| \|R_j\|.$$

Using the fact that $\|R_j\| = 1/\operatorname{dist}(X_j, R_j)$, the claim follows. □

In practice, once k gets moderately large (e.g. $k = n^c$ for some small $c > 0$), one can control the expressions $\|\pi_i(X_j)\|$ appearing here by Talagrand's inequality [38], and so this inequality tells us that once $\operatorname{dist}(X_j, V_j)$ is bounded away from zero for $j = 1, \ldots, k$, it is bounded away from zero for all other k also. This turns out to be enough to get enough uniform control on the R_j to make Exercise 1.15 useful, and ultimately to complete the proof of Theorem 1.13.

2. The circular law

In this section, we leave the realm of self-adjoint matrix ensembles, such as Wigner random matrices, and consider instead the simplest examples of non-self-adjoint ensembles, namely the iid matrix ensembles.

The basic result in this area is

Theorem 2.1 (Circular law)**.** *Let M_n be an $n \times n$ iid matrix, whose entries ξ_{ij}, $1 \leqslant i, j \leqslant n$ are iid with a fixed (complex) distribution $\xi_{ij} \equiv \xi$ of mean zero and variance one. Then the spectral measure $\mu_{\frac{1}{\sqrt{n}} M_n}$ converges both in probability and almost surely to the circular law $\mu_{\mathrm{circ}} := \frac{1}{\pi} 1_{|x|^2+|y|^2 \leqslant 1}\, dxdy$, where x, y are the real and imaginary coordinates of the complex plane.*

This theorem has a long history; it is analogous to the semicircular law, but the non-Hermitian nature of the matrices makes the spectrum so unstable that key techniques that are used in the semicircular case, such as truncation and the moment method, no longer work; significant new ideas are required. In the case of random Gaussian matrices, this result was established by Mehta [31] (in the complex case) and by Edelman [11] (in the real case), using the explicit formulae for the joint distribution of eigenvalues available in these cases. In 1984, Girko [21] laid out a general strategy for establishing the result for non-gaussian matrices, which formed the base of all future work on the subject; however, a key ingredient in the argument, namely a bound on the least singular value of shifts $\frac{1}{\sqrt{n}} M_n - zI$, was not fully justified at the time. A rigorous proof of the circular law was then established by Bai [3], assuming additional moment and boundedness conditions on the individual entries. These additional conditions were then slowly removed in a sequence of papers [23, 22, 34, 41], with the last moment

condition being removed in [46]. There have since been several further works in which the circular law was extended to other ensembles [1, 2, 4, 5, 6, 10, 32, 33, 49], including several models in which the entries are no longer jointly independent; see also the surveys [7, 41].

2.1. **Spectral instability.** One of the basic difficulties present in the non-Hermitian case is *spectral instability*: small perturbations in a large matrix can lead to large fluctuations in the spectrum. In order for any sort of analytic technique to be effective, this type of instability must somehow be precluded.

The canonical example of spectral instability comes from perturbing the right shift matrix

$$U_0 := \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 0 \end{pmatrix}$$

to the matrix

$$U_\varepsilon := \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \varepsilon & 0 & 0 & \dots & 0 \end{pmatrix}$$

for some $\varepsilon > 0$.

The matrix U_0 is nilpotent: $U_0^n = 0$. Its characteristic polynomial is $(-\lambda)^n$, and it thus has n repeated eigenvalues at the origin. In contrast, U_ε obeys the equation $U_\varepsilon^n = \varepsilon I$, its characteristic polynomial is $(-\lambda)^n - \varepsilon(-1)^n$, and it thus has n eigenvalues at the n^{th} roots $\varepsilon^{1/n} e^{2\pi i j/n}$, $j = 0, \dots, n-1$ of ε. Thus, even for exponentially small values of ε, say $\varepsilon = 2^{-n}$, the eigenvalues for U_ε can be quite far from the eigenvalues of U_0, and can wander all over the unit disk. This is in sharp contrast with the Hermitian case, where eigenvalue inequalities such as the Weyl inequalities or Wielandt-Hoffman inequalities ensure stability of the spectrum.

One can explain the problem in terms of *pseudospectrum*[7]. The only spectrum of U is at the origin, so the resolvents $(U - zI)^{-1}$ of U are finite for all non-zero z. However, while these resolvents are finite, they can be extremely large. Indeed, from the nilpotent nature of U_0 we have the Neumann series

$$(U_0 - zI)^{-1} = -\frac{1}{z} - \frac{U_0}{z^2} - \dots - \frac{U_0^{n-1}}{z^n}$$

so for $|z| < 1$ we see that the resolvent has size roughly $|z|^{-n}$, which is exponentially large in the interior of the unit disk. This exponentially large size of resolvent is consistent with the exponential instability of the spectrum:

[7]The *pseudospectrum* of an operator T is the set of complex numbers z for which the operator norm $\|(T - zI)^{-1}\|_{\text{op}}$ is either infinite, or larger than a fixed threshold $1/\varepsilon$. See [48] for further discussion.

Exercise 2.2. Let M be a square matrix, and let z be a complex number. Show that $\|(M - zI)^{-1}\|_{\mathrm{op}} \geqslant R$ if and only if there exists a perturbation $M + E$ of M with $\|E\|_{\mathrm{op}} \leqslant 1/R$ such that $M + E$ has z as an eigenvalue.

This already hints strongly that if one wants to rigorously prove control on the spectrum of M near z, one needs some sort of upper bound on $\|(M - zI)^{-1}\|_{\mathrm{op}}$, or equivalently one needs some sort of lower bound on the least singular value $\sigma_n(M - zI)$ of $M - zI$.

Without such a bound, though, the instability precludes the direct use of the *truncation method*, which was so useful in the Hermitian case. In particular, there is no obvious way to reduce the proof of the circular law to the case of bounded coefficients, in contrast to the semicircular law for Wigner matrices where this reduction follows easily from the Wielandt-Hoffman inequality. Instead, we must continue working with unbounded random variables throughout the argument (unless, of course, one makes an additional decay hypothesis, such as assuming certain moments are finite; this helps explain the presence of such moment conditions in many papers on the circular law).

2.2. **Incompleteness of the moment method.** In the Hermitian case, the moments

$$\frac{1}{n} \operatorname{tr}(\frac{1}{\sqrt{n}} M)^k = \int_{\mathbb{R}} x^k \, d\mu_{\frac{1}{\sqrt{n}} M_n}(x)$$

of a matrix can be used (in principle) to understand the distribution $\mu_{\frac{1}{\sqrt{n}} M_n}$ completely (at least, when the measure $\mu_{\frac{1}{\sqrt{n}} M_n}$ has sufficient decay at infinity. This is ultimately because the space of real polynomials $P(x)$ is dense in various function spaces (the Weierstrass approximation theorem).

In the non-Hermitian case, the spectral measure $\mu_{\frac{1}{\sqrt{n}} M_n}$ is now supported on the complex plane rather than the real line. One still has the formula

$$\frac{1}{n} \operatorname{tr}(\frac{1}{\sqrt{n}} M)^k = \int_{\mathbb{C}} z^k \, d\mu_{\frac{1}{\sqrt{n}} M_n}(z)$$

but it is much less useful now, because the space of complex polynomials $P(z)$ no longer has any good density properties[8]. In particular, the moments no longer uniquely determine the spectral measure.

This can be illustrated with the shift examples given above. It is easy to see that U and U_ε have vanishing moments up to $(n-1)^{\text{th}}$ order, i.e.

$$\frac{1}{n} \operatorname{tr} \left(\frac{1}{\sqrt{n}} U \right)^k = \frac{1}{n} \operatorname{tr} \left(\frac{1}{\sqrt{n}} U_\varepsilon \right)^k = 0$$

for $k = 1, \ldots, n-1$. Thus we have

$$\int_{\mathbb{C}} z^k \, d\mu_{\frac{1}{\sqrt{n}} U}(z) = \int_{\mathbb{C}} z^k \, d\mu_{\frac{1}{\sqrt{n}} U_\varepsilon}(z) = 0$$

[8]For instance, the uniform closure of the space of polynomials on the unit disk is not the space of continuous functions, but rather the space of holomorphic functions that are continuous on the closed unit disk.

for $k = 1, \ldots, n-1$. Despite this enormous number of matching moments, the spectral measures $\mu_{\frac{1}{\sqrt{n}}U}$ and $\mu_{\frac{1}{\sqrt{n}}U_\varepsilon}$ are dramatically different; the former is a Dirac mass at the origin, while the latter can be arbitrarily close to the unit circle. Indeed, even if we set *all* moments equal to zero,

$$\int_{\mathbb{C}} z^k \, d\mu = 0$$

for $k = 1, 2, \ldots$, then there are an uncountable number of possible (continuous) probability measures that could still be the (asymptotic) spectral measure μ: for instance, any measure which is rotationally symmetric around the origin would obey these conditions.

If one could somehow control the mixed moments

$$\int_{\mathbb{C}} z^k \overline{z}^l \, d\mu_{\frac{1}{\sqrt{n}}M_n}(z) = \frac{1}{n}\sum_{j=1}^n \left(\frac{1}{\sqrt{n}}\lambda_j(M_n)\right)^k \left(\frac{1}{\sqrt{n}}\overline{\lambda_j}(M_n)\right)^l$$

of the spectral measure, then this problem would be resolved, and one could use the moment method to reconstruct the spectral measure accurately. However, there does not appear to be any easy way to compute this quantity; the obvious guess of $\frac{1}{n}\operatorname{tr}(\frac{1}{\sqrt{n}}M_n)^k(\frac{1}{\sqrt{n}}M_n^*)^l$ works when the matrix M_n is *normal*, as M_n and M_n^* then share the same basis of eigenvectors, but generically one does not expect these matrices to be normal.

Remark 2.3. The failure of the moment method to control the spectral measure is consistent with the instability of spectral measure with respect to perturbations, because moments are stable with respect to perturbations.

Exercise 2.4. Let $k \geqslant 1$ be an integer, and let M_n be an iid matrix whose entries have a fixed distribution ξ with mean zero, variance 1, and with k^{th} moment finite. Show that $\frac{1}{n}\operatorname{tr}(\frac{1}{\sqrt{n}}M_n)^k$ converges to zero as $n \to \infty$ in expectation, in probability, and in the almost sure sense. Thus we see that $\int_{\mathbb{C}} z^k \, d\mu_{\frac{1}{\sqrt{n}M_n}}(z)$ converges to zero in these three senses also. This is of course consistent with the circular law, but does not come close to establishing that law, for the reasons given above.

Remark 2.5. The failure of the moment method also shows that methods of free probability do not work directly. For instance, observe that for fixed ε, U_0 and U_ε (in the noncommutative probability space $(\mathrm{Mat}_n(\mathbb{C}), \frac{1}{n}\operatorname{tr})$) both converge in the sense of $*$-moments as $n \to \infty$ to that of the right shift operator on $\ell^2(\mathbb{Z})$ (with the trace $\tau(T) = \langle e_0, Te_0\rangle$, with e_0 being the Kronecker delta at 0); but the spectral measures of U_0 and U_ε are different. Thus the spectral measure cannot be read off directly from the free probability limit.

2.3. **The logarithmic potential.** With the moment method out of consideration, attention naturally turns to the Stieltjes transform

$$s_n(z) = \frac{1}{n}\operatorname{tr}\left(\frac{1}{\sqrt{n}}M_n - zI\right)^{-1} = \int_{\mathbb{C}} \frac{d\mu_{\frac{1}{\sqrt{n}}M_n}(w)}{w - z}.$$

This is a rational function on the complex plane. Its relationship with the spectral measure is as follows:

Exercise 2.6. Show that

$$\mu_{\frac{1}{\sqrt{n}}M_n} = \frac{1}{\pi}\partial_{\bar{z}} s_n(z)$$

in the sense of distributions, where

$$\partial_{\bar{z}} := \frac{1}{2}(\frac{\partial}{\partial x} + i\frac{\partial}{\partial y})$$

is the Cauchy-Riemann operator, and the Stieltjes transform is interpreted distributionally in a principal value sense.

One can control the Stieltjes transform quite effectively away from the origin. Indeed, for iid matrices with subgaussian entries, one can show that the operator norm of $\frac{1}{\sqrt{n}}M_n$ is $1+o(1)$ almost surely; this, combined with (2.4) and Laurent expansion, tells us that $s_n(z)$ almost surely converges to $-1/z$ locally uniformly in the region $\{z : |z| > 1\}$, and that the spectral measure $\mu_{\frac{1}{\sqrt{n}}M_n}$ converges almost surely to zero in this region (which can of course also be deduced directly from the operator norm bound). This is of course consistent with the circular law, but is not sufficient to prove it (for instance, the above information is also consistent with the scenario in which the spectral measure collapses towards the origin). One also needs to control the Stieltjes transform inside the disk $\{z : |z| \leqslant 1\}$ in order to fully control the spectral measure.

For this, many existing methods for controlling the Stieltjes transform are not particularly effective in this non-Hermitian setting (mainly because of the spectral instability, and also because of the lack of analyticity in the interior of the spectrum). Instead, one proceeds by relating the Stieltjes transform to the *logarithmic potential*

$$f_n(z) := \int_{\mathbb{C}} \log|w-z| d\mu_{\frac{1}{\sqrt{n}}M_n}(w).$$

It is easy to see that $s_n(z)$ is essentially the (distributional) gradient of $f_n(z)$:

$$s_n(z) = \left(-\frac{\partial}{\partial x} + i\frac{\partial}{\partial y}\right) f_n(z),$$

and thus g_n is related to the spectral measure by the distributional formula[9]

$$\mu_{\frac{1}{\sqrt{n}}M_n} = \frac{1}{2\pi}\Delta f_n \tag{2.1}$$

where $\Delta := \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$ is the Laplacian.

The following basic result relates the logarithmic potential to probabilistic notions of convergence.

Theorem 2.7 (Logarithmic potential continuity theorem)**.** *Let M_n be a sequence of random matrices, and suppose that for almost every complex number z, $f_n(z)$ converges almost surely (resp. in probability) to*

$$f(z) := \int_{\mathbb{C}} \log|z-w| d\mu(w)$$

[9]This formula just reflects the fact that $\frac{1}{2\pi}\log|z|$ is the *Newtonian potential* in two dimensions.

for some probability measure μ. Then $\mu_{\frac{1}{\sqrt{n}}M_n}$ converges almost surely (resp. in probability) to μ in the vague topology.

Proof. We prove the almost sure version of this theorem, and leave the convergence in probability version as an exercise.

On any bounded set K in the complex plane, the functions $\log|\cdot - w|$ lie in $L^2(K)$ uniformly in w. From Minkowski's integral inequality, we conclude that the f_n and f are uniformly bounded in $L^2(K)$. On the other hand, almost surely the f_n converge pointwise to f. From the dominated convergence theorem this implies that $\min(|f_n - f|, M)$ converges in $L^1(K)$ to zero for any M; using the uniform bound in $L^2(K)$ to compare $\min(|f_n - f|, M)$ with $|f_n - f|$ and then sending $M \to \infty$, we conclude that f_n converges to f in $L^1(K)$. In particular, f_n converges to f in the sense of distributions; taking distributional Laplacians using (2.1) we obtain the claim. □

Exercise 2.8. Establish the convergence in probability version of Theorem 2.7.

Thus, the task of establishing the circular law then reduces to showing, for almost every z, that the logarithmic potential $f_n(z)$ converges (in probability or almost surely) to the right limit $f(z)$.

Observe that the logarithmic potential

$$f_n(z) = \frac{1}{n}\sum_{j=1}^{n} \log\left|\frac{\lambda_j(M_n)}{\sqrt{n}} - z\right|$$

can be rewritten as a log-determinant:

$$f_n(z) = \frac{1}{n}\log\left|\det(\frac{1}{\sqrt{n}}M_n - zI)\right|.$$

To compute this determinant, we recall that the determinant of a matrix A is not only the product of its eigenvalues, but also has a magnitude equal to the product of its *singular* values:

$$|\det A| = \prod_{j=1}^{n}\sigma_j(A) = \prod_{j=1}^{n}\lambda_j(A^*A)^{1/2}$$

and thus

$$f_n(z) = \frac{1}{2}\int_0^\infty \log x \; d\nu_{n,z}(x)$$

where $d\nu_{n,z}$ is the spectral measure of the matrix $\left(\frac{1}{\sqrt{n}}M_n - zI\right)^*\left(\frac{1}{\sqrt{n}}M_n - zI\right)$.

The advantage of working with this spectral measure, as opposed to the original spectral measure $\mu_{\frac{1}{\sqrt{n}}M_n}$, is that the matrix $(\frac{1}{\sqrt{n}}M_n - zI)^*(\frac{1}{\sqrt{n}}M_n - zI)$ is *self-adjoint*, and so methods such as the moment method or free probability can now be safely applied to compute the limiting spectral distribution. Indeed, Girko [21] established that for almost every z, $\nu_{n,z}$ converged both in probability and almost surely to an explicit

(though slightly complicated) limiting measure ν_z in the vague topology. Formally, this implied that $f_n(z)$ would converge pointwise (almost surely and in probability) to

$$\frac{1}{2}\int_0^\infty \log x \; d\nu_z(x).$$

A lengthy but straightforward computation then showed that this expression was indeed the logarithmic potential $f(z)$ of the circular measure μ_{circ}, so that the circular law would then follow from the logarithmic potential continuity theorem.

Unfortunately, the vague convergence of $\nu_{n,z}$ to ν_z only allows one to deduce the convergence of $\int_0^\infty F(x)\; d\nu_{n,z}$ to $\int_0^\infty F(x)\; d\nu_z$ for F continuous and compactly supported. The logarithm function $\log x$ has singularities at zero and at infinity, and so the convergence

$$\int_0^\infty \log x \; d\nu_{n,z}(x) \to \int_0^\infty \log x \; d\nu_z(x)$$

can fail if the spectral measure $\nu_{n,z}$ sends too much of its mass to zero or to infinity.

The latter scenario can be easily excluded, either by using operator norm bounds on M_n (when one has enough moment conditions) or even just the Frobenius norm bounds (which require no moment conditions beyond the unit variance). The real difficulty is with preventing mass from going to the origin.

The approach of Bai [3] proceeded in two steps. Firstly, he established a polynomial lower bound

$$\sigma_n\left(\frac{1}{\sqrt{n}}M_n - zI\right) \geqslant n^{-C}$$

asymptotically almost surely for the least singular value of $\frac{1}{\sqrt{n}}M_n - zI$. This has the effect of capping off the $\log x$ integrand to be of size $O(\log n)$. Next, by using Stieltjes transform methods, the convergence of $\nu_{n,z}$ to ν_z in an appropriate metric (e.g. the Levi distance metric) was shown to be polynomially fast, so that the distance decayed like $O(n^{-c})$ for some $c > 0$. The $O(n^{-c})$ gain can safely absorb the $O(\log n)$ loss, and this leads to a proof of the circular law assuming enough boundedness and continuity hypotheses to ensure the least singular value bound and the convergence rate. This basic paradigm was also followed by later works [23, 34, 41], with the main new ingredient being the advances in the understanding of the least singular value (Section 1).

Unfortunately, to get the polynomial convergence rate, one needs some moment conditions beyond the zero mean and unit variance rate (e.g. finite $2 + \eta^{\mathrm{th}}$ moment for some $\eta > 0$). In [46] the additional tool of the Talagrand concentration inequality [38] was used to eliminate the need for the polynomial convergence. Intuitively, the point is that only a small fraction of the singular values of $\frac{1}{\sqrt{n}}M_n - zI$ are going to be as small as n^{-c}; most will be much larger than this, and so the $O(\log n)$ bound is only going to be needed for a small fraction of the measure. To make this rigorous, it turns out to be convenient to work with a slightly different formula for the determinant magnitude $|\det(A)|$ of a square matrix than the product of the eigenvalues, namely the base-times-height

formula

$$|\det(A)| = \prod_{j=1}^{n} \mathrm{dist}(X_j, V_j)$$

where X_j is the j^{th} row and V_j is the span of $X_1, \ldots, X_{j-1}$.

Exercise 2.9. Establish the inequality

$$\prod_{j=n+1-m}^{n} \sigma_j(A) \leqslant \prod_{j=1}^{m} \mathrm{dist}(X_j, V_j) \leqslant \prod_{j=1}^{m} \sigma_j(A)$$

for any $1 \leqslant m \leqslant n$. (*Hint:* the middle product is the product of the singular values of the first m rows of A, and so one should try to use the Cauchy interlacing inequality for singular values.) Thus we see that $\mathrm{dist}(X_j, V_j)$ is a variant of $\sigma_j(A)$.

The least singular value bounds, translated in this language (with $A := \frac{1}{\sqrt{n}} M_n - zI$), tell us that $\mathrm{dist}(X_j, V_j) \geqslant n^{-C}$ with high probability; this lets ignore the most dangerous values of j, namely those j that are equal to $n - O(n^{0.99})$ (say). For low values of j, say $j \leqslant (1-\delta)n$ for some small δ, one can use the moment method to get a good lower bound for the distances and the singular values, to the extent that the logarithmic singularity of $\log x$ no longer causes difficulty in this regime; the limit of this contribution can then be seen by moment method or Stieltjes transform techniques to be *universal* in the sense that it does not depend on the precise distribution of the components of M_n. In the medium regime $(1-\delta)n < j < n - n^{0.99}$, one can use Talagrand's inequality [38] to show that $\mathrm{dist}(X_j, V_j)$ has magnitude about $\sqrt{n-j}$, giving rise to a net contribution to $f_n(z)$ of the form $\frac{1}{n} \sum_{(1-\delta)n<j<n-n^{0.99}} O(\log \sqrt{n-j})$, which is small. Putting all this together, one can show that $f_n(z)$ converges to a universal limit as $n \to \infty$ (independent of the component distributions); see [46] for details. As a consequence, once the circular law is established for one class of iid matrices, such as the complex Gaussian random matrix ensemble, it automatically holds for all other ensembles also.

3. The Lindeberg exchange method

The *central limit theorem* asserts that if $X_1, X_2, \ldots$ are a sequence of iid real random variables of mean zero and variance one, then the normalised averages means

$$\frac{X_1 + \cdots + X_n}{\sqrt{n}}$$

converge in distribution to the normal distribution $N(0,1)$ as $n \to \infty$, thus

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{X_1 + \cdots + X_n}{\sqrt{n}} \leqslant t\right) = P(G \leqslant t)$$

for any $t \in \mathbb{R}$, where G is a normally distributed random variable of mean zero. Equivalently, if $F : \mathbb{R} \to \mathbb{R}$ is any smooth, compactly supported function, then one has

$$\mathbb{E}F\left(\frac{X_1 + \cdots + X_n}{\sqrt{n}}\right) = \mathbb{E}F(G) + o(1) \tag{3.1}$$

as $n \to \infty$, where $o(1)$ denotes a quantity that goes to zero as $n \to \infty$.

Exercise 3.1. Why are these two claims equivalent?

One consequence of the central limit theorem is that the statistics $\mathbb{E}F(\frac{X_1+\cdots+X_n}{\sqrt{n}})$ are asymptotically *universal* in the sense that they do not depend on the precise distribution of the individual random variables. In particular, if $Y_1, \ldots, Y_n$ are another sequence of iid random variables, with a different distribution than the $X_1, \ldots, X_n$, then we have the asymptotic universality property

$$\mathbb{E}F\left(\frac{X_1 + \cdots + X_n}{\sqrt{n}}\right) = \mathbb{E}F\left(\frac{Y_1 + \cdots + Y_n}{\sqrt{n}}\right) + o(1) \tag{3.2}$$

as $n \to \infty$.

The central limit theorem is traditionally proven by Fourier analytic methods, and then the universality (3.2) obtained as a corollary. However, one could also argue in the reverse direction, establishing the central limit theorem (3.1) as a consequence. Indeed, in 1922 Lindeberg [29] established the central limit theorem by establishing the following three claims:

(1) If $Y_1, Y_2, \ldots$ were an iid sequence of *gaussian* random variables of mean zero and variance one, then

$$\mathbb{E}F\left(\frac{Y_1 + \cdots + Y_n}{\sqrt{n}}\right) = \mathbb{E}F(G) + o(1).$$

(2) If $X_1, X_2, \ldots$ were an iid sequence of random variables mean zero and variance one, and finite third moment $\mathbb{E}|X_i|^3 < \infty$, then

$$\mathbb{E}F\left(\frac{X_1 + \cdots + X_n}{\sqrt{n}}\right) = \mathbb{E}F\left(\frac{Y_1 + \cdots + Y_n}{\sqrt{n}}\right) + o(1). \tag{3.3}$$

(3) If the central limit theorem (3.1) was true for iid random variables of mean zero, variance one, and finite third moment, it would also be true without the finite third moment condition.

Clearly, the central limit theorem is immediate from the above three claims.

The first claim is an easy computation: since the sum of any finite number of independent gaussian random variables is still gaussian, the variable $\frac{Y_1+\cdots+Y_n}{\sqrt{n}}$ is also gaussian. Since this variable has mean zero and variance one, the claim follows (indeed we don't even have the $o(1)$ error in this case).

The third claim is also an easy consequence of a standard *truncation argument*. Suppose that $X_1, X_2, \ldots$ were iid copies of a random variable X of mean zero and variance one, but possibly infinite third moment. For any cutoff parameter M, the truncation $X1_{|X|\leqslant M}$ will have finite third moment; it might not have mean zero and variance one, but from the dominated convergence theorem we see that the mean of this random variable goes to zero, and the variance goes to one, as $M \to \infty$. Similarly, the tail $X1_{|X|>M}$ has variance going to zero as $M \to \infty$. By adjusting the former random

variable slightly to normalise the mean and variance, we thus see that for any $\varepsilon > 0$, one can obtain a decomposition

$$X = X'_\varepsilon + X''_\varepsilon$$

where X'_ε is a random variable of mean zero, variance one, and finite third moment, while X''_ε is a random variable of variance at most ε (and necessarily of mean zero, by linearity of expectation). The random variables X'_ε and X''_ε may be coupled to each other, but this will not concern us. We can therefore split each X_i as $X_i = X'_{i,\varepsilon} + X''_{i,\varepsilon}$, where $X'_{1,\varepsilon}, \ldots, X'_{n,\varepsilon}$ are iid copies of X'_ε, and $X''_{1,\varepsilon}, \ldots, X''_{n,\varepsilon}$ are iid copies of X''_ε. We then have

$$\frac{X_1 + \cdots + X_n}{\sqrt{n}} = \frac{X'_{1,\varepsilon} + \cdots + X'_{n,\varepsilon}}{\sqrt{n}} + \frac{X''_{1,\varepsilon} + \cdots + X''_{n,\varepsilon}}{\sqrt{n}}.$$

If the central limit theorem holds in the case of finite third moment, then the random variable $\frac{X'_{1,\varepsilon}+\cdots+X'_{n,\varepsilon}}{\sqrt{n}}$ converges in distribution to G. Meanwhile, the random variable $\frac{X''_{1,\varepsilon}+\cdots+X''_{n,\varepsilon}}{\sqrt{n}}$ has mean zero and variance at most ε. From this, we see that (3.1) holds up to an error of $O(\varepsilon)$; sending ε to zero, we obtain the claim.

The heart of the Lindeberg argument is in the second claim. Without loss of generality we may take the tuple $(X_1, \ldots, X_n)$ to be independent of the tuple $(Y_1, \ldots, Y_n)$. The idea is not to swap the $X_1, \ldots, X_n$ with the $Y_1, \ldots, Y_n$ all at once, but instead to swap them one at a time. Indeed, one can write the difference

$$\mathbb{E}F\left(\frac{X_1 + \cdots + X_n}{\sqrt{n}}\right) - \mathbb{E}F\left(\frac{Y_1 + \cdots + Y_n}{\sqrt{n}}\right)$$

as a telescoping series formed by the sum of the n terms

$$\mathbb{E}F\left(\frac{Y_1 + \cdots + Y_{i-1} + X_i + X_{i+1} + \cdots + X_n}{\sqrt{n}}\right) - \mathbb{E}F\left(\frac{Y_1 + \cdots + Y_{i-1} + Y_i + \cdots + Y_n}{\sqrt{n}}\right) \tag{3.4}$$

for $i = 1, \ldots, n$, each of which represents a single swap from X_i to Y_i. Thus, to prove (3.3), it will suffice to show that each of the terms (3.4) is of size $o(1/n)$ (uniformly in i).

We can write (3.4) as

$$\mathbb{E}F(Z_i + \frac{1}{\sqrt{n}}X_i) - \mathbb{E}F(Z_i + \frac{1}{\sqrt{n}}Y_i)$$

where Z_i is the random variable

$$Z_i := \frac{Y_1 + \cdots + Y_{i-1} + X_{i+1} + \cdots + X_n}{\sqrt{n}}.$$

The key point here is that Z_i is independent of both X_i and Y_i. To exploit this, we use the smoothness of F to perform a Taylor expansion

$$F(Z_i + \frac{1}{\sqrt{n}}X_i) = F(Z_i) + \frac{1}{\sqrt{n}}X_i F'(Z_i) + \frac{1}{2n}X_i^2 F''(Z_i) + O\left(\frac{1}{n^{3/2}}|X_i|^3\right)$$

and hence on taking expectations (and using the finite third moment hypothesis and independence of X_i from Z_i)

$$\mathbb{E}F(Z_i + \frac{1}{\sqrt{n}}X_i) = \mathbb{E}F(Z_i) + \frac{1}{\sqrt{n}}\mathbb{E}X_i\mathbb{E}F'(Z_i) + \mathbb{E}\frac{1}{2n}\mathbb{E}X_i^2\mathbb{E}F''(Z_i) + O\left(\frac{1}{n^{3/2}}\right).$$

Similarly

$$\mathbb{E}F(Z_i + \frac{1}{\sqrt{n}}Y_i) = \mathbb{E}F(Z_i) + \frac{1}{\sqrt{n}}\mathbb{E}Y_i\mathbb{E}F'(Z_i) + \mathbb{E}\frac{1}{2n}\mathbb{E}Y_i^2\mathbb{E}F''(Z_i) + O\left(\frac{1}{n^{3/2}}\right).$$

Now observe that as X_i and Y_i both have mean zero and variance one, their first two moments match: $\mathbb{E}X_i = \mathbb{E}Y_i$ and $\mathbb{E}X_i^2 = \mathbb{E}Y_i^2$. As such, the first three terms of the above two right-hand sides match, and thus (3.4) is bounded by $O(n^{-3/2})$, which is $o(1/n)$ as required. Note how important it was that we had two matching moments; if we only had one matching moment, then the bound for (3.4) would only be $O(1/n)$, which is not sufficient. (And of course the central limit theorem would fail as stated if we did not correctly normalise the variance.) One can therefore think of the central limit theorem as a *two moment theorem*, asserting that the asymptotic behaviour of the statistic $\mathbb{E}F(\frac{X_1+\cdots+X_n}{\sqrt{n}})$ for iid random variables $X_1, \ldots, X_n$ depends only on the first two moments of the X_i.

Exercise 3.2 (Lindeberg central limit theorem)**.** Let $m_1, m_2, \ldots$ be a sequence of natural numbers. For each n, let $X_{n,1}, \ldots, X_{n,m_n}$ be a collection of independent real random variables of mean zero and total variance one, thus

$$\mathbb{E}X_{n,i} = 0 \forall 1 \leqslant i \leqslant m_n$$

and

$$\sum_{i=1}^{m_n} \mathbb{E}X_{n,i}^2 = 1.$$

Suppose also that for every fixed $\delta > 0$ (not depending on n), one has

$$\sum_{i=1}^{m_n} \mathbb{E}X_{n,i}^2 1_{|X_{n,i}|\geqslant\delta} = o(1)$$

as $n \to \infty$. Show that the random variables $\sum_{i=1}^{m_n} X_{n,i}$ converge in distribution as $n \to \infty$ to a normal variable of mean zero and variance one.

Exercise 3.3 (Martingale central limit theorem)**.** Let $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \mathcal{F}_2 \subset \ldots$ be an increasing collection of σ-algebras in the ambient sample space. For each n, let X_n be a real random variable that is measurable with respect to $\mathcal{F}_n$, with conditional mean and variance one with respect to $\mathcal{F}_0$, thus

$$\mathbb{E}(X_n|\mathcal{F}_{n-1}) = 0$$

and

$$\mathbb{E}(X_n^2|\mathcal{F}_{n-1}) = 1$$

almost surely. Assume also the bounded third moment hypothesis

$$\mathbb{E}(|X_n|^3|\mathcal{F}_{n-1}) \leqslant C$$

almost surely for all n and some finite C independent of n. Show that the random variables $\frac{X_1+\cdots+X_n}{\sqrt{n}}$ converge in distribution to a normal variable of mean zero and variance

one. (One can relax the hypotheses on this martingale central limit theorem substantially, but we will not explore this here.)

Exercise 3.4 (Weak Berry-Esséen theorem)**.** Let $X_1, \ldots, X_n$ be an iid sequence of real random variables of mean zero and variance 1, and bounded third moment: $\mathbb{E}|X_i|^3 = O(1)$. Let G be a gaussian random variable of mean zero and variance one. Using the Lindeberg exchange method, show that

$$\mathbb{P}(\frac{X_1 + \cdots + X_n}{\sqrt{n}} \leqslant t) = \mathbb{P}(G \leqslant t) + O(n^{-1/8})$$

for any $t \in \mathbb{R}$. (The full Berry-Esséen theorem improves the error term to $O(n^{-1/2})$, but this is difficult to establish purely from the Lindeberg exchange method; one needs alternate methods, such as Fourier-based methods or Stein's method, to recover this improved gain.) What happens if one assumes more matching moments between the X_i and G, such as matching third moment $\mathbb{E}X_i^3 = \mathbb{E}G^3$ or matching fourth moment $\mathbb{E}X_i^4 = \mathbb{E}G^4$?

One can think of the Lindeberg method as having the schematic form of the telescoping identity

$$X^n - Y^n = \sum_{i=1}^n Y^{i-1}(X-Y)X^{n-i}$$

which is valid in any (possibly non-commutative) ring. It breaks the symmetry of the indices $1, \ldots, n$ of the random variables $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$, by performing the swaps in a specified order. A more symmetric variant of the Lindeberg method was introduced recently by Knowles and Yin [26], and has the schematic form of the fundamental theorem of calculus identity

$$X^n - Y^n = \int_0^1 \sum_{i=1}^n ((1-\theta)X + \theta Y)^{i-1}(X-Y)((1-\theta)X + \theta Y)^{n-i}\, d\theta,$$

which is valid in any (possibly non-commutative) real algebra, and can be established by computing the θ derivative of $((1-\theta)X + \theta Y)^n$. We illustrate this method by giving a slightly different proof of (3.3). We may again assume that the $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ are independent of each other. We introduce auxiliary random variables $t_1, \ldots, t_n$, drawn uniformly at random from $[0,1]$, independently of each other and of the $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$. For any $0 \leqslant \theta \leqslant 1$ and $1 \leqslant i \leqslant n$, let $X_i^{(\theta)}$ denote the random variable

$$X_i^{(\theta)} = 1_{t_i \leqslant \theta} X_i + 1_{t_i > \theta} Y_i,$$

thus for instance $X_i^{(0)} = Y_i$ and $X_i^{(1)} = X_i$ almost surely. One then has the following key derivative computation:

Exercise 3.5. With the notation and assumptions as above, show that

$$\frac{d}{d\theta}\mathbb{E}F\left(\frac{X_1^{(\theta)} + \cdots + X_n^{(\theta)}}{\sqrt{n}}\right) = \sum_{i=1}^n \mathbb{E}F\left(Z_i^{(\theta)} + \frac{1}{\sqrt{n}}X_i\right) - \mathbb{E}F\left(Z_i^{(\theta)} + \frac{1}{\sqrt{n}}Y_i\right) \quad (3.5)$$

where

$$Z_i^{(\theta)} := \frac{X_1^{(\theta)} + \cdots + X_{i-1}^{(\theta)} + X_{i+1}^{(\theta)} + \cdots + X_n^{(\theta)}}{\sqrt{n}}.$$

In particular, the derivative on the left-hand side of (3.5) exists and depends continuously on θ.

From the above exercise and the fundamental theorem of calculus, we can write the left-hand side of (3.3) as

$$\int_0^1 \sum_{i=1}^n \mathbb{E}F(Z_i^{(\theta)} + \frac{1}{\sqrt{n}}X_i) - \mathbb{E}F(Z_i^{(\theta)} + \frac{1}{\sqrt{n}}Y_i)\ d\theta.$$

Repeating the Taylor expansion argument used in the original Lindeberg method, we see that

$$\mathbb{E}F\left(Z_i^{(\theta)} + \frac{1}{\sqrt{n}}X_i\right) - \mathbb{E}F\left(Z_i^{(\theta)} + \frac{1}{\sqrt{n}}Y_i\right) = O(n^{-3/2})$$

thus giving an alternate proof of (3.3).

Remark 3.6. In this particular instance, the Knowles-Yin version of the Lindeberg method did not offer any significant advantages over the original Lindeberg method. However, the more symmetric form of the former is useful in some random matrix theory applications due to the additional cancellations this induces. See [26] for an example of this.

The Lindeberg exchange method was first employed to study statistics of random matrices in [44]; the method was subsequently simplified in [27]. To illustrate this method, we will (for simplicity of notation) restrict our attention to real Wigner matrices $M_n = \frac{1}{\sqrt{n}}(\xi_{ij})_{1 \leqslant i,j \leqslant n}$, where ξ_{ij} are real random variables of mean zero and variance either one (if $i \neq j$) or two (if $i = j$), with the symmetry condition $\xi_{ij} = \xi_{ji}$ for all $1 \leqslant i, j \leqslant n$, and with the upper-triangular entries ξ_{ij}, $1 \leqslant i \leqslant j \leqslant n$ jointly independent. For technical reasons we will also assume that the ξ_{ij} are uniformly subgaussian, thus there exist constants $C, c > 0$ such that

$$\mathbb{P}(|\xi_{ij}| \geqslant t) \leqslant C \exp(-ct^2)$$

for all $t > 0$. In particular, the k^{th} moments of the ξ_{ij} will be bounded for any fixed natural number k. These hypotheses can be relaxed somewhat, but we will not aim for the most general results here. One could easily consider Hermitian Wigner matrices instead of real Wigner matrices, after some minor modifications to the notation and discussion below (e.g. replacing the Gaussian Orthogonal Ensemble (GOE) with the Gaussian Unitary Ensemble (GUE)). The $\frac{1}{\sqrt{n}}$ term appearing in the definition of M_n is a standard normalising factor to ensure that the spectrum of M_n (usually) stays bounded (indeed it will almost surely obey the famous *semicircle law* restricting the spectrum almost completely to the interval $[-2, 2]$).

The most well known example of a real Wigner matrix ensemble is the *Gaussian Orthogonal Ensemble* (GOE), in which the ξ_{ij} are all real gaussian random variables (with the mean and variance prescribed as above). This is a highly symmetric ensemble, being invariant with respect to conjugation by any element of the orthogonal group $O(n)$, and as such many of the sought-after spectral statistics of GOE matrices can be computed explicitly (or at least asymptotically) by direct computation of certain multidimensional integrals (which, in the case of GOE, eventually reduces to computing the integrals of

certain Pfaffian kernels). We will not discuss these explicit computations further here, but view them as the analogue to the simple gaussian computation used to establish Claim 1 of Lindeberg's proof of the central limit theorem. Instead, we will focus more on the analogue of Claim 2 - using an exchange method to compare statistics for GOE to statistics for other real Wigner matrices.

We can write a Wigner matrix $\frac{1}{\sqrt{n}}(\xi_{ij})_{1 \leqslant i,j \leqslant n}$ as a sum

$$M_n = \frac{1}{\sqrt{n}} \sum_{(i,j) \in \Delta} \xi_{ij} E_{ij}$$

where Δ is the upper triangle

$$\Delta := \{(i,j) : 1 \leqslant i \leqslant j \leqslant n\}$$

and the real symmetric matrix E_{ij} is defined to equal $E_{ij} = e_i^T e_j + e_j^T e_i$ for $i < j$, and $E_{ii} = e_i^T e^i$ in the diagonal case $i = j$, where $e_1, \ldots, e_n$ are the standard basis of $\mathbb{R}^n$ (viewed as column vectors). Meanwhile, a GOE matrix G_n can be written as

$$G_n = \sum_{(i,j) \in \Delta} \eta_{ij} E_{ij}$$

where $\eta_{ij}, (i,j) \in \Delta$ are jointly independent gaussian real random variables, of mean zero and variance either one (for $i \neq j$) or two (for $i = j$). For any natural number m, we say that M_n and G_n have m *matching moments* if we have

$$\mathbb{E}\xi_{ij}^k = \mathbb{E}\eta_{ij}^k$$

for all $k = 1, \ldots, m$. Thus for instance, we always have two matching moments, by our definition of a Wigner matrix.

If $S(M_n)$ is any (deterministic) statistic of a Wigner matrix M_n, we say that we have a m *moment theorem* for the statistic $S(M_n)$ if one has

$$S(M_n) - S(G_n) = o(1) \tag{3.6}$$

whenever M_n and G_n have m matching moments. In most applications m will be very small, either equal to 2, 3, or 4.

One can prove moment theorems using the Lindeberg exchange method. For instance, consider statistics of the form $S(M_n) = \mathbb{E}F(M_n)$, where $F : V \to \mathbb{C}$ is a bounded measurable function on the space V of real symmetric matrices. Then we can write the left-hand side of (3.6) as the sum of $|\Delta| = \frac{n(n+1)}{2}$ terms of the form

$$\mathbb{E}F\left(M_{n,i,j} + \frac{1}{\sqrt{n}}\xi_{ij}E_{ij}\right) - \mathbb{E}F\left(M_{n,i,j} + \frac{1}{\sqrt{n}}\eta_{ij}E_{ij}\right)$$

where for each $(i,j) \in \Delta$, $M_{n,i,j}$ is the real symmetric matrix

$$M_{n,i,j} := \sum_{(i',j') < (i,j)} \frac{1}{\sqrt{n}}\eta_{i'j'}E_{i'j'} + \sum_{(i',j') > (i,j)} \frac{1}{\sqrt{n}}\xi_{i'j'}E_{i'j'}$$

where one imposes some arbitrary ordering $<$ on Δ (e.g. the lexicographical ordering). Strictly speaking, the $M_{n,i,j}$ are not Wigner matrices, because their ij entry vanishes and thus has zero variance, but their behaviour turns out to be almost identical to that

of a Wigner matrix (being a rank one or rank two perturbation of such a matrix). Thus, to prove an m moment theorem for $\mathbb{E}F(M_n)$, it thus suffices by the triangle inequality to establish a bound of the form

$$\mathbb{E}F\left(M_{n,i,j} + \frac{1}{\sqrt{n}}\xi_{ij}E_{ij}\right) - \mathbb{E}F\left(M_{n,i,j} + \frac{1}{\sqrt{n}}\eta_{ij}E_{ij}\right) = o(n^{-2}) \tag{3.7}$$

for all $(i,j) \in \Delta$, whenever ξ_{ij} and η_{ij} have m matching moments. (For the diagonal entries $i = j$, a bound of $o(n^{-1})$ will in fact suffice, as the number of such entries is n rather than $O(n^2)$.)

As with the Lindeberg proof of the central limit theorem, one can establish (3.7) for various statistics F by performing a Taylor expansion with remainder. To illustrate this, we consider the expectation $\mathbb{E}s(M_n, z)$ of the *Stieltjes transform*

$$s(M_n, z) := \mathbb{E}\frac{1}{n}\mathrm{tr}R(M_n, z)$$

for some complex number $z = E + i\eta$ with $\eta > 0$, where $R(M_n, z) := (M_n - z)^{-1}$ denotes the resolvent (also known as the *Green's function*), and we identify z with the matrix zI_n. The application of the Lindeberg exchange method to quantities relating to Green's functions is also referred to as the *Green's function comparison method*. The Stieltjes transform is closely tied to the behavior of the eigenvalues $\lambda_1, \ldots, \lambda_n$ of M_n (counting multiplicity), thanks to the spectral identity

$$s(M_n, z) = \frac{1}{n}\sum_{j=1}^{n}\frac{1}{\lambda_j - z}. \tag{3.8}$$

On the other hand, the resolvents $R(M_n, z)$ are particularly amenable to the Lindeberg exchange method, thanks to the *resolvent identity*

$$R(M_n + A_n, z) = R(M_n, z) - R(M_n, z)A_nR(M_n + A_n, z)$$

whenever M_n, A_n, z are such that both sides are well defined (e.g. if M_n, A_n are real symmetric and z has positive imaginary part); this identity is easily verified by multiplying both sides by $M_n - z$ on the left and $M_n + A_n - z$ on the right. One can iterate this to obtain the Neumann series

$$R(M_n + A_n, z) = \sum_{k=0}^{\infty}(-R(M_n, z)A_n)^kR(M_n, z), \tag{3.9}$$

assuming that the matrix $R(M_n, z)A_n$ has spectral radius less than one. Taking normalised traces and using the cyclic property of trace, we conclude in particular that

$$s(M_n + A_n, z) = s(M_n, z) + \sum_{k=1}^{\infty}\frac{(-1)^k}{n}\mathrm{tr}\left(R(M_n, z)^2A_n(R(M_n, z)A_n)^{k-1}\right). \tag{3.10}$$

To use these identities, we invoke the following useful facts about Wigner matrices:

Theorem 3.7. *Let M_n be a real Wigner matrix, let $\lambda_1, \ldots, \lambda_n$ be the eigenvalues, and let $u_1, \ldots, u_n$ be an orthonormal basis of eigenvectors. Let $A > 0$ be any constant. Then with probability $1 - O_A(n^{-A})$, the following statements hold:*

(i) *(Weak local semi-circular law) For any interval $I \subset \mathbb{R}$, the number of eigenvalues in I is at most $n^{o(1)}(1 + n|I|)$.*
(ii) *(Eigenvector delocalisation) All of the coefficients of all of the eigenvectors $u_1, \ldots, u_n$ have magnitude $O(n^{-1/2+o(1)})$.*

The same claims hold if one replaces one of the entries of M_n, together with its transpose, by zero.

Proof. See for instance [44, Theorem 60, Proposition 62, Corollary 63]; related results are also given in the lectures of Erdős. Estimates of this form were introduced in the work of Erdős, Schlein, and Yau [16, 17, 18]. One can sharpen these estimates in various ways (e.g. one has improved estimates near the edge of the spectrum), but the form of the bounds listed here will suffice for the current discussion. □

This yields some control on resolvents:

Exercise 3.8. Let M_n be a real Wigner matrix, let $A > 0$ be a constant, and let $z = E + i\eta$ with $\eta > 0$. Show that with probability $1 - O_A(n^{-A})$, all coefficients of $R(M_n, z)$ are of magnitude $O(n^{o(1)}(1 + \frac{1}{n\eta}))$, and all coefficients of $R(M_n, z)^2$ are of magnitude $O(n^{o(1)}\eta^{-1}(1 + \frac{1}{n\eta}))$. (*Hint:* use the spectral theorem to express $R(M_n, z)$ in terms of the eigenvalues and eigenvectors of M_n. You may wish to treat the $\eta > 1/n$ and $\eta \leqslant 1/n$ cases separately.) The bounds here are not optimal regarding the off-diagonal terms of $R(M_n, z)$ or $R(M_n, z)^2$; see [20] for some stronger estimates.

On the exceptional event where the above exercise fails, we can use the crude estimate (from spectral theory) that $R(M_n, z)$ has operator norm at most $1/\eta$.

We are now ready to establish (3.7) for the Stieltjes transform statistic $F(M_n) := s(M_n, z)$ for certain choices of spectral parameter $z = E + i\eta$, and certain choices of Wigner ensemble M_n. Let (i, j) be an element of Δ. By Exercise 3.8, we have with probability $1 - O(n^{-100})$ (say) that all coefficients of $R(M_{n,i,j}, z)$ are of magnitude $O(n^{o(1)}(1 + \frac{1}{n\eta}))$, and all coefficients of $R(M_{n,i,j}, z)^2$ are of magnitude $O(\eta^{-1}n^{o(1)}(1 + \frac{1}{n\eta}))$; also from the subgaussian hypothesis we may assume that $\xi_{ij} = O(n^{o(1)})$ without significantly increasing the failure probability of the above event. Among other things, this implies that $R(M_{n,i,j}, z)E_{ij}$ has spectral radius $O(n^{o(1)}(1 + \frac{1}{n\eta}))$. Conditioning to this event, and assuming that $\eta \geqslant n^{-3/2+\varepsilon}$ for some fixed $\varepsilon > 0$ (to keep the spectral radius of $R(M_{n,i,j}, z)E_{ij}$ less than one), we then see from (3.10) that

$$F(M_{n,i,j} + \frac{1}{\sqrt{n}}\xi_{ij}E_{ij}) = F(M_{n,i,j}) + \sum_{k=1}^{\infty} \frac{(-\xi_{ij})^k}{n^{1+k/2}} \mathrm{tr}\left(R(M_{n,i,j}, z)^2 E_{ij}(R(M_{n,i,j}, z)E_{ij})^{k-1}\right).$$

From the coefficient bounds, we see that the trace here is of size $O(\eta^{-1}(n^{o(1)}(1 + \frac{1}{n\eta}))^k)$ (where the $n^{o(1)}$ expression, or the implied constant in the $O()$ notation, does not depend

on k). Thus we may truncate the sum at any stage to obtain

$$F\left(M_{n,i,j} + \frac{1}{\sqrt{n}}\xi_{ij}E_{ij}\right) = F(M_{n,i,j}) + \sum_{k=1}^{m} \frac{(-\xi_{ij})^k}{n^{1+k/2}} \mathrm{tr}\left(R(M_{n,i,j},z)^2 E_{ij}(R(M_{n,i,j},z)E_{ij})^{k-1}\right) + O_m\left(\eta^{-1} n^{-\frac{m+3}{2}+o(1)}(1+\frac{1}{n\eta})^{m+1}\right)$$

with probability $1-O(n^{-100})$. On the exceptional event, we can bound all terms on the left and right-hand side crudely by $O(n^{10})$ (say). Taking expectations, and using the independence of ξ_{ij} from $M_{n,i,j}$, we conclude that

$$\mathbb{E}F\left(M_{n,i,j} + \frac{1}{\sqrt{n}}\xi_{ij}E_{ij}\right) = \mathbb{E}F(M_{n,i,j}) + \sum_{k=1}^{m} \frac{\mathbb{E}(-\xi_{ij})^k}{n^{1+k/2}} \mathbb{E}\mathrm{tr}\left(R(M_{n,i,j},z)^2 E_{ij}(R(M_{n,i,j},z)E_{ij})^{k-1}\right) + O\left(\eta^{-1} n^{-\frac{m+3}{2}+o(1)}(1+\frac{1}{n\eta})^{m+1}\right)$$

for any $1 \leqslant m \leqslant 10$ (say). Similarly with ξ_{ij} replaced by η_{ij}. If we assume that ξ_{ij} and η_{ij} have m matching moments in the sense that

$$\mathbb{E}\xi_{ij}^k = \mathbb{E}\eta_{ij}^k$$

for all $k = 1, \ldots, m$, we conclude on subtracting that

$$\mathbb{E}F\left(M_{n,i,j} + \frac{1}{\sqrt{n}}\xi_{ij}E_{ij}\right) - \mathbb{E}F\left(M_{n,i,j} + \frac{1}{\sqrt{n}}\eta_{ij}E_{ij}\right) = O\left(\eta^{-1} n^{-\frac{m+3}{2}+o(1)}(1+\frac{1}{n\eta})^{m+1}\right).$$

Comparing this with (3.7), we arrive at the following conclusions for the statistic $F(M_n) = s(M_n, z)$:

- By definition of a Wigner matrix, we already have 2 matching moments. Setting $m = 2$, we conclude that $\mathbb{E}s(M_n, z)$ enjoys a two moment theorem whenever $\eta \geqslant n^{-1/2+\varepsilon}$ for some fixed $\varepsilon > 0$.
- If we additionally assume a third matching moment $\mathbb{E}\xi_{ij}^3 = \mathbb{E}\eta_{ij}^3$, then we may set $m = 3$, and we conclude that $\mathbb{E}s(M_n, z)$ enjoys a three moment theorem whenever $\eta \geqslant n^{-1+\varepsilon}$ for a fixed $\varepsilon > 0$.
- If we assume third and fourth matching moments $\mathbb{E}\xi_{ij}^3 = \mathbb{E}\eta_{ij}^3$, $\mathbb{E}\xi_{ij}^4 = \mathbb{E}\eta_{ij}^4$, then we may set $m = 4$, and we conclude that $\mathbb{E}s(M_n, z)$ enjoys a four moment theorem whenever $\eta \geqslant n^{-\frac{13}{12}+\varepsilon}$ for some $\varepsilon > 0$.

Thus we see that the expected Stieltjes transform $\mathbb{E}s(M_n, z)$ has some universality, although the amount of universality provided by the Lindeberg exchange method degrades as z approaches the real axis. Additional matching moments beyond the fourth will allow one to approach the real axis even further, although with the bounds provided here, one cannot get closer than $n^{-3/2}$ due to the potential divergence of the Neumann series beyond this point. Some of the exponents in the above results can be improved by using more refined control on the Stieltjes transform, and by using the Knowles-Yin variant of the Lindeberg exchange method; see for instance [26].

One can generalise these arguments to more complicated statistics than the Stieltjes transform $s(M_n, z)$. For instance, one can establish a four moment theorem for multilinear averages of the Stieltjes transform:

Exercise 3.9. Let k be a fixed natural number, and let $\psi : \mathbb{R}^k \to \mathbb{C}$ be a smooth compactly supported function, both of which are independent of n. Let $z = E + i\eta$ be a complex number with $\eta \geqslant n^{-1-\frac{1}{100k}}$. Show that the statistic

$$\mathbb{E} \int_{\mathbb{R}^k} \psi(t_1, \ldots, t_k) \prod_{l=1}^{k} s\left(M_n, z + \frac{t_l}{n}\right)\, dt_1 \ldots dt_k$$

enjoys a four moment theorem. Similarly if one replaces one or more of the $s(M_n, z + \frac{t_l}{n})$ with their complex conjugates.

Using this, one can then obtain analogous four moment theorems for k-point correlation functions. Recall that for any fixed $1 \leqslant k \leqslant n$, the k-point correlation function $\rho^{(k)} : \mathbb{R}^k \to \mathbb{R}^+$ of a random real symmetric (or Hermitian) matrix M_n is defined via duality by requiring that $\rho^{(k)}$ be symmetric and obey the relation

$$\int_{\mathbb{R}^k} \rho^{(k)}(x_1, \ldots, x_k) F(x_1, \ldots, x_k)\, dx_1 \ldots dx_k = \mathbb{E} \sum_{1 \leqslant i_1 < \cdots < i_k \leqslant n} F(\lambda_{i_1}, \ldots, \lambda_{i_k})$$

for all continuous compactly supported symmetric functions $F : \mathbb{R}^k \to \mathbb{C}$, where $\lambda_1 \leqslant \cdots \leqslant \lambda_n$ denote the eigenvalues of M_n arranged in increasing order (and counting multiplicity). From the Riesz representation theorem we see that this defines $\rho^{(k)}$ as a Radon measure at least; if M_n has an absolutely continuous distribution (as is the case for instance with the GOE ensemble) then ρ will in fact be a locally integrable function. Setting $k = 1$ and $F(x) = \frac{1}{x-z}$, we see in particular that

$$\int_{\mathbb{R}} \rho^{(1)}(x)\, \frac{dx}{x - z} = n \mathbb{E} s(M_n, z). \tag{3.11}$$

Similarly, setting $k = 2$ and $F(x_1, x_2) = \frac{1}{x_1 - z_1} \frac{1}{x_2 - z_2} + \frac{1}{x_1 - z_2} \frac{1}{x_2 - z_1}$, then setting $k = 1$ and $F(x) = \frac{1}{(x-z_1)(x-z_2)}$, and adding, we see that

$$2 \int_{\mathbb{R}^2} \rho^{(2)}(x_1, x_2)\, \frac{dx_1 dx_2}{(x_1 - z_1)(x_2 - z_2)} + \int_{\mathbb{R}^2} \rho^{(1)}(x)\, \frac{dx}{(x - z_1)(x - z_2)} = n^2 \mathbb{E} s(M_n, z_1) s(M_n, z_2).$$

By combining these sorts of identities with Exercise 3.9, one can obtain four moment theorems for correlation functions. For instance, we have

Proposition 3.10. *Let E be a real number, and let $\psi : \mathbb{R} \to \mathbb{R}$ be a smooth compactly supported function. Then the statistic*

$$\frac{1}{n} \int_{\mathbb{R}} \rho^{(1)}(E + \frac{s}{n}) \psi(s)\, ds$$

enjoys a four moment theorem.

Proof. Set $\eta := n^{-1-\frac{1}{100}}$. From Exercise 3.9, the statistic

$$\mathbb{E} \int_{\mathbb{R}} \psi(t) s(M_n, E + \frac{t}{n} + i\eta)\, dt$$

enjoys a four moment theorem. Applying (3.11), we conclude that

$$\frac{1}{n}\int_{\mathbb{R}}\int_{\mathbb{R}} \rho^{(1)}(x)\psi(t)\ \frac{dxdt}{x-E-\frac{t}{n}-i\eta}$$

enjoys a four moment theorem. Making the change of variables $x = E + \frac{s}{n}$, this becomes

$$\frac{1}{n}\int_{\mathbb{R}} \rho^{(1)}(E+\frac{s}{n})\left(\int_{\mathbb{R}} \frac{\psi(t)\ dt}{s-t-in\eta}\right)\ ds.$$

Taking imaginary parts and dividing by π, we conclude that

$$\frac{1}{n}\int_{\mathbb{R}} \rho^{(1)}(E+\frac{s}{n})\left(\int_{\mathbb{R}} \frac{n\eta\psi(t)\ dt}{\pi((s-t)^2+(n\eta)^2)}\right)\ ds$$

enjoys a four moment theorem. However, from the smoothness and compact support of ψ, and the fact that the Cauchy distribution $\frac{n\eta}{\pi((s-t)^2+(n\eta)^2)}$ has mean one, one can establish a bound of the form

$$\int_{\mathbb{R}} \frac{n\eta\psi(t)\ dt}{\pi((s-t)^2+(n\eta)^2)} = \psi(s) + O\left(\frac{n^{-1/100}}{1+s^2}\right)$$

(Exercise!). On the other hand, from (3.11) and Theorem 3.7 one can show that

$$\frac{1}{n}\int_{\mathbb{R}} \rho^{(1)}\left(E+\frac{s}{n}\right)\frac{ds}{1+s^2} \ll n^{o(1)}$$

(Exercise!). The claim now follows from the triangle inequality. □

Exercise 3.11. Justify the two steps marked (Exercise!) in the above proof.

Exercise 3.12. If k is a fixed natural number, $E_1, \ldots, E_k$ are real numbers, and $\psi : \mathbb{R}^k \to \mathbb{R}$ is a smooth compactly supported function, show that the statistic

$$\frac{1}{n^k}\int_{\mathbb{R}} \rho^{(k)}\left(E_1+\frac{s_1}{n}, \ldots, E_k+\frac{s_k}{n}\right)\psi(s_1,\ldots,s_k)\ ds_1\ldots s_k$$

enjoys a four moment theorem.

With a bit more effort (using now the real part of the Stieltjes transform, in addition to the imaginary part), one can also use Exercise 3.9 to establish a four moment theorem for statistics involving the individual eigenvalues and eigenvectors of M_n; see [26] for details. Such theorems were also established by direct Taylor expansion of the eigenvalues and eigenvectors; see [44], [45].

Of course, one would like to also establish universality results for classes of random matrices with fewer than four matching moments. At the scale of the mean eigenvalue spacing of $1/n$, it appears (in the bulk, at least) that the Lindeberg exchange method is not powerful enough on its own to accomplish this task. However, the Lindeberg exchange method combines well with other universality results involving ensembles in which the third and fourth moments are allowed to vary. For instance, a different way to obtain universality between a real Wigner matrix M_n and a GOE matrix G_n (which we take to be independent of M_n) is not to exchange the elements from M_n to G_n

one at a time, but instead to consider an Ornstein-Uhlenbeck type process that flows continuously from M_n to G_n from time $t = 0$ to time $t = 1$, by the formula

$$M_n^t := (1-t)^{1/2} M_n + t^{1/2} G_n.$$

Clearly M_n^t equals M_n when $t = 0$ and G_n when $t = 1$. For any intermediate time t, M_n^t is a real Wigner matrix (of a special type known as a *Gauss divisible* Wigner matrix), and the third and fourth moments of the components of M_n^t vary in some explicit polynomial fashion from those of ξ_{ij} to those of η_{ij}. A key feature of this flow is that the eigenvalues of M_n^t evolve by the laws of *Dyson Brownian motion*. Using techniques such as the method of *local relaxation flow* as discussed in the lecture notes of Erdős, one can obtain good universality results[10] that compare M_n^t to G_n for t as low as $n^{-1+\varepsilon}$ for any fixed $\varepsilon > 0$. Meanwhile, the Lindeberg exchange method can be used to compare[11] M_n^t to M_n for $t = n^{-1+\varepsilon}$. Combining these two claims, one can obtain universality results for Wigner matrices that do not require additional matching moments beyond the second; see e.g. [15] or [20]. On the other hand, there are some spectral statistics, particularly those involving a fixed eigenvalue of a Wigner matrix, for which the matching of the fourth moment is in fact necessary: see [47], [13].

References

[1] R. Adamczak, *On the Marchenko-Pastur and circular laws for some classes of random matrices with dependent entries*, Electronic Journal of Probabability, **16** (2011), 1068–1095.

[2] R. Adamczak, D. Chafaï, P. Wolff, *Circular law for random matrices with exchangeable entries*, Random Structures & Algorithms, **48** (2016), 454–479.

[3] Z. D. Bai, *Circular law*, Ann. Probab. **25** (1997), no. 1, 494–529.

[4] A. Basak, N. Cook, O. Zeitouni, *Circular law for the sum of random permutation matrices*, preprint.

[5] A. Basak, M. Rudelson, *Circular law for sparse matrices*, preprint.

[6] C. Bordenave, P. Caputo, D. Chafaï, *Circular law theorem for random markov matrices*, Probability Theory and Related Fields, **152** (2012), 751–779.

[7] C. Bordenave, D. Chafaï, *Around the circular law*, Probability surveys, **9** (2012), 1–89.

[8] P. Bourgade, L. Erdős, H.-T. Yau, J. Yin, *Fixed energy universality for generalized Wigner matrices*, Comm. Pure Appl. Math. **69** (2016), no. 10, 1815–1881.

[9] J. Bourgain, V. Vu, P. M. Wood, *On the singularity probability of discrete random matrices*, J. Funct. Anal. **258** (2010), no. 2, 559–603.

[10] N. Cook, *The circular law for random regular digraphs*, preprint.

[11] A. Edelman, *The probability that a random real Gaussian matrix has k real eigenvalues, related distributions, and the circular law*, J. Multivariate Anal. **60** (1997), no. 2, 203–232.

[12] A. Edelman, *Eigenvalues and condition numbers of random matrices*, SIAM J. Matrix Anal. Appl. **9** (1988), 543–560.

[13] A. Edelman, A. Guionnet, S. Péché, *Beyond universality in random matrix theory*, Ann. Appl. Probab. **26** (2016), no. 3, 1659–1697.

[14] P. Erdős, *On a lemma of Littlewood and Offord*, Bull. Amer. Math. Soc. **51**, (1945) 898–902.

[10] Strictly speaking, the method of local relaxation flow requires an additional averaging in the energy parameter E in order to obtain this claim. However, more recent methods are now available to avoid this averaging in energy: see [8].

[11] There is a minor additional issue because the third and fourth moments of the coefficients of M_n and M_n^t are not quite identical, however it turns out that the additional error terms caused by this are negligible when $t = n^{-1+\varepsilon}$ for ε small enough.

[15] L. Erdős, J. Ramírez, B. Schlein, T. Tao, V. Vu, H.-T. Yau, *Bulk universality for Wigner hermitian matrices with subexponential decay*, Math. Res. Lett. **17** (2010), 793–794.
[16] L. Erdős, B. Schlein and H.-T. Yau, *Semicircle law on short scales and delocalization of eigenvectors for Wigner random matrices*, Ann. Probab. **37** (2009), no. 3, 815–852.
[17] L. Erdős, B. Schlein and H.-T. Yau, *Local semicircle law and complete delocalization for Wigner random matrices*, Comm. Math. Phys. **287** (2009), no. 2, 641–655.
[18] L. Erdős, B. Schlein and H.-T. Yau, *Wegner estimate and level repulsion for Wigner random matrices*, . Int. Math. Res. Not. IMRN 2010, no. 3, 436–479.
[19] L. Erdős, H.-T. Yau, and J. Yin, *Bulk universality for generalized Wigner matrices*, Probab. Theory Related Fields **154** (2012), no. 1-2, 341–407.
[20] L. Erdős, H.-T. Yau, and J. Yin, *Rigidity of Eigenvalues of Generalized Wigner Matrices*, Adv. Math. **229** (2012), no. 3, 1435–1515.
[21] V. L. Girko, *The circular law*, Teor. Veroyatnost. i Primenen. **29** (1984), no. 4, 669–679.
[22] V. L. Girko, *The strong circular law. Twenty years later. II*, Random Oper. Stochastic Equations 12 (2004), no. 3, 255–312.
[23] F. Götze, A. Tikhomirov, *On the circular law*, `arXiv:math/0702386` .
[24] G. Halász,*Estimates for the concentration function of combinatorial number theory and probability*, Period. Math. Hungar. **8** (1977), no. 3–4, 197–211.
[25] J. Kahn, J. Komlós, E. Szemerédi, *On the probability that a random* ± 1*-matrix is singular*, J. Amer. Math. Soc. **8** (1995), no. 1, 223–240.
[26] A. Knowles, J. Yin, *Anisotropic local laws for random matrices*, `arXiv:1410.3516`
[27] A. Knowles, J. Yin, *Eigenvector Distribution of Wigner Matrices*, Probab. Theory Related Fields 155 (2013), no. 3-4, 543–582.
[28] J. Komlós, *On the determinant of (0,1) matrices*, Studia Sci. Math. Hungar **2** (1967), 7–21.
[29] J. W. Lindeberg, *Eine neue Herleitung des Exponentialgesetzes in der Wahrscheinlichkeitsrechnung*, Math. Z. **15** (1922), 211–225.
[30] A. E. Litvak, A. Pajor, M. Rudelson, N. Tomczak-Jaegermann, *Smallest singular value of random matrices and geometry of random polytopes*, Adv. Math. **195** (2005), no. 2, 491–523.
[31] M. Mehta, Random matrices. Third edition. Pure and Applied Mathematics (Amsterdam), 142. Elsevier/Academic Press, Amsterdam, 2004.
[32] H. Nguyen, *Random doubly stochastic matrices: the circular law*, The Annals of Probability, **42** (2014), 1161–1196.
[33] H. Nguyen, V. Vu, *Circular law for random discrete matrices of given row sum*, Journal of Combinatorics, **4** (2013), 1–30.
[34] G. Pan, W. Zhou, *Circular law, extreme singular values and potential theory*, J. Multivariate Anal. **101** (2010), no. 3, 645–656.
[35] M. Rudelson, *Invertibility of random matrices: norm of the inverse*, Ann. of Math. (2) **168** (2008), no. 2, 575–600.
[36] M. Rudelson, R. Vershynin, *The Littlewood-Offord problem and invertibility of random matrices*, Adv. Math. **218** (2008), no. 2, 600–633.
[37] M. Rudelson, R. Vershynin, *The least singular value of a random square matrix is* $O(n^{-1/2})$, C. R. Math. Acad. Sci. Paris **346** (2008), no. 15–16, 893–896.
[38] M. Talagrand, *Concentration of measure and isoperimetric inequalities in product spaces*, Inst. Hautes Études Sci. Publ. Math. No. **81** (1995), 73–205.
[39] T. Tao, Topics in Random Matrix Theory. Graduate Studies in Mathematics, 132. American Mathematical Society, Providence, RI, 2012.
[40] T. Tao, V. Vu, *On the singularity probability of random Bernoulli matrices*, J. Amer. Math. Soc. **20** (2007), no. 3, 603–628.
[41] T. Tao, V. Vu, *Random matrices: the circular law*, Commun. Contemp. Math. **10** (2008), no. 2, 261–307.
[42] T. Tao, V. Vu, *Inverse Littlewood-Offord theorems and the condition number of random discrete matrices*, Ann. of Math. (2) **169** (2009), no. 2, 595–632.
[43] T. Tao, V. Vu, *Random matrices: the distribution of the smallest singular values*, Geom. Funct. Anal. **20** (2010), no. 1, 260–297.

[44] T. Tao, V. Vu, *Random matrices: universality of local eigenvalue statistics*, Acta Math **206** (2011), 127–204
[45] T. Tao, V. Vu, *Random matrices: Universal properties of Eigenvectors*, Random matrices: Theory and Applications **1** (2012), 1150001.
[46] T. Tao, V. Vu, M. Krishnapur, *Random matrices: Universality of ESDs and the circular law*, Annals of Probability **38** (2010), no. 5, 2023–206.
[47] T. Tao, V. Vu, *Random matrices: Localization of the eigenvalues and the necessity of four moments*, Acta Mathematica Vietnamica 36 (2011), 431–449.
[48] L. N. Trefethen, *Pseudospectra of matrices*, Numerical analysis 1991 (Dundee, 1991), 234–266, Pitman Res. Notes Math. Ser., 260, Longman Sci. Tech., Harlow, 1992.
[49] P. M. Wood, *Universality and the circular law for sparse random matrices*, The Annals of Applied Probability, **22** (2012), 1266–1300.

UCLA Department of Mathematics, Los Angeles, CA 90095-1555.

E-mail address: tao@math.ucla.edu